PREFACE

This volume constitutes the Proceedings of the 9th International Congress of Logic, Methodology and Philosophy of Science arranged by the Division of Logic, Methodology and Philosophy of Science of the International Union of History and Philosophy of Science. The logical sections of the Congress also constituted the European Logic Colloqium '91. The Congress took place in Uppsala, Sweden, from August 7 to August 14, 1991. It was held at the invitation of the Swedish National Committee for Logic, Methodology and Philosophy under the auspices of the Royal Swedish Academy of Sciences, the Royal Academy of Letters, History and Antiquities, and the University of Uppsala.

The scientific program of the Congress consisted of 54 invited lectures and about 650 contributed papers scheduled within 10 ordinary working sessions, one opening and one closing session, and one evening session. Most of the work of the Congress was divided into 15 sections, listed below. In addition there were 2 plenary lectures, 4 special symposia, and 2 affiliated meetings. During the Congress the General Assembly of the Division of Logic, Methodology and Philosophy of Science met in Stockholm at the the Royal Swedish Academy of Sciences.

This volume contains the text, sometimes revised, of most of the invited lectures; some of the speakers were unfortunately not able to turn in their manuscripts in time. The names of all the invited lecturers appear below arranged by sections, symposia or plenary sessions. The titles of all accepted contributed papers are listed at the end of the volume. A small selection of papers contributed to sections 1–5 has been published as a special issue of *Annals of Pure and Applied Logic*; it appeared as number 1 of volume 63, 1993. A selection of papers contributed to the other sections will appear in a volume *Logic and Philosophy of Science in Uppsala* published in Synthese Library by Kluwer Academic Publishers.

Appended to this preface is a list of the members of the General Program Committee, which had the overall responsibility for the scientific program. The members of the Section Program Committees are named in the list of sections. Also appended to the preface is a list of the officers of the Division of Logic, Methodology and Philosophy of Science for 1987–91, a list of the members of the Organizing Committee, and a list of those who sponsored the Congress financially.

The volume has been produced from camera ready manuscripts. In most respects the papers appear in the typographical form choosen by the authors. However, the final preparation of the manuscripts has been the responsibility of the editors. The production of the camera ready manuscripts has been carried out by Mrs. Freyja Hreinsdóttir and Mrs. Siv Sandvik. We thank them for their patient and careful work. Their work was financially supported by a generous grant from the Swedish Council for Research in the Humanities and Social Sciences.

Stockholm and Irvine, March 1994.

DAG PRAWITZ

BRIAN SKYRMS

DAG WESTERSTÅHL

APPENDIX TO THE PREFACE

Sections, Plenary Lectures and Special Symposia

(Titles of invited lectures are given below only when the corresponding paper does not appear in this volume.)

LOGIC

Section 1. Proof Theory and Categorial Logic
Section program committee: G. Takeuti (USA, chair), M. Makkai (Canada), W. Pohlers (USA).
Speakers: S. Buss (USA), J. Lambek (Canada), G. Mints (Estonia), M. Rathjen (Germany).

Section 2. Model Theory, Set Theory and Formal Systems
Section program committee: A. MacIntyre (UK, chair), G.L. Cherlin (USA), D.A. Martin (USA).
Speakers: A. Louveau (France), W. Mitchell (USA), A. Woods (Australia, "Counting finite models").

Section 3. Recursion Theory and Constructivism
Section program committee: R.I. Soare (USA, chair), A. Lachlan (Canada), S. Wainer (UK).
Speakers: M. Arslanov (Russia), B. Cooper (UK), M. Lerman (USA).

Section 4. Logic and Computer Science
Section program committee: J. van Benthem (The Netherlands, chair), D. Gabbay (UK), Yu. Gurevich (USA).
Speakers: P. Aczel (UK, "Structured objects"), J. Makowsky (Israel), D. Nute (USA).

Section 5. Philosophical Logic
Section program committee: K. Fine (USA, chair), M.J. Cresswell (New Zealand), T. Smiley (England).
Speakers: J. Burgess (USA), K. Segerberg (Sweden), A.I.F. Urquhart (New Zealand).

GENERAL PHILOSOPHY OF SCIENCE

Section 6. Methodology
Section program committee: B. van Fraassen (USA, chair), M.-L. Dalla Chiara (Italy), R. Cooke (The Netherlands).
Speakers: C. Glymour (USA, "Reliability"—joint paper with K.M. Kelly, read at the Congress by P. Spirtes), P. Maddy (USA), P.M. Williams (UK).

Section 7. Probability, Induction and Decision Theory
Section program committee: H.E. Kyburg (USA, chair), I. Levi (USA), W. Spohn (Germany).
Speakers: P. Gärdenfors (Sweden), T. Seidenfeld (USA).

Section 8. History of Logic, Methodology and Philosopy of Science
Section program committee: C. Thiel (Germany, chair), B.V. Birjukov (Russia), G. Lolli (Italy).
Speakers: N. Nagorny (Russia), J. Wolenski (Poland).

Section 9. Ethics of Science and Technology
Section program committee: R. Haller (Austria, chair), J. Feinberg (USA), K.-E. Tranøy (Norway).
Speakers: L. Bergström (Sweden), A. Gibbard (USA), Qiu Renzong (P. R. China).

PHILOSOPHICAL AND FOUNDATIONAL PROBLEMS ABOUT THE SCIENCES

Section 10. Logic, Mathematics and Computer Science
Section program committee: P. Martin-Löf (Sweden, chair), P. Aczel (UK), S. Feferman (USA).
Speakers: T. Coquand (France), E. Nelson (USA), R. Tieszen (USA).

Section 11. Physical Sciences
Section program committee: A. Shimony (USA, chair), J. Butterfield (UK), J. von Plato (Finland).
Speakers: M. Berry (UK), H. Primas (Switzerland), H. Stein (USA).

Section 12. Biological Sciences
Section program committee: E. Sober (USA, chair), J. Hodges (UK), J. Mosterin (Spain).
Speakers: G. Vollmer (Germany), J. Wicken (USA, "Extending Darwinism: perspectives from the physical sciences").

Section 13. Cognitive Science and Artifical Intelligence (including Computational Perspectives in Psychology)
Section program committee: P. Gärdenfors (Sweden, chair), A. Clark (UK).
Speakers: D.C. Dennett (USA), R. Penrose (UK, discussion with Dennett on artificial intelligence and physics).

Section 14. Linguistics
Section program committee: T. Parsons (USA, chair).
Speakers: S. Bromberger and M. Halle (USA), M.J. Cresswell (New Zealand), J. van Benthem (The Netherlands).

Section 15. Social Sciences (including Non-Computational Psychology)
Section program committee: C. Glymour (USA, chair), M.A. Wylie (Canada).
Speakers: C. Granger (USA), J. Pearl (USA), P. Spirtes (USA).

PLENARY SPEAKERS

Opening speaker: G.H. von Wright (Finland).
Closing speaker: S. Kripke (USA, "Logicism and de re belief about natural numbers").

SPECIAL SYMPOSIA

Symposium on Prediction
Chair: P. Suppes (USA).
Speakers: J. Crutchfield (USA, "Thermodynamics of the artificial"), P. Diaconis (USA, "The problems of thinking too much"), W.D. Sudderth (USA).

Carnap and Reichenbach Centennial Symposium
Chair: R. Marcus (USA).
Speakers: R. Jeffrey (USA), H. Putnam (USA).

Stig Kanger Memorial Symposium on the Logic of Rights and Choices
Chair: D. Føllesdal (Norway).
Speakers: D. Føllesdal (Norway), L. Lindahl (Sweden), A. Sen (USA).

Symposium on Game Theory
Chair: E. Kalai (USA).
Speakers: R. Aumann (Israel, "Some thoughts on the foundations of game theory"), K. Binmore (USA), J.C. Harsanyi (USA).

Executive Committee of the Division of Logic, Methodology and Philosophy of Science, International Union of History and Philosophy of Science, 1987–1991

L. Jonathan Cohen (UK, president), Ivan T. Frolov (Russia, 1st vice president), Dirk van Dalen (The Netherlands, 2nd vice president), Risto Hilpinen (Finland, secretary general), Helmut Pfeiffer (Germany, treasurer), Dana S. Scott (USA, past president).

General Program Committee

Brian Skyrms (USA, chair), Margaret Boden (UK), V. A. Lektorsky (Russia), Graham Nerlich (Australia), Dag Prawitz (Sweden).

Local Organizing Committee

Dag Prawitz (chair), Dag Westerståhl (congress secretary), Martin Edman, Aant Elzinga, Peter Gärdenfors, Sten Lindström, Per Martin-Löf, Wlodzimierz Rabinowicz, Sören Stenlund, Viggo Stoltenberg-Hansen, Claes Åberg. Secretary: Jane Schultzberg.

Financial Sponsors of the Congress

UNESCO, International Council of Scientific Unions, Swedish Ministry of Education and Cultural Affairs, Swedish Council for Planning and Coordination of Research (FRN), Swedish Council for Research in the Humanities and Social Sciences (HSFR), Swedish Natural Science Research Council (NFR), Swedish National Board of Technical Development (STU), Royal Swedish Academy of Sciences and its foundation In Memory of Jakob and Marcus Wallenberg, Royal Academy of Letters, History and Antiquities, University of Uppsala, University of Stockholm, City of Uppsala, City of Stockholm, Anders Karitz Foundation, Royal Scientific Society in Uppsala, Ericsson Telecom AB.

CONTENTS

4. LOGIC AND COMPUTER SCIENCE

5. PHILOSOPHICAL LOGIC

6. METHODOLOGY

7. PROBABILITY, INDUCTION AND DECISION THEORY

8. HISTORY OF LOGIC, METHODOLOGY AND PHILOSOPHY OF SCIENCE

9. ETHICS OF SCIENCE AND TECHNOLOGY

10. FOUNDATIONS OF LOGIC, MATHEMATICS AND COMPUTER SCIENCE

11. FOUNDATIONS OF PHYSICAL SCIENCES

12. FOUNDATIONS OF BIOLOGICAL SCIENCES

13. FOUNDATIONS OF COGNITIVE SCIENCE AND AI (including Computational Perspectives in Psychology)

14. FOUNDATIONS OF LINGUISTICS

15. FOUNDATIONS OF SOCIAL SCIENCES (including Non-Computational Psychology)

INTERSECTIONAL SYMPOSIUM: PREDICTION

INTERSECTIONAL SYMPOSIUM: CARNAP AND REICHENBACH CENTENNIAL SYMPOSIUM

INTERSECTIONAL SYMPOSIUM: STIG KANGER MEMORIAL SYMPOSIUM ON THE LOGIC OF RIGHTS AND CHOICES

INTERSECTIONAL SYMPOSIUM: GAME THEORY

Logic, Methodology and Philosophy of Science IX
D. Prawitz, B. Skyrms and D. Westerståhl (Editors)

PRESIDENT'S ADDRESS

L. JONATHAN COHEN

The Queen's College, Oxford

Ladies and gentlemen: We have been most courteously welcomed by Dr. Gustavsson, as representative of the Government of this beautiful country, where the National Committee for Logic Methodology and Philosophy of Science — of the Royal Swedish Academy of Science — has very kindly invited us to hold our Congress. And we have been welcomed with equal warmth by Prof. Strömholm, the Vice-Chancellor of this famous university, the facilities of which have been so generously put at our disposal. We have been told also about the structure of the Programme, which promises to make this Congress one of the most rewarding that has so far taken place in its field. And we have been told about the local organisational arrangements, which are always so important for the success of a congress. So in this context it is appropriate for me to add — not only on behalf of the Executive Committee of the Division for Logic, Methodology and Philosophy of Science but also on behalf of all the other participants in the Congress — how grateful we are to the various national and international, governmental and non-governmental, bodies that have made the Congress financially possible. Without their most generous assistance the Congress could not have taken place. I hope tht in each case Prof. Prawitz will convey our deeply sincere gratitude to the appropriate authorities. And we must also express our sincerest gratitude to the chairman and members of the Programme Committee, and to the chairman and members of the Organising Committee, for all the work that they have put into making the arrangements for the Congress. Their long, careful and conscientious labours have been essential to it, and we must congratulate them on the dedication to their task that they have all exhibited.

All that I have said so far undoubtedly needs to be said in the Congress' introductory proceedings, and I take pleasure in saying it. But it concerns only the 'how' of the Congress. It relates to how the Congress has

been made financially viable, how it has been accommodated, how its programme has been put together, and how its membership and living arrangements have been organised. It does not say anything about the 'why' of the Congress. Why has it taken place at all? What justifies it? Why should people bother to attend a congress on logic, methodology and philosophy of science, instead of just publishing their own papers, and reading the papers of others, in the relevant books and journals? What justifies such an extensive commitment of time, money and other resources?

Now these are questions to which the difference between a specialised international conference and a comprehensive international congress is highly important. The reason why those working in the relevant areas need to attend international conferences in appropriate areas of logic or philosophy of science are largely the same as those applying to historians, for example, or to physiologists, or to any other scholars and scientists in relation to *their* conferences. What is fundamental is that one needs to sharpen one's perception of what is going on at the frontiers of one's subject by not only listening to what is said by others, but also by participating in formal discussions at the actual meetings or in informal discussions outside them. Indeed this is also, of course, the best way of making sure that you fully understand the complexities of the issues. In addition people need from time to time to submit their own standards of rigour and clarity to recalibration, as it were, in case their normal academic environment is insufficiently stimulating to maintain these standards. If you come from a relatively small country, or a country in which your own intellectual interests are not widely shared, it is obvious that you may sometimes need this — sometimes, indeed, without being fully aware of it. One needs, too, to form some conscious or unconscious criteria of relative importance in regard to the various issues that one may be minded to investigate in one's research, and attendance at relevant international conferences is one good way to help achieve this. The participants at such international academic conferences, especially if supplied beforehand with abstracts of the papers to be delivered, constitute a kind of jury that votes with its feet, both in regard to the importance of a topic and in regard to the ability of a speaker to say something interesting about it.

But is it necessary to organise a whole congress for this purpose, with fifteen different sections, covering a wide range of different issues? Why not just attend conferences in the area of the subject in which you yourself are working? Why not cover just one or two sections rather than fifteen?

I can suggest two main reasons for this. The first is an obvious one, but nevertheless worth bearing constantly in mind. There may well be

stages in the progress of one's research when one has to concentrate attention on a fairly narrow range of issues. But most of us also encounter other stages when analogies or connections between one problem-area and another can be highly illuminating. For example, there are well-known problems about the existence or non-existence of a correlation between psychological laws and neurological uniformities. But is such a correlation made easier to accept by adherence to the thesis so common in the philosophy of physics, that laws of nature are always idealisations, rather than descriptions, of the actual world? Or does the idealisation thesis make such a correlation — a correlation between psychological laws and neurological uniformities — more difficult to accept? Or again, many difficult issues arise in philosophical logic about the relationship between sentences in natural languages and sentences in artificial ones: for example, how is context-dependent disambiguation — so common in everyday speech — to be formalised on a systematic basis? Without a fairly comprehensive theory of linguistic syntax, as at least a starting-point, no solid progress can be made here. So philosophical logic can profit by not being wholly insulated from the philosophy of linguistics. These are just two examples — and there must be many others — of how the broadness of the Congress's coverage may enable participants to get into touch at first hand with the latest thinking in relevant areas other than those in which they themselves normally specialise. So you might well be wasting an important opportunity for the cross-fertilisation of ideas if you spend all your time at the Congress attending meetings on a too narrowly construed range of interests. Nor can you always be sure in advance just which other sections may produce ideas that are important for your own work and which cannot. Sometimes a quite idle curiosity about what the neighbours are up to pays big dividends. Indeed you may even find yourself intrigued by other people's problems or excited at suddenly having an idea about how to solve them.

But there is also another reason why from time to time it is important to hold general congresses like the present one, rather than just more specialised conferences. A problem exists about the nature of what we are doing when, as philosophers of science, we argue about, say, the structure of scientific theories or the fundamental assumptions of particular sciences. If these topics demand philosophical concern, then so too does the philosophy of science itself. You may think that you, as a practising philosopher of science, know what the philosophy of science is. But *do* you — any more than the ordinary practising scientist knows reflectively and self-conciously what science is? You may think you know how the structure and methods of the philosophy of science differ from one type

of issue to another. But *do* you — any more than the ordinary practising scientist can articulate just how scientific reasoning differs from one area of science to another? Or, just as some practising scientists regard the philosophy of science as a worthless and unproductive activity, so too you, as a practising philosopher of science, may despise the metaphilosophy of science. But is the latter attitude any more defensible than the former? And where better to consider these problems than at a Congress where examples of every current branch of the subject are displayed and a high level of professional quality is ensured by the discriminating labours of an international panel of referees?

So the question naturally arises: why is there not a special section of the Congress devoted to characterising what the philosophy of science and the philosophies of mathematics, physics and the other sciences are all about — do they have a unitary problematic, how do they make progress, how are their conclusions justified, what is their relationship with other branches of philosophy, and so on? Surely the planners and organisers of the Congress, you may say, must have realised that this is a serious issue and must have been capable of doing something about it? But behind that question lies the possibility of a paradox, which I shall call the paradox of the essential incompleteness of philosophy congress programmes. Even if there had been enough activity in the area to justify adding a sixteenth section, devoted to papers in the metaphilosophy of science, we might still have heard cries for a seventeenth section, devoted to the meta-meta-philosophy of science. Nor would that necessarily be a frivolous, shallow, or inconsequential request. It is far too soon to be sure that recursively higher and higher levels of meta-philosophy have no interesting problems of their own. So however diligent are the planners and organisers of the Congress in encouraging the submission of papers, their task may be essentially incompletable within a finite framework.

But the point I want to make is just that, if you are a philosopher of science, you should not leave the Congress without having taken at least some advantage of the opportunity that the Congress affords to reflect on the nature of philosophy of science. And there are certainly some important questions to be answered here. For example, here — if anywhere — it should at least be evident what the philosophy of science is not. It is not primarily oriented towards the description or explanation of scientists' activities, as is the history, or the sociology, of science. Nor is it a purely *a priori* enterprise, like the construction of formal systems and the study of their logical or mathematical properties. Nor is it just a tapestry of intuitions, woven together into an inspirational commentary. Yet it has affinities with each of these. Nor is it to be identified with

the ethics of science, though the latter may be part of it. And there are also some more positive questions to be answered. For example, whence does the philosophy of science derive its premisses? From science itself, or from the history or sociology or psychology of science, or from the intuitions of scientists, or from the intuitions of philosophers? Again, when philosophers of science seek to resolve antinomies, like Zeno's paradoxes of motion, or Russell's set-theoretical paradox, or Hempel's paradox of the ravens, or the Prisoner's Dilemma, is the type of resolution at which they aim — i.e. the kind of thing that would count as a resolution of the paradox — the same in each case, or not? More generally — we may ask — does dialogue have a special part to play in the philosophy of science, whereby apparently opposed positions may be seen to be reconcilable? Or is the philosophy of science continuous with science itself, as Russell once claimed? Should philosophers of science aim to establish the validity of selected propositions about scientific reasoning or scientific explanation, or is the existence of such a special category of philosophical propositions an illusion, as Wittgenstein came to think?

There is no better place to reflect on all these issues — in an up-to-date, state-of-the-art setting — than at the present Congress. The only regret that you may come to have is that the Congress does not last longer. So let us now spend no more time on preliminaries, but move forward into the substantive programme of the Congress. We have all got much to learn from it.

Logic, Methodology and Philosophy of Science IX
D. Prawitz, B. Skyrms and D. Westerståhl (Editors)

LOGIC AND PHILOSOPHY IN THE TWENTIETH CENTURY

GEORG HENRIK VON WRIGHT

Academy of Finland, Helsinki, Finland

1. In my talk I shall try to evaluate the place of logic in the philosophy of our century. The attempt is necessarily subjective. Its outcome may be different depending upon whether the evaluator is primarily a logician or primarily a philosopher. I think of myself as a philosopher who, over a period of almost sixty years, has at close quarters been watching and also, to some extent, participated in the development of logic.

As I see things, the most distinctive feature of 20th century philosophy has been the revival of logic and the fermenting role which this has played in the overall development of the subject. The revival dates from the turn of the century. Its entrance on the philosophical stage was heralded by movements which had their original centres at Cambridge and in Vienna, and which later fused and broadened to the multibranched current of thought known as analytical philosophy. As the century is approaching its end we can notice, I think, signs of decline in the influence of logic on developments in philosophy.

Our era was not the first in history which saw logic rise to prominence in philosophy. In the orbit of Western civilization this happened at least twice before. First it happened in Ancient Greece, in the 4th and 3rd centuries B.C. The second great epoch of logical culture was in the Christian Middle Ages. This was connected with the rediscovery of Aristotle mediated by the Arabs, and it lasted, roughly, from the middle of the 12th to the middle of the 14th century.

In between the peaks logic "hibernated". Its latest winter sleep lasted nearly half a millennium — from the mid-fourteenth to the mid-nineteenth century. In this period, there were also logicians of great ability and power. The greatest of them was Leibniz. But his influence *as a logician* on the philosophic climate of the time was small. It was not until the beginning of our century, when Louis Couturat published his *La logique de Leib-*

niz and a number of unedited fragments that Leibniz the logician was discovered.

Logic in the state of hibernation was respected for its past achievements, but not thought capable of significant further development. This attitude is epitomized in Kant's well known *dictum* that logic after Aristotle "keinen Schritt vorwärts hat tun können, und also allem Ansehen nach geschlossen und vollendet zu sein scheint". [1]

2. What we nowadays commonly understand by "logic" was not always referred to with that name.

Although the word derives from a Greek root, Aristotle did not use it for what we think of as his works in logic. Initially, they had no common label at all. The name for them, *Organon* ("instrument") dates from the first century B.C. The Stoics used, with some consistency, the term *dialectics* for what we would call logical study. This term was transmitted to the Middle Ages through the Latin tradition of late Antiquity. One of the earliest works which signalizes the revival of logic is Abelard's *Dialectica*. The same author, however, also used the name "logica" which then became current during the Golden Age of Scholasticism — only to yield ground once more to the rival "dialectica" in the period of the Renaissance. Later, also the name "Organon" was revived. [2] In German writings of the 18th and 19th centuries the terms "Vernunfts-" and "Wissenschaftslehre" were largely used. [3]

For the rehabilitation of the name "logic" the once influential *Logique ou l'art de penser* (1662), also known as the Logic of Port Royal, appears to have been of decisive importance. This revival, however, was concurrent with a deprecation of the medieval tradition and with efforts to create something more in tune with the emerging new science of nature. The logic of Port Royal is not "logic" in our sense. It is more like what we would call "methodology", an "aid to thinking" as the title says.

Kant, who thought Aristotelian logic incapable of development, wanted to renew the subject by creating what he called a *transscendental* logic. This was to deal with "the origin, scope, and objective validity" of *a priori* or "purely rational" knowledge. [4] And Hegel who, it is said, [5] more than anybody else is responsible for the final establishment of the term "logic", says in so many words that the time has come when the conceptions previously associated with the subject "should completely vanish and the position of this science (*sc.* logic) be utterly changed". [6]

Hegel was not entirely unsuccessful in his reformist zeal. What has since been known as Hegelian or dialectical logic has had a foothold in philosophy up to the present day. But it is not *this* which I had in mind

when extolling the role of logic in contemporary philosophic culture. Far from it!

It is characteristic of the terminological vacillations that when the true *logica rediviva* entered the philosophic stage in the early decades of our century, it too wanted to appear under a name of its own. Couturat proposed for it the neologism *logistique*; [7] in German it bacame *Logistik*. The idea was to emphasize, not only its novelty, but also its difference both from the corrupted logic of the immediately preceding centuries and from the Aristotelian and the Scholastic traditions thought obsolete. [8] It was in this "spirit of modernity" that I, for example, was trained in logic as a young student. That the term "logistic" never acquired wide currency in English is probably due to the fact that the plural form of the word already had an established use with a different connotation in this language. [9] Instead, the attributes "mathematical" and "symbolic" were long used to distinguish the new logic from its ancestral forms.

3. In view of the confusion in terminology and multiplicity of traditions, it is necessary to say a few words about what I — and I believe most of us at this congress — understand by logic.

Kant appears to have been first to use the term "formal" for logic in the tradition of Aristotle and the School. [10] Logic studies the *structural* aspects of the ratiocinative processes we call argument, inference, or proof. It lays down rules for judging the correctness of the transition from premisses to conclusions — not rules for judging the truth of the premisses and conclusions themselves. This gives to logic its *formal* character — and it was with a view to it that both Kant and Hegel complained of the subject's "barrenness" and lack of *content*.

The "content" of formal logical study are *concepts*, one could say. Logic studies them, not in their external relation to the world, but in their internal relationships of coherence or its opposite. This is what we call "conceptual analysis". In the simplest cases it takes the form of Aristotelian definitions through specific differences within proximate genera. In more complex and interesting cases it consists of the construction of conceptual networks or "fields", the structural properties of which give meaning to the entities involved. Formalized axiomatic systems are examples of such constructs. Hilbert aptly called them "implicit definitions".

The study of inference and of meaning relations between concepts are the two main pursuits of the discipline of logic. Some would perhaps wish to separate the two aspects more sharply from one another and distinguish them as "formal logic" and "conceptual analysis" respectively. Both attitudes can be justified. The fact remains that it is the close

alliance of the two aspects which has given to philosophy in our century its strong "logical colouring".

4. When one of the many subdivisions of philosophy — be it metaphysics or ethics or logic — assumes distinctive prominence, this is usually connected with some *other* characteristic features of the cultural physiognomy of the time. This holds also for the three epochs in Western culture when the study of logic excelled.

In the history of philosophy, the 4th and 3rd centuries B.C. succeeded the period usually named after the Sophists. This had been an era of childish delight in the newly discovered power of *words* (the λογόι) in the uses and misuses of *arguments* for settling disputes in courts or in the market. The challenge to reflect critically on these early eruptions of untamed rationality gave rise to the tradition in philosophy known as Socratic and, within it, to the more specialized study of the forms of thought we call logic. This was also the time of the first attempts to systematize knowledge of mathematics — as witness Eudoxos's doctrine of proportions and the pre-Euclidean efforts to axiomatize the elements of geometry.

The cultural setting in which medieval Scholasticism flourished was very different. Mathematics and the study of nature were in low waters. The rational efforts of the times were turned toward elucidating and interpreting the *logos* of the Christian scriptures. In its deteriorated forms this activity acquired a reputation for hairsplitting. But it should be remembered that the "hairs" split were *concepts* and that their "splitting", when skilfully done, was conceptual analysis of an acuteness which rivals the best achievements of our century.

With the calamities that befell Europe in the 14th century, the intellectual culture of the Christian Middle Ages also declined. Gradually, a new picture of the world and of man's place in it took shape. It was based on the study of natural phenomena and the use of mathematical tools for theorizing about them. Scholasticism fell in disrepute, and on logic dawned the halfmillennial slumber to which we have already alluded.

What was the cause for the revival of logic in the late 19th century? One might see it in the fact that Western science had by then reached a maturity which made it ripe to reflect critically on its own rational foundations. The organ of the new scientific world-picture being mathematics, it was but natural that the reflexion should start with people who were themselves primarily mathematicians like the two founding fathers of modern logic: Boole and Frege.

Their respective approaches to the subject, however, were rather differ-

ent. [11] Boole, like his contemporary Augustus de Morgan, was concerned with the application of mathematical tools to traditional logic. Their trend was continued, among others, by Peirce and Schröder. Frege's objective was different. He wanted to secure for mathematics a foundation in pure logic. To this end he had not only to revive but also to reshape it.

5. The revitalization of logic thus took its origin from foundation research in mathematics.

The line first taken by Frege and then continued by Russell was, however, but one of a number. In the light of later developments, Frege's and Russell's approach is perhaps better characterized as an attempt to give to mathematics a set-theoretic foundation rather than to derive mathematics from a basis in pure logic. Cantor's figure looms heavily in the background of the logicists' efforts.

Another approach to the foundation problems was Hilbert's conception of mathematics as a family of axiomatized formal calculi to be investigated for consistency, completeness, independence, and other "perfection properties" in a *meta-mathematics*. Hilbert's program is in certain ways a revival of Leibniz's conception of a *calculus ratiocinator*, operating within a *characteristica unversalis*.

A third venture into the foundations of mathematics, finally, was Brouwer's intuitionism. It had forerunners in Kronecker's constructivism and the "semi-intuitionism" of Borel and Poincaré. Brouwer's view of the role of logic was very different both from that of Frege and Russell and from that of Hilbert. [12] The bitter polemics between "intuitionists" and "formalists" bear witness to this. By raising doubts about one of the cornerstones of traditional logic, *viz.* the Law of Excluded Third (or Middle), Brouwer and his followers were also pioneers of what is nowadays known as Deviant or Non-Standard or Non-Classical Logic(s).

Logicism, formalism, and intuitionism were the three main schools which, rivals among themselves, dominated the stage during what I propose to call "the heroic age" in the reborn study of logic. It lasted about half a century, from Frege's *Begriffsschrift* (1879) and *Grundlagen der Arithmetik* (1884) to the appearance of the first volume of Hilbert's and Bernay's monumental *Grundlagen der Mathematik* in 1934. As one who was brought up in the aftermath of this era, I cannot but look back on it with a certain amount of nostalgia. It came to an end in a dramatic climax. I shall shortly return to this. But first, we must take a look at the more immediate repercussions on philosophy which the new logic had had.

6. In earlier days it used to be said that logic studies "the laws of thought". This has been the title of Boole's *magnum opus*. But it was also said that logic was not concerned with (the laws of) psychological thought processes. So what aspect of thought did logic study then? One could answer: *the articulation of thought in language*. Language is, so to speak, the raw material with which logic works. (The Greek *logos* means, ambiguously, both speech and ratiocination.) A time when logic holds a place in the foreground of philosophy is also one in whose intellectual culture language is bound to be prominent.

This is eminently true of the Golden Age of logic in antiquity. The Sophist movement had been an outburst of exuberant delight in the discovery of language as *logos*, i.e. as an instrument of argument, persuasion, and proof. The disciplines of logic and of grammar were the twin offsprings of this attitude.

The logic of the School, too, has been described as a *Sprachlogik* or logic of language. [13] An excessive interest in the linguistic leg-pulling known as "sophismata" seems to have been a contributory cause of the disrepute into which Scholasticism fell in its later days.

The "linguistic turn", [14] which philosophy has taken in our century, has become commonplace. So much so that one may feel tempted to view logic as one offshot among many of the study of language, other branches being theoretical linguistics, computer science, and the study of artificial intelligence and information processing. But this would be a distortion of the historical perspective. Unlike what was the case with the Ancients, with whom logic grew out of an interest in language, it was the revival of logic which, with us, made language central to philosophy. Here Frege's work became a seminal influence. But it is noteworthy that Frege the philosopher of language was "discovered" very much later than Frege the philosopher of logic. This renaissance of Frege's influence and of Fregean studies took place only with "the turn to semantics" in logic in the mid-century.

Hilbert's concern with the language fragments we call calculi did not much influence developments in the philosophy of language. [15] Nor did Brouwer's work do this directly. But Brouwer's attack on formalism is, interestingly, also a critique of language as an articulation of the intuitions underlying mathematical thinking. With his thoughts on the limits of language as well as with some other ideas of his, Brouwer is a precursor of the philosoper who, more than anybody else, has contributed to making language a major concern of contemporary thinking.

7. Even though Wittgenstein never adhered to the logicist position in the philosophy of mathematics, he stands in the *Tractatus* firmly on the shoulders of Frege and Russell. The place of this book in the picture we are here drawing is peculiar.

It would be quite wrong to think of Wittgenstein's contribution to logic as limited to the discovery of the truth-table method for propositional logic and the conception of logical truths as truth-functional tautologies. (The truth-table idea has a long tradition going back way before Wittgenstein.)

Foremostly, *Tractatus* is an inquiry into the possibility of language. How can signs mean? The answer Wittgenstein gave was his picture theory about the isomorphic reflection of the configurations of things in the world, in the configurations of names (words) in the sentence. The essence of language is the essence of the world — their common logical form. This, however, is veiled by the grammatical surface structure of actual speech. The logical deep structure of language is a postulated ideal which shows itself in meaningful discourse but which, since presupposed, cannot be itself described in language.

If we abstract from the peculiarities, not to say eccentricities, of the picture theory and the mysticism of the saying-showing distinction, the *Tractatus* view of logic reflects what I think are common and deep-rooted conceptions of the nature of logical form, necessity, and truth. Indirect confirmation of this may be seen in the coolness, and even hostility, with which logicians and mathematicians, until recently, have received the partly devastating criticism to which Wittgenstein later submitted, not only his own earlier views of logic, but foundation research in general.

The "metaphysics of logic" — as I would like to call it — of the *Tractatus* has survived and, moreover, experienced revivals in more recent times. I am thinking of developments in linguistic theory and in the partly computer-inspired philosophy of mind represented by cognitive science and the study of artificial intelligence.

The "never-never language" [16] which Wittgenstein had postulated in order to explain how language, as we mean it, is possible, has been resurrected in equally speculative ideas about innate grammatical structures or about an ineffable language of thought ("mentalese"), deemed necessary for explaining the child's ability to assimilate with the language community where it belongs. Chomsky's revived *grammaire universelle* or "Cartesian linguistics" is another "crystalline structure" of the kind Wittgenstein in the *Tractatus* had postulated for logic. [17]

For these reasons alone, I think that Wittgenstein's criticism has a message worthy of attention also for contemporary philosophy of language and philosophy of mind. The similarity between the *Tractatus* views and

these latter-day phenomena has not escaped notice. [18] But it has, so far, hardly been deservedly evaluated from a critical point of view. [19] The present situation in cognitive and linguistic research offers interesting parallels to the search for "foundations" which earlier in the century made logic central to the philosophy of mathematics, and which reached what I would call its self-defeating climax in Wittgenstein's *Tractatus*.

8. "Every philosophical problem", Russell wrote on the eve of the First World War, "when it is subjected to the necessary analysis and purification, is found either to be not really philosophical at all, or else to be - - - logical." [20] But he also said that the type of philosophy he was advocating and which had "crept into philosophy through the critical scrutiny of mathematics" had "not as yet many whole-hearted adherents". [21] In this respect a great change was brought about in the post-war decades by the movement known as logical positivism, stemming from the activities of the Wiener Kreis and some kindred groups of science-oriented philosophers and philosophy-oriented scientists in Central Europe. One saw a new era dawning in the intellectual history of man when philosophy too, at long last, had attained *den sicheren Gang einer Wissenschaft*.

According to an influential formulation by Carnap, philosophy was to become the logical syntax of the language of science. This was an extreme position and was in origin associated with views, inherited from earlier positivist and sensualist philosophy, of how a logical constitution of reality, a *logischer Aufbau der Welt*, was to be accomplished.

It is nowadays commonplace to declare logical positivism dead and gone. It should be remembered, however, that the movement was conquered and superseded largely thanks to self-criticism generated in its own circle. This combination of self-destruction with self-development is perhaps unique in the history of thought. At least I know no comparable case. As a result, a narrow conception of philosophy as the logic of science gradually gave place to a conception of it as logical analysis of all forms of discourse. For a just assessment of logical positivism, it is necessary to see the movement as the fountain-head which eventually grew into the broad current of analytic philosophy with its multifarious bifurcations. No one would deny that this has been a mainstream — I should even say *the* mainstream — of philosophy in our century. It is in these facts about its origins: first with foundation research in mathematics, and then with the extension of the use of logical tools to the conceptual analysis of scientific and, eventually, also everyday language, that I found my claim that logic has been the distinctive hallmark of philosophy in our era.

9. What I called "the heroic age" in the history of modern logic came to an end in the 1930s. The turn of a new era [22] was marked by two events, themselves of "heroic" magnitude. The one was Gödel's discovery of the incompleteness properties of formalized calculi; the second Tarski's semantic theory of truth. There is, moreover, an intrinsic connection between the two achievements. [23]

Gödel's incompleteness theorem had serious repercussions on the formalist program of axiomatization, consistency proof, and decidability. It set limits to the idea, ultimately of Leibnizian origin, of the formalization of all ratiocinative thought in syntactic structures and of reasoning as a *jeu de charactères*, a game of signs ignoring their meaning. The related achievement of Tarski meant a transcendence of the syntactic point of view and its supplementation by a semantic one. Therewith it made the relation of language *structure* to language *meaning* amenable to exakt treatment. The immensely fertile field of model theory is an outgrowth of this opening up of the semantic dimension of logic. For its further investigation, Tarski's later work was also of decisive, seminal importance. His pioneering role is in no way minimized by the fact that, seen in the perspective of history, basic ideas in model theory go back to the earlier work of Skolem and Löwenheim.

Gödel's impact on the formalist program, although devastating for the more ambitious, philosophic aspirations of the Hilbert school, also greatly furthered its less ambitious aims. Proof-theory crystallized in the arithmetization of metamathematics and in the theory of computable and recursive functions.

Something similar happened to the line in logic stemming from Frege and Russell and continued through the 1930s, most conspicuously in the work of the young Quine. The antinomies turned out to be a more serious stumbling block than it had seemed after the early efforts of Russell's to conquer the difficulties which had threatened to wreck Frege's system. The semantic antinomies, like the Liar, required extensions beyond type-theory which in none of their suggested forms can be said to have gained universal recognition. The sought for basis of mathematics in pure logic gradually took the shape of a foundation in set-theory. Set-theory, being itself a controversial branch of mathematics, gave prominence to another challenge, *viz.* that of clarifying the axiomatic and conceptual foundations of Cantor's paradise. Even though the difficulties which the logicist approach encountered can be said to have ruined the original aspirations of its initiators, this heir to their program remains, in my opinion, the philosophically most challenging aspect of foundation research in mathematics today. Not surprisingly Gödel, the perhaps most philosophic-minded

mathematical logician of the century, devoted his later efforts mainly to work in this area.

The third mainstream in the early foundation research, intuitionism, also changed its course. In 1930 Heyting codified, in a formal system, the logical rules which were thought acceptable from the intuitionist point of view. Thereby he created an instrument which has turned out to be very useful in the mathematical study of proof, and thus for vindicating that part of Hilbert's program which remained unaffected by Gödel's discoveries. In view of the acrimony which once embittered the fight between formalists and intuitionists and not least the relations between the founders of the two schools, [24] their reconciliation in the later developments of proof-theoretic study may even appear a little ironic.

Brouwer himself was of the opinion that no system of formal rules can encompass the entire range of mathematically sound intuitions. He could therefore not attach great importance to Heyting's achievement. Of Gödel's results he is reported to have said that their gist had been obvious to him long before Gödel presented his proofs. [25]

In his rebuttal of the idea that logic could provide a foundation for mathematics, Brouwer can be said to anticipate the attitude of the later Wittgenstein. Wittgenstein also shared the constructivist leanings of the intuitionists and their critical reflection on some basic principles of classical logic.

The change of climate in logic after the 1930s I would describe as a "disenchantment" (*Entzauberung*) in Max Weber's sense. When the grand dreams and visions of the formalist, intuitionist, and logicist schools had lost their philosophic fascination, what remained and grew out of them was sober, solid science. The discipline which had been the mother of the new logic, *viz.* mathematics, took back its offspring to its sheltered home.

The homecoming did not fail to raise suspicions among the settled members of the family, however. Early in the century, Poincaré had objected to the *logisticiens*, that they pretended to give "wings" (*ailes*) to mathematics but had in fact provided it only with a "hand-rail" (*lisière*) and, moreover, not a very reliable one. [26] On my first encounter with Tarski a few years after the war, Tarski told me of the difficulties and frustrations he had experienced trying to make mathematical logic respected in the mathematics department at Berkeley. I recall something similar from the mathematical establishment in my own country in the form of complaints that some of the most promising students had left the subject and migrated to philosophy. Now, forty years or more later, this attitude no longer prevails in the mathematical profession, except maybe in corners of the world not yet much touched by modern developments.

10. When viewing the history of modern logic as a process of "rational disenchantment" in areas of conceptual crisis or confusion, one is entitled to the judgement that the most exciting development in logical theory after the second world war has been the rebirth of modal logic. The study of modal concepts had flourished in the Aristotelian tradition — not only with its founder, but also with its medieval continuation. In the renaissance starting with Boole and Frege, this study, however, long remained neglected. When eventually it was revived in the work of Łukasiewicz and C. I. Lewis, its rebirth was something of a miscarriage. This was so because it took the form of a critique of Russellian logic. Modal logic was thought of as a "non-classical" alternative or even rival to it.

It was only with the conception of modal logic, not as an alternative to Russell's but rather as a "superstructure" standing on its basis, that the logical study of modalities got a good start in modern times. This conception did not gain ground until after the second world war, although it had had precursors in the 1930s with Gödel and Feys.

A result of the new start was something that could be called a General Theory of Modality. Instead of "General Theory" one could also speak of a *family* of related "logics" of a similar formal structure. These offshoots of an old stem of traditional modal logic have become known as epistemic, doxastic, prohairetic, deontic, and interrogative logic. Historical research has revealed ancestors of many of them either in ancient and medieval logic or with Leibniz, this prodigious logical genius, whose seeds mainly fell in the barren soil of his own time.

One thing which made the study of modal concepts controversial is that it problematized one of the basic principles of logic — it too of Leibnizian ancestry — known as the law of intersubstitutivity *salva veritate* of identities. Such substitutivity in sentential contexts is the hallmark of what is known as *extensionality* in logic. A system of logic which disputes or limits the validity of Leibniz's principle is called *intensional*. Modal logic may therefore be regarded as a province within the broader study of *intensional logic*.

Already Frege had drawn attention to limits of extensionality in doxastic and epistemic contexts. Formal operations in intensional contexts, particularly the use in them of quantifiers, have seemed doubtful and unsound to many logicians of a conservative bent of mind. Above all, Quine has been an acute and staunch critic of modal and other forms of intensional logic. But his criticism has also been a challenge and source of inspiration for a younger generation of logicians, partly following in Quine's footsteps, to clear the jungle of modal and intensional concepts and make their study respectable. To this has contributed the invention

of the very powerful techniques known as possible worlds semantics. The Leibnizian echo in the name is not mere accident.

With these later developments the study of modal and intensional logic has become progressively less "philosophical" and technically more refined. Another process of "disenchantment" is taking place, an initially controversial subject being handed over by philosophically-minded logicians to logically-minded mathematicians.

11. Modal logic, also intensional logic in general, is still in some quarters called "non-classical". There is no received view of what should count as "classical", or not, in logic. As long as modal logic was regarded as an *alternative* to some already canonized structure, the name might have been justified. But modal logic is *not* an "alternative" to the logic systematized by Frege and Russell — at least not to that part of it which is known as first order logic and which consists of the two layers of the propositional and the predicate calculus.

A way of distinguishing classical from non-classical logic, which cuts deeper both historically and systematically, is the following: Classical logic accepts as unrestrictedly valid the two basic principles, first stated by Aristotle and subsequently known as the Law of (Excluded) Contradiction and Law of Excluded Middle (or Third). Both are also fundamental in the logic of Frege and Russell. To question the one or the other is tantamount to doubting the division of what is sometimes called *logical space* in two jointly exhaustive and mutually exclusive parts.

Doubts about the exhaustive nature of the partition were already entertained by the founding father of logic himself. (Yet I do not think it right to interpret Aristotle's discussion of the "Sea-Battle Problem" in the ninth chapter of *Peri Hermeneias* as a denial of the universal validity of the *tertium non datur*.) The same doubts reappeared in the Middle Ages — together with groping attempts to construct a many-valued logic for coping with them. Within modern logic these efforts were renewed by Łukasiewicz. His grand vision of polyvalent logic as a generalization of classical logic did not turn out as fertile as its originator had imagined it to be. The idea of polyvalence has useful technical applications. But the conception of it as a grating of logical space finer than the true-false dichotomy encounters interpretational difficulties. It is therefore doubtful whether many-valued logic should even count as non-classical in the sense which I have in mind here.

A more consequential onslaught on the Law of Excluded Third and some other "classical" ideas associated with it, such as the Principle of Double Negation, came from Brouwer and the intuitionists. As already

noted, formalized intuitionist logic has turned out to be a useful conceptual tool for proof-theoretic study. It provides the logical backbone for a constructivist approach to the notion of existence in mathematics and is also helpful for efforts to clarify the concept of the actual infinite. To count with truth-value "gaps" has become standard in many fields of formal study where one deals with concepts of restricted definability or of an open texture. The Law of Excluded Middle can hardly any longer be regarded as a controversial topic in the philosophy of logic.

More firm and less assailed, until recently, has been the second pillar of classical logic, the Law of Contradiction, which prohibits truth-value "overlaps". Therefore, doubts about it, once they are raised, cut much deeper into the foundations of logic than doubts relating to the *tertium non datur*.

In fact, already Aristotle realized that there might be problems here. First among the moderns to see the possibility of a non-classical opening were Łukasiewicz and the Russian Vasiliev. [27]

Throughout the history of thought, antinomies have been a headache of philosophers — and since the origin of set-theory also of mathematicians. Antinomies exemplify seemingly impeccable logical inference terminating in conclusions contradicting each other. If this is thought unacceptable, one has to look for some error in the reasoning — and lay down rules for how to avoid the calamity. This was what Russell did with his Type-Theory and Vicious Circle Principle.

Moreover, the appearance of a contradiction in a context of reasoning, such as for example an axiomatic system, seems to have the vitiating consequence of making everything derivable within the system, thus trivializing or, as one also says, "exploding" it. Hilbert's efforts were partly aimed at proving that sound systems are immune to such disasters. This presupposed that the logic of the meta-proofs has the required immunity. Hilbert saw a warrant of this in what he called the *finite Einstellung* ("finitist stand"), allowing only *finite Schlussweisen*.

Another way of meeting the challenge presented by contradictions is to scrutinize the idea of logical consequence itself. Contradictions may have to be rejected as false, but must they have the catastrophic consequences which "classical" logic seems to allow by virtue of what is sometimes referred to as Duns Scotus' Law after the *doctor subtilis* of the School? Efforts to modify the classical view of logical consequence or entailment have been the motivating force behind the venture called Relevance Logic. A more recent and more radical step in the same direction is known as Paraconsistent Logic. One of its aims is to show how contradictions can be "accomodated" within contexts of reasoning without fear of trivialization

or collapse.

These non-classical developments in logic, of the past decades, have found an unexpected, but I think not very thrustworthy, ally in Dialectical Logic, ultimately of Hegelian inspiration. The best one can hope for is that the treatment of dialectics with the formal tools of paraconsistent and related "deviant" logics will contribute to a demystification of those features of it which have made it little palatable to rational understanding. A similar service which these new tools may render is that of reducing to its right proportions what Wittgenstein called "the superstitious dread and veneration by mathematicians in face of contradiction". [28]

Just as classical logic, *i.e.* the logic of Frege and Russell, can be called the sub-structure on which stand the several branches of modal and intensional logic — in a similar way the two main varieties of non-classical logic: the intuitionist-like ones which admit truth-value gaps and the paraconsistent-like ones which admit truth-value overlaps, will serve as sub-structures from which a variety of alternative epistemic, deontic and other logics will grow out and be further cultivated. But these developments are still in early infancy.

12. I have tried to review the development of logic in this century as a gradual progress *from* the philosophic fascination of a foundation crisis in mathematics and the confusions excited by the rediscovery of fields of study long lying fallow *to* increased clarity, exactness, and conceptual sobriety. But logic thus transformed ceases to be philosophy and becomes science. It either melts into one of the old sciences or contributes to the formation of a new one. What happened to logic was that it fused with the multifarious study of mathematics, but also with newcomers on the scientific stage such as computer science and cognative study, cybernetics and information theory, general linguistics — all being fields with a strong mathematical slant.

Transformations of parts of philosophy into independent branches of scientific study are well known from history. The phenomenon has gained for philosophy the name "mother of the sciences". Physics was born of natural philosophy; in some English and Scottish universities it still bears that name. The second half of the 19th century witnessed the birth of psychology and sociology through a transformation of predominantly speculative thinking into experimental and empirical research. In our century something similar happened with logic. [29]

Already in the early days of the developments which we have here been following, Russell wrote: "Mathematical logic - - - is not *directly* of philosophical importance except in its beginnings. After the beginnings, it

belongs rather to mathematics than to philosophy." [30] And in an unpublished typescript of Wittgenstein's we read: "Die formale Logik — ein Teil der Mathematik." [31]

Philosophy, I would say, thrives in the twilight of unclarity, confusion, and crisis in fields which in their "normal" state do not bewilder those who cultivate them or cause excitement in their intellectual surroundings. From time to time, however, philosophic storms will occur even in the seemingly calmest of waters. We can be certain that there will always remain obscure corners in logic too, thus assuring for it a permanent place among the concerns of philosophers. And I can well imagine that individual thinkers will find in logic the raw material for bold metaphysical constructions. As an example might be cited Gödel's conceptual realism with echos of Plato and Leibniz. But it seems to me unlikely that logic will continue to play the prominent role in the overall picture of an epoch's philosophy which it has held in the century now approaching its end. This will be so partly because of logic's own success in integrating itself into the neighbouring sciences just mentioned. But it will also be due to the rise on the philosophical horizon of new clouds calling for the philosophers' attention and craving for clarification.

Big shifts in the centre of philosophy signalize changes in the general cultural atmosphere which in their turn reflect changes in political, economic and social conditions. The optimistic mood and belief in progress, fostered by scientific and technological developments, which has been our inheritance from the time of the Enlightenment, is giving way to a sombre mood of self-critical scrutiny of the achievements and foundation of our civilization. No attempt to survey the overall situation in contemporary philosophy can fail to notice this and to ponder over its significance.

I shall not try to predict what will be the leading trends in the philosophy of the first century of the 2000s. But I think they will be markedly different from what they have been in this century, and that logic will *not* be one of them. If I am right, the twentieth century will even clearer than now stand out as another Golden Age of Logic in the history of those protean forms of human spirituality we call Philosophy.

NOTES

1. KANT, *Kritik der reinen Vernuft*, p. 7 (Pagination of the second edition, 1787.)
2. Most notably with FRANCIS BACON'S *Novum Organum* (1620); later also with LAMBERT'S *Neues Organon* (1764); and once again with WILLIAM WHEWELL'S *Novum Organum Renovatum* (1858).

3. Thus, for example, by BOLZANO whose *Wissenschaftslehre* (1837) is one of the early precursors of logic in its modern form.

4. KANT, *op.cit.*, p. 78.

5. HEINRICH SCHOLZ, *Geschichte der Logik*, p. 12. Junker und Dünnhaupt, Berlin 1931.

6. HEGEL, *Wissenschaft der Logik*, Teil I, p. 36: "Allein - - - sind überhaupt die Vorstellungen, auf denen der Begriff der Logik bisher beruhte, teils bereits untergegangen, teils ist es Zeit, dass sie vollends verschwinden, dass der Standpunkt dieser Wissenschaft höher gefasst werde und dass sie eine völlig veränderte Gestalt gewinne." (Quoted from *Werkausgabe*, Suhrkamp Verlag, Frankfurt am Main 1969.)

7. See the article "Logistique" in LALANDE'S *Vocabulaire technique et critique de la philosophie.*

8. WHITEHEAD, in his Foreword to QUINE's early work *A System of Logistic* (1934), wrote: "In the modern development of Logic, the traditional Aristolelian Logic takes its place as a simplification of the problem presented by the subject. In this there is an analogy to arithmetic of primitive tribes compared to modern mathematics."

9. Cf. comments on the term "logistic" in C. I. LEWIS, *A Survey of Symbolic Logic* (1918), p. 3ff. Dover Publications, New York 1960.

10. SCHOLZ, *op.cit.*, p. 14. KANT, *op.cit.*, p. 76ff.

11. The difference is interestingly reflected in the titles of the works with which they embarked on their respective tasks. BOOLE'S was called *The Mathematical Analysis of Logic, Being an Essay towards a Calculus of Deductive Reasoning.* FREGE'S pioneering work had the title *Begriffsschrift, eine der arithmetischen nachgebildete Formelsprache des reinen Denkens.*

12. A contemporary account of the state of foundation research in mathematics, still very worth reading is A. HEYTING'S *Mathematische Grundlagenforschung, Intuitionismus, Beweistheorie*, Julius Springer, Berlin 1934.

13. The term presumably first used by MARTIN GRABMANN in his renowned work *Geschichte der scholastischen Methode* I–II, Freiburg i B. 1909–1911.

14. The phrase borrowed from the title of RICHARD RORTY'S book *The Linguistic Turn*, Chicago 1967. Rorty attributes the invention of the phrase to Gustav Bergmann.

15. I would conjecture, however, that WITTGENSTEIN'S notion of "language game" and his ideas from the early 1930s of language as calculus have a remote source of inspiration in the influence of Hilbertian formalism on the discussions about logic and the philosophy of mathematics among members of the Vienna Circle. Cf. *Ludwig Wittgenstein und der Wiener Kreis. Gespräche, aufgezeichnet von Friedrich Waismann.* Aus dem Nachlass herausgegeben von B.F. McGuinness. In *Ludwig Wittgenstein, Schriften 3*, Suhrkamp Verlag, Frankfurt am Main 1967.

16. The phrase was invented by the late Professor MAX BLACK. See his *A Companion to Wittgenstein's Tractatus*, p. 11. Cambridge University Press, Cambridge 1964.

17. WITTGENSTEIN, *Philosophische Untersuchungen* (1953), $ 107: "Die Kristallenreinheit der Logik hatte sich mir ja nicht *ergeben*, sondern sie war eine *Forderung*."

18. See R. M. MCDONOUGH, *The Argument of the "Tractatus". Its Relevance to Contemporary Theories of Logic, Language, Mind, and Philosophical Truth.* State University of New York Press, Albany, N. Y. 1986. Particularly pp. 172–183.

19. The best attempt known to me of such critical evaluation is that of NORMAN MALCOLM. See in particular his book *Nothing is Hidden, Wittgenstein's Criticism of His Early Thought*, Basil Blackwell, Oxford 1986.

20. RUSSELL, *Our Knowledge of the External World, As a Field for Scientific Method in Philosophy* (1914), p. 42. Quoted from the edition by Allen & Unwin, London 1949.

21. RUSSELL, *op.cit.*, p. 14.

22. On this turn and its repercussions on foundation research in mathematics, see the excellent account by ANDRZEJ MOSTOWSKI, *Thirty Years of Foundation Studies, Lectures on the Development of Mathematical Logic and the Study of the Foundations of Mathematics in 1930–1964*. Basil Blackwell, Oxford 1966.

23. TARSKI, *"Der Wahrheitsbegriff in den formalisierten Sprachen", Studia Philosophica I*, 1935. Postscript (Nachwort), p. 404f.

24. Cf. HEYTING, *op.cit.*, p. 53f. Also D. VAN DALEN, "The War of the Frogs and the Mice, or the Crisis of the *Mathematische Annalen*", The Mathematical Intelligence 12, 1990.

25. HAO WANG, *Reflections on Kurt Gödel*, p. 57 and p. 88. The MIT Press; Cambridge, Mass. 1987.

26. POINCARÉ, *Science et Méthode* (1909), p. 193f. The references are to the Edition Flamarion, Paris 1924. Cf. also RUSSELL, *op.cit.*, p. 68.

27. N. A. VASILIEV, *Voobražaemaja logika. Izbrannye Trudy*, Ed. by V. A. Smirnov. Nauka, Moscow 1989.

28. WITTGENSTEIN, *Remarks on the Foundations of Mathematics*, Third Edition, Basil Blackwell, Oxford 1978, p. 122. In German: "Die abergläubige Angst und Verehrung der Mathematiker vor dem Widerspruch."

29. In a well-known simile, John Langshaw Austin compared this process to philosophy perpetually being "kicked upstaris" — and he envisaged that the "linguistic turn" in philosophy will eventually also result in the birth of an independent descriptive study of conceptual features of linguistic uses, in a "linguistic phenomenology". J. L. AUSTIN, "Ifs and Cans", *Proceedings of the British Academy*, Vol. XLII, Oxford University Press, Oxford 1956.

30. RUSSELL, *op.cit.*, p. 50.

31. WITTGENSTEIN, TS 219. Wittgenstein's relegation of formal logic to mathematics is not in conflict with the fact that he calls his own investigations in philosophy "logical". The adjective then means roughly the same as *conceptual* or, in Wittgenstein's somewhat excentric terminology, *grammatical*.

Logic, Methodology and Philosophy of Science IX
D. Prawitz, B. Skyrms and D. Westerståhl (Editors)

THE WITNESS FUNCTION METHOD AND PROVABLY RECURSIVE FUNCTIONS OF PEANO ARITHMETIC

SAMUEL R. BUSS*

Department of Mathematics
University of California, San Diego, La Jolla, CA 92093-0112, USA
sbuss@ucsd.edu

1. Introduction

The witness function method has been used with great success to characterize some classes of the provably total functions of various fragments of bounded arithmetic [2, 4, 18, 23, 17, 16, 5, 6, 1, 7, 8]. In this paper, it is shown that the witness function method can be applied to the fragments $I\Sigma_n$ of Peano arithmetic to characterize the functions which are provably recursive in these fragments. This characterization of provably recursive functions has already been performed by a variety of methods; including: via Gentzen's assignment of ordinals to proofs [9, 27], with the Gödel Dialectica interpretation [12, 13], and by model-theoretic methods (see [20, 15, 26]). The advantage of the methods in this paper is, firstly, that they provide a simple, elegant and purely proof-theoretic method of characterizing the provably total functions of $I\Sigma_n$ and, secondly, that they unify the proof methods used for fragments of Peano arithmetic and for bounded arithmetic.

The witness function method is related to the classical proof-theoretic methods of Kleene's recursive realizability, Gödel's Dialectica interpretation and the Kreisel no-counterexample interpretation; however, the witness function method does not require the use of functionals of higher type. We feel that the witness function method provides an advantage over the other methods in that it leads to a more direct and intuitive understanding of many formal systems. The classical methods are somewhat more general but are also more cumbersome and more difficult to understand (consider the difficulty of comprehending the Dialectica interpretation or no-counterexample interpretation of a formula with more than three alternations of quantifiers,

*Supported in part by NSF grants DMS-8902480 and INT-8914569.

for instance). On the other hand, the more direct and intuitive witness function method has been extremely valuable for the understanding of *why* the provably total functions of a theory are what they are and also for the formulation of new theories for desired classes of computational complexity and, conversely, for the formulation of conjectures about the provably total functions of extant theories. The main support for our favorable opinion of the witness function method is, firstly, its successes for bounded arithmetic and, secondly, the results of this paper showing its applicability to Peano arithmetic.

While checking references for this paper, the author read Mints [19] for the first time — it turns out that Mints's proof that the provably recursive functions of $I\Sigma_1$ are precisely the primitive recursive functions is based on what is essentially the witness function method. This theorem of Mints is, in essence, Theorem 9 below. Mints's use of the witness function method predates its independent development by this author for applications to bounded arithmetic. The present paper expands the applicability of the witness function method to all of Peano arithmetic.

The outline of this paper is as follows: section 2 develops the necessary background material on Peano arithmetic, the subtheories $I\Sigma_n$, transfinite induction axioms, least ordinal principle axioms, the sequent calculus and the correct notion of free-cut free proof for transfinite induction/least number principle axioms. In section 3, the central notions of the witness function method and witness oracles are developed and the Σ_n-definable functions of $I\Sigma_n$ and $I\Delta_0 + TI(\omega_m, \Pi_n)$ are characterized. This includes the definition of α-primitive recursive (in Σ_k) functions and normal forms for such functions. Then the provably recursive (i.e., Σ_1-defined) functions of $I\Sigma_n$ are characterized by proving a conservation theorem for $TI(\omega_m, \Pi_n)$ over $TI(\omega_{m+1}, \Pi_{n-1})$. Section 4 outlines a proof of Parsons's theorem on the conservativity of the Π_{n+1}-induction rule over the Σ_n-induction axiom. Section 5 contains a proof of the Π_{n+1}-conservativity of $B\Sigma_{n+1}$ over $I\Sigma_n$. Section 6 concludes with a discussion of the analogies between the methods of this paper and the methods used for bounded arithmetic.

2. Preliminaries

2.1. Arithmetic and ordinals

Peano arithmetic (PA) is formulated[2] in the language 0, S, $+$, $\cdot$ and $\leq$.

[2] Our formulation of PA is similar to the usual one in [21] except that it has different non-induction axioms and has $\leq$ instead of $<$. It is easily seen that our definition of $I\Sigma_n$ and PA is equivalent to the usual one apart from the inessential replacement of $<$ by $\leq$.

It has induction axioms

$$A(0) \wedge (\forall x)(A(x) \rightarrow A(S(x))) \rightarrow (\forall x)A(x)$$

for all formulas A, plus it has a finite base set of axioms, namely, Robinson's theory Q of seven axioms defining 0, S, $+$ and $\cdot$ and, in addition, the axiom

$$(\forall x)(\forall y)(x \leq y \leftrightarrow (\exists z)(x + z = y))$$

which defines $\leq$. A *bounded quantifier* is of the form $(\exists x \leq t)$ or $(\forall x \leq t)$ where t is any term not involving x. The usual quantifiers, $(\forall x)$ and $(\exists x)$, are called *unbounded quantifiers*. The Δ_0-formulas, or *bounded formulas*, are the formulas in which every quantifier is bounded. The classes Σ_n and Π_n of formulas are defined by induction on n, so that $\Sigma_0 = \Pi_0 = \Delta_0$ and so that Σ_{n+1} is the set of formulas of the form $(\exists \vec{x})B$ where $B \in \Pi_n$ and so that Π_{n+1} is defined dually. The theory $I\Sigma_n$ is defined to be the theory in the language of Peano arithmetic with the same eight non-induction axioms as PA and with induction axioms for all formulas $A \in \Sigma_n$.

The collection axioms provide an alternative way to define fragments of Peano arithmetic. A collection axiom is of the form

$$(\forall x \leq t)(\exists y)A(x, y) \rightarrow (\exists z)(\forall x \leq t)(\exists y \leq z)A(x, y).$$

We let $B\Sigma_n$ denote the set of collection axioms for all $A \in \Sigma_n$; $B\Pi_n$ is defined similarly. It is well-known that $I\Delta_0 + B\Sigma_{n+1} \vdash I\Sigma_n$ and $I\Sigma_n \vdash B\Sigma_n$. It is also well-known that $I\Delta_0 + B\Sigma_{n+1}$ is Π_{n+1}-conservative over $I\Sigma_n$ and we shall reprove this in section 5 below. An important feature of the collection axioms is that it provides a 'quantifier exchange' principle that allows moving bounded quantifiers inside the scope of unbounded quantifiers. The classes Σ_n and Π_n can be generalized to classes Σ_n^G and Π_n^G by allowing bounded quantifiers to appear anywhere in the formula (instead of only in the Δ_0 matrix) but counting only the alternations of unbounded quantifiers. For example, the hypothesis and conclusion of the collection axiom above are Σ_n^G-formulas if $A \in \Sigma_n$. The theory $I\Delta_0 + B\Sigma_n$, and hence $I\Sigma_n$, can prove that every Σ_n^G-formula is equivalent to a Σ_n-formula.

Remark: Some authors include function symbols for all primitive recursive functions in the language of PA. We do not adopt this convention; however, as is well-known, every primitive recursive function is provably recursive (Σ_1-definable, see below) in $I\Sigma_1$ and hence the theories $I\Sigma_n$, for $n \geq 1$ are not significantly affected by the addition of symbols for primitive recursive functions. Thus the theorems and proofs of this paper also apply to theories with symbols for primitive recursive functions.

Definition Let T be a subtheory of PA and $f : \mathbb{N}^k \to \mathbb{N}$. The function f is Σ_i-definable in T iff there is a formula $A(x_1, \ldots, x_k, y) \in \Sigma_i$ such that

(1) $T \vdash (\forall \vec{x})(\exists! y) A(\vec{x}, y)$, and

(2) $\{(\vec{n}, m) : \mathbb{N} \models A(\vec{n}, m)\}$ is the graph of f, i.e., $A(\vec{n}, m)$ holds iff $f(\vec{n}) = m$ for all integers $\vec{n}, m$.

The function f is *provably recursive in* T iff f is Σ_1-definable in T.

The intuitive idea of 'provably recursive' is that the theory T should prove that some Turing machine M, which computes f, halts on all appropriate inputs. Since $A(\vec{x}, y)$ can be taken to be a Σ_1-formula expressing "there is a w which codes a halting M-computation with input $\vec{x}$ and output y", it is clear that any function which is provably recursive in this intuitive sense is also Σ_1-definable. Conversely, if f is Σ_1-definable in T, then there is Turing machine M which computes f, provably in T. Namely, M performs a brute-force search for values of y and the unboundedly existentially quantified variables of A. Thus 'Σ_1-definable' coincides with the intuitive notion of 'provably recursive'.

One reason that the provably recursive functions of T are of particular significance is that if f is provably recursive in T, then T may be conservatively extended by adding f as a new function symbol with $f(\vec{x}) = y \leftrightarrow A(\vec{x}, y)$ as a new axiom. If T is a fragment $I\Sigma_n$ then f may be used freely in induction formulas (without affecting quantifier complexity). Similarly, if T can prove that a Π_1-formula and a Σ_1-formula are equivalent then T can be conservatively extended by adding a new predicate symbol with arguments including the free variables of the two formulas and adding a new axiom defining the predicate symbol to be equivalent to the formulas. The new predicate may also be used freely in induction formulas. Such new predicates are called Δ_1*-defined predicates of* T.

Recall that $I\Sigma_1$ (and even $I\Delta_0$) can formalize many metamathematical notions; of particular importance are the sequence coding functions $\langle x_0, \ldots, x_k \rangle$, $(\langle x_0, \ldots, x_k \rangle)_i = x_i$, and $Len(\langle x_0, \ldots, x_k \rangle) = k + 1$.

The ordinals are set-theoretically defined to be those sets which are transitive and well-founded by α. We write $\prec$ for the ordering of ordinals, so $\alpha \prec \beta$ means $\alpha \in \beta$. It is well-known how to define ordinal addition, multiplication and exponentiation. The Cantor normal form for an ordinal α is the unique expression

$$\alpha \quad = \quad \omega^{\gamma_1} \cdot n_1 + \omega^{\gamma_2} \cdot n_2 + \cdots \omega^{\gamma_r} \cdot n_r$$

where $\gamma_1 \succ \gamma_2 \succ \cdots \succ \gamma_r$ are ordinals and $n_1, \ldots, n_r$ are positive integers (i.e., nonzero, finite ordinals). Here ω is the first infinite ordinal; we let

$\omega_0 = 1$, $\omega_1 = \omega$ and, generally, $\omega_{n+1} = \omega^{\omega_n}$. Thus ω_n is a stack of n ω's. The limit of ω_n as $n \to \omega$ is called ϵ_0; hence ϵ_0 is the least ordinal such that $\epsilon_0 = \omega^{\epsilon_0}$. For $\alpha \prec \epsilon_0$, the Cantor normal form can be extended so that the exponents γ_i are also written in Cantor normal form, and with exponents in the latter Cantor normal forms also in Cantor normal form, etc. (eventually the process must stop). For example,

$$\omega^{\omega^{\omega^0 \cdot 3} + \omega^{\omega^0 \cdot 2}} \cdot 4 + \omega^0$$

is a Cantor normal form; usually this is expressed more succinctly as $\omega^{\omega^3 + \omega^2} \cdot 4 + 1$. In this paper, we shall always use ordinals $\preccurlyeq \epsilon_0$ and by Cantor normal form always means the extended version with exponents also in Cantor normal form. ϵ_0 is its own Cantor normal form.

By using Gödel numbering, integers can encode Cantor normal forms and this can be intensionally formalized[3] in $I\Sigma_1$; with care, these can even be formalized in $I\Delta_0$. In particular, $I\Delta_0$ can define the relation *IsOrdinal*(x) expressing that x is the Gödel number of an ordinal, the relation $x \prec y$, and the functions for ordinal addition, multiplication and exponentiation. To avoid excessive notation, we use the same notation for actual and for metamathematical operations; for example, $\omega + 1$ also denotes its own Gödel number. However, there will occasionally be situations where context is not sufficient to distinguish between ordinals and their Gödel numbers: this occurs when n may be either an integer or a finite ordinal; to resolve ambiguity, we write $\ulcorner n \urcorner$ for the Gödel number of the ordinal n and we write n for the integer n. To improve readability, we use $\alpha, \beta, \gamma, \ldots$ and $\rho, \sigma, \tau, \ldots$ as variables that range over Gödel numbers of ordinals. For example, the formula $(\forall \sigma \prec \beta)(\cdots)$ abbreviates the first-order formula

$$\mathit{IsOrdinal}(\beta) \wedge (\forall x)(\mathit{IsOrdinal}(x) \wedge x \prec \beta \to \cdots).$$

Note that $\forall \sigma \prec \beta$ corresponds to an *unbounded* quantifier unless β is known to code a finite ordinal.

Transfinite induction on ordinals may be used to provide alternate axiomatizations for fragments of Peano arithmetic:

Definition Let Ψ be a set of formulas and let $\kappa \preccurlyeq \epsilon_0$. Then $TI(\kappa, \Psi)$ is the set of axioms

$$(\forall \gamma \preccurlyeq \kappa)[(\forall \beta \prec \gamma) A(\beta) \to A(\gamma)] \to A(\kappa) \tag{1}$$

where A is a formula in Ψ, possibly with other free variables as parameters.

[3] 'Intensionally formalized' means that $I\Sigma_1$ can prove simple syntactic facts about ordinal encodings and about operations on encoded ordinals.

The *least ordinal principle* axioms $LOP(\kappa, \Psi)$ are

$$A(\kappa) \rightarrow (\exists \gamma \preccurlyeq \kappa)[A(\gamma) \wedge (\forall \beta \prec \gamma)(\neg A(\beta))] \tag{2}$$

where $A \in \Psi$ and A may have parameter variables. For a fixed formula A, the axioms (1) and (2) are called $TI(\kappa, A)$ and $LOP(\kappa, A)$, respectively.

$TI(\prec \kappa, \Psi)$ is the theory $\cup_{\mu \prec \kappa} TI(\mu, \Psi)$.

$LOP(\prec \kappa, \Psi)$ is the theory $\cup_{\mu \prec \kappa} LOP(\mu, \Psi)$.

A slight variation on the least ordinal principle and transfinite induction axioms is

$$TI^*(\kappa, \Psi): \quad (\forall \gamma \preccurlyeq \kappa)[(\forall \beta \preccurlyeq \gamma) A(\beta) \rightarrow A(\gamma)] \rightarrow (\forall \gamma \preccurlyeq \kappa) A(\gamma)$$

$$LOP^*(\kappa, \Psi): \quad (\exists \gamma \preccurlyeq \kappa) A(\gamma) \rightarrow (\exists \gamma \preccurlyeq \kappa)[A(\gamma) \wedge (\forall \beta \prec \gamma)(\neg A(\beta))].$$

For Ψ one of the classes Σ_n or Π_n, $TI^*(\kappa, \Psi)$ is equivalent to $TI(\kappa, \Psi)$ since the former obviously implies the latter and since $TI^*(\kappa, A)$ may be inferred from $TI(\kappa, B)$ where $B(\alpha)$ is $A(\alpha) \vee (\alpha \succ \gamma' \wedge A(\gamma'))$, where γ' is a new variable acting as a parameter. Similarly, LOP^* and LOP are equivalent for Ψ one of the classes Σ_n or Π_n.

This paper is concerned primarily with the axioms $TI(\prec \omega_m, \Sigma_n)$ and $LOP(\prec \omega_m, \Sigma_n)$ where $m \geq 2$ and $n \geq 0$. The next two propositions give equivalences among such axioms (see [26] for generalizations of these propositions).

PROPOSITION 1 *Let $m \geq 2$ and $n \geq 0$.*

(a) $I\Delta_0 + TI(\prec \omega_m, \Sigma_n) \equiv I\Delta_0 + LOP(\prec \omega_m, \Pi_n)$

(b) $I\Delta_0 + TI(\prec \omega_m, \Pi_n) \equiv I\Delta_0 + LOP(\prec \omega_m, \Sigma_n)$

(c) $I\Delta_0 + LOP(\prec \omega_m, \Pi_n) \equiv I\Delta_0 + LOP(\prec \omega_m, \Sigma_{n+1})$

(d) $I\Delta_0 + TI(\prec \omega_m, \Sigma_n) \equiv I\Delta_0 + TI(\prec \omega_m, \Pi_{n+1})$

Proof (a) and (b) are trivial since $TI(\kappa, A)$ and $LOP(\kappa, \neg A)$ are logically equivalent (essentially, contrapositives). For (c), if $A \in \Sigma_{n+1}$ then $A(\rho)$ must be $(\exists \vec{y}) B(\rho, \vec{y})$ where $B \in \Pi_n$. Now, $LOP(\kappa, A)$ follows from $LOP^*(\omega \cdot \kappa + \omega, C)$ where $C(\rho)$ is the Π_n-formula expressing

> "ρ encodes an ordinal $\omega \cdot \kappa + \langle \vec{y} \rangle$, with $\vec{y}$ integers, such that $B(\kappa, \vec{y})$ holds."

Also, if $\kappa \prec \omega_m$, then $\omega \cdot \kappa + \omega \prec \omega_m$; so (c) is proved. Finally, (d) follows immediately from (a), (b) and (c). □

It is important to note that Proposition 1 holds for $n = 0$; it is easy to see that the proof of (c) is valid for $n = 0$ since C is a Δ_0-formula if B is. This has as consequence that $I\Delta_0 + TI(\prec \omega_m, \Sigma_0)$ is equivalent to $I\Delta_0 + TI(\prec \omega_m, \Pi_1)$ and since $I\Delta_0$ can express every primitive recursive predicate as a Π_1 formula, it follows that transfinite $(\prec \omega_m)$ induction on Δ_0-formulas implies the same amount of transfinite induction on primitive recursive predicates. In addition, relative to $I\Delta_0$, $TI(\prec \omega_m, \Sigma_0)$ is equivalent to $LOP(\prec \omega_m, \Sigma_0)$, which in turn is equivalent to $LOP(\prec \omega_m, \Sigma_1)$. Since every primitive recursive predicate can be expressed as a Σ_1-formula, it follows that transfinite $(\prec \omega_m)$ induction on Δ_0-formulas implies the $\prec \omega_m$ least ordinal principle for primitive recursive predicates. We shall, in section 3, frequently informally argue that various complicated metamathematical constructions can be formalized in theories $I\Delta_0 + TI(\prec \omega_m, \Sigma_{n-1})$; since $m \geq 2$ always holds, these theories can prove the usual induction and least number principles for primitive recursive predicates, which is sufficient for formalizing all the metamathematical constructions in section 3.

PROPOSITION 2 *Let* $n \geq 1$.

$$\begin{aligned} I\Sigma_n \equiv I\Pi_n \quad &\equiv \quad I\Delta_0 + TI(\omega, \Sigma_n) \equiv I\Delta_0 + TI(\omega, \Pi_n) \\ &\equiv \quad I\Delta_0 + LOP(\prec \omega_2, \Sigma_n) \\ &\equiv \quad I\Delta_0 + TI(\prec \omega_2, \Sigma_{n-1}) \end{aligned}$$

Proof It is clear that $I\Sigma_n \equiv I\Delta_0 + TI(\omega, \Sigma_n)$ and by standard techniques these are equivalent to $I\Pi_n$ and $I\Delta_0 + TI(\omega, \Pi_n)$. In light of Proposition 1, it suffices to show that $LOP(\prec \omega_2, \Sigma_n)$ follows from $I\Delta_0 + TI(\omega, \Pi_n)$. To accomplish this, we show, by induction on k, that $LOP(\prec \omega^k, \Sigma_n)$ follows from the latter theory. For $k = 1$ this is proved by the kind of reasoning used to prove Proposition 1(a),(b). To show $LOP(\prec \omega^{k+1}, \Sigma_n)$; let $A(\alpha) \in \Sigma_n$, let $\alpha_0 \prec \omega^{k+1}$ and reason informally with the assumptions $TI(\omega, \Pi_n)$ and $LOP(\prec \omega^k, \Sigma_n)$: further set $C(\alpha)$ to be the formula $(\exists i)A(\omega \cdot \alpha + i)$, so $C(\alpha) \in \Sigma_n$. Now assume $A(\alpha_0)$ holds; since $\alpha_0 = \omega \cdot \alpha_1 + i_1$ for some $\alpha_1 \prec \omega^k$ and some finite i_1, $C(\alpha_1)$ holds also. By $LOP(\prec \omega^k, \Sigma_n)$, there is a least α_2 such that $C(\alpha_2)$ holds and now by $TI(\omega, \Pi_n)$, there is a least i_2 such that $A(\omega \cdot \alpha_2 + i_2)$. Clearly $\alpha = \omega \cdot \alpha_2 + i_2$ is the least ordinal such that $A(\alpha)$ holds. □

2.2. Arithmetic and the sequent calculus

This section describes how the sequent calculus and free cut elimination are applied to the fragments of arithmetic defined above. The reader is presumed to be familiar with the sequent calculus (refer to [27] or Chapter 4 of [2] for the

necessary background material). We shall assume the language of first-order logic contains symbols $\neg$, $\wedge$, $\vee$, $\rightarrow$, $\exists$ and $\forall$; this leads to a large number of rules of inference but we shall omit most cases from our proofs in any event. It will be assumed that bounded quantifiers are part of the syntax of first-order logic with the sequent calculus containing the four appropriate rules of inference for bounded quantifiers.[4] See [2] for the full definition of the sequent calculus LKB with bounded quantifier rules of inference.

To formalize the proof theory of arithmetic with the sequent calculus, it is customary to use special induction inferences in place of induction axioms. An *induction inference* is of the form

$$\frac{\Gamma, A(a) \longrightarrow A(Sa), \Delta}{\Gamma, A(0) \longrightarrow A(t), \Delta}$$

where t may be any term, a is a free variable called the *eigenvariable* and a must not appear in the lower sequent. The induction inference for A is equivalent to the induction axiom for A, because the side formulas Γ and Δ are allowed. Thus $I\Sigma_k$ is formalized in the sequent calculus with a finite set of axiom schemes plus the induction inferences for Σ_k formulas. The finite set of axiom schemes for $I\Sigma_k$ consists of the following initial sequents:

$$\begin{array}{ll} Sr = St \longrightarrow r = t & \longrightarrow r \cdot 0 = 0 \\ St = 0 \longrightarrow & \longrightarrow r \cdot (St) = r \cdot t + r \\ \longrightarrow r + 0 = r & \longrightarrow r = 0, (\exists x \leq r)(Sx = r) \\ \longrightarrow r + St = S(r + t) & r \leq t \longrightarrow (\exists x \leq t)(r + x = t) \\ & r + s = t \longrightarrow r \leq t \end{array}$$

where r, s and t are allowed to be any terms. Of course the usual logical initial sequents $A \longrightarrow A$ with A atomic and the initial sequents for equality are also allowed. It is important for us that every initial sequent consists of only Δ_0 formulas.

The theory $I\Delta_0 + TI(\prec \omega_m, \Sigma_n)$ is formalized in the sequent calculus with the same initial sequents, with induction inferences for Δ_0-formulas and for transfinite induction, with the $LOP(\prec \omega_m, \Pi_n)$ inferences defined below.

Let τ be a **closed** term with value the Gödel number of an ordinal and let $B(\alpha)$ be a formula; the $LOP(\tau, B)$ inference is

$$LOP(\tau, B): \qquad \frac{\alpha \preccurlyeq \tau, B(\alpha), \Gamma \longrightarrow \Delta, (\exists \beta \prec \alpha) B(\beta)}{B(\tau), \Gamma \longrightarrow \Delta}$$

[4] This assumption is not absolutely necessary and the reader may prefer to think of the bounded quantifiers as abbreviations — in this case the proofs by induction on the number of inferences in a free-cut free proof must be slightly modified.

where α is an eigenvariable and may occur only as indicated. It is not hard to see that the inference rule $LOP(\tau, B)$ is equivalent to the axiom form of $LOP(\tau, B)$: to derive the inference rule from the axiom, recall that the axiom $LOP(\tau, B)$ is

$$B(\tau) \longrightarrow (\exists \alpha \preccurlyeq \tau)[B(\alpha) \wedge (\forall \beta \prec \alpha)(\neg B(\beta))], \tag{3}$$

and use the derivation

$$\dfrac{\dfrac{\alpha \preccurlyeq \tau, B(\alpha), \Gamma \longrightarrow \Delta, (\exists \beta \prec \alpha) B(\beta)}{(\exists \alpha \preccurlyeq \tau)(B(\alpha) \wedge (\forall \beta \prec \alpha)(\neg B(\beta))), \Gamma \longrightarrow \Delta}\,(3)}{B(\tau), \Gamma \longrightarrow \Delta}$$

where the double horizontal line indicates omitted inferences. Conversely, to see that the $LOP(\tau, B)$ follows from the inference rule, use

$$\dfrac{\alpha \preccurlyeq \tau, B(\alpha) \longrightarrow (\exists \gamma \preccurlyeq \tau)[B(\gamma) \wedge (\forall \beta \prec \gamma)(\neg B(\beta))], (\exists \beta \prec \alpha) B(\beta)}{B(\tau) \longrightarrow (\exists \gamma \preccurlyeq \tau)[B(\gamma) \wedge (\forall \beta \prec \gamma)(\neg B(\beta))]}$$

where the upper sequent is, of course, provable in $I\Delta_0$.

The $LOP(\prec \omega_m, \Psi)$ inferences are the set of inferences $LOP(\tau, B)$ for $\tau \prec \omega_m$ and $B \in \Psi$. The *principal* formula of an LOP inference is the formula $B(\tau)$ in the lower sequent; the *auxiliary* formulas are the three formulas in the upper sequent other than Γ and Δ. An important property of the $LOP(\prec \omega_m, \Pi_{n-1})$ inferences is that the principal formula and the auxiliary formulas are all in Σ_n.[5]

Below we shall extensively study the theory $I\Delta_0 + TI(\prec \omega_m, \Sigma_{n-1})$, which is equivalent to $I\Delta_0 + LOP(\prec \omega_m, \Pi_{n-1})$ and henceforth is to be formalized in the sequent calculus with initial sequents given above, the $I\Delta_0$-induction rule and the $LOP(\prec \omega_m, \Pi_{n-1})$ inference rule. This theory enjoys the important property of *free-cut elimination*. We say that a cut in a sequent calculus proof is *free* unless one of its cut formulas is a direct descendent of a formula in an axiom (initial sequent) or of a principal formula of an $I\Delta_0$ inference or of a principal formula of an $LOP(\prec \omega_m, \Pi_{n-1})$ inference. The free-cut elimination theorem implies that if $I\Delta_0 + LOP(\prec \omega_m, \Pi_{n-1})$ proves a sequent then there is a proof (in the same theory and of the same sequent) which contains no free cuts. Such a proof is called *free-cut free*.

[5] This is the reason we use LOP inferences instead of TI inferences. The $TI(\tau, \Sigma_{n-1})$ inferences would be

$$\dfrac{\alpha \preccurlyeq \tau, (\forall \beta \prec \alpha) B(\beta), \Gamma \longrightarrow \Delta, B(\alpha)}{\Gamma \longrightarrow \Delta, B(\tau)}$$

where $B \in \Sigma_{n-1}$ and α is an eigenvariable. These inferences contain a Π_n auxiliary formula.

This free-cut elimination theorem is proved by a elementary triple induction argument (equivalently, induction to ω^3) by the same argument used for the cut elimination theorem for first-order logic. In particular, the free-cut elimination theorem can be proved in $I\Sigma_1$.

A formula A is a *subformula of B in the wide sense* if A can be obtained from some subformula C of B by substituting freely terms for variables in C. In a free-cut free proof, each formula A is either (1) a direct descendent of a formula in an axiom or of a principal formula of an $I\Delta_0$ or LOP inference, or (2) a subformula in the wide sense of such a formula, or (3) a subformula in the wide sense of an auxiliary formula of an $I\Delta_0$ inference or an LOP inference, or (4) a subformula in the wide sense of a formula in the endsequent of the proof. This is because each formula in the proof has a (not necessarily direct) descendent which is a cut formula (so (1) or (2) applies), or which is an auxiliary formula of an induction or LOP inference (so (3) applies), or which is in the endsequent (so (4) applies).

The above gives the following important proposition:

PROPOSITION 3 *($n \geq 1$) Let T be a theory $I\Sigma_n$ or $I\Delta_0 + TI(\prec \omega_m, \Sigma_{n-1})$. Suppose $\Gamma \longrightarrow \Delta$ is a consequence of T and every formula in Γ and Δ is in Σ_n. Then there is a T-proof of $\Gamma \longrightarrow \Delta$ in which every formula is in Σ_n.*

3. Definable functions of $I\Sigma_n$

3.1. Witness functions and ordinal primitive recursion

A witness oracle for an existential property $(\exists x)A(x, \vec{z})$ is an oracle which when queried with values for $\vec{z}$ responds either with a value for x such that $A(x, \vec{z})$ or with the statement that there is no such value for x. If A is a decidable predicate then a witness oracle for A is clearly equivalent to an oracle for the function

$$U_{\exists x A}(\vec{z}) \quad = \quad \begin{cases} 1 + (\mu x)A(x, \vec{z}) & \text{if } (\exists x)A(x, \vec{z}) \\ 0 & \text{otherwise} \end{cases}$$

where $(\mu x)A(x, \vec{z})$ is the least value for x such that $A(x, \vec{z})$ holds. The advantage of viewing a witness oracle as a function is that it allows the definition of being primitive recursive relative to a witness oracle:

Definition Let $n \geq 1$. The set of functions which are *primitive recursive in Σ_n* is defined inductively by:

(1) The constant function 0, the successor function $S(x) = x + 1$, and the projection functions $\pi_i^n(x_1, \ldots, x_n) = x_i$ are primitive recursive in Σ_n.

(2) The set of functions primitive recursive in Σ_n is closed under composition.

(3) If $g : \mathbb{N}^k \to \mathbb{N}$ and $h : \mathbb{N}^{k+2} \to \mathbb{N}$ are primitive recursive in Σ_n then so is the function f defined by

$$\begin{aligned} f(0, \vec{z}) &= g(\vec{z}) \\ f(m+1, \vec{z}) &= h(m, \vec{z}, f(m, \vec{z})). \end{aligned}$$

(4) If $A(\vec{z})$ is a formula $(\exists x)B(x, \vec{z})$ where $B \in \Pi_{n-1}$ then U_A is primitive recursive in Σ_n.

The set of functions primitive recursive in Σ_0 is just the set of primitive recursive functions, and is defined, as usual, by (1), (2) and (3).

It is important for the definition of primitive recursive in Σ_n that the functions U_A are included instead of just the characteristic functions of A. For example, if $n = 1$, these two functions are Turing equivalent; however, for primitive recursive processes these are not equivalent since even if $(\exists x)B$ is guaranteed to be true and if B is primitive recursive, a primitive recursive process can not find a value for x making B true without knowing (at least implicitly) an upper bound on the least value for x.

A primitive recursive in Σ_n function may ask any (usual) query to a Π_n or a Σ_n oracle. This is because, for example, if $A(\vec{z}) \in \Sigma_n$, then A is equivalent to a formula $(\exists x)B$ where $B \in \Pi_{n-1}$ and a witness oracle $U_{(\exists x)B}$ can be used to determine if $A(\vec{z})$ is true.

Definition Let α be (the Gödel number of) an ordinal. The set of α-*primitive recursive functions* is defined inductively by the closure properties of (1), (2) and (3) above and by

(5) If $g : \mathbb{N}^k \to \mathbb{N}$, $h : \mathbb{N}^{k+1} \to \mathbb{N}$ and $\kappa : \mathbb{N}^k \to \mathbb{N}$ are α-primitive recursive then so is the function f defined by

$$f(\beta, \vec{z}) = \begin{cases} h(\beta, \vec{z}, f(\kappa(\beta, \vec{z}), \vec{z})) & \text{if } \kappa(\beta, \vec{z}) \prec \beta \preccurlyeq \alpha \\ g(\beta, \vec{z}) & \text{otherwise} \end{cases}$$

where $\kappa(\beta, \vec{z}) \prec \beta \preccurlyeq \alpha$ means that β and $\kappa(\beta, \vec{z})$ are the Gödel numbers of ordinals obeying the inequalities.

A function is said to be $\prec \alpha$-*primitive recursive* iff it is γ-primitive recursive for some $\gamma \prec \alpha$.

Combining the notions of witness oracles and ordinal primitive recursion gives:

Definition Let $n \geq 0$ and α be (the Gödel number of) an ordinal. The set of functions which are *α-primitive recursive in Σ_n* is defined inductively by the closure properties of (1)-(5) above (omitting (4) if $n = 0$).

A function is said to be *$\prec \alpha$-primitive recursive in Σ_n* iff it is γ-primitive recursive in Σ_n for some $\gamma \prec \alpha$.

It is well-known, and not too hard to show, that a function is primitive recursive in Σ_n iff it is ω-primitive recursive and iff it is $\prec \omega^\omega$-primitive recursive in Σ_n.

3.2. Normal forms for ordinal primitive recursive functions

This section presents three normal forms for the definitions of $\prec \omega_m$-primitive recursive functions. These are called the zeroth, first and second normal forms and will be helpful for the proofs of the characterization of provably total functions of various fragments of Peano arithmetic.

Recall that that the set of functions $\prec \omega_m$-primitive recursive in Σ_n is, by definition, the smallest set of functions satisfying the closure properties (1)-(5): the Zeroth Normal Form Theorem states that the closure (3) under primitive recursion may be dropped at the expense of adding more base functions.

THEOREM 4 (ZEROTH NORMAL FORM) *Let $m \geq 2$ and $n \geq 0$. The functions $\prec \omega_m$-primitive recursive in Σ_n can be inductively defined by*

(0.1) *Every primitive recursive function is $\prec \omega_m$-primitive recursive in Σ_n.*

(0.2) *The set of functions $\prec \omega_m$-primitive recursive in Σ_n is closed under composition.*

(0.3) *If $n \geq 1$ and $A(\vec{z})$ is $(\exists x)B(x, \vec{z})$ where $B \in \Pi_{n-1}$, then U_A is $\prec \omega_m$-primitive recursive in Σ_n.*

(0.4) *If $\kappa_0 \prec \omega_m$ and if $g : \mathbb{N}^k \to \mathbb{N}$, $h : \mathbb{N}^{k+1} \to \mathbb{N}$ and $\kappa : \mathbb{N}^k \to \mathbb{N}$ are $\prec \omega_m$-primitive recursive in Σ_n then so is the function f defined by*

$$f(\beta, \vec{z}) \quad = \quad \begin{cases} h(\beta, \vec{z}, f(\kappa(\beta, \vec{z}), \vec{z})) & \text{if } \kappa(\beta, \vec{z}) \prec \beta \preccurlyeq \kappa_0 \\ g(\beta, \vec{z}) & \text{otherwise.} \end{cases}$$

Proof The fact that $\prec \omega_m$-primitive recursive in Σ_n functions satisfy conditions (0,1)-(0.4) is obvious. The idea for the other direction is quite simple; namely, that ω-primitive recursion may be used to simulate ordinary

primitive recursion. For example, if f is defined by primitive recursion from g and h by

$$\begin{aligned} f(0,\vec{z}) &= g(\vec{z}) \\ f(m+1,\vec{z}) &= h(m,\vec{z},f(m,\vec{z})) \end{aligned}$$

then f can also be defined via ω-primitive recursion as follows. For $n \in \mathbb{N}$, let $\ulcorner n \urcorner$ denote the Gödel number of the finite ordinal n. Define

$$F(\alpha,\vec{z}) = \begin{cases} g(\vec{z}) & \text{if } \alpha = \ulcorner 0 \urcorner \\ H(\alpha,\vec{z},F(Pred(\alpha),\vec{z})) & \text{otherwise} \end{cases}$$

where

$$Pred(\alpha) = \begin{cases} \alpha - 1 & \text{if } \alpha \text{ is (the Gödel number of) a successor ordinal} \\ \alpha & \text{otherwise} \end{cases}$$

and

$$H(\alpha,\vec{z},w) = \begin{cases} h(m,\vec{z},w) & \text{if } \alpha = \ulcorner m+1 \urcorner \text{ with } m \in \mathbb{N} \\ \text{arbitrary} & \text{otherwise.} \end{cases}$$

Now $Pred$ is primitive recursive and H is definable by composition from h and primitive recursive functions; furthermore, $f(m,\vec{z}) = F(\ulcorner m \urcorner,\vec{z})$. Thus f is defined from g and h and some primitive recursive functions using composition and ω-primitive recursion. □

Note that the proof of Theorem 4 shows that (0.1) could be weakened to include only the usual base functions (1) and a few specific primitive recursive functions for manipulating Gödel numbers of finite ordinals.

THEOREM 5 (FIRST NORMAL FORM) *Let $m \geq 2$ and $n \geq 0$. The set of functions $\prec \omega_m$-primitive recursive in Σ_n is the smallest set of functions satisfying the four conditions (1.1)-(1.4):*

(1.1)-(1.3): *same as (0.1)-(0.3).*

(1.4) *If $\kappa_0 \prec \omega_m$ and if g and κ are unary functions which are $\prec \omega_m$-primitive recursive in Σ_n then so is the function f defined by*

$$f(\alpha) = \begin{cases} f(\kappa(\alpha)) & \text{if } \kappa(\alpha) \prec \alpha \preccurlyeq \kappa_0 \\ g(\alpha) & \text{otherwise.} \end{cases}$$

In (1.4), we say that f *is defined by parameter-free* κ_0*-primitive recursion from* g *and* κ.

Proof For this proof only, let $\mathcal{F}$ denote the smallest set of functions which satisfies the closure conditions of (1.1)-(1.4). Obviously, the Zeroth Normal Form implies that every function in $\mathcal{F}$ is $\prec \omega_m$-primitive recursive in Σ_n. To show that $\mathcal{F}$ contains every function $\prec \omega_m$-primitive recursive in Σ_n, it will suffice to show that $\mathcal{F}$ is closed under the $\prec \omega_m$-primitive recursion of (0.4). For this, suppose f is defined by

$$f(\beta, \vec{z}) \quad = \quad \begin{cases} h(\beta, \vec{z}, f(\kappa(\beta, \vec{z}), \vec{z})) & \text{if } \kappa(\beta, \vec{z}) \prec \beta \preccurlyeq \kappa_0 \\ g(\beta, \vec{z}) & \text{otherwise.} \end{cases}$$

To give a definition of f using parameter-free $\prec \omega_m$-primitive recursion, we shall use ordinals that code the parameters $\vec{z}$ and which code a history of the computation of $f(\beta)$ with $\beta \preccurlyeq \kappa_0$. In order to code the history of the computation of f, we need ordinals $\beta_0, \beta_1, \ldots, \beta_s$ so that $\beta_0 = \beta$ and $\beta_{i+1} = \kappa(\beta_i, \vec{z}) \prec \beta_i$ and so that $\kappa(\beta_s, \vec{z}) \not\prec \beta_s$; also we need values $a_s, \ldots, a_0$ so that $a_s = g(\beta_s, \vec{z})$ and $a_i = h(\beta_i, \vec{z}, a_{i+1})$ for all $i < s$; it will follow that $f(\beta, \vec{z})$ is equal to a_0. We shall code and index this computation by the following scheme. We use ordinals of the form $\omega^2 \cdot \beta_i + \langle \vec{z}, \beta_0, \ldots, \beta_{i-1} \rangle$ to code the first phase of the computation of f, where $\langle \vec{z}, \beta_0, \ldots, \beta_{i-1} \rangle$ denotes the finite ordinal equal to the Gödel number of the sequence containing the entries $\vec{z}$ and the Gödel numbers $\beta_0, \ldots, \beta_{i-1}$. To code the second phase of the computation we use ordinals of the form $\omega \cdot i + \langle \vec{z}, \beta_0, \ldots, \beta_{i-1}, a_i \rangle$. Since $\kappa_0 \prec \omega_m$ there is an ordinal $\sigma_0 \prec \omega_{m-1}$ such that $\kappa_0 \prec \omega^{\sigma_0}$. Define

$$F(\alpha) \quad = \quad \begin{cases} F(K(\alpha)) & \text{if } K(\alpha) \prec \alpha \preccurlyeq \omega^{2+\sigma_0} \\ G(\alpha) & \text{otherwise} \end{cases}$$

where K and G are defined so that

$$K(\omega^2 \cdot \beta_i + \langle \vec{z}, \beta_0, \ldots, \beta_{i-1} \rangle) = \omega^2 \cdot \kappa(\beta_i, \vec{z}) + \langle \vec{z}, \beta_0, \ldots, \beta_i \rangle$$
$$\text{if } i \geq 0 \text{ and } \kappa(\beta_i, \vec{z}) \prec \beta_i$$

$$K(\omega^2 \cdot \beta_i + \langle \vec{z}, \beta_0, \ldots, \beta_{i-1} \rangle) = \omega \cdot i + \langle \vec{z}, \beta_0, \ldots, \beta_{i-1}, g(\beta_i, \vec{z}) \rangle$$
$$\text{where } 0 \leq i \in \mathbb{N} \text{ and } \kappa(\beta_i, \langle \vec{z} \rangle) \not\prec \beta_i$$

$$K(\omega \cdot (i+1) + \langle \vec{z}, \beta_0, \ldots, \beta_i, a \rangle) = \omega \cdot i + \langle \vec{z}, \beta_0, \ldots, \beta_{i-1}, h(\beta_i, \vec{z}, a) \rangle$$
$$\text{for } i \in \mathbb{N}$$

$$K(\ulcorner \langle \vec{z}, a \rangle \urcorner) = \ulcorner \langle \vec{z}, a \rangle \urcorner$$

$$G(\ulcorner \langle \vec{z}, a \rangle \urcorner) = a$$

where, in the last two equations, $\ulcorner\langle\vec{z}, a\rangle\urcorner$ denotes the Gödel number of the finite ordinal $\langle\vec{z}, a\rangle$. K and G may be arbitrarily defined for other inputs. Clearly F is defined by $\omega^{2+\sigma_0}$-primitive recursion from G and K. And f is definable in terms of F and g using composition:

$$f(\beta, \vec{z}) = \begin{cases} F(\omega^2 \cdot \beta + \langle\vec{z}\rangle) & \text{if } \beta \preccurlyeq \kappa_0 \\ g(\beta, \vec{z}) & \text{otherwise} \end{cases}$$

We have used only $\prec \omega_m$-primitive recursion (since $\omega^{2+\sigma_0} \prec \omega_m$) and composition to define f from g, h, κ and primitive recursive functions. Hence $f \in \mathcal{F}$.
Q.E.D. Theorem 5

The final and best normal form for $\prec \omega_m$-primitive recursive in Σ_n functions is not an inductive definition, but is a true normal form.

THEOREM 6 (SECOND NORMAL FORM)

(a) *Let $m \geq 2$ and $n \geq 1$. A function $F(\vec{z})$ is $\prec \omega_m$-primitive recursive in Σ_n iff there are a $\kappa_0 \prec \omega_m$, a $A(\vec{y})$ of the form $(\exists x)B(x, y)$ with $B \in \Pi_{n-1}$, and primitive recursive functions τ, g and κ so that $F(\vec{z}) = f(\tau(\vec{z}))$ where $f(\beta)$ is defined by*

$$f(\beta) = \begin{cases} f(\kappa(\beta, U_A(\beta))) & \text{if } \kappa(\beta, U_A(\beta)) \prec \beta \preccurlyeq \kappa_0 \\ g(\beta) & \text{otherwise} \end{cases}$$

(b) *Let $m \geq 2$. A function $F(\vec{z})$ is $\prec \omega_m$-primitive recursive iff there are a $\kappa_0 \prec \omega_m$ and primitive recursive functions τ, g and κ so that $F(\vec{z}) = f(\tau(\vec{z}))$ where*

$$f(\beta) = \begin{cases} f(\kappa(\beta)) & \text{if } \kappa(\beta) \prec \beta \preccurlyeq \kappa_0 \\ g(\beta) & \text{otherwise} \end{cases}$$

An important feature of the second normal form theorem is that κ is now required to be primitive recursive, instead of merely $\prec \omega_m$-primitive recursive in Σ_n.

Proof We shall prove (a); the proof of (b) is essentially identical. First, every primitive recursive function can be expressed in the form (a): to prove this, if F is primitive recursive, let $\kappa_0 = 0$, let $\tau(\vec{z}) = \ulcorner\langle\vec{z}\rangle\urcorner$, let $\kappa(\beta, a) = 0$ and $g(\ulcorner\langle\vec{z}\rangle\urcorner) = F(\vec{z})$. The functions τ and κ are clearly primitive recursive and g is primitive recursive since F is. Second, if $A(y)$ is $(\exists x)B(x, y)$ where $B \in \Pi_{n-1}$, then the function U_A can be expressed in the form (a) by letting

$\kappa_0 = \omega \cdot 2$, letting $\tau(y) = \omega + y$, letting $\kappa(\omega + y, i) = \ulcorner i \urcorner$ and $\kappa(\ulcorner i \urcorner, a) = \ulcorner i \urcorner$ and letting $g(\ulcorner i \urcorner) = i$.

Next we show that the set of functions definable in the form (a) is closed under composition. Suppose F_1 and F_2 are defined by $F_1(v, \vec{z}) = f_1(\tau_1(v, \vec{z}))$ and $F_2(\vec{z}) = f_2(\tau_2(\vec{z}))$ where

$$f_i(\beta) = \begin{cases} f_i(\kappa_i(\beta, U_{A_i}(\beta))) & \text{if } \kappa_i(\beta, U_{A_i}(\beta)) \prec \beta \preccurlyeq \kappa_{0,i} \\ g_i(\beta) & \text{otherwise} \end{cases}$$

for $i = 1, 2$. By assumption, τ_i, κ_i and g_i are primitive recursive functions. We must show $F(\vec{z}) = F_1(F_2(\vec{z}), \vec{z})$ is also definable in this way. Pick $\sigma \prec \omega_{m-1}$ to be an ordinal such that $\kappa_{0,1}, \kappa_{0,2} \prec \omega^\sigma$. We set $F(\vec{z}) = f(\omega^{1+\sigma} \cdot 2 + \langle \vec{z} \rangle)$ and define $f(\beta)$ as in (a) with $\kappa_0 = \omega^{1+\sigma} \cdot 3$ and with κ defined so that, if $\beta \prec \omega^\sigma$,

$$\kappa(\omega^{1+\sigma} \cdot 2 + \langle \vec{z} \rangle) = \begin{cases} \omega^{1+\sigma} + \omega \cdot \tau_2(\vec{z}) + \langle \vec{z} \rangle & \text{if } \tau_2(\vec{z}) \preccurlyeq \kappa_{0,2} \\ \tau_1(g_2(\tau_2(\vec{z}))) & \text{if } \tau_2(\vec{z}) \not\preccurlyeq \kappa_{0,2} \text{ and } \tau_1(g_2(\tau_2(\vec{z}))) \preccurlyeq \kappa_{0,1} \\ \omega^{1+\sigma} \cdot 3 & \text{otherwise} \end{cases}$$

$$\kappa(\omega^{1+\sigma} + \omega \cdot \beta + \langle \vec{z} \rangle) = \begin{cases} \omega^{1+\sigma} + \omega \cdot \kappa_2(\beta, U_{A_2}(\beta)) + \langle \vec{z} \rangle & \\ \qquad \text{if } \kappa_2(\beta, U_{A_2}(\beta)) \prec \beta \preccurlyeq \kappa_{0,2} & \\ \tau_1(g_2(\beta), \vec{z}) & \text{if not } \kappa_2(\beta, U_{A_2}(\beta)) \prec \beta \preccurlyeq \kappa_{0,2} \\ & \quad \text{and } \tau_1(g_2(\beta), \vec{z}) \preccurlyeq \kappa_{0,1} \\ \omega^{1+\sigma} \cdot 3 & \text{otherwise} \end{cases}$$

and, if $\beta \prec \kappa_{0,1}$, $\kappa(\beta) = \kappa_1(\beta, U_{A_1}(\beta))$. Also define g so that, for all $\beta \prec \omega^\sigma$,

$$\begin{aligned} g(\omega^{1+\sigma} \cdot 2 + \langle \vec{z} \rangle) &= g_1(\tau_1(g_2(\tau_2(\vec{z})))) \\ g(\omega^{1+\sigma} + \omega \cdot \beta + \langle \vec{z} \rangle) &= g_1(\tau_1(g_2(\beta))) \\ g(\beta) &= g_1(\beta) \end{aligned}$$

This almost defines $f(\vec{z})$ in the desired form (a); however, there is a problem since $\kappa(\alpha)$ is defined using both U_{A_1} and U_{A_2} (and not using them in correct manner either). To fix this, we define a new $A(y) = (\exists x) B(x, y)$ so that $\kappa(\alpha)$ is a primitive recursive function of only α and $U_A(\alpha)$. For this, suppose $A_i = (\exists x) B_i(x, y)$ where $B_i \in \Pi_{n-1}$. Define B by

$$B(x, \alpha) \Leftrightarrow \begin{cases} B_2(x, \beta) & \text{if } \alpha = \omega^{1+\sigma} + \omega \cdot \beta + m \\ B_1(x, \alpha) & \text{if } \alpha \prec \omega^{1+\sigma} \\ \text{arbitrary} & \text{otherwise.} \end{cases}$$

Since $B_1, B_2 \in \Pi_{n-1}$, so is B. That completes the proof that the set of functions definable in the form (a) are closed under composition.

Finally, we must show that the functions definable in the form (a) are closed under parameter-free $\prec \omega_m$-primitive recursion. For this, suppose f is defined from functions g and κ, which are defined in form (a), and from an ordinal $\kappa_0 \prec \omega_m$ as in (1.4) and further suppose that κ is defined in the normal form (a) by

$$\begin{aligned} \kappa(\alpha) &= f_1(\tau_1(\alpha)) \\ f_1(\beta) &= \begin{cases} f_1(\kappa_1(\beta, U_{A_1}(\beta))) & \text{if } \kappa_1(\beta, U_{A_1}(\beta)) \prec \beta \preccurlyeq \kappa_{0,1} \\ g_1(\beta) & \text{otherwise} \end{cases} \end{aligned}$$

where $\kappa_{0,1} \prec \omega_m$ and τ_1, κ_1 and g_1 are primitive recursive functions. Pick σ_0, σ_1 to be the least ordinals such that $\kappa_0 \prec \omega^{\sigma_0}$ and $\kappa_{0,1} \prec \omega^{\sigma_1}$; hence $\sigma_0, \sigma_1 \prec \omega_{m-1}$. We now define $F(\alpha) = f'(\omega^{1+\sigma_1+\sigma_0} + {}^{\ulcorner}\alpha^{\urcorner})$ where f' will be defined in the second normal form (a) with primitive recursive functions κ', g' and ordinal κ_0' where κ' is defined by

$$\kappa'(\omega^{1+\sigma_1+\sigma_0} + {}^{\ulcorner}\alpha^{\urcorner}) = \begin{cases} \omega^{1+\sigma_1} \cdot \alpha + \omega^{\sigma_1} & \text{if } \alpha \preccurlyeq \kappa_0 \\ \omega^{1+\sigma_1+\sigma_0} + \omega & \text{otherwise} \end{cases}$$

$$\kappa'(\omega^{1+\sigma_1} \cdot \beta + \omega^{\sigma_1}) = \begin{cases} \omega^{1+\sigma_1} \cdot \beta + \tau_1(\beta) & \text{if } \tau_1(\beta) \preccurlyeq \kappa_{0,1} \\ \omega^{1+\sigma_1} \cdot g_1(\tau_1(\beta)) & \text{if } \tau_1(\beta) \not\preccurlyeq \kappa_{0,1} \\ & \quad \text{and } g_1(\tau_1(\beta)) \prec \beta \\ \omega^{1+\sigma_1+\sigma_0} & \text{otherwise} \end{cases}$$

$$\kappa'(\omega^{1+\sigma_1} \cdot \beta + \gamma) = \begin{cases} \omega^{1+\sigma_1} \cdot \beta + \kappa_1(\gamma, U_{A_1}(\gamma)) & \\ \qquad \text{if } \kappa_1(\gamma, U_{A_1}(\gamma)) \prec \gamma \preccurlyeq \kappa_{0,1} & \\ \omega^{1+\sigma_1} \cdot g_1(\gamma) + \omega^{\sigma_1} & \\ \qquad \text{if } \kappa_1(\gamma, U_{A_1}(\gamma)) \not\prec \gamma \text{ and } g_1(\gamma) \prec \beta & \\ \omega^{1+\sigma_1+\sigma_0} & \text{if } \kappa_1(\gamma, U_{A_1}(\gamma)) \not\prec \gamma \text{ and } g_1(\gamma) \not\prec \beta \end{cases}$$

(provided $\gamma \preccurlyeq \kappa_{0,1}$), and g' is defined by

$$\begin{aligned} g'(\omega^{1+\sigma_1+\sigma_0} + {}^{\ulcorner}\alpha^{\urcorner}) &= \alpha \\ g'(\omega^{1+\sigma_1} \cdot \beta + \gamma) &= \beta \qquad \text{if } \beta \preccurlyeq \kappa_0 \text{ and } \gamma \preccurlyeq \omega^{\sigma_1} \end{aligned}$$

and $\kappa_0' = \omega^{1+\sigma_1+\sigma_0} + \omega$. Any values of κ' and g' left unspecified may be arbitrary. Now, inspection shows that

$$F(\alpha) = \begin{cases} F(\kappa(\alpha)) & \text{if } \kappa(\alpha) \prec \alpha \preccurlyeq \kappa_0 \\ \alpha & \text{otherwise} \end{cases}$$

and, by construction, F is definable in form (a). Now the function f is definable by $f(\alpha) = g(F(\alpha))$ and since g and F are expressible in form (a) it follows by the earlier part of this proof that their composition f is too. Q.E.D. Theorem 6

One further refinement can be made to the second normal form theorem: instead of allowing arbitrary U_A's with $A \in \Sigma_n$, it is possible to allow only a single, fixed, suitably chosen U_A. Of course, such an A is many-one complete for Σ_n. It is necessary to modify the ordinal coding methods in the above proof to establish this refinement — the details are left to the reader.

3.3. Some definability theorems

The next theorems characterize the Σ_n definable functions of $I\Sigma_n$; their proof will be a straightforward use of the witness function method.

THEOREM 7 *Let $m \geq 2$ and $n \geq 1$. The Σ_n-definable functions of the theory $I\Delta_0 + TI(\prec \omega_m, \Sigma_{n-1})$ are precisely the functions which are $\prec \omega_m$-primitive recursive in Σ_{n-1}.*

THEOREM 8 *Let $n \geq 1$. The Σ_n-definable functions of the theory $I\Sigma_n$ are precisely the functions which are primitive recursive in Σ_{n-1}.*

THEOREM 9 *The Σ_1-definable (provably recursive) functions of $I\Sigma_1$ are precisely the primitive recursive functions.*

There are (at least) three prior prooftheoretic proofs of Theorem 9. Parsons [22] gave a proof based on the Gödel Dialectica interpretation, Mints [19] gave a proof which uses a method very close to the witness function method except presented with a functional language, and Takeuti [27] gives a proof based on Gentzen-style assignment of ordinals to proofs.

Proof Theorems 8 and 9 are corollaries of Theorem 7 since $I\Sigma_n$ and $I\Delta_0 + TI(\prec \omega_m, \Sigma_{n-1})$ are the same theory. Although only the proof of Theorem 7 is given below, it should be remarked that the other two theorems can be proved directly by a similar and easier argument.

The easier half of the proof is to show that every $\prec \omega_m$-primitive recursive in Σ_{n-1} function is Σ_n-definable in $I\Delta_0 + TI(\prec \omega_m, \Sigma_{n-1})$. Recall that every primitive recursive function is Σ_1-definable in $I\Sigma_1$ so this half of the $m = 2$ and $n = 1$ case of Theorem 7 follows. For other values of m and n, suppose F is $\prec \omega_m$-primitive recursive in Σ_{n-1} and that F is defined by $F(\vec{z}) = f(\tau(\vec{z}))$ where

$$f(\beta) = \begin{cases} f(\kappa(\beta, U_A(\beta))) & \text{if } \kappa(\beta, U_A(\beta)) \prec \beta \preccurlyeq \kappa_0 \\ g(\beta) & \text{otherwise} \end{cases}$$

in accordance with the Second Normal Form, so g, κ and τ are primitive recursive functions, $\kappa_0 \prec \omega_m$ and $A(y)$ is $(\exists x)B(x,y)$ where $B \in \Pi_{n-2}$ (in the simpler case where $n = 1$, $\kappa(\beta, U_A(\beta))$ is replaced by $\kappa(\beta)$ and U_A is not used at all). Obviously it will suffice to show that f is Σ_n-definable by $I\Delta_0 + TI(\prec \omega_m, \Sigma_{n-1})$.

A sequence of ordinals $\beta_0, \ldots, \beta_k$ is an *f-computation series* if $\beta_{i+1} \prec \kappa_0$, $\beta_{i+1} = \kappa(\beta_i, U_A(\beta_i))$ and $\beta_{i+1} \prec \beta_i$, for all $0 \leq i < k$. To metamathematically define an f-computation series, we use (if $n > 1$),

"w codes an f-computation series" $\Leftrightarrow$

w is a sequence of Gödel numbers of ordinals of length $k+1$

and $(\forall i < k)\Big[\Big((\exists y)[B((w)_i, y) \wedge (\forall y' < y)(\neg B((w)_i, y))$

$\wedge (w)_{i+1} = \kappa((w)_i, y+1)]\Big)$

$\vee\Big((\forall y)(\neg B((w)_i, y)) \wedge (w)_{i+1} = \kappa((w)_i, 0)\Big)\Big]$

and $(\forall i < k)((w)_{i+1} \prec (w)_i \wedge (w)_{i+1} \prec \kappa_0)$.

(Recall that if $w = \langle \beta_0, \ldots, \beta_k \rangle$, then $(w)_i = \beta_i$.) Since $I\Delta_0 + TI(\prec \omega_m, \Sigma_{n-1})$ contains $I\Sigma_n$, it also contains the collection axiom $B\Sigma_n$. Thus the subformula $(\forall y' < y)(\cdots)$ above is equivalent to a Σ_{n-2} formula, and by applying prenex operations, the formula "w codes an f-computation series" is equivalent to a Π_n formula. By applying prenex operations in a different order, and using $B\Sigma_n$, this formula is also equivalent to a Σ_n-formula. If $n = 1$, then instead define

"w codes an f-computation series" $\Leftrightarrow$

w is a sequence of Gödel numbers of ordinals of length $k+1$

and $(\forall i < k)(\beta_{i+1} = \kappa(\beta_i) \prec \beta_i \wedge \beta_{i+1} \prec \kappa_0)$,

so, in this case, it is a primitive recursive property.[6]

The graph of the function $f(\beta)$ can now be defined by using the fact that $y = f(\beta)$ iff $y = g(\beta')$ where β' is the least ordinal such that there is an f-computation series $\langle \beta, \ldots, \beta' \rangle$. More formally, letting $fCS(w)$ be the formula "w is an f-computation series",

$$y = f(\beta) \Leftrightarrow (\exists \langle \beta, \ldots, \beta' \rangle)\Big[y = g(\beta') \wedge fCS(\langle \beta, \ldots, \beta' \rangle) \wedge$$
$$\wedge \neg(\kappa(\beta', U_A(\beta')) \prec \beta' \wedge \beta' \preccurlyeq \kappa_0)\Big].$$

Since $fCS(\cdots)$ is equivalent to a Σ_n-formula and since $z = U_A(\beta')$ can be expressed as a Π_{n-1}-formula, the relation $y = f(\beta)$ is a Σ_n-property,

[6] It is possible to strengthen the second normal form theorem to make this a Δ_0-formula.

provably in $I\Delta_0 + TI(\prec \omega_m, \Sigma_{n-1})$. The theory also proves

$$\forall\beta\ \exists \text{ a least } \beta' \text{ s.t. } \exists\langle\beta,\ldots,\beta'\rangle(fCS(\langle\beta,\ldots,\beta'\rangle))$$

since $fCS(\langle\beta\rangle)$ and by $LOP(\prec \omega_m, \Sigma_n)$ since $\kappa_0 \prec \omega_m$.[7] Thus $I\Delta_0 + TI(\prec \omega_m, \Sigma_{n-1})$ can Σ_n-define the function f as it proves $(\forall\beta)(\exists!y)(y = f(\beta))$ where $y = f(\beta)$ denotes the Σ_n-formula defining the graph of f. Likewise,

$$(\forall\vec{z})(\exists!y)(\exists\beta)(\beta = \tau(\vec{z}) \wedge y = f(\beta))$$

is also provable and Σ_n-defines the function F. That completes the first half of the proof of Theorem 7.

To prove the rest of Theorem 7, assume that $I\Delta_0 + TI(\prec \omega_m, \Sigma_{n-1})$ proves $(\forall x)(\exists!y)A(x,y)$, with $A \in \Sigma_n$ — we must show that $x \mapsto y$ is a $\prec \omega_m$-primitive recursive in Σ_{n-1} function. Since $I\Delta_0 + TI(\prec \omega_m, \Sigma_{n-1})$ proves $(\forall x)(\exists y)A$, there must be a free-cut free proof in the theory $I\Delta_0 + TI(\prec \omega_m, \Sigma_{n-1})$ of the sequent

$$\longrightarrow (\exists y)A(c,y)$$

where c is a new free variable. Only Σ_n formulas can appear in this free-cut free proof. The general idea of the proof is to show that this free-cut free proof embodies an algorithm for computing y from c. Indeed, the free-cut free proof can be interpreted as explicitly containing a $\prec \omega_m$-primitive recursive in Σ_{n-1} algorithm. Since the proofs of the normal form theorems were constructive, the free-cut free proof also contains an implicit description of a $\prec \omega_m$-primitive recursive in Σ_{n-1} algorithm in the second normal form. Our proof below that an algorithm can be extracted from the free-cut free proof is quite constructive and can be formalized in $I\Delta_0 + TI(\prec \omega_m, \Sigma_{n-1})$ — the upshot is that there is a $\prec \omega_m$-primitive recursive in Σ_{n-1} function f which is Σ_n-defined by $I\Delta_0 + TI(\prec \omega_m, \Sigma_{n-1})$ in the form given by the Second Normal Form Theorem such that $I\Delta_0 + TI(\prec \omega_m, \Sigma_{n-1}) \vdash (\forall x)A(x, f(x))$. As a corollary to the proof method, if $I\Delta_0 + TI(\prec \omega_m, \Sigma_{n-1})$ proves $(\forall x)(\exists y)B(x,y)$ with $B \in \Sigma_n$ then there is a $B^*(x,y) \in \Sigma_n$ such that $(\forall x)(\exists!y)B^*(x,y)$ and $B^*(x,y) \rightarrow B(x,y)$ are provable.[8]

We shall see later that the proof is formalizable, not only in $I\Delta_0 + TI(\prec \omega_m, \Sigma_{n-1})$, but also in $I\Delta_0 + TI(\prec \omega_{m+1}, \Sigma_{n-2})$, provided $n > 1$.

---

[7] $LOP(\prec \omega_m, \Sigma_n)$ is a consequence of $I\Delta_0 + TI(\prec \omega_m, \Sigma_{n-1})$ by Proposition 1.

[8] This fact is readily proved directly anyway. If $B \in \Pi_{n-1}$ then let B^* be the formula $B(x,y) \wedge (\forall y' < y)(\neg B(x,y'))$, which is equivalent to a Σ_n formula by $B\Sigma_n$. For general $B \in \Sigma_n$, incorporate outermost existential quantifiers of B into the the $(\exists y)$ and proceed similarly.

Rather than just considering the free-cut free proof of $\longrightarrow(\exists y)A$, we more generally consider proofs of sequents $\Gamma\longrightarrow\Delta$ of Σ_n-formulas. Since every principal and auxiliary formula of a $LOP(\prec\omega_m,\Pi_{n-1})$ inference is in Σ_n and every formula in the endsequent is in Σ_n, it follows that every formula in the free-cut free proof is in Σ_n. For convenience, assume also that the proof is in free variable normal form (so free variables are not reused).

Definition Let $i\geq 1$ and $A(\vec{x})\in\Sigma_i$. If $A\in\Pi_{i-1}$ then Wit^i_A is defined to be the formula A. Otherwise, A is uniquely expressible in the form $(\exists y_0)\cdots(\exists y_k)B(\vec{x},\vec{y})$ where $B\in\Pi_{i-1}$. Then $Wit^i_A(w,\vec{x})$ is the formula

$$B(\vec{x},(w)_0,\ldots,(w)_k).$$

Note that $Wit^i_A\in\Pi_{i-1}$. If $Wit^i_A(w,\vec{x})$ holds, we say w *witnesses* the truth of $A(\vec{x})$.

MAIN LEMMA 10 *($n\geq 1$, $m\geq 2$) Suppose $I\Delta_0+TI(\prec\omega_m,\Sigma_{n-1})$ proves the sequent $A_1,\ldots,A_k\longrightarrow B_1,\ldots,B_\ell$ and that each A_i and B_j is in Σ_n and that $\vec{c}$ are all the variables free in the sequent. Then there are functions $f_1,\ldots,f_\ell$ which are $\prec\omega_m$-primitive recursive in Σ_{n-1} and are Σ_n-definable in $I\Delta_0+TI(\prec\omega_m,\Sigma_{n-1})$ such that $I\Delta_0+TI(\prec\omega_m,\Sigma_{n-1})$ proves*

$$Wit^n_{A_1}(w_1,\vec{c}),\ldots,Wit^n_{A_k}(w_k,\vec{c})\longrightarrow Wit^n_{B_1}(f_1(\vec{w},\vec{c}),\vec{c}),\ldots,Wit^n_{B_\ell}(f_\ell(\vec{w},\vec{c}),\vec{c}).$$

Informally, the $f_1,\ldots,f_\ell$ will, given witnesses for all of $A_1,\ldots,A_k$, produce a witness for at least one of $B_1,\ldots,B_\ell$.

The proof of the Main Lemma is by induction on the number of inferences in a free-cut free proof of the sequent. In the base case, there are zero inferences, so the sequent is an axiom and consists of Δ_0-formulas — for these axioms, the lemma is trivial. For the induction step, the proof splits into cases depending in the final inference of the proof. Most of the cases are straightforward; for example, if the last inference is an $\exists$*:left* inference then the proof ends with

$$\frac{A_1,\ldots,A_k\longrightarrow B_0(\vec{c},s),B_2,\ldots,B_\ell}{A_1,\ldots,A_k\longrightarrow(\exists z_0)B_0(\vec{c},z_0),B_2,\ldots,B_\ell}$$

where $s=s(\vec{c})$ is a term with free variables from $\vec{c}$ only and where B_1 is $(\exists z_0)B_0$ and is of the form $(\exists z_0)\cdots(\exists z_r)B'(\vec{z},\vec{c})$ with $B'\in\Pi_{n-1}$ (possibly $r=0$). The induction hypothesis is that

$$\begin{aligned}Wit^n_{A_1}(w_1,\vec{c}),\ldots,Wit^n_{A_k}(w_k,\vec{c})\\ \longrightarrow Wit^n_{B_0(\vec{c},s)}(f_0(\vec{w},\vec{c}),\vec{c}),\ldots,Wit^n_{B_\ell}(f_\ell(\vec{w},\vec{c}),\vec{c})\end{aligned}\qquad(4)$$

is provable in $I\Delta_0 + TI(\prec \omega_m, \Sigma_{n-1})$ for appropriate functions $f_0, f_2, \ldots, f_\ell$. If $B_1 \in \Sigma_{n-2}$ then $Wit^n_{B_0}$ is just B_0 and $Wit^n_{B_1}$ is just B_1; and a single $\exists$*:right* inference applied to (4) gives

$$Wit^n_{A_1}(w_1, \vec{c}), \ldots, Wit^n_{A_k}(w_k, \vec{c}) \longrightarrow Wit^n_{B_1}(f_1(\vec{w}, \vec{c}), \vec{c}), \ldots, Wit^n_{B_\ell}(f_\ell(\vec{w}, \vec{c}), \vec{c}) \quad (5)$$

where f_1 is arbitrary. Otherwise, let $f_1(\vec{w}, \vec{c})$ be defined so that

$$f_1(\vec{w}, \vec{c}) = \langle s(\vec{c}), a_1, \ldots, a_r \rangle \quad \text{where} \quad f_0(\vec{w}, \vec{c}) = \langle a_1, \ldots, a_r \rangle$$

if $r > 0$, and $f(\vec{w}, \vec{c}) = \langle s(\vec{c}) \rangle$ if $r = 0$. Clearly f_1 is $\prec \omega_m$-primitive recursive in Σ_{n-1} since f_0 is and, also clearly, $I\Delta_0 + TI(\prec \omega_m, \Sigma_{n-1})$ proves (5) for this f_1.

We leave the rest of the simpler cases to the reader and consider only the two substantial cases of $\forall$*:right* and $LOP(\prec \omega_m, \Pi_{n-1})$ as last inference. (Part of the $\exists$*:left* case is also substantial, but is very similar to $\forall$*:right.*)

($\forall$*:right*) Suppose the last inference is

$$\frac{A_1, \ldots, A_k \longrightarrow B_0(b, \vec{c}), B_2, \ldots, B_\ell}{A_1, \ldots, A_k \longrightarrow (\forall z_0) B_0(z_0, \vec{c}), B_2, \ldots, B_\ell}$$

where the free variable b does not occur except as indicated and B_1 is $(\forall z_0) B_0(\vec{z}, \vec{c})$. Since B_1 is in Σ_n and has outermost quantifier universal, it must therefore actually be in Π_{n-1} and be of the form $(\forall z_0) \cdots (\forall z_r) B'(\vec{z}, \vec{c})$ where $B' \in \Sigma_{n-2}$. Also $Wit^n_{B_0}$ and $Wit^n_{B_1}$ are just B_0 and B_1, respectively. The induction hypothesis is that $I\Delta_0 + TI(\prec \omega_m, \Sigma_{n-1})$ proves

$$Wit^n_{A_1}(w_1, \vec{c}), \ldots, Wit^n_{A_k}(w_k, \vec{c}) \longrightarrow B_0(b, \vec{c}), Wit^n_{B_2}(g_2(\vec{w}, b, \vec{c}), \vec{c}), \ldots, Wit^n_{B_\ell}(g_\ell(\vec{w}, b, \vec{c}), \vec{c})$$

for functions $g_2, \ldots, g_\ell$ which are $\prec \omega_m$-primitive recursive in Σ_{n-1}. The difficulty is that these functions take b as an argument, but b is not free in the endsequent so we can not just set $f_i = g_i$. The solution to this difficulty is to let $C(v, \vec{c})$ be the Π_{n-2}-formula $\neg B'((v)_0, \ldots, (v)_r, \vec{c})$ and use the function $U_{\exists v C}$ to find a value, if any, for b such that $B_0(b, \vec{c})$ holds: define

$$f_i(\vec{w}, \vec{c}) \quad = \quad g_i(\vec{w}, (U_{\exists v C}(\vec{c}) - 1)_0, \vec{c}).$$

When $B_1(\vec{c})$ is false, $U_{\exists v C}(\vec{c}) - 1$ codes a sequence $\langle b_0, \ldots, b_r \rangle$ such that $\neg B_0(b_0, \ldots, b_r)$ and $(U_{\exists v C}(\vec{c}) - 1)_0$ equals b_0. Thus $I\Delta_0 + TI(\prec \omega_m, \Sigma_{n-1})$ proves

$$Wit^n_{A_1}(w_1, \vec{c}), \ldots, Wit^n_{A_k}(w_k, \vec{c}) \longrightarrow B_1(\vec{c}), Wit^n_{B_2}(f_2(\vec{w}, \vec{c}), \vec{c}), \ldots, Wit^n_{B_\ell}(f_\ell(\vec{w}, \vec{c}), \vec{c})$$

and $f_2, \ldots, f_k$ are $\prec \omega_m$-primitive recursive in Σ_{n-1} since $g_2, \ldots, g_k$ are and since $(\exists v)C$ is in Σ_{n-1}.

$LOP(\prec \omega_m, \Pi_{n-1})$: Suppose the last inference is

$$\frac{\alpha \preccurlyeq \kappa_0, A_1(\alpha, \vec{c}), A_2, \ldots, A_k \longrightarrow B_1, \ldots, B_\ell, (\exists \beta \prec \alpha) A_1(\beta, \vec{c})}{A_1(\kappa_0, \vec{c}), A_2, \ldots, A_k \longrightarrow B_1, \ldots, B_\ell}$$

where $A_1 \in \Pi_{n-1}$, where κ_0 is a closed term with value a Gödel number of an ordinal $\prec \omega_m$, where α is a free variable, which appears only as indicated, and where $(\exists \beta \prec \alpha) A_1(\beta)$ is an abbreviation for the formula $(\exists \beta)(\beta \prec \alpha \wedge A_1(\beta))$. The induction hypothesis states that $I\Delta_0 + TI(\prec \omega_m, \Sigma_{n-1})$ proves

$$\begin{aligned} \alpha \preccurlyeq \kappa_0, A_1(\alpha, \vec{c}), Wit^n_{A_2}(w_2, \vec{c}), \ldots, Wit^n_{A_k}(w_k, \vec{c}) \\ \longrightarrow Wit^n_{B_1}(g_1(\vec{w}, \alpha, \vec{c}), \vec{c}), \ldots, Wit^n_{B_\ell}(g_\ell(\vec{w}, \alpha, \vec{c}), \vec{c}), \\ g_{\ell+1}(\vec{w}, \alpha, \vec{c}) \prec \alpha \wedge A_1(g_{\ell+1}(\vec{w}, \alpha, \vec{c}), \vec{c}) \end{aligned}$$

for appropriate functions $g_1, \ldots, g_{\ell+1}$. Define

$$H(\beta, \vec{c}) \quad = \quad \begin{cases} \beta & \text{if } A_1(\beta, \vec{c}) \\ \kappa_0 & \text{otherwise} \end{cases}$$

H is $\prec \omega_m$-primitive recursive in Σ_{n-1} since $A_1 \in \Pi_{n-1}$. Now define

$$F(\vec{w}, \beta, \vec{c}) \quad = \quad \begin{cases} F(\vec{w}, H(g_{\ell+1}(\vec{w}, \beta, \vec{c}), \vec{c}), \vec{c}) & \text{if } H(g_{\ell+1}(\vec{w}, \beta, \vec{c}), \vec{c}) \prec \beta \preccurlyeq \kappa_0 \\ \beta & \text{otherwise.} \end{cases}$$

Clearly F is also $\prec \omega_m$-primitive recursive in Σ_{n-1}. Finally set

$$f_i(\vec{w}, \vec{c}) = g_i(\vec{w}, F(\vec{w}, \kappa_0, \vec{c}), \vec{c});$$

it is easy to check that $I\Delta_0 + TI(\prec \omega_m, \Sigma_{n-1})$ proves

$$\begin{aligned} A_1(\kappa_0, \vec{c}), Wit^n_{A_2}(w_2, \vec{c}), \ldots, Wit^n_{A_k}(w_k, \vec{c}) \\ \longrightarrow B_1(\vec{c}), Wit^n_{B_2}(f_2(\vec{w}, \vec{c}), \vec{c}), \ldots, Wit^n_{B_\ell}(f_\ell(\vec{w}, \vec{c}), \vec{c}) \end{aligned}$$

since $F(\vec{w}, \kappa_0, \vec{c})$ gives the ordinal at which $g_{\ell+1}$ fails to give a smaller ordinal satisfying A_1 and with this ordinal, one of $g_1, \ldots, g_\ell$ must produce a witness for the corresponding $B_1, \ldots, B_\ell$.

Q.E.D. Lemma 10 and Theorems 7, 8 and 9

The above proof did not consider the case where the last inference of the proof is an induction inference: since induction is restricted to Δ_0-formulas and the witness formula for a Δ_0-formula is just the formula itself, that

case is completely trivial. However, $I\Sigma_n$ is, by Proposition 2 a consequence of $I\Delta_0 + TI(\prec \omega_m, \Sigma_{n-1})$ and it must, a priori, be possible to handle $I\Sigma_n$ induction inferences by the witness function method as above. In fact, it is quite simple — an $I\Sigma_n$-induction inference is handled by primitive recursion in Σ_{n-1}. This leads to a direct proof of Theorems 8 and 9; we leave the details of this direct proof to the reader.

We have now finished the characterization of the Σ_n-definable functions of $I\Delta_0 + TI(\prec \omega_m, \Sigma_{n-1})$ and of $I\Sigma_n$. It remains to characterize the Σ_k-definable functions of these theories when $k < n$. (In section 6, we discuss the case $k > n$ too). The central result needed for this characterization is that the theory $I\Delta_0 + TI(\prec \omega_m, \Sigma_{n-1})$ is Π_{n+1}-conservative over $I\Delta_0 + TI(\prec \omega_{m+1}, \Sigma_{n-2})$:

THEOREM 11 *Let $m \geq 2$ and $n \geq 1$.*

(a) $I\Delta_0 + TI(\prec \omega_m, \Sigma_{n-1}) \vdash TI(\prec \omega_{m+1}, \Sigma_{n-2})$.

(b) *If $I\Delta_0 + TI(\prec \omega_m, \Sigma_{n-1}) \vdash A$ where $A \in \Pi_{n+1}$, then $I\Delta_0 + TI(\prec \omega_{m+1}, \Sigma_{n-2}) \vdash A$.*

Part (a) of this theorem is due to Gentzen [10]; the proof can be found in Lemma 3.4 of [26] or Theorem 12.3 of [27] and is also repeated below. Part (b) extends the prior result of Schmerl [24] that $I\Delta_0 + TI(\prec \omega_m, \Sigma_{n-1})$ is Π_{n-1}-conservative over $I\Delta_0 + TI(\prec \omega_{m+1}, \Sigma_{n-2})$; Schmerl's proof was based on reflection principles. A weaker version of (b) with Π_2-conservativity in place of Π_{n+1}-conservativity can be found in [26].

Proof (a) By Proposition 1, it will suffice to show that the theory $I\Delta_0 + TI(\prec \omega_m, \Pi_n)$ can prove $TI(\prec \omega_{m+1}, \Pi_{n-1})$. Let $A(\alpha) \in \Pi_{n-1}$ and let HYP_A be the formula $(\forall\beta)[(\forall\gamma \prec \beta)A(\gamma) \rightarrow A(\beta)]$ and let $\kappa \prec \omega_{m+1}$. We reason inside $I\Delta_0 + TI(\prec \omega_m, \Pi_n)$ to prove $A(\kappa)$ assuming HYP_A. Let $A^*(\alpha)$ be the formula $(\forall\gamma \prec \alpha)A(\gamma)$; by HYP_A, $A^*(\alpha) \rightarrow A^*(\alpha+1)$. Let $J(\beta)$ be the formula

$$(\forall\alpha)\left(A^*(\alpha) \rightarrow A^*(\alpha + \omega^\beta)\right).$$

Clearly, $J \in \Pi_n$. We shall use transfinite induction on J to prove $J(\kappa_0)$ for some fixed $\kappa_0 \prec \omega_m$ such that $\kappa \prec \omega^{\kappa_0}$. Since $A^*(0)$ holds trivially, $J(\kappa_0)$ implies $A^*(\omega^{\kappa_0})$ which, in turn implies $A(\kappa)$. Thus it suffices to prove HYP_J:

$$(\forall\beta)[(\forall\gamma \prec \beta)J(\gamma) \rightarrow J(\beta)]$$

since, using $TI(\prec \omega_m, \Pi_n)$, this implies $J(\kappa_0)$ holds for this particular κ_0. First note that $J(0)$ holds by our observation that $A^*(\alpha) \rightarrow A^*(\alpha+1)$.

Now let β be an arbitrary non-zero ordinal and suppose $(\forall\gamma \prec \beta)J(\beta)$: we must prove $J(\beta)$. If β is a successor ordinal, $\beta = \beta' + 1$, it suffices to show $J(\beta') \to J(\beta'+1)$, i.e.,

$$(\forall\alpha)\left(A^*(\alpha) \to A^*(\alpha + \omega^{\beta'})\right) \to (\forall\alpha')\left(A^*(\alpha') \to A^*(\alpha' + \omega^{\beta'+1})\right).$$

Assume $J(\beta')$ holds and let α' be arbitrary such that $A^*(\alpha')$ and let $\gamma \prec \alpha' + \omega^{\beta'+1}$; we must show $A^*(\gamma)$. By consideration of Cantor normal forms, $\gamma \prec \alpha' + \omega^{\beta'} \cdot n$ for some finite n. From $J(\beta')$, it follows that

$$(\forall\alpha)\Big(A^*(\alpha) \to A^*(\alpha + \omega^{\beta'} \cdot k)\Big) \to (\forall\alpha)\Big(A^*(\alpha) \to A^*(\alpha + \omega^{\beta'} \cdot (k+1))\Big)$$

holds for all (finite) k. By ordinary Π_n-induction, this implies that

$$(\forall\alpha)\Big(A^*(\alpha) \to A^*(\alpha + \omega^{\beta'} \cdot k)\Big)$$

holds for all finite k. Thus $A^*(\gamma)$ holds. Finally, suppose β is a limit ordinal and assume $(\forall\delta \prec \beta)J(\delta)$ and assume $A^*(\alpha)$. If $\gamma \prec \alpha + \omega^\beta$ then $\gamma \prec \alpha + \omega^\delta$ for some $\delta \prec \beta$ so $A(\gamma)$ holds by $J(\delta)$. Since γ was arbitrary, $J(\beta)$ follows. That completes the proof of (a).

The proof of (b) consists of a partial formalization of the Main Lemma 10 in the theory $I\Delta_0 + TI(\prec \omega_{m+1}, \Sigma_{n-2})$. First an important lemma is necessary:

LEMMA 12 *Let $m \geq 2$ and $n \geq 2$. $I\Delta_0 + TI(\prec \omega_{m+1}, \Sigma_{n-2})$ can Σ_n-define precisely the $\prec \omega_m$-primitive recursive in Σ_{n-1} functions.*

Proof By the just established part (a) of Theorem 11, every Σ_n-definable function of $I\Delta_0 + TI(\prec \omega_{m+1}, \Sigma_{n-2})$ is also Σ_n-defined by $I\Delta_0 + TI(\prec \omega_m, \Sigma_{n-1})$ and hence, by Theorem 7, is $\prec \omega_m$-primitive recursive in Σ_{n-1}. To show the converse, suppose $F(\vec{z})$ is defined from primitive recursive functions g, τ, and κ, from $A(y) = (\exists x)B(x)$ with $B \in \Pi_{n-2}$, and from an ordinal $\kappa_0 \prec \omega_m$ as in the Second Normal From Theorem; so $F(\vec{z}) = f(\tau(\vec{z}))$ where

$$f(\beta) \quad = \quad \begin{cases} f(\kappa(\beta, U_A(\beta))) & \text{if } \kappa(\beta, U_A(\beta)) \prec \beta \preccurlyeq \kappa_0 \\ g(\beta) & \text{otherwise.} \end{cases}$$

Recall the definition of an f-computation series $\beta_0, \ldots, \beta_k$ used in the proof of Theorem 7 to code a partial computation of f. In the proof of Theorem 7, the existence of a maximal length f-computation series beginning with $\beta_0 = \tau(\vec{z})$ was proved by finding the least β_k such that there exists an f-computation series from β_0 to β_k. The existence of β_k was proved via $LOP(\prec \omega_m, \Sigma_n)$: this was the key step in Σ_n-defining F in $I\Delta_0 + TI(\prec \omega_m, \Sigma_{n-1})$.

To Σ_n-define f and F in $I\Delta_0 + TI(\prec \omega_{m+1}, \Sigma_{n-2})$ requires a more subtle argument. The basic motivation for this argument is that one could try to minimize the ordinals of the form

$$\omega^{\beta_0} + \omega^{\beta_1} + \cdots + \omega^{\beta_{k-1}} + \omega^{\beta_k} \cdot 2$$

with $\beta_0, \ldots, \beta_k$ an f-computation series — but this is too simplistic because of the presence of the U_A function. Instead, we encode partial computations of f by a sequence of ordinals

$$\beta_0, \alpha_0, \beta_1, \alpha_1, \ldots, \beta_k, \alpha_k$$

where $\beta_0, \ldots, \beta_k$ is an f-computation series and where each $\alpha_i \preccurlyeq \omega$ and encodes the value of $U_A(\beta_i)$:

Definition Let α be the Gödel number of an ordinal $\preccurlyeq \omega$. Then $D(\alpha)$ is the integer defined by

$$D(\alpha) = \begin{cases} 0 & \text{if } \alpha = \omega \\ n+1 & \text{if } \alpha = \ulcorner n \urcorner \end{cases}$$

Definition An *f-computation ordinal (fCO)* is (the Gödel number of) an ordinal of the form

$$\omega^{\omega^2 \cdot \beta_0 + \alpha_0} + \omega^{\omega^2 \cdot \beta_1 + \alpha_1} + \cdots \omega^{\omega^2 \cdot \beta_{k-1} + \alpha_{k-1}} + \omega^{\omega^2 \cdot \beta_k + \alpha_k} + \omega^{\omega^2 \cdot \beta_k + \alpha_k}$$

(only the final summand is repeated), where

(*i*) $\beta_{i+1} \prec \beta_i \preccurlyeq \kappa_0$, for $0 \leq i < k$,

(*ii*) $\alpha_i \preccurlyeq \omega$, for $0 \leq i < k$,

(*iii*) $\beta_{i+1} = \kappa(\beta_i, D(\alpha_i))$, for $0 \leq i < k$,

(*iv*) For $0 \leq i \leq k$,

- if $\alpha_i = \ulcorner n \urcorner$, then $B(\beta_i, n)$ and for all $m < n$, $\neg B(\beta_i, m)$
- if $\alpha_i = \omega$, then $(\forall m) \neg B(\beta_i, m)$,

(*v*) It is not the case that $\kappa(\beta_k, D(\alpha_k)) \prec \beta_k \preccurlyeq \kappa_0$.

A *psuedo-f-computation ordinal (PfCO)* is defined exactly like an f-computation ordinal except that (*v*) is omitted and (*iv*) is replaced by

(*iv'*) For $0 \leq i \leq k$, if $\alpha_i = \ulcorner n \urcorner$ then $B(\beta_i, n)$.

We write $fCO(\alpha, \vec{z})$ and $PfCO(\alpha, \vec{z})$ for formulas expressing the condition that α is an fCO or PfCO, respectively, with $\beta_0 = \tau(\vec{z})$.

The quantifier complexity of $PfCO$ is easily analyzed since (i)-(iii) are primitive recursive and (iv') is Π_{n-2} since $B \in \Pi_{n-2}$ and by $B\Pi_{n-2}$-collection (which is a consequence of $I\Delta_0 + TI(\prec \omega_{m+1}, \Sigma_{n-2})$ since this theory contains $I\Sigma_{n-1}$). Thus $PfCO$ is a Π_{n-2} formula. Letting $\kappa_1 = \omega^{\omega^2 \cdot \kappa_0 + \omega + 1}$ we have that $\kappa_1 \prec \omega_{m+1}$ and, therefore, if $\tau(\vec{z}) \preccurlyeq \kappa_0$ and $PfCO(\alpha, \vec{z})$, then $\alpha \prec \kappa_1$. We henceforth assume w.l.o.g. that $\tau(\vec{z}) \preccurlyeq \kappa_0$. Now, there exists α such that $PfCO(\alpha, \vec{z})$; namely, $\omega^{\omega^2 \cdot \tau(\vec{x}) + \omega} \cdot 2$. Hence, by $LOP(\prec \omega_{m+1}, \Pi_{n-2})$, there is a minimum ordinal denoted α_{min} such that $PfCO(\alpha_{min}, \vec{z})$. We claim that $fCO(\alpha_{min}, \vec{z})$ also holds. To prove this, suppose

$$\alpha_{min} = \omega^{\omega^2 \cdot \beta_0 + \alpha_0} + \cdots + \omega^{\omega^2 \cdot \beta_k + \alpha_k} + \omega^{\omega^2 \cdot \beta_k + \alpha_k};$$

the only way $fCO(\alpha_{min})$ can fail is if condition (iv) or (v) is violated. First suppose (iv) fails for some value of i. Then, if $\alpha_i = \omega$ but $B(\beta_i, m)$ holds, then

$$\omega^{\omega^2 \cdot \beta_0 + \alpha_0} + \cdots + \omega^{\omega^2 \cdot \beta_{i-1} + \alpha_{i-1}} + \omega^{\omega^2 \cdot \beta_i + m} + \omega^{\omega^2 \cdot \beta_i + m} \tag{6}$$

is a psuedo f-computation ordinal $\prec \alpha_{min}$ violating the choice of α_{min}. Likewise, if $\alpha_i = \ulcorner n \urcorner$ but $B(\beta_i, m)$ holds with $m < n$, then the same ordinal (6) is a psuedo f-computation ordinal $\prec \alpha_{min}$. Hence (iv) must hold. Now suppose (v) fails. Then,

$$\omega^{\omega^2 \cdot \beta_0 + \alpha_0} + \cdots + \omega^{\omega^2 \cdot \beta_{k-1} + \alpha_{k-1}} + \omega^{\omega^2 \cdot \beta_k + \alpha_k} + \omega^{\omega^2 \cdot \beta_{k+1} + \omega} + \omega^{\omega^2 \cdot \beta_{k+1} + \omega}$$

where $\beta_{k+1} = \kappa(\beta_k, D(\alpha_k))$ is a psuedo f-computation ordinal $\prec \alpha_{min}$, which is again a contradiction. Hence (v) must also hold and α_{min} is an fCO.

Thus $I\Delta_0 + TI(\prec \omega_{m+1}, \Sigma_{n-2})$ can define $F(\vec{z})$ by proving

$$(\forall \vec{z})(\exists! y)\Big[(\exists \alpha)\Big\{PfCO(\alpha, \vec{z}) \wedge (\forall \alpha')(\alpha' \prec \alpha \rightarrow \neg PfCO(\alpha', \vec{z})) \wedge \alpha = \omega^{\omega^2 \cdot \beta_0 + \alpha_0} + \cdots + \omega^{\omega^2 \cdot \beta_k + \alpha_k} \cdot 2 \wedge y = g(\beta_k)\Big\}\Big]. \tag{7}$$

$PfCO$ is a Π_{n-2}-formula so the subformula $(\forall \alpha')(\cdots)$ is in Π_{n-1} and the subformula $(\exists \alpha)(\cdots)$ is a Σ_n-formula; thus this is a Σ_n-definition of $F(\vec{z})$ in $I\Delta_0 + TI(\prec \omega_{m+1}, \Sigma_{n-2})$.

Q.E.D. Lemma 12

Lemma 12 stated that the Σ_n-definable functions of $I\Delta_0 + TI(\prec \omega_{m+1}, \Sigma_{n-2})$ are precisely the $\prec \omega_m$-primitive recursive in Σ_{n-1} functions; the lemma was proved using the second normal form for such functions. However, this use of the second normal form was not essential for the proof: $I\Delta_0 + TI(\prec \omega_{m+1}, \Sigma_{n-2})$ can also prove that the $\prec \omega_m$-primitive recursive in Σ_{n-1} functions are closed under composition and under $\prec \omega_m$-primitive recursion. These closure properties are proved in $I\Delta_0 + TI(\prec \omega_{m+1}, \Sigma_{n-2})$ by formalizing the proofs of the three normal form theorems. Since the proofs of the normal form theorem were completely constructive, this formalization is straightforward (and left to the reader).

We are now ready to return to the proof of part (b) of Theorem 11, for which it suffices to prove that if $B(\vec{c})$ is a Σ_n-formula and $I\Delta_0 + TI(\prec \omega_m, \Sigma_{n-1})$ proves the sequent $\longrightarrow B(\vec{c})$, then so does $I\Delta_0 + TI(\prec \omega_{m+1}, \Sigma_{n-2})$. In fact, more than this is true: a sequent $\Gamma \longrightarrow \Delta$ of Σ_n-formulas is a consequence of $I\Delta_0 + TI(\prec \omega_m, \Sigma_{n-1})$ if and only if it is a consequence of $I\Delta_0 + TI(\prec \omega_{m+1}, \Sigma_{n-2})$ — this is a corollary of the next lemma.

MAIN LEMMA 13 *($n \geq 2$, $m \geq 2$) Suppose $I\Delta_0 + TI(\prec \omega_m, \Sigma_{n-1})$ proves the sequent $A_1, \ldots, A_k \longrightarrow B_1, \ldots, B_\ell$ and that each A_i and B_j is in Σ_n and that $\vec{c}$ are all the variables free in the sequent. Then there are functions $f_1, \ldots, f_\ell$ which are $\prec \omega_m$-primitive recursive in Σ_{n-1} and are Σ_n-definable in $I\Delta_0 + TI(\prec \omega_{m+1}, \Sigma_{n-2})$ such that $I\Delta_0 + TI(\prec \omega_{m+1}, \Sigma_{n-2})$ proves*

$$Wit^n_{A_1}(w_1, \vec{c}), \ldots, Wit^n_{A_k}(w_k, \vec{c}) \longrightarrow Wit^n_{B_1}(f_1(\vec{w}, \vec{c}), \vec{c}), \ldots, Wit^n_{B_\ell}(f_\ell(\vec{w}, \vec{c}), \vec{c}).$$

The proof of Lemma 13 is exactly like the proof of Lemma 10 except that now the definitions of the functions $f_1, \ldots, f_k$ and the proofs that they produce the correct witnesses are now carried out in $I\Delta_0 + TI(\prec \omega_{m+1}, \Sigma_{n-2})$ — the reader should refer back to the earlier proof to verify that it works out as claimed. □

Now suppose $A_1, \ldots, A_k \longrightarrow B_1, \ldots, B_\ell$ is a sequent of Σ_n-formulas which is provable in $I\Delta_0 + TI(\prec \omega_m, \Sigma_{n-1})$. By the just stated lemma and from the definition of Wit, $I\Delta_0 + TI(\prec \omega_m, \Sigma_{n-1})$ proves

$$Wit^n_{A_1}(w_1, \vec{c}), \ldots, Wit^n_{A_k}(w_k, \vec{c}) \longrightarrow B_1(\vec{c}), \ldots, B_\ell(\vec{c})$$

which, via $\exists$*:left* inferences gives

$$A_1(\vec{c}), \ldots, A_k(\vec{c}) \longrightarrow B_1(\vec{c}), \ldots, B_\ell(\vec{c}).$$

Q.E.D. Theorem 11

THEOREM 14 *Let* $m \geq 2$ *and* $n \geq 1$ *and* $1 \leq k \leq n-1$. *Then* $I\Delta_0 + TI(\prec \omega_m, \Sigma_{n-1}) \vdash TI(\prec \omega_{m+k}, \Sigma_{n-1-k})$ *and* $I\Delta_0 + TI(\prec \omega_m, \Sigma_{n-1})$ *is conservative over the theory* $I\Delta_0 + TI(\prec \omega_{m+k}, \Sigma_{n-1-k})$ *with respect to* Π_{n+2-k}*-consequences.*

Proof Apply Theorem 11 k times. □

COROLLARY 15 *Let* $n \geq 1$. *The theory* $I\Sigma_n$ *contains and is* Π_3*-conservative over the theory* $I\Delta_0 + TI(\prec \omega_{n+1}, \Delta_0)$.

Proof Take $m = 2$; since $I\Sigma_n$ is equal to $I\Delta_0 + TI(\prec \omega_2, \Sigma_{n-1})$ the previous theorem with $k = n-1$ yields the corollary. □

Now we are ready to prove the theorem characterizing the Σ_j-definable functions of $I\Delta_0 + TI(\prec \omega_m, \Sigma_{n-1})$ and of $I\Sigma_n$ for all $1 \leq j \leq n$.

THEOREM 16 *Let* $m \geq 2$ *and* $1 \leq j \leq n$.

(a) *If* $j > 1$ *then the* Σ_j*-definable functions of* $I\Delta_0 + TI(\prec \omega_m, \Sigma_{n-1})$ *are precisely the functions which are* $\prec \omega_{m+n-j}$*-primitive recursive in* Σ_{j-1}.

(b) *(For* $j = 1$.*) The* Σ_1*-definable functions (i.e., the provably recursive functions) of* $I\Delta_0 + TI(\prec \omega_m, \Sigma_{n-1})$ *are precisely the functions which are* $\prec \omega_{m+n-1}$*-primitive recursive.*

THEOREM 17 *Suppose* $1 \leq j \leq n$. *The functions which are* Σ_j*-definable in* $I\Sigma_n$ *are precisely the functions which are* $\prec \omega_{n-j+2}$*-primitive recursive in* Σ_{j-1}.

THEOREM 18 *Let* $n \geq 1$. *The provably total functions of* $I\Sigma_n$ *are precisely the* $\prec \omega_{n+1}$*-primitive recursive functions.*

Proof The proof of Theorem 16 is phrased for $j > 1$, but applies equally well to the $j = 1$ case. Suppose $F(\vec{z})$ is Σ_j-defined by $I\Delta_0 + TI(\prec \omega_m, \Sigma_{n-1})$ proving $(\forall \vec{z})(\exists! y)A(y, \vec{z})$ where $A \in \Sigma_j$. By Theorem 14 with $k = n-j$, $I\Delta_0 + TI(\prec \omega_{m+n-j}, \Sigma_{j-1})$ also proves the Π_{j+1}-sentence $(\forall \vec{z})(\exists! y)A$; that is, it also Σ_j-defines f. Hence, by Theorem 7, $F(\vec{z})$ is $\prec \omega_{m+n-j}$-primitive recursive in Σ_{j-1}. Conversely, every $\prec \omega_{m+n-j}$-primitive recursive in Σ_{j-1} function is Σ_j-definable in $I\Delta_0 + TI(\prec \omega_{m+n-j}, \Sigma_{j-1})$, and hence in $I\Delta_0 + TI(\prec \omega_m, \Sigma_{n-1})$, by Theorems 7 and 14. That proves Theorem 16. Theorems 17 and 18 are corollaries of Theorem 16, since $I\Sigma_n$ is the same theory as $I\Delta_0 + TI(\prec \omega_2, \Sigma_{n-1})$. □

Theorem 18 immediately implies the well-known fact that the provably total functions of Peano arithmetic are precisely the $\prec \epsilon_0$-primitive recursive functions.

4. Π_{n+1}-induction rule versus Σ_n induction axiom

This section presents a sketch for a proof of Parsons's theorem on the conservativity of a restricted Π_{n+1}-induction rule over the usual Σ_n-induction axiom — this proof is based on the witness function method. For reasons of length we omit the details of the proof.

The Π_{n+1}*-strict induction rule* allows inferences of the form

$$\frac{\longrightarrow A(0) \qquad A(b) \longrightarrow A(b+1)}{\longrightarrow A(t)}$$

where b is the eigenvariable and occurs only as indicated, t is any term and A is in Π_{n+1}. Note that no side formulas are allowed (otherwise it would be equivalent to the Π_{n+1}-induction axiom). The strict induction rule is equivalent to what Parsons calls the "induction rule" modified only slightly to fit in the framework of the sequent calculus. By free-cut elimination any sequent of Π_{n+1}-formulas which is provable in $I\Delta_0$ plus the Π_{n+1}-strict induction rule has a proof in which every formula is in Π_{n+1}.

Notation Π_{n+1}-IR denotes the theory of arithmetic $I\Delta_0$ plus the Π_{n+1}-strict induction rule. This system is always presumed to be formalized in the sequent calculus.

It is not too difficult to see that Π_{n+2}-IR proves the Σ_n induction axioms, for all $n \geq 0$. To prove this, if $A(b) \in \Sigma_n$, use the strict induction rule on the formula

$$[A(0) \wedge (\forall x)(A(x) \rightarrow A(x+1))] \rightarrow A(b)$$

with respect to the variable b.

THEOREM 19 *(Parsons [22]) Let $n \geq 1$. A Π_{n+1}-sentence is a theorem of $I\Sigma_n$ iff it is a consequence of Π_{n+1}-IR.*

Parsons's proof of Theorem 19 was based on the Gödel Dialectica interpretation; other proof-theoretic proofs of Theorem 19 have been given in [19, 25]. The main novelty of our proof outlined below is that it uses the witness function method directly.

Proof (Outline): The easy direction is that if $I\Sigma_n \vdash A$ where $A \in \Pi_{n+1}$, then Π_{n+1}-IR also proves A. Since $A \in \Pi_{n+1}$, A is expressible as $(\forall \vec{x})B(\vec{x})$ where $B \in \Sigma_n$; it suffices to show that Π_{n+1}-$IR \vdash B(\vec{c})$. By free-cut elimination, there is a $I\Sigma_n$-proof P of $B(\vec{c})$ such that every formula occuring in P is a Σ_n-formula. We now can prove by induction on the number of inferences in

this proof that every sequent in P is a consequence of Π_{n+1}-IR. The only difficult case is the induction inferences, which are of the form

$$\frac{\Gamma, A(b) \longrightarrow A(b+1), \Delta}{\Gamma, A(0) \longrightarrow A(t), \Delta}$$

Letting $D(b)$ be the formula $(\bigwedge \Gamma \wedge A(b)) \vee (\bigvee \Delta)$, the upper sequent is logically equivalent to $D(b) \rightarrow D(b+1)$ and the lower sequent is logically equivalent to $D(0) \rightarrow D(t)$. And if Π_{n+1}-IR proves the upper sequent, then it also proves the lower sequent by use of the strict induction rule on the formula $D(0) \rightarrow D(b)$, which, as a Boolean combination of Σ_n-formulas is logically equivalent to a Π_{n+1}-formula.

For the hard direction of Theorem 19, we need the next lemma. We let PRA_n be a set of function symbols for the functions which are primitive recursive in Σ_n. By Theorem 8, each function symbol in PRA_{n-1} represents a function which is Σ_n-definable in $I\Sigma_n$ — we may augment the language of $I\Sigma_n$ with these function symbols, provided we are careful not to use them in induction formulas. In the next lemma, the notation $\vec{x}_i$ denotes a vector of variables and $||\vec{x}_i||$ denotes the number (possibly zero) of variables in the vector.

LEMMA 20 *Suppose $A_i(\vec{x}_i, \vec{c})$ and $B_j(\vec{y}_j, \vec{c})$ are Σ_n-formulas, for $1 \leq i \leq k$ and $1 \leq j \leq \ell$, and that Π_{n+1}-IR proves the sequent*

$$(\forall \vec{x}_1) A_1(\vec{x}_1, \vec{c}), \ldots, (\forall \vec{x}_k) A_k(\vec{x}_k, \vec{c}) \longrightarrow (\forall \vec{y}_1) B_1(\vec{y}_1, \vec{c}), \ldots, (\forall \vec{y}_\ell) B_\ell(\vec{y}_\ell, \vec{c}). \quad (8)$$

Let $f_1, \ldots, f_k$ be new function symbols so that f_i has arity $||\vec{x}_i|| + ||c||$. Then there are terms $t_i(\vec{y}_1, \ldots, \vec{y}_\ell, \vec{c})$ in the language $PRA_{n-1} \cup \{f_1, \ldots, f_k\}$, for $1 \leq i \leq \ell$, such that $I\Sigma_n$ proves

$$\begin{aligned}(\forall \vec{x}_1) Wit^n_{A_1}(f_1(\vec{x}_1, \vec{c}), \vec{x}, \vec{c}), \ldots, (\forall \vec{x}_k) Wit^n_{A_k}(f_k(\vec{x}_k, \vec{c}), \vec{x}, \vec{c}) \\ \longrightarrow Wit^n_{B_1}(t_1, \vec{y}_1, \vec{c}), \ldots, Wit^n_{B_\ell}(t_\ell, \vec{y}_\ell, \vec{c}). \quad (9)\end{aligned}$$

Theorem 19 follows immediately from Lemma 20 with $k = 0$ and $\ell = 1$ and from the fact that every PRA_{n-1}-function is definable in $I\Sigma_n$. For reasons of length, we omit the proof of Lemma 20: the general idea of the proof is a relatively straightforward use of the witness function method; however, it requires the development of some deep facts about primitive recursive (in Σ_n) functions. An important feature of the lemma is that each term t_i may involve all of $\vec{y}_1, \ldots, \vec{y}_\ell$.

A second theorem of Parsons is that Theorem 19 also holds with the addition of the $B\Sigma_n$-collection axiom:

THEOREM 21 *(Parsons [22]) Let $n \geq 1$. The Π_{n+1}-consequences of Π_{n+1}-$IR + B\Sigma_n$ are the same as the Π_{n+1}-consequences of $I\Sigma_n$.*

Proof (Outline) Recall that $B\Pi_{n-1}$ is equivalent to $B\Sigma_n$, relative to the base theory $I\Delta_0$. The $B\Pi_{n-1}$ axioms contain unbounded quantifiers in the scope of bounded quantifiers, so it is not possible to use free-cut elimination to force a proof in Π_{n+1}-$IR + B\Sigma_n$ to contain only Π_{n+1}-formulas. We let Π_n^+ denote the set of formulas which have n blocks of like unbounded quantifiers, starting with a block of universal quantifiers, allowing arbitrary bounded quantifiers to be included in the first block of unbounded quantifiers (see the next section for a careful definition of the analogous class Σ_n^+). Now, temporarily define the set of Σ_n^* formulas to be the formulas which are of one of the following forms: (1) $(\exists \vec{y})B(\vec{x})$ where $B \in \Pi_{n-1}^+$ or (2) $(\forall z \leq t)(\exists y_1)B(y_1, z, \vec{c})$ where $B \in \Pi_{n-1}$. We also define the Π_{n+1}^* formulas to be the formulas which are either Π_{n+1} or Σ_n^*. Since the $B\Pi_{n-1}$ axioms can be formulated in the form $A \longrightarrow A'$ with A and A' both in Σ_n^*, the free-cut elimination theorem implies that if $\Gamma \longrightarrow \Delta$ is a sequent of Π_{n+1}^*-formulas provable in Π_{n+1}-$IR + B\Pi_n$, then this sequent has a proof in which every formula is a Π_{n+1}^*-formula. The notion of "witness" can be generalized as follows: if $A(\vec{c})$ is a Σ_n^*-formula in one of the above forms; then, if A is of form (1), $Wit_A^{*n}(w, \vec{c})$ is defined just like $Wit_A^n(w, \vec{c})$ was and, if A is of form (2) then $Wit_A^{*n}(w, \vec{c})$ is defined to be the formula

$$(\forall z \leq t) Wit^n_{(\exists y_1)B}((w)_z, z, \vec{c}).$$

LEMMA 22 *Suppose $A_i(\vec{x}_i, \vec{c})$ and $B_j(\vec{y}_j, \vec{c})$ are Σ_n^*-formulas, for $1 \leq i \leq k$ and $1 \leq j \leq \ell$, and that Π_{n+1}-$IR + B\Sigma_n$ proves the sequent*

$$(\forall \vec{x}_1)A_1(\vec{x}_1, \vec{c}), \ldots, (\forall \vec{x}_k)A_k(\vec{x}_k, \vec{c}) \longrightarrow (\forall \vec{y}_1)B_1(\vec{y}_1, \vec{c}), \ldots, (\forall \vec{y}_\ell)B_\ell(\vec{y}_\ell, \vec{c}).$$

Let $f_1, \ldots, f_k$ be new function symbols so that f_i has arity $||\vec{x}_i|| + ||c||$. Then there are terms $t_i(\vec{y}_1, \ldots, \vec{y}_\ell, \vec{c})$ in the language $PRA_{n-1} \cup \{f_1, \ldots, f_k\}$, for $1 \leq i \leq \ell$, such that $I\Sigma_n$ proves

$$(\forall \vec{x}_1) Wit^{*n}_{A_1}(f_1(\vec{x}_1, \vec{c}), \vec{x}, \vec{c}), \ldots, (\forall \vec{x}_k) Wit^{*n}_{A_k}(f_k(\vec{x}_k, \vec{c}), \vec{x}, \vec{c}) \longrightarrow Wit^{*n}_{B_1}(t_1, \vec{y}_1, \vec{c}), \ldots, Wit^{*n}_{B_\ell}(t_\ell, \vec{y}_\ell, \vec{c}). \quad (10)$$

We omit the proof of the lemma and the rest of Theorem 21.

Finally, it should be remarked that Π_{n+1}-$IR + B\Sigma_n$ does not contain $I\Sigma_n$. This can be proved by noting that Π_{n+1}-$IR + I\Sigma_n$ is not Π_{n+2}-conservative over $I\Sigma_n$. For example, with $n = 1$, let $A(k, m)$ be the Ackermann function so that the functions $f_k(m) = A(k, m)$ are all primitive recursive and so

that each primitive recursive function is eventually dominated by f_k for sufficiently large k. Let $A^*(k,m,y)$ be the graph of the Ackermann function; it is well-known that $A^*(k,m,y)$ is Δ_0 (for us it is sufficient that it is Σ_1). Now, it is easy to see that $I\Sigma_1$ proves $(\forall x)(\exists y)A^*(0,x,y)$ and

$$(\forall x)(\exists y)A^*(b,x,y) \longrightarrow (\forall x)(\exists y)A^*(b+1,x,y).$$

Thus Π_2-$IR + I\Sigma_1 \vdash (\forall k)(\forall x)(\exists y)A^*(k,x,y)$. But the Ackermann function is not primitive recursive, hence not Σ_1-definable in $I\Sigma_1$. Thus Π_2-$IR + I\Sigma_1$ is not Π_2-conservative over $I\Sigma_1$ and thus not equal to Π_2-IR and not a subtheory of Π_2-$IR + B\Sigma_1$.

To show Π_{n+1}-$IR + B\Sigma_n \not\vdash I\Sigma_n$ for $n > 1$, use essentially the same argument, but use 'primitive recursive in Σ_{n-1}' in place of 'primitive recursive' and use a suitable replacement of the Ackermann function that dominates the functions primitive recursive in Σ_{n-1}.

5. Conservativity of collection over induction

In this section we prove the well-known theorem that the $B\Sigma_{n+1}$-collection axioms are Π_{n+2}-conservative over $I\Sigma_n$. The proof method does not use the witness function method per se, but it involves an induction on the length of free-cut free proofs similar to the methods of earlier sections. Earlier proofs of this theorem include Parsons [22] and Paris-Kirby [21]; see in addition, [3, 25]. The advantage of our proof below is that it gives a direct and elementary proof-theoretic proof.

Recall that the $B\Sigma_{n+1}$-collection axioms are equivalent to the $B\Pi_n$-collection axioms. In the sequent calculus, the $B\Pi_n$-collection axioms are of the form

$$(\forall x \leq a)(\exists y)A(x,y) \longrightarrow (\exists z)(\forall x \leq a)(\exists y \leq z)A(x,y)$$

where $A \in \Pi_n$ and may contain free variables besides x, y. In the above sequent there are bounded quantifiers outside of unbounded quantifiers so the formulas are not, strictly speaking, Σ_{n+1}-formulas. Accordingly, we define a generalized form of Σ_{n+1}-formulas that will be allowed to appear in free-cut free proofs.

Definition The class Σ^+_{n+1} of formulas is defined inductively by

(1) $\Pi_n \subseteq \Sigma^+_{n+1}$,

(2) If $A \in \Sigma^+_{n+1}$, then $(\exists x)A$, $(\exists x \leq t)A$ and $(\forall x \leq t)A$ are in Σ^+_{n+1}, where t is any term not involving x.

If s is a term and A is a Σ^+_{n+1}-formula, then $A^{\leqslant s}$ is the formula obtained by bounding unbounded existential quantifiers in the outermost block of quantifiers of A by the term s; namely,

Definition Fix n and suppose $A \in \Sigma^+_{n+1}$.

(1) If $A \in \Pi_n$, then $A^{\leqslant s}$ is A.

(2) If A is $(\exists x)B$ and $A \notin \Pi_n$, then $A^{\leqslant s}$ is $(\exists x \leq s)B$.

(3) If A is $(Qx \leq t)B$ then $A^{\leqslant s}$ is $(Qx \leq t)(B^{\leqslant s})$.

Let $\Gamma \longrightarrow \Delta$ be a sequent $A_1, \ldots, A_k \longrightarrow B_1, \ldots, B_\ell$ of Σ^+_{n+1}-formulas. Then $\Gamma^{\leqslant s}$ is the formula $\bigwedge_{i=1}^{k} A_i^{\leqslant s}$ and $\Delta^{\leqslant s}$ is the formula $\bigvee_{j=1}^{\ell} B_j^{\leqslant s}$. This notation should cause no confusion since antecedents and succedents are always clearly distinguished.

If $\vec{c} = c_1, \ldots c_s$ is a vector of free variables, then $\vec{c} \leq u$ abbreviates the formula $c_1 \leq s \wedge \cdots \wedge c_s \leq u$. $(\forall \vec{c} \leq u)$ and $(\exists \vec{c} \leq u)$ abbreviate the corresponding vectors of bounded quantifiers.

THEOREM 23 *($n \geq 1$) Suppose $\Gamma \longrightarrow \Delta$ is a sequent of Σ^+_{n+1}-formulas that is provable in $I\Delta_0 + B\Sigma_{n+1}$. Let $\vec{c}$ include all the free variables occurring in $\Gamma \longrightarrow \Delta$. Then*

$$I\Sigma_n \vdash (\forall u)(\exists v)(\forall \vec{c} \leq u)\Big(\Gamma^{\leqslant u} \rightarrow \Delta^{\leqslant v}\Big).$$

Intuitively, the theorem is saying that given a bound u on the sizes of the free variables and on the sizes of the witness for the formulas in Γ, there is a bound v for the values of a witness for a formula in Δ.

Theorem 23 immediately implies the main theorem of this section:

THEOREM 24 *$I\Delta_0 + B\Sigma_{n+1}$ is Π_{n+2}-conservative over $I\Sigma_n$.*

Recall that $I\Delta_0 + B\Sigma_{n+1} \vdash I\Sigma_n$. Before proving Theorem 23, we establish the following lemma (due to Clote and Hájek).

LEMMA 25 *($n \geq 1$) Let $B(\vec{c}, d) \in \Pi_n$. Then*

$$I\Sigma_n \vdash (\forall u)(\exists v)(\forall \vec{c} \leq u)[(\forall x)B(\vec{c}, x) \leftrightarrow (\forall x \leq v)B(\vec{c}, x)].$$

The formula of Lemma 25 is called the Σ_n-strong replacement principle.

Proof Let s be the length of the vector $\vec{c}$. We reason inside $I\Sigma_n$. Let $C(\vec{c}, d)$ be the Σ_n-formula $\neg B(\vec{c}, d)$. Let $Num(u, \ell)$ be the formula expressing

$\exists\langle \vec{c}_1, d_1, \ldots, \vec{c}_\ell, d_\ell\rangle$ s.t. $\vec{c}_1, \ldots, \vec{c}_\ell$ are distinct s-tuples $\leq u$ and $C(\vec{c}_i, d_i)$ holds for all $1 \leq i \leq \ell$.

Of course, this asserts that there are $\geq \ell$ distinct values of $\vec{c} \leq u$ for which $(\exists x)C(\vec{c}, x)$ holds. Now Num is a Σ_n-formula and $Num(\vec{c}, (u+1)^s + 1)$ is clearly false; so by $I\Sigma_n$, there is a value ℓ_0 such that $Num(\vec{c}, \ell_0)$ but not $Num(\vec{c}, \ell_0 + 1)$. Given $\vec{c}_1, d_1, \ldots, \vec{c}_{\ell_0}, d_{\ell_0}$ witnessing $Num(\vec{c}, \ell_0)$, let $v = \max\{d_1, \ldots, d_{\ell_0}\}$. It follows that

$$(\forall \vec{c} \leq u)\Big((\exists x)C(\vec{c}, x) \leftrightarrow (\exists x \leq v)C(\vec{c}, x)\Big)$$

which is what we needed to prove. □

Proof of Theorem 23: By free-cut elimination, $\Gamma \longrightarrow \Delta$ has a sequent calculus proof P in which every formula is a Σ^+_{n+1}-formula. (Since we allow bounded quantifiers in Σ^+_{n+1}-formulas, it is convenient to work in the sequent calculus LKB with inference rules for bounded quantifiers [2].) We prove the theorem by induction on the number of inferences in P. The proof splits into cases depending on the last inference of P. The hardest case, $\forall$*:right* is saved for last.

Case (1): If P has no inferences and $\Gamma \longrightarrow \Delta$ is an initial sequent, then either $\Gamma \longrightarrow \Delta$ is a logical, equality or arithmetic axiom, containing only Δ_0-formulas, and the theorem is trivial, or $\Gamma \longrightarrow \Delta$ is a $B\Sigma_{n+1}$ axiom. In the latter case, taking $v = u$, it is immediate that $I\Sigma_n$ proves

$$(\forall x \leq a)(\exists y \leq u)A(x, y) \rightarrow (\exists z \leq u)(\forall x \leq a)(\exists y \leq z)A(x, y)$$

and the theorem holds.

Case(2): Suppose the last inference of P is a structural inference, a propositional inference or a $\forall$*:left* or $\forall \leq$*:left* inference. The inference may have either one or two premisses:

$$\frac{\Pi \longrightarrow \Lambda}{\Gamma \longrightarrow \Delta} \qquad \text{or} \qquad \frac{\Pi_1 \longrightarrow \Lambda_1 \quad \Pi_2 \longrightarrow \Lambda_2}{\Gamma \longrightarrow \Delta}$$

It is easily checked that, in the first case we have that $I\Sigma_n$ proves $\Gamma^{\leqslant u} \rightarrow \Pi^{\leqslant u}$ and $\Lambda^{\leqslant v} \rightarrow \Delta^{\leqslant v}$ and, in the second case we have that $I\Sigma_n$ proves $\Gamma^{\leqslant u} \rightarrow \Pi_1^{\leqslant u} \wedge \Pi_2^{\leqslant u}$ and $\Lambda_1^{\leqslant v} \wedge \Lambda_2^{\leqslant v} \rightarrow \Delta^{\leqslant v}$. In the first case, the induction hypothesis states that $I\Sigma_n$ proves

$$(\exists v)(\forall \vec{c} \leq u)\Big(\Pi^{\leqslant u} \rightarrow \Lambda^{\leqslant v}\Big)$$

from which $(\exists v)(\forall \vec{c} \leq u)(\Gamma^{\leqslant u} \to \Delta^{\leqslant v})$ follows. In the second case, by the induction hypothesis, $I\Sigma_n$ proves

$$(\exists v_i)(\forall \vec{c} \leq u)\big(\Pi_i^{\leqslant u} \to \Lambda_i^{\leqslant v_i}\big)$$

for $i = 1, 2$. Taking $v = \max\{v_1, v_2\}$ and noting that $I\Sigma_n$ proves $v_i \leq v \wedge \Lambda_i^{\leqslant v_i} \to \Lambda_i^{\leqslant v}$, we get that $I\Sigma_n$ proves $(\exists v)(\forall \vec{c} \leq u)(\Gamma^{\leqslant u} \to \Delta^{\leqslant v})$.

Case (3): Suppose the final inference of P is an $\exists$*:right* inference:

$$\frac{\Gamma \longrightarrow B(\vec{c}, t(\vec{c})), \Lambda}{\Gamma \longrightarrow (\exists x)B(\vec{c}, x), \Lambda}$$

We reason inside $I\Sigma_n$ as follows: given arbitrary u, there is (by the induction hypothesis) a v' such that

$$(\forall \vec{c} \leq u)\big(\Gamma^{\leqslant u} \to B^{\leqslant v'}(\vec{c}, t(\vec{c})) \vee \Lambda^{\leqslant v'}\big).$$

Letting $v = \max\{v', t(u, \ldots, u)\}$ we have that $t(\vec{c}) \leq v$ for all $\vec{c} \leq u$ (since the language has 0, S, $+$ and $\cdot$ as the only function symbols). This v makes the theorem true. The case where the last inference of P is a $\exists \leq$*:right* is similar.

Case (4): Suppose the last inference of P is an $\exists$*:left*:

$$\frac{A(\vec{c}, d), \Gamma \longrightarrow \Delta}{(\exists x)A(\vec{c}, x), \Gamma \longrightarrow \Delta}$$

where d is the eigenvariable occuring only where indicated. The induction hypothesis is that $I\Sigma_n$ proves

$$(\forall u)(\exists v)(\forall \vec{c}, d \leq u)\big(A^{\leqslant u}(\vec{c}, d) \wedge \Gamma^{\leqslant u} \to \Delta^{\leqslant v}\big).$$

This is equivalent to

$$(\forall u)(\exists v)(\forall \vec{c} \leq u)\big((\exists d \leq u)A^{\leqslant u}(\vec{c}, d) \wedge \Gamma^{\leqslant u} \to \Delta^{\leqslant v}\big)$$

which is what we needed to prove.

Case (5): The $\exists \leq$*:left* inference is a little more subtle. If the final inference of P is

$$\frac{d \leq t(\vec{c}), A(\vec{c}, d), \Gamma \longrightarrow \Delta}{(\exists x \leq t(\vec{c}))A(\vec{c}, x), \Gamma \longrightarrow \Delta}$$

we reason inside $I\Sigma_n$ as follows. Let u be arbitrary, there is a v' such that

$$(\forall \vec{c}, d \leq u)\big(d \leq t(\vec{c}) \wedge A^{\leqslant u}(\vec{c}, d) \wedge \Gamma^{\leqslant u} \to \Delta^{\leqslant v'}\big). \tag{11}$$

Let $u' = \max\{u, t(\vec{u})\}$; by the induction hypothesis, there is a v such that (11) holds with u', v in place of u, v'. Now let $\vec{c} \leq u$ and suppose $(\exists x \leq t)A^{\leq u}(\vec{c}, x) \wedge \Gamma^{\leq u}$. Clearly, this implies $(\exists x \leq u')(x \leq t \wedge A^{\leq u'} \wedge \Gamma^{\leq u'})$. Taking d to be this x, we have $\Delta^{\leq v}$ holds.

Case (6): Suppose the last inference of P is a *Cut*:

$$\frac{\Gamma_1 \longrightarrow \Delta_1, A \qquad A, \Gamma_2 \longrightarrow \Delta_2}{\Gamma_1, \Gamma_2 \longrightarrow \Delta_1, \Delta_2}$$

We reason inside $I\Sigma_n$. Suppose u is arbitrary and $\Gamma_1^{\leq u} \wedge \Gamma_2^{\leq u}$. Pick v_1, depending only on u by the induction hypothesis, so that $\Delta^{\leq v_1} \vee A^{\leq v_1}$. Let $u_2 = \max\{v_1, u\}$. By the induction hypothesis, there is a $v \geq v_1$ depending only on u_2 so that if $A^{\leq v_1}$ holds, then $\Delta_2^{\leq v}$ holds. Now clearly either $\Delta_1^{\leq v}$ or $\Delta_2^{\leq v}$ holds. Since v depends only on u, this proves this case.

Case (7): Suppose the final inference of P is a $\forall$*:right*:

$$\frac{\Gamma \longrightarrow B(\vec{c}, d), \Lambda}{\Gamma \longrightarrow (\forall x)B(\vec{c}, x), \Lambda}$$

Note $B \in \Pi_n$ since $(\forall x)B$ must be a Σ_{n+1}^+-formula. We reason inside $I\Sigma_n$. Let u be arbitrary. By Σ_n-strong replacement (Lemma 25) there is a $u' \geq u$ such that

$$(\forall \vec{c} \leq u)\Big((\forall x)B(\vec{c}, x) \leftrightarrow (\forall x \leq u')B(\vec{c}, u')\Big).$$

Let $v \geq u'$ be given by the induction hypothesis so that

$$(\forall \vec{c}, d \leq u')\Big(\Gamma^{\leq u'} \rightarrow B(\vec{c}, d) \vee \Delta^{\leq v}\Big). \tag{12}$$

Now let $\vec{c} \leq u$ be arbitrary such that $\Gamma^{\leq u}$. We need to show $(\forall x)B(\vec{c}, x) \vee \Delta^{\leq v}$. Suppose not, then there is a $d \leq u'$ such that $\neg B(\vec{c}, d)$, and by (12), $\Delta^{\leq v}$ holds, which is a contradiction.

The case where the final inference of P is a $\forall \leq$*:left* inference is similar, although Lemma 25 is not needed.
Q.E.D. Theorem 23

It would be interesting to give a similar proof that Π_{n+1}-$IR + B\Sigma_n$ is Π_{n+1}-conservative over Π_{n+1}-IR, in place of the more complicated and omitted proof of Theorem 21 above.

6. Analogies between bounded and Peano arithmetic

The witness function method has been extensively used characterizing definable functions of fragments of bounded arithmetic — the work in section 3

above gives an approach to Peano arithmetic which is very similar to some of the proofs used earlier in bounded arithmetic.

First, Theorem 8, which characterized the Σ_n-definable functions of $I\Sigma_n$ is analogous to the main theorem of Buss [2] which characterized the Σ_n^b-definable functions of S_2^n (which is axiomatized with Σ_n^b-PIND axioms). In $I\Sigma_n$, the Σ_n-definable functions are precisely the functions primitive recursive in Σ_{n-1}; whereas, in S_2^n, the Σ_n^b-definable functions are precisely the functions polynomial time computable with respect to a (usual) Σ_{n-1}^p-oracle. It should be noted that a usual Σ_{n-1}^p-oracle is equivalent to a witness oracle for Σ_{n-1}^p with respect to polynomial time computation, since there is an a-priori bound on the size of a witness and a witness value may be queried bit-by-bit. The proofs of these two theorems are analogous as well.

Second, Theorem 11, which stated that $I\Delta_0 + TI(\prec \omega_m, \Sigma_{n-1})$ is Π_{n+1}-conservative over $I\Delta_0 + TI(\prec \omega_{m+1}, \Sigma_{n-2})$ is analogous to the result of [4] that S_2^n is $\forall\Sigma_n^b$-conservative over T_2^{n-1}. To see the analogy more sharply, note on one hand $I\Delta_0 + TI(\prec \omega_m, \Sigma_{n-1})$ and $I\Delta_0 + TI(\prec \omega_{m+1}, \Sigma_{n-2})$ are equivalent to $I\Delta_0 + TI(\prec \omega_m, \Pi_n)$ and $I\Delta_0 + TI(\prec \omega_{m+1}, \Pi_{n-1})$ (respectively), which are axiomatized with transfinite induction on Π_n-formulas up to ordinals $\prec \omega_m$ and on Π_{n-1} formulas up to ordinals $\prec \omega^{\omega_m}$; and on the other hand, S_2^n may be axiomatized by induction (PIND) on Π_n^b-formula up to lengths $|x|$ and T_2^{n-1} may be axiomatized by induction on Π_n^b-formulas up to $2^{|x|}$. So both conservation theorems give situations where the complexity of induction formulas may be reduced by one block of quantifier alternation in exchange for "exponentiating" the length of induction. Another theorem of this type is the result of [6] that R_3^n is $\forall\Sigma_n^b$-conservative over S_3^{n-1}.

Witness oracles have been applied to bounded arithmetic in [18] and in [6]. Another area of contact between bounded arithmetic and Peano arithmetic may be found in Kaye [14] who gives a proof that $I\Sigma_n \neq B\Sigma_{n+1}$ based on methods used earlier by [18] to show that if $T_2^{n+1} = S_2^{n+1}$ then the polynomial time hierarchy collapses.

We conclude with a partial characterization of the Σ_j-definable functions of $I\Sigma_n$ when $j > n$:

Definition Let A be a formula; w.l.o.g. all negations in A are on atomic formulas. The *counterexample oracles of* A are the witness oracles $U_{(\exists x)\neg B}$ for $(\forall x)B$ a subformula of A.

THEOREM 26 *Let $j > n \geq 1$. Suppose $I\Sigma_n \vdash (\forall x)(\exists! y)A(x,y)$ where $A \in \Sigma_j$. Then the function $f : x \mapsto y$, such that $(\forall x)A(x, f(x))$, is primitive recursive in Σ_{n-1} and in the counterexample oracles for A.*

The same holds for $I\Delta_0 + TI(\prec \omega_m, \Sigma_{n-1})$ with "primitive recursive" replaced by "$\prec \omega_m$-primitive recursive".

The proof of this theorem is analogous to the proof of Theorems 7 and 8 except that the $\forall$*:right* cases of the proof now have to accommodate the fact that a $\forall$*:right* quantifier may be an ancestor of a quantifier in $(\exists y)A(c,y)$. Of course a counterexample oracle for A is exactly what is needed for this case.

Theorem 26 can be extended to partially characterize the Σ^b_j-definable functions of T^{n-1}_2 or S^n_2 when $j > n$; namely,

THEOREM 27 *(See [18, 23, 17]) Let* $j > n \geq 1$.

(a) *Suppose* $A \in \Sigma^b_j$ *and* $S^n_2 \vdash (\forall x)(\exists! y)A(x,y)$. *Then the function* f *such that* $(\forall x)A(x,f(x))$ *can be computed by a polynomial time Turing machine with an oracle for* Σ^p_{n-1} *and with the counterexample oracles of* A.

(b) *Suppose* $A \in \Sigma^b_j$ *and* $T^{n-1}_2 \vdash (\forall x)(\exists! y)A(x,y)$. *Then the function* f *such that* $(\forall x)A(x,f(x))$ *can be computed by a polynomial time Turing machine which makes a constant number of queries to an oracle for* Σ^p_{n-1} *and to the counterexample oracles of* A.

References

[1] B. ALLEN, *Arithmetizing uniform NC.* Ph.D. thesis, University of Hawaii, 1989.

[2] S. R. BUSS, *Bounded Arithmetic*, Bibliopolis, 1986. Revision of 1985 Princeton University Ph.D. thesis.

[3] ——, *A conservation result concerning bounded theories and the collection axiom*, Proceedings of the American Mathematical Society, 100 (1987), pp. 709–716.

[4] ——, *Axiomatizations and conservation results for fragments of bounded arithmetic*, in Logic and Computation, proceedings of a Workshop held Carnegie-Mellon University, 1987, vol. 106 of Contemporary Mathematics, American Mathematical Society, 1990, pp. 57–84.

[5] S. R. BUSS AND J. KRAJÍČEK, *An application of Boolean complexity to separation problems in bounded arithmetic.* in preparation.

[6] S. R. BUSS, J. KRAJÍČEK, AND G. TAKEUTI, *Provably total functions in bounded arithmetic theories* R^i_3, U^i_2 *and* V^i_2. To appear in Arithmetic, Proof theory and Computational Complexity, ed. P. Clote and J. Krajíček, Oxford University Press.

[7] P. CLOTE, *A first-order theory for the parallel complexity class NC*, Tech. Rep. BCCS-89-01, Boston College, Jan. 1989. To be published in expanded form jointly with G. Takeuti.

[8] P. CLOTE AND G. TAKEUTI, *Exponential time and bounded arithmetic (extended abstract)*, in Structure in Complexity, Lecture Notes in Computer Science #223, Springer Verlag, 1986, pp. 77–103.

[9] G. GENTZEN, *Neue Fassung des Widerspruchsfreiheitbeweis für der reinen Zahlentheorie*, Forschungen zur Logik end zur Grundlegung der exacten Wissenschaften, New Series, 4 (1938), pp. 19–44. English translation in [11], pp. 252-286.

[10] ——, *Beweisbarkeit und Unbeweisbarkeit von Anfangsfällen der transfiniten Induktion in der reinen Zahlentheorie*, Mathematische Annalen, 119 (1943), pp. 140–161. English translation in [11], pp. 287-308.

[11] ——, *Collected Papers of Gerhard Gentzen*, North-Holland, 1969. Edited by M. E. Szabo.

[12] K. GÖDEL, *Über eine bisher noch nicht benützte Erweiterung des finiten Standpunktes*, Dialectica, 12 (1958), pp. 280–287.

[13] W. A. HOWARD, *Assignment of ordinals to terms for primitive recursive functionals of finite type*, in Intuitionism and Proof Theory, North-Holland, 1970, pp. 443–458.

[14] R. KAYE, *Using Herbrand-type theorems to separate strong fragments of arithmetic.* Typeset manuscript, August 1991.

[15] J. KETONEN AND R. SOLOVAY, *Rapidly growing Ramsey functions*, Annals of Mathematics, 113 (1981), pp. 267–314.

[16] J. KRAJÍČEK, *Fragments of bounded arithmetic and bounded query classes.* To appear in Transactions of the A.M.S.

[17] ——, *No counter-example interpretation and interactive computation.* In Logic From Computer Science, Proc. of a Workshop held Nov. 13-17, 1989, Springer-Verlag, 1992, pp. 287-293.

[18] J. KRAJÍČEK, P. PUDLÁK, and G. TAKEUTI, *Bounded arithmetic and the polynomial hierarchy*, Annals of Pure and Applied Logic, 52 (1991), pp. 143–153.

[19] G. E. MINTS, *Quantifier-free and one-quantifier systems*, Journal of Soviet Mathematics, 1 (1973), pp. 71–84.

[20] J. PARIS AND L. HARRINGTON, *A Mathematical Incompleteness in Peano Arithmetic* in Handbook of Mathematical Logic, North-Holland, 1977, pp. 1133–1142.

[21] J. B. PARIS AND L. A. S. KIRBY, Σ_n*-collection schemes in arithmetic*, in Logic Colloquium '77, North-Holland, 1978, pp. 199–210.

[22] C. PARSONS, *On a number-theoretic choice schema and its relation to induction*, in Intuitionism and Proof Theory: proceedings of the summer conference at Buffalo N.Y. 1968, North-Holland, 1970, pp. 459–473.

[23] P. PUDLÁK, *Some relations between subsystems of arithmetic and the complexity of computations.* In Logic From Computer Science, Proc. of a Workshop held Nov. 13-17, 1989, Springer-Verlag, 1992, pp. 499-519.

[24] U. R. SCHMERL, *A fine structure generated by reflection formulas over primitive recursive arithmetic*, in Logic Colloquium 78, North-Holland, 1979, pp. 335–350.

[25] W. SIEG, *Fragments of arithmetic*, Annals of Pure and Applied Logic, 28 (1985), pp. 33–71.

[26] R. SOMMER, *Transfinite Induction and Hierarchies Generated by Transfinite Recursion within Peano Arithmetic*, PhD thesis, U.C. Berkeley, 1990.

[27] G. TAKEUTI, *Proof Theory*, North-Holland, 2nd ed., 1987.

Logic, Methodology and Philosophy of Science IX
D. Prawitz, B. Skyrms and D. Westerståhl (Editors)

SOME ASPECTS OF CATEGORICAL LOGIC

J. LAMBEK

Department of Mathematics and Statistics,
McGill University, Montreal, Quebec, Canada

0. Introduction

Logic, like Caesar's Gaul, used to be divided into three parts: proof theory, recursion theory and model theory. To all these category theory can make some contributions, although more in analyzing basic concepts and determining direction for further development than in addressing specific problems or pursuing technical details. On the other hand, techniques developed by logicians have proved very fruitful when lifted to the categorical level. Here I shall largely confine myself to proof theory, mentioning recursive functions only briefly and barely touching the theory of models.

The present account is largely a survey of results to which the author feels he has made some contributions; it tends to ignore parallel developments by others, in particular the work by Makkai and Reyes [1977] on coherent and infinitary logic and the study of models of elementary theories by Makkai and Paré [1990]. Nor do the many profound contributions to categorical logic by André Joyal fall within the scope of this paper. However, credit must be given to Lawvere [e.g.1969,1970] for the basic insight to encode logical concepts into categorical language. Most of the ideas discussed here have already been treated elsewhere, but a few, such as those concerned with the subcongruence completion of deductive systems, are still under investigation.

1. Categorical proof theory

Proofs can be described in various ways, one of them being in terms of *deductive systems*. A deductive system deals with *deductions*, also called "entailments", $f : A \to B$, where A and B are formulas. In labeling the deduction f, we have already been influenced by category theory. Before the advent of computer science, logicians were usually just interested in

the existence of a deduction $A \to B$ and did not consider the necessity of sometimes distinguishing several such deductions.

The deduction symbol (arrow) is assumed to be reflexive and transitive, so we always have the deduction

$$1_A : A \to A$$

and the rule

$$\frac{f : A \to B \quad g : B \to C}{gf : A \to C} .$$

As categorists we must also study the equality relation on the set of all deductions $A \to B$; in particular, we insist on the following equations:

$$f1_A = f = 1_B f , \quad (hg)f = h(gf) ,$$

where $h : C \to D$, say. The deductive system then becomes a *category*.

To get away from generalities, let us look at a particular deductive system, the *positive intuitionistic propositional calculus*. Among its formulas there is $\top$ (*true*) and there are given two ways of forming new formulas from old: $A \wedge B$(*A and B*) and $A \Rightarrow B$ (*if A then B*). We postulate the following *axioms* and rules of inference, duly labeled in the spirit of category theory:

$$\circ_A : A \to \top \qquad \pi_{AB} : A \wedge B \to A \qquad \epsilon_{CB} : (C \Rightarrow B) \wedge C \to B$$

$$\pi'_{AB} : A \wedge B \to B$$

$$\frac{f : C \to A \quad g : C \to B}{< f, g > : C \to A \wedge B} \qquad \frac{f : A \wedge C \to B}{f^* : A \to C \Rightarrow B}$$

The logician will recognize all these, for example, ϵ_{CB} as *modus ponens* and the transition from f to f^* as a form of the *deduction rule*.

A categorist will now insist on suitable equations between deductions. In fact, Lawvere [1963] introduced the notion of a *cartesian closed category*, henceforth abbreviated CCC, which can be described by imposing appropriate equations in a positive intuitionistic propositional calculus:

$$k = \circ_A ; \qquad \pi_{AB} < f, g >= f , \qquad \epsilon_{CB} < f^* \pi_{AC}, \pi'_{AC} >= f ,$$

$$\pi'_{AB} < f, g >= g , \qquad (\epsilon_{CB} < g\pi_{AC}, \pi'_{AC} >)^* = g .$$

$$< \pi_{AB} h, \pi'_{AB} h >= h ;$$

It is here assumed that

$$k : A \to \top\ , \qquad h : C \to A \wedge B\ , \qquad g : A \to (C \Rightarrow B)\ .$$

Cartesian closed categories abound in mathematics. To mention just one example, in the familiar category of sets we have:

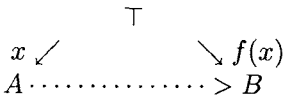
$$\begin{aligned} \top &= \{*\}\ , && \text{a typical one-element set,} \\ A \wedge B &= A \times B\ , && \text{the usual cartesian product,} \\ C \Rightarrow B &= B^C\ , && \text{the set of all functions from } C \text{ to } B\ . \end{aligned}$$

The idea that one should look at an equivalence relation (equality) between proofs was introduced, quite independently of category theory, by Prawitz [1965]. A comparison between his method and ours [L1972] was carried out by Mann [1975]. In his review of Mann's paper, Feferman [1976] suggested that two things should still be incorporated: quantifiers and equalizers. As for quantifiers, over either individual or propositional variables, this is by now well understood, see Lawvere [1970], Seely [1984,1987] and also Section 3 below. The possibility of introducing equalizers, though not perhaps in the most economic way, will be discussed in Section 5. (For a more economic way, see Seely [1984].)

2. Introducing variable arrows

A typical metatheorem in elementary logic is the *deduction theorem*, which may be stated in the present context as asserting the following: if we assume A, or rather $\top \to A$, and if, from this assumption, we can prove B, or rather $\top \to B$, then there must exist a deduction $A \to B$ not depending on this assumption.

We shall label the assumption $x : \top \to A$, then the conclusion $\varphi(x) : \top \to B$ may be viewed as a polynomial in x and the theorem asserts the existence of a deduction $f : A \to B$ not depending on x. The categorist is now inclined to draw a commutative diagram:



In fact, he would insist that there is a unique deduction f such that $fx \underset{x}{=} \varphi(x)$, where $\underset{x}{=}$ is the equality relation in the CCC of all polynomials in x.

This property is known from combinatory logic as *functional completeness* (see Rosenbloom [1950], Curry and Feys [1958], not to be confused with a different notion in universal algebra by the same name). If we write $f = \lambda_{x \in A} \varphi(x)$, it may be summarized by the following two equations:

$$(\beta) \qquad \lambda_{x \in A} \varphi(x) x \underset{x}{=} \varphi(x) \ ,$$

$$(\eta) \qquad \lambda_{x \in A}(fx) = f \ .$$

Using the general algebraic principle of substituting for free variables, we may specialize (β) thus:

$$\lambda_{x \in A} \varphi(x) a = \varphi(a) \ ,$$

where $a : \top \to A$ is any deduction of A from $\top$.

A word of warning: what we have described here is not the usual λ-calculus, where $\lambda_{x \in A} \varphi(x)$ is written not for the deduction $f : A \to B$ but for the deduction $\ulcorner f \urcorner : \top \to B \Rightarrow A$. Still, it serves to illustrate the proof theoretic interpretation of the λ-calculus, known as the Curry-Howard isomorphism, according to which formulas correspond to types and deductions to λ-terms.

For more details on the material of this and the previous section, the reader is referred to the book [LS1986].

3. Introducing variable formulas

What happens if we also allow variable formulas X, Y,etc.? If we also admit universal quantification over variable formulas, we obtain *second order* positive intuitionistic propositional logic or *second order polymorphic λ-calculus*, with the following additional axiom and inference rule:

$$\pi(X) : \forall_X F(X) \to F(X) \ ,$$

$$\frac{\varphi(X) : C \to F(X)}{\Lambda_X \varphi(X) : C \to \forall_X F(X)} \ ,$$

usually called *universal specification* and *generalization* respectively. According to the usual algebraic principle of substituting for free variables, namely

$$\frac{F(X) \to G(X)}{F(A) \to G(A)} \ ,$$

the axiom of universal specification gives rise to

$$\pi(A) : \forall_X F(X) \to F(A) \ .$$

We shall also insist on the following additional equations:

$$\pi(X)\Lambda_X\varphi(X) = \varphi(X) \ ,$$

where we have suppressed the subscript X on the equality sign, and

$$\Lambda_X(\pi(X)f) = f \ .$$

Moreover, the usual principle

$$\frac{\varphi(X) = \psi(X)}{\varphi(A) = \psi(A)}$$

allows us to specialize the first of these equations to

$$\pi(A)\Lambda_X\varphi(X) = \varphi(A) \ .$$

With this additional structure, we obtain a *cartesian closed category* with formal *products*, to be abbreviated as CCCP. However, a proper treatment of these involves simultaneous consideration of the categories $\mathcal{C}, \mathcal{C}[X], \mathcal{C}[X,Y], \ldots$ together with functors $\forall_X : \mathcal{C}[X] \to \mathcal{C}, \forall_Y \mathcal{C}[X,Y] \to \mathcal{C}[X]$, and so on (see Seely [1987]).

By confining attention to positive intuitionistic logic, we have completely ignored $\bot$ (*false*), $A \vee B$(*A or B*) and $\exists_X F(X)$. Here, for example, are the axioms, rule of inference and equations to be satisfied by disjunction if it is to agree with the categorical *coproduct*:

$$\kappa_{AB} : A \to A \vee B \ , \qquad \kappa'_{AB} : B \to A \vee B \ , \qquad \frac{f : A \to C \quad g : B \to C}{[f,g] : A \vee B \to C} \ ,$$

$$[f,g]\kappa_{AB} = f \ , \qquad [f,g]\kappa'_{AB} = g \ , \qquad [h\kappa_{AB}, h\kappa'_{AB}] = g \ ,$$

where it is assumed that $h : A \vee B \to C$.

In the presence of universal quantifiers, it is easily shown that

$$A \vee B \leftrightarrow \forall_X F_{AB}(X) \ ,$$

where

$$F_{AB}(X) = ((A \Rightarrow X) \wedge (B \Rightarrow X)) \Rightarrow X \ .$$

Indeed, Prawitz [1965] has suggested this as a definition of $A \vee B$. Unfortunately, $\leftrightarrow$ does not always translate into categorical isomorphism. The best known counter-example is

$$((A \Rightarrow C) \Rightarrow C) \Rightarrow C \leftrightarrow A \Rightarrow C \ ,$$

whereas, in the category of sets,

$$C^{C^{C^A}} \not\cong C^A \ .$$

At best we can say that the right hand side is a *retract* of the left hand side. Similarly, $A \vee B$, when interpreted as a categorical coproduct, turns out to be only a retract of $\forall_X F_{AB}(X)$ in a CCCP.

4. Introducing equalizers

Not all is lost however. One easily obtains an arrow $\varphi_{AB}(X) : A \to F_{AB}(X)$ in $\mathcal{C}[X]$, where $\mathcal{C}$ is any CCC, hence an arrow

$$k_{AB} = \Lambda_X \varphi_{AB}(X) : A \to \forall_X F_{AB}(X)$$

in any CCCP, and similarly an arrow k'_{AB} starting from B. As in [L1991b], it is now easily shown that

$$A \overset{k_{AB}}{\to} \forall_X F_{AB}(X) \overset{k'_{AB}}{\leftarrow} B$$

is a *weakly* initial object in the category of all $A \overset{f}{\to} C \overset{g}{\leftarrow} B$ (meaning that there is an arrow from the former to the latter which is not necessarily unique). If follows that we have a *weak coproduct* of A and B (meaning that all but the last equation in the definition of a coproduct in the previous section are satisfied.)

If our CCCP also has joint equalizers of families of parallel arrows [*loc.cit.*], we can define the coproduct $A \vee B$ as the subobject of $\forall_X F_{AB}(X)$ which equalizes all pairs of arrows (h, h') from the latter into any object C such that

$$hk_{AB} = h'k_{AB} \quad , \quad hk'_{AB} = h'k'_{AB} \ .$$

We recall from general category theory that the *equalizer* of two arrows $f, g : A \rightrightarrows B$ is an arrow $m : E(f,g) \to A$ such that $fm = gm$ and m is universal with respect to this property: if also $m' : C \to A$ is such that $fm' = gm'$ then there is a unique arrow $k : C \to E(f,g)$ such that

$m' = mk$. In the category of sets, equalizers are easily constructed: let $E(f,g) = \{a \in A | fa = ga\}$ and take m to be the inclusion.

What could we possibly mean by $E(f,g)$ in a deductive system? Martin-Löf [1984] might speak of a "judgement" that the deductions f and g are equal, but such a judgement does not come equipped with an arrow into A. To imitate the above set-theoretic construction we should rather carry out the *Brouwer-Heyting-Kolmogorov* interpretation of intuitionistic logic, in which formulas are replaced by sets (see Troelstra and van Dalen [1988]), say the formula A by the set $\{A\}$ of all *reasons* for A. I say "reason" rather than "proof", since otherwise $\{A\}$ would be empty whenever A is not a theorem.

Without enlarging our ontology, we may take as a reason for A any deduction $C \to A$ from an arbitrary formula C. (Categorists have called this a "generalized element" of A.) Thus $\{A\}$ is the union of all sets $Hom(C,A)$, where C ranges over all objects of the category $\mathcal{C}$.

We can then introduce $E(f,g)$ as the set of all reasons $a : C \to A$ such that $fa = ga$. Then $E(f,g)$ is not a formula, or object of $\mathcal{C}$, but a *right ideal* or *sieve* $\mathfrak{a}$ of A, which assigns to each formula C a subset $\mathfrak{a}_C \subseteq Hom(C,A)$ such that, whenever $a \in \mathfrak{a}_C$ and $c : D \to C$, then also $ac \in \mathfrak{a}_D$. (If $\mathcal{C}$ is a monoid, that is, a one-object category, then this agrees with the usual definition of right ideal in a monoid.) In categorical language, $\mathfrak{a}$ is nothing else than a subfunctor of the representable functor $Hom(-,A)$.

We can now form a new category whose objects are right ideals of $\mathcal{C}$ and whose arrows are induced from those in $\mathcal{C}$. This new category will have equalizers, but unfortunately it won't inherit the CCC structure of $\mathcal{C}$. It turns out that we can remedy the situation by considering right ideals modulo congruence relations instead. In other words, we pass from subfunctors of representables to quotients of such. By an application of Occam'a razor, we shall replace the right ideals modulo congruence relations by the congruence relations on the right ideals.

5. A subcongruence completion

A (*right*) *subcongruence* α of A assigns to each formula C a partial equivalence relation α_C on $Hom(C,A)$ such that, whenever $a, a' : C \to A$ and $c : D \to C$, then

$$a \alpha_C a' \text{ implies } (ac) \alpha_D (a'c) .$$

It is easily seen that the domain $Dom\ \alpha$ of any (right) subcongruence of A is a right ideal of A.

We now form the category $\mathcal{C}^R$: its objects are subcongruences of objects of $\mathcal{C}$ and its arrows $(\beta, f, \alpha) : \alpha \to \beta$ are induced by arrows $f : A \to B$ in $\mathcal{C}$ (assuming α to be a subcongruence of A and β one of B) such that, for each object C and all $a, a' : C \to A$,

$$(*) \qquad a\alpha_C a' \text{ implies } (fa)\beta_C(fa') .$$

We shall often just write $f : \alpha \to \beta$. If also $g : \alpha \to \beta$, we shall say that $(\beta, f, \alpha) = (\beta, g, \alpha)$ provided, under the same conditions,

$$(**) \qquad a\alpha_C a' \text{ implies } (fa)\beta_C(ga') .$$

Here $(*)$ was chosen to ensure that f induces a mapping $Dom\ \alpha_C/\alpha_C \to Dom\ \beta_C/\beta_C$. Then $(**)$ just means that f and g induce the same mapping. However, there is another way of interpreting $(**)$. Observe that α gives rise to a functor $F_\alpha : \mathcal{C}^{op} \to$ Set such that

$$F_\alpha(C) = Dom\ \alpha_C/\alpha_C .$$

If β similarly gives rise to $F_\beta(C)$, then (β, f, α) gives rise to a natural transformation $t_f : F_\alpha \to F_\beta$. Then $(**)$ means precisely that $t_f = t_g$.

However, not all natural transformations $F_\alpha \to F_\beta$ are obtained in this fashion. To describe all of them we would have to consider (*right homomorphic*) *relations* ρ between A and B, which to each object C assign a binary relation ρ_C between $Hom(C, A)$ and $Hom(C, B)$ such that, whenever $a : C \to A, b : C \to B$ and $c : D \to C$, then

$$b\rho_C a \text{ implies } (bc)\rho_D(ac) .$$

Then ρ induces a mapping $Dom\ \alpha/\alpha \to Dom\ \beta/\beta$, or a natural transformation $F_\alpha \to F_\beta$, if and only if, for all objects C,

$$(\dagger) \qquad \alpha_C \subseteq \rho_C^{\cup}\beta\rho_C \text{ and } \rho_C\alpha_C\rho_C^{\cup} \subseteq \beta_C ,$$

where juxtaposition denotes the *relative product* and $\rho_C^{\cup}$ is the *converse* of ρ_C. Moreover two relations ρ and σ between A and B induce the same mapping, or natural transformation, if and only if, for all objects C,

$$(\dagger\dagger) \qquad \alpha_C \subseteq \rho_C^{\cup}\beta\sigma_C \text{ or } \rho_C\alpha_C\sigma_C^{\cup} \subseteq \beta_C .$$

(In view of $(\dagger)$, the two clauses of $(\dagger\dagger)$ are equivalent.)

In this way one obtains all natural transformations $F_\alpha \to F_\beta$, hence a full subcategory $\mathcal{C}^{(R)}$ of the functor category $\text{Set}^{\mathcal{C}^{op}}$. (I believe that the

transition from $\mathcal{C}^R$ to $\mathcal{C}^{(R)}$ resembles a construction recently described by McLarty.) We shall not consider the category $\mathcal{C}^{(R)}$ any further in this article, except for drawing the following picture, in which all arrows represent faithful functors, but not all of them are full:

$$\begin{array}{ccc} \mathcal{C} & \overset{\text{Yoneda}}{\underset{\text{full}}{\longrightarrow}} & \mathrm{Set}^{\mathcal{C}^{\mathrm{op}}} \\ \text{full}\downarrow & & \uparrow \text{full} \\ \mathcal{C}^R & \underset{\text{not full}}{\longrightarrow} & \mathcal{C}^{(R)} \end{array}$$

Of course, the functor $\mathcal{C}^R \to \mathcal{C}^{(R)}$ sends the arrow $f : \alpha \to \beta$ onto $\rho_f : \alpha \to \beta$, where $b(\rho_f)_C a$ means $b = fa$, for all $a : C \to A$ and $b : C \to B$. I hope to return to a consideration of $\mathcal{C}^{(R)}$ on another occasion.

6. Some properties of $\mathcal{C}^R$

Readers not too familiar with category theory may wish to skip this section, some of which describes work in progress.

If α and α' are subcongruences of A such that $\alpha \subseteq \alpha'$, then $(\alpha', 1_A, \alpha)$ will be an arrow $\alpha \to \alpha'$ in $\mathcal{C}^R$. Two extreme cases are of special interest.

(a) If α is the restriction of α' to $Dom\ \alpha$, that is, if $\alpha\alpha'\alpha = \alpha, 1_A : \alpha \to \alpha'$ is a (canonical) *subobject* of α'. The subobjects of α' form a complete lattice.

(b) If $Dom\ \alpha = Dom\ \alpha'$, that is, if $\alpha\alpha'\alpha = \alpha'$, α' induces a congruence relation on $Dom\ \alpha$ and $1_A : \alpha \to \alpha'$ is a (canonical) *quotient* of α. The quotient objects of α form a complete lattice.

The original motivation for introducing $\mathcal{C}^R$ was to construct equalizers. Indeed, the equalizer of $f, g : \alpha \rightrightarrows \beta$ is $1_A : \kappa \to \alpha$, where

$$\kappa_C = \alpha_C \cap f^{\cup}\beta_C g\ .$$

(We identify f with the relation ρ_f mentioned at the end of Section 5.) Since the lattice of subobjects is complete, we also have equalizers of families of parallel arrows. (Perhaps, we have achieved too much; we would have been happy with joint equalizers of, in some sense, "definable" families.) We also have coequalizers: the coequalizer of $f, g : \alpha \rightrightarrows \beta$ is $\beta \to \lambda$, where λ is the smallest congruence on $Dom\ \beta$ extending $\beta \cup f\alpha g^{\cup}$. By completeness, we also have joint coequalizers of families of parallel arrows.

We shall discuss some further properties of $\mathcal{C}^R$ under various assumptions about $\mathcal{C}$. In doing so, we shall switch from logical to categorical notation and terminology, replacing $\top, A \wedge B, A \Rightarrow B, \bot, A \vee B$ by

$1, A \times B, B^A, 0, A + B$ respectively, speaking of terminal object, product, exponentiaion, initial object and coproduct.

(1) If $\mathcal{C}$ has finite products, then so does $\mathcal{C}^R$. In fact $\mathcal{C}^R$ is then a *regular* category. (I am endebted to Duško Pavlović for pointing this out.) This is so since regular epimorphisms are isomorphic to quotient objects and are preserved under pullbacks.

(2) If $\mathcal{C}$ is a CCC, then so is $\mathcal{C}^R$ and the functor $\mathcal{C} \to \mathcal{C}^R$ preserves the CCC structure. (See [L1991b, Proposition 10.1].)

(3) If $\mathcal{C}$ is a CCCP, then $\mathcal{C}^R$ is also cocartesian and any finite coproduct of definable endofunctors of $\mathcal{C}$ has a least fixpoint in $\mathcal{C}^R$. (See loc.cit., Proposition 10.2, where a precise definition of "definable" is given. Suffice it here to point out that e.g. $1, X, X^2$ and C^{C^X} are definable.)

(4) If $\mathcal{C}$ is a poset, then $\mathcal{C}^R$ is the complete lattice of downward closed subsets of $\mathcal{C}$. (Unfortunately, this is not the same as the Dedekind-MacNeille completion, see Birkhoff [1967].)

(5) If $\mathcal{C}$ is the monoid of partial recursive funtions $\mathbb{N} \to \mathbb{N}$, that is, partial functions whose graphs are recursively enumerable, $\mathcal{C}^R$ contains the category PER, whose objects are partial equivalence relations on $\mathbb{N}$. To see this, one associates with every partial equivalence relation A on $\mathbb{N}$ the subcongruence $\alpha(A)$ such that, for all partial recursive funtions $f, g : \mathbb{N} \to \mathbb{N}$,

$$f\alpha(A)g \text{ if and only if, for all } n \in \mathbb{N}, \ (fn)A(gn) ,$$

and with every subcongruence α the partial equivalence relation $A(\alpha)$ such that

$$mA(\alpha)n \text{ if and only if } k_m \alpha k_n ,$$

where k_m is the constant function with value m. Every finite coproduct of definable endofunctors of PER has a least fixpoint (see loc.cit. Proposition 8.1.). Hyland[1988] and Moggi have shown PER to be a model of polymorphic λ-calculus, though not in the usual universe of sets.

(6) *Weak binary products* are defined like binary products, except that the equation $< \pi_{AB} h, \pi'_{AB} h >= h$ is missing. If $\mathcal{C}$ has weak binary products, then $\mathcal{C}^R$ has binary products; if α and β are subcongruences of A and B respectively, we define

$$\alpha \times \beta = \pi_{AB}^{\cup} \alpha \pi_{AB} \cap \pi'^{\cup}_{AB} \beta \pi'_{AB}$$

and check that $\pi_{AB} : \alpha \times \beta \to \alpha$ and $\pi'_{AB} : \alpha \times \beta \to \beta$. Moreover, if $a : \gamma \to \alpha$ and $b : \gamma \to \beta$, then $< a, b >: \gamma \to \alpha \times \beta$ and, if $h : \gamma \to \alpha \times \beta$, then $< \pi_{AB}h, \pi'_{AB}h >$ induces the same arrow as h. A similar result holds for nullary products, that is, terminal objects, and for exponentiation.

(7) If $\mathcal{C}$ has weak coproducts, then $\mathcal{C}^R$ has coproducts; if $\mathcal{C}$ has a weak natural numbers object, then $\mathcal{C}^R$ has a natural numbers object. These and similar results are buried in the proof of (3), bearing in mind that a natural numbers object is a least fixpoint of the endofunctor $1 + X$.

(8) We say that $\mathcal{C}$ has a *weak subobject classifier* $t : I \to \Omega$ provided I is a weak terminal object (that is, for each C there is an arrow $i_C : C \to I$) and, for each right ideal $\mathfrak{a}$ of any object A, there is an arrow $h : A \to \Omega$ such that $\mathfrak{a} = Ker\ h$, where

$$(Ker\ h)_C = \{a : C \to A | \exists_{i:C \to I} ti = ha\} .$$

Let $i \iota_C i'$ for all $i, i' : C \to I$ and let

$$h \omega_A h' \text{ if and only if } Ker\ h = Ker\ h' ,$$

then ι is a congruence on I and ω is a congruence on Ω. One easily checks that ι is a terminal object (just take $\circ_A = i_A$) and we shall see that ω is a subobject classifier in the usual sense.

In fact, it is easily verified that (ω, t, ι) is an arrow. Now let $1_A : \alpha' \to \alpha$ be any subobject of α, we pick $h : A \to \Omega$ such that $Dom\ \alpha' = Ker\ h$ and verify that $h : \alpha \to \omega$ and $\alpha' = \alpha \cap h^{\cup} \omega t \circ_A$. Moreover, it follows that (ω, h, α) is unique with this property.

7. Gentzen style proof theory (intuitionistic)

A different way of presenting proofs was discovered by Gentzen. For dealing with intuitionistic logic he introduced generalized deductions, usually called *sequents*, of the form

$$f : A_1 \cdots A_m \to B ,$$

where the A_i and B are formulas. When the set of sequents $\Gamma \to B$ (we use capital Greek letters to denote strings of formulas) is subjected to appropriate equations, we obtain a *Gentzen multicategory*. It is not so obvious how to describe these equations directly [Szabo 1978], we shall

here follow [L1989a] and pass to the internal language of a multicategory, in which the sequents act as operations, while equations between sequents are best described through equations between terms.

Here then is the language: for each object A of the multicategory we introduce countably many variables. Rather than burdening our terminology by labeling these variables once and for all, we shall just say that x is a variable of type A and write $x \in A$. Terms of given types are defined inductively: every variable is a term of its type and, if $f : A_1 \cdots A_m \to B$ is a sequent and a_i is a term of type A_i, for $i = 1, \cdots, m$, then $fa_1 \cdots a_m$ is a term of type B. In particular, if $x_i \in A_i$, then $fx_1 \cdots x_m \in B$. There can also be constants of type B, namely any sequent $b : \ \to B$, with $m = 0$.

Algebraically speaking, $fx_1 \cdots x_m$ is a *polynomial*; we think of x_i as an indeterminate of type A_i and of f as an m-ary operation. In fact, a Gentzen multicategory is nothing else than a *many-sorted algebraic theory* (see Higgins [1963], Birkhoff and Lipson [1970]).

Logically speaking, $fx_1 \cdots x_m$ is a *proof* from certain hypotheses; we think of x_i as an occurrence of the hypothesis A_i and of f as a deduction of B from these hypotheses. Thus a Gentzen multicategory may also be viewed as a *natural deduction system* as promulgated by Prawitz [1965].

Gentzen postulated for each formula A the sequent

$$1_A : A \to A$$

and he admitted four rules for deriving new sequents from old, the *cut rule*

$$\frac{f : \Lambda \to A \quad g : \Gamma A \Delta \to B}{g < f > : \Gamma \Lambda \Delta \to B}$$

and three *structural rules*:

$$\frac{f : \Gamma AB\Delta \to C}{f^i : \Gamma BA\Delta \to C} \qquad \text{(interchange)} ,$$

$$\frac{f : \Gamma AA\Delta \to B}{f^c : \Gamma A\Delta \to B} \qquad \text{(contraction)} ,$$

$$\frac{f : \Gamma\Delta \to B}{f^w : \Gamma A\Delta \to B} \qquad \text{(weakening)} .$$

The intended algebraic interpretation requires that the internal language of a multicategory be subject to the following equations and, of

course, others derivable from them by the usual rules of equality:

$$1_A x = x \ ,$$

$$g < f > \overline{uwv} = g\overline{u}f\overline{wv} \ ,$$

$$f^i\overline{u}yx\overline{v} = f\overline{u}xy\overline{v} \ ,$$

$$f^c\overline{u}x\overline{v} = f\overline{u}xx\overline{v} \ ,$$

$$f^w\overline{u}x\overline{v} = f\overline{uv} \ .$$

Here $\overline{u} = x_1 \cdots x_m$, where $x_i \in A_i$ and $\Gamma = A_1 \cdots A_m$. Similarly $\overline{v}$ is related to Δ and $\overline{w}$ to Λ, while $x \in A$ and $y \in B$. (Our account is somewhat incomplete, as we have not bothered to *declare* variables; e.g. the first equation should really be written $1_A x \underset{x}{=} x$.)

To return to the question: what are the equations of a multicategory, we are now in a position to define equality between operations $f, g : A_1 \cdots A_m \to B$ to mean that $fx_1 \cdots x_m = gx_1 \cdots x_n$ is provable in the internal language.

There is quite an industry involved in studying *substructural* logics, in which some or all of the structural rules are absent. For example, the weakening rule is omitted in relevance logic, the weakening and contraction rules are both omitted in Girard's [1987] linear logic and all three rules are omitted in the syntactic calculus [L1958], a form of bidirectional categorial grammar.

Gentzen's original motivation was to show that, for example, in the positive intuitionistic propositional calculus freely generated from a given multicategory, the cut rule (and incidentally also, for compound A, the axiom $1_A : A \to A$) is redundant, provided the connectives $\top$, $\wedge$ and $\Rightarrow$ are subjected to appropriate introduction rules. Thus, for given Γ and B, he was able to find all provable sequents $\Gamma \to B$. His method can also be used for deciding when two sequents from Γ to B are equal. (See e.g. Szabo [1978], [L1958] and [L1991a] for the case without structural rules.) At least one of Gentzen's introduction rules was discovered independently by Bourbaki [1948] in his study of multilinear operations.

8. The algebraic theory of primitive recursive functions

I learned about Gentzen's sequence calculus from Kleene [1952]. While Kleene does not discuss many-sorted algebraic theories, equations very much like the above occur in his treatment of primitive recursive functions. The question thus arises: can these be viewed as realizations of operations of an algebraic theory? Indeed, there is such an algebraic theory, even a single-sorted one.

Consider a sort N, operations

$$0 : \rightarrow N(\text{zero}) \quad , \quad S : N \rightarrow N(\text{successor})$$

and the following rule for forming new operations from old:

$$\frac{a : N^n \rightarrow N \quad h : N^{n+2} \rightarrow N}{R_{ah} : N^{n+1} \rightarrow N} \ (\text{recursion}) \ .$$

These operations are subject to the following equations, which already appear in Kleene's book:

$$(E_{ah}) \qquad \begin{aligned} R_{ah}\overline{x}0 &= a\overline{x} \\ R_{ah}\overline{x}Sy &= h\overline{x}yR_{ah}\overline{x}y \ , \end{aligned}$$

where $\overline{x} = x_1 \cdots x_n$.

Since the operation defined by recursion is supposed to be uniquely determined, one would also like to satisfy the conditions:

$$(U_{fh}) \qquad \text{if} \quad f\overline{x}Sy = h\overline{x}yf\overline{x}y \quad \text{then} \quad f = R_{f<0>h} \ ,$$

where $f : N^{n+1} \rightarrow N$ and $f < 0 > \overline{x} = f\overline{x}0$, an example of the cut rule. It had been an embarrassment for some time that the conditions (U_{fh}) were not presented in equational form (see Gödel [1958], Sanchis [1967], [LS1986]). Yet they can be so presented [L1988], although the argument is a little tricky.

A ternary operation $m : N^3 \rightarrow N$ is called a *Mal'cev operation* if it satisfies:

$$mxyy = x \quad , \quad myyz = z \ .$$

(Mal'cev [1954] had shown that the existence of such an operation is necessary and sufficient for congruence relations in every model of the algebraic theory to permute.) There are many Mal'cev operations in primitive recursive arithmetic, e.g.

$$mxyz = (x + z) \dot{-} y \ ,$$

where $x+y$ and $x \dot{-} y$, the naive difference, are easily defined by recursion.

With the help of a Mal'cev operation m, we can construct a new $n+2$-ary operation H_{mfh} from h and f:

$$H_{mfh}\bar{x}yz = m(h\bar{x}yz)(h\bar{x}yf\bar{x}y)(f\bar{x}Sy) \ .$$

We can now replace (U_{fh}) by

$$(M_{mfh}) \qquad\qquad R_{f<0>H_{mfh}} = f \ .$$

More precisely, one notices that

$$(M_{mfh}) \Rightarrow (U_{fh}) \ ,$$
$$(U_{fH_{mfh}}) \Rightarrow (M_{mfh}) \ .$$

Once stated, these implications are easily verified.

If we choose $mxyz = (x + z) \dot{-} y$, we finally arrive at the following equational presentation of primitive recursive arithmetic:

$$(E_{ah}) \ , \ (M_{mfh}) \ , \ (x + y) \dot{-} y = x \ .$$

The last equation must be postulated, because its usual proof depends on (U_{fh}), whereas here it is used to derive (U_{fh}) from (M_{mfh}).

It is quite easy to find a primitive recursive function $fxyzt$ such that Fermat's conjecture asserts that f is the function with constant value 0. It is not impossible that $fxyz4 = 0$ already in the algebraic theory discussed here, in view of Fermat's method of descent. (A similar question was raised conversationally by Joyal.) We can, of course, replace f by a function of one argument here. Work on reducing such polynomials to normal form is being done by Okada and Scott. Unfortunately, as they point out, one cannot expect to have both strong normalization and the Church-Rosser property, in view of Gödel's result that provability in Peano arithmetic is not decidable.

Note added in proof: Gregory Mints informs me that Mal'cev and his students have also studied algebras of primitive recursive functions.

9. Gentzen style proof theory (classical)

Gentzen devised a different sequent calculus for classical logic, with sequents of the form

$$f : A_1 \cdots A_m \rightarrow B_1 \cdots B_n \ .$$

These were supposed to be interpreted as

$$f : A_1 \wedge \cdots \wedge A_m \to B_1 \vee \cdots \vee B_n \ ,$$

in categorical language

$$f : A_1 \times \cdots \times A_m \to B_1 + \cdots + B_n \ .$$

Having convinced oneself that Gentzen's intuitionistic sequents are nothing else than operations in a many-sorted algebraic theory, one wonders why his classical sequents have not surfaced in algebra.

Take, for example, a single-sorted "bi-operation"

$$f : A \times A \to A + A \ .$$

Up to isomorphism, this may be written

$$f : A^2 \to A \times 2 \ ,$$

where $2 = \{\top, \bot\}$ is a representative two-element set. We can write $f =< f_0, f_1 >$, where

$$f_0 : A^2 \to A \quad , \quad f_1 : A^2 \to 2 \ ,$$

the former being a binary operation, the latter a binary relation. It would therefore appear that, in the single-sorted case, bi-operations may be analyzed into operations and relations. Indeed, algebraists study ordered groups and similar structures.

The classical Gentzen calculus can also be extended to the substructural situation, where one should interpret $f : A_1 \cdots A_m \to B_1 \cdots B_n$ as

$$f : A_1 \otimes \cdots \otimes A_m \to B_1 \oplus \cdots \oplus B_n \ .$$

The tensor product $\otimes$ comes up in multilinear algebra. Its dual $\oplus$ appears in Girard's linear logic, although in a different notation. It so happens that in Hopf algebras $\oplus$ coincides with $\otimes$, and the same is true in production grammars. However, it is easy to give examples in which $\oplus$ and $\otimes$ are different.

In the algebra of all binary relations on a set X, we may write:

$$a(R \otimes S)b \quad \text{for} \quad \exists_{x \in X}(aRx \wedge xSb) \ ,$$
$$a(R \oplus S)b \quad \text{for} \quad \forall_{x \in X}(aRx \vee xSb) \ .$$

The former is the usual relative product. The latter may be described classically as its De Morgan dual by

$$R \oplus S = \neg(\neg R \otimes \neg S) ,$$

but it also exists intuitionistically. Unfortunately, intuitionistic logic does not suffice to prove that $\oplus$ is associative. (See [L 1991 c].)

The categorical treatment of classical Gentzen style proof theory requires something new in place of multicategories, namely "polycategories". These have been discussed by Szabo [1975] and others, but not yet in sufficient generality to handle the last example above.

10. Completeness of higher order logic

Categories may also be used to clarify and sharpen some basic ideas in model theory, though these may be far removed from the combinatorial preoccupations of specialists in that area. We shall only briefly sketch the situation as it concerns higher order logic.

On the one hand, there are certain intuitionistic higher order languages or type theories; they form a category, whose arrows may be called "translations". On the other hand, there are elementary toposes, namely cartesian closed categories with subobject classifier and, for the present purpose, also a natural numbers object; they form a category too, whose arrows are called "logical functors".

For each higher order language $\mathcal{L}$ we may construct a topos $T(\mathcal{L})$, the *topos generated* by $\mathcal{L}$, essentially what logicians call the "term model", which might also have been called the "Tarski-Lindenbaum topos". With each topos we may associate a higher order language $L(\mathcal{T})$, its *internal language*. It turns out that L and T are functors between the two categories under consideration and that T is left adjoint to L. (Actually, this assumes that toposes have "canonical" subobjects, a technical detail which we shall ignore here, see [LS1986] for a fuller account.) In particular, this means that we have a one-to-one correspondence between logical functors

$$T(\mathcal{L}) \to \mathcal{T}$$

and translations

$$\mathcal{L} \to L(\mathcal{T}) ,$$

either of which may be called an *interpretation* of $\mathcal{L}$ in $\mathcal{T}$.

If we allow every such interpretation as a model, the completeness theorem would be quite trivial, as a single model would suffice, namely

$\mathcal{L} \rightarrow LT(\mathcal{L})$. We shall, however, define a *model topos* as one whose internal language has very special properties:

(1) $\bot$ is not true in $\mathcal{T}$, that is, not provable in $L(\mathcal{T})$;

(2) if $p \vee q$ is true in $\mathcal{T}$, then p is true or q is true;

(3) if $\exists_{x \in A} \varphi(x)$ is true in $\mathcal{T}$, then $\varphi(a)$ is true for some closed term of type A in $L(\mathcal{T})$.

Peter Freyd observed that these properties can be translated into algebraic properties of the terminal object 1 of $\mathcal{T}$:

(1) $1 \neq 0$;

(2) 1 is indecomposable;

(3) 1 is projective.

Here "projective" means exactly the same as in module theory, where the term originated. It turns out that, when $L(\mathcal{T})$ is classical, a model topos is precisely a non-standard model in the sense of Henkin [1950].

The completeness theorem for intuitionistic higher order logic, in the tradition of Gödel [1931], Henkin [1950], Aczel [1969] and others, may now be stated as follows:

every higher order language $\mathcal{L}$ has enough models, meaning that a formula p is provable in $\mathcal{L}$ if and only if it is true under every interpretation of $\mathcal{L}$ in a model topos $\mathcal{T}$.

Algebraically, this result asserts:

every topos is equivalent to a subtopos of a product of model toposes.

This last statement bears a formal resemblance to the following result about commutative rings:

every commutative ring is isomorphic to a subring of a product of local rings.

But actually we know more. According to Grothendieck and Dieudonné [1960]:

every commutative ring is isomorphic to the ring of continuous sections of a sheaf of local rings.

A similar result holds for toposes in which *all subobjects of 1 are projective.* This is so for toposes of the form $\top(\mathcal{L})$ when $\mathcal{L}$ satisfies Hilbert's rule (see Hilbert and Bernays [1970]):

for any inhabited type A (i.e. $\exists_{x \in A} \top$ is provable in $\mathcal{L}$), *if $\alpha = \{x \in A | \varphi(x)\}$ is a closed term of type PA, then there is a closed term $e_\alpha = \varepsilon_{x \in A} \varphi(x)$ of type A such that $\exists_{x \in A} \varphi(x) \vdash \varphi(e_\alpha)$ in $\mathcal{L}$.*

In [L1989] I wrongly asserted also the converse of this; I am endebted to John Bell for pointing out the error.

One can get rid of the restriction to Hilbert's rule, provided one looks at toposes generated by languages with sufficiently many constants. The idea goes back to Henkin [1949]. One may think of the constants as variables held constant: if V is a set of variables, $\mathcal{L}(V)$ is the same language as $\mathcal{L}$, except that now a formula is called "closed" if it contains no free variables other than those from V. Here is the final result:

if $\mathcal{L}$ is any higher order language and V is a sufficiently large set of variables, then $T(\mathcal{L}(V))$ is the topos of continuous sections of a sheaf of model toposes.

I have been told that Grothendieck actually used the expression "local topos" for what has here been called "model topos", although in a different context, making the analogy with commutative rings even more striking.

For successive versions of this sheaf representation the reader is referred to [LM1982], [LS1986] and [L1989].

REFERENCES

P. H. ACZEL, *Saturated intuitionistic theories*, in: H.A. Schmidt et al. (eds), Contributions to mathematical logic, North Holland Publ. Co., Amsterdam 1969, 1–11.

G. BIRKHOFF, *Lattice theory*, AMS Coll. Publ. 25, 1967.

G. BIRKHOFF and J. D. LIPSON, *Heterogeneous algebras*, J. Combinatorial Theory 8 (1970), 115–133.

N. BOURBAKI, *Algèbre multilinéaire*, , Hermann, Paris 1948.

H. B. CURRY, *Some logical aspects of grammatical structures*, AMS Proceedings Symposium Applied Math. 12 (1961), 56–68.

H. B. CURRY and R. FEYS, *Combinatory Logic I*, North Holland Publ. Co., Amsterdam 1958.

S. FEFERMAN, *Review of Mann* [1975], Math. Reviews 52 (1976), 7869.

J.-Y. GIRARD, *Linear logic*, J. Theoretical Computer Science 50 (1987), 1–102.

K. GÖDEL, *Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I*, Monatshefte für Mathematik und Physik 38 (1931), 173-198.

K. GÖDEL, *Über eine bisher noch nicht benutzte Erweiterung des finiten Standpunktes*, Dialectica 12 (1958), 280-287.

A. GROTHENDIECK and J. DIEUDONNÉ, *Eléments de géométrie algébrique I*, IHES Publ. Math. 4, Paris 1960.

L. A. HENKIN, *The completeness of first-order functional calculus*, J. Symbolic Logic 14 (1949), 159–166.

L. A. HENKIN, *Completeness in the theory of types*, J. Symbolic Logic 15 (1950), 81–91.

P. J. HIGGINS, *Algebras with a scheme of operators*, Math. Nachrichten 27 (1963), 115–132.

D. HILBERT and P. BERNAYS, *Grundlagen der Mathematik II*, Springer Verlag, Berlin 1939, 1970.

M. HYLAND, *A small complete category*, Ann. Pure and Applied Logic 40 (1988), 135–165.

S. C. KLEENE, *Introduction to metamathematics*, Van Nostrand, New York 1952.

J. LAMBEK, *The mathematics of sentence structure*, Amer. Math. Monthly 65 (1958), 154–169.

J. LAMBEK, *Deductive systems and categories III*, Springer LNM 274 (1972), 57–82.

J. LAMBEK, *On the unity of algebra and logic*, in: F.Borceux (ed.), Categorical algebra and its applications, Springer LNM 1348 (1988), 221–229.

J. LAMBEK, *Multicategories revisited*, Contemporary Math. 92 (1988), 217–339.

J. LAMBEK, *On the sheaf of possible worlds*, in: J. Adámek and S. MacLane (eds.), Categorical topology, World Scientific, Singapore 1989, 36–53.

J. LAMBEK, *Logic without structural rules*, Reports from the Dept. of Math. 91-02, McGill University, Montreal 1991; to appear in; K. Došen and P. Schroeder-Heister, Substructural logics, Oxford University Press.

J. LAMBEK, *Least fixpoints of endofunctors of cartesian closed categories*, Report from the Dept. of Math. 91-11, McGill University, Montreal 1991; to appear in *Mathematical Structures in Computer Science.*

J. LAMBEK, *From categorial grammar to bilinear logic*, Report from the Dept. of Math. 91-14, McGill University, Montreal 1991; to appear in: K. Došen and P. Schroeder-Heister, Substructural logics, Oxford University Press.

J. LAMBEK and I. MOERDIJK, *Two sheaf representations of elementary toposes*, in: A.S. Troelstra and D. van Dalen (eds.), The L.E.J. Brouwer Centenary Symposium, Studies in Logic and Foundations of Math. 110, North Holland Publ. Co., Amsterdam 1982, 275–295.

J. LAMBEK and P. J. SCOTT, *Introduction to higher order categorical logic*, Cambridge University Press, Cambridge 1986.

F. W. LAWVERE, *Functorial semantics of algebraic theories*, Dissertation, Columbia University, New York 1963.

F. W. LAWVERE, *Adjointness in foundations*, Dialectica 23 (1969), 281–296.

F. W. LAWVERE, *Equality in hyperdoctrines and comprehension schema as an adjoint functor*, in: A. Heller (ed.), Proc. New York Symp. on Applications of Categorical Algebra, Amer. Math. Soc., Providence, R.I. (1970), 1–14.

M. MAKKAI and R. PARÉ, *Accessible categories; the foundations of categorical model theory*, Contemporary Math. 104 (1990).

M. MAKKAI and G. E. REYES, *First order categorical logic*, Springer LNMG11 (1977).

A. I. MAL'CEV, *On a general theory of algebraic systems*, Math. Sbornik NS35 (1954), 3–20.

C. R. MANN, *The connection between equivalence of proofs and cartesian closed categories*, Proc. London Math. Soc. 31 (1975), 289–310.

P. MARTIN-LÖF, *Intuitionistic type theory*, Studies in Proof Theory I, Bibliopolis, Naples 1984.

D. PRAWITZ, *Natural deduction*, Almquist and Wiksell, Stockholm 1965.

G. E. REYES and H. ZOLFAGHARI, *Bi-Heyting algebras, toposes and modalities*, Rapports de récherches 91-9, Université de Montréal, 1991.

P. C. ROSENBLOOM, *The elements of mathematical logic*, Dover Publ., New York, 1950.

L. E. SANCHIS, *Functionals defined by recursion*, Notre Dame J. of Formal Logic 8 (1967), 161–174.

R. A. G. SEELY, *Locally cartesian closed categories and type theory*, Math. Proc. Cambridge Philosophical Soc. 95 (1984), 33–48.

R. A. G. SEELY, *Categorical semantics for higher order polymorphic lambda calculus*, J. Symbolic Logic 52 (1987), 969–989.

M. E. SZABO (ed.) The collected papers of Gerhard Gentzen, Studies in Logic and the Foundations of Math., North-Holland Publ. Co., Amsterdam 1969.

M. E. SZABO, *Polycategories*, Communications in Algebra 3 (1975), 663–689.

M. E. SZABO, *Algebra of proofs*, Studies in Logic and the Foundations of Math. 88, North-Holland Publ. Co., Amsterdam 1978.

A. S. TROELSTRA and D. VAN DALEN, *Constructivism in Mathematics I*, Studies in Logic and the Foundations of Math. 121, Elsevier Science Publishers, Amsterdam 1988.

Logic, Methodology and Philosophy of Science IX
D. Prawitz, B. Skyrms and D. Westerståhl (Editors)

GENTZEN-TYPE SYSTEMS AND HILBERT'S EPSILON SUBSTITUTION METHOD. I

G.E.MINTS

Dept. of Philosophy, Stanford University, Stanford, CA 94305, USA

Introduction

The substitution method was suggested by Hilbert within the framework of his program for the foundations of mathematics. It is a successive approximation method for finding a finite function solution of a system of equations derived from a proof in a formal system. The problem of convergence, i.e. termination of the process after a finite number of steps was treated by von Neumann [1927] for quantifier free induction, by Ackermann [1940] for first order arithmetic, and by other authors including Kreisel [1951,1952] and Tait [1965a]. Related material is in Scanlon [1973], Goldfarb and Scanlon [1974]. We present here a new proof of Ackermann's result allowing extension to analysis (second order arithmetic). This extension, to be described in a sequel to this paper, settles one of the problems stated by Hilbert in [1929].

The proof consists of the following parts:

1. The formalization in the infinitary sequent calculus of a non-effective proof of the existence of a solution.
2. A standard normalization proof.
3. A proof that the normal form after this normalization is a convergence protocol for the epsilon substitution process.

Note that in the case of first order arithmetic there is a simple non-effective proof of convergence (cf. Kreisel [1952], Tait [1965a], Mints [1982, 1989]), but no generalization of this proof to the second order case is known.

The metamathematical means used in our convergence proof are the same as in other normalization proofs for the systems considered: epsilon-0 induction (on quantifier-free formulas) for first order arithmetic, and Girard's method of computability predicates in the second order case.

Hence it is difficult to claim that it has a foundational significance, and one should look for its applications elsewhere.

This work started at Steklov Institute of Mathematics (Leningrad), continued at the Institute of Cybernetics of the Estonian Academy of Sciences and was finished at Stanford University. The author is grateful to participants of logic and proof theory seminars of these institutions for attention and valuable discussions. Especially important was correspondence with G. Kreisel who drew the author's attention to the problem of convergence of the epsilon substitution method and had systematically pointed out various related problems (much more significant in his opinion than the problem investigated here). Special thanks are to S. Tupailo, who went through the first draft of this paper and helped to find and correct several discrepancies and to improve the presentation.

1. Language and the description of the substitution method

We use standard terminology of the substitution method from Hilbert & Bernays [1970].

There are two kinds of individuum variables: *free* variables, denoted by $a, b, a_1, \ldots$, and *bound* variables denoted by $x, y, z, x_1, \ldots$. Primitive recursive (PR) terms are constructed from free individuum variables and the constant 0 by means of fixed supply of PR function symbols including suc (successor), addition and multiplication. *Numerals* $0, \mathrm{suc}(0), \mathrm{suc}(\mathrm{suc}(0)), \ldots$ are sometimes also treated as constants. *PR formulas* are constructed from equations $t = u$ between PR terms by boolean connectives $\&, \vee, \sim$ etc. $A[v := t]$ (abbreviated by $A[t]$ when v is obvious) denotes the result of substituting the expression t for a variable v in the expression A.

Terms and *formulas* are defined by simultaneous recursion beginning with PR terms and formulas. If A is a formula which does not contain the bound variable x, and a is a free variable then $\epsilon x A[a := x]$ is an (epsilon) term. The set of terms is closed under function symbols of our language, and the set of formulas is closed under boolean connectives.

So our formulas are formally quantifier free. In fact the epsilon symbol ϵ plays the role of a quantifier via the translation

$$(1) \qquad (\exists x)F \mapsto F[\epsilon x F]; \quad (\forall x)F \mapsto F[\epsilon x \sim F]$$

which agrees with the interpretation of $\epsilon x A$ as the least natural number x satisfying $A[x]$ and 0 if $\sim A[x]$ for all x.

Terms and finite sequences of terms are denoted by $s, t, u, s_1, \ldots$, formulas are denoted by $A, B, C, D, F, A_1, \ldots$.

The *rank* $rk(t)$ of a term t is the measure of nesting of bound variables defined in a standard way. $rk(t) = 0$, if t is PR, rank is not changed by primitive recursive functions, and $rk(\epsilon xA) = \max\{rk(\epsilon yB) : \epsilon yB$ is a subterm of $A[x := a]$ containing $a\} + 1$, where a is a free variable which does not occur in ϵxA.

The *degree* $\deg(t)$ is the familiar nesting of epsilon-terms. $deg(t) = 0$ for PR term t, and $\deg(\epsilon xA) = \max\{\deg(\epsilon yB) : \epsilon yB$ is a proper subterm of $\epsilon xA\} + 1$. An *ϵ-matrix* or simply matrix is an ϵ-term having only variables as proper subterms with no variable occuring twice.

An *ϵ-substitution* for an ϵ-matrix $\epsilon xA[x, a_1, \ldots, a_n]$ with all free variables (arguments) explicitly listed is a finite numerical function $f(a_1, \ldots, a_n)$, i.e. a function of finite support with explicitly given finite domain. By the definition $f = 0$ (trivial zero value) outside this domain. The ϵ-substitution for a finite set of ϵ-matrices is the set of ϵ-substitutitons for each matrix. Each ϵ-term t is uniquely representable in the form $t = \epsilon xA[x, t_1, \ldots, t_k]$ where $t_1, \ldots, t_k$ are terms and $\epsilon xA[x, a_1, \ldots, a_k]$ is a matrix called the matrix of t. The *value* $S(t)$ of a constant term t under a given ϵ-substitution S is defined by recursion in a natural way with the main step: if $S(t_1) = N_1, \ldots, S(t_k) = N_k$ then $S(\epsilon xA[x, t_1, \ldots, t_k]) = f(N_1, \ldots, N_k)$ where $f = S(\epsilon xA[x, a_1, \ldots, a_k])$ is the ϵ-substitution for the matrix $\epsilon xA\ [x, a_1, \ldots, a_k]$. This also determines the values of all constant formulas.

The *value of a matrix* $\epsilon xA[x, a_1, \ldots, a_k]$ for a tuple $N_1, \ldots, N_k$ of numeric arguments is the same as the value of the constant term $\epsilon xA\ [x, N_1, \ldots, N_k]$ of degree 1. Such terms $\epsilon xA[x, N_1, \ldots, N_k]$ will be called *canonical*.

The role of quantifier axioms in Hilbert's ϵ-calculus is played by *critical formulas*

$$A[t] \to A[\epsilon xA] \tag{2}$$

Formula (2) is said to belong to the rank $rk(\epsilon xA)$.

The goal of Hilbert's ϵ-substitution method is to find a *satisfying* (or *solving*) *substitution* for any system E of closed critical formulas , i.e. a substitution S such that $S(C) =$ true for any critical formula C in E. Such a substitution provides finitistic proofs of constant combinatorial identities and numerical realizations of Σ_1^0 sentences provable from E in a free variable equation calculus.

It is assumed that critical formulas in the system E are numbered

$$Cr_i : A[t] \to A[\epsilon xA] \tag{3}$$

in such a way that formulas belonging to bigger ranks have greater numbers.

The *epsilon substitution method* as outlined by Hilbert and made precise by von Neumann [1927] and Ackermann [1940] applied to a given system E of constant critical formulas (2) generates successive epsilon substitutions $S_0, S_1, \ldots, S_m$ for matrices in E until all critical formulas Cr in E are satisfied: $S_m(Cr_i) = \text{true}$ for all i.

By the definition, the initial substitution S_0 is the *trivial* zero *substitutiton* assigning a function with empty domain to every matrix. In other words $S_0(M)(N) = 0$ for every matrix M and every tuple N of numeric arguments.

The *successor* S' for any given epsilon substitution S is defined as follows. If all critical formulas in E are satisfied by S, i.e. S is a solution, then S has no successor. Otherwise pick up the first critical formula (3) in E which is false under S:

$$S(A[t]) = \text{true and } S(A[\epsilon xA]) = \text{false} \tag{4}$$

Change the value of ϵxA to be equal to the least natural number N for which $S(A[N]) = \text{true}$:

$$S'(\epsilon xA) = \text{ the least } N \leq S(t) \text{ such that } S(A[N]) = \text{true} \tag{5}$$

and make the domains empty for all matrices of greater rank. More precisely, if the matrix of ϵxA is $M_x = \epsilon xB[x, a]$ then

$$S'(M) = S(M) \text{ for all matrices } M \text{ with } rk(M) \leq rk(M_x) \tag{6}$$

except M_x itself. If $\epsilon xA = \epsilon xB[x, u]$ then letting $t^* = S(t), u^* = S(u)$ etc. put

$$S'(M_x)(u^*) = \text{ the least } N \leq t^* \text{ such that } S(A[N]) = \text{true} \tag{7}$$

$$S'(M_x)(K) = S(M_x)(K) \text{ for all remaining tuples } K. \tag{8}$$

Put

$$S'(M) = 0 \text{ for all matrices } M \text{ with } rk(M) > rk(M_x) \tag{9}$$

If the substitution S_i has already been defined and is not a solution for the given system E of critical formulas, put

$$S_{i+1} = (S_i)' \tag{10}$$

The epsilon substitution method *terminates* (or *converges*) at the step i for the system E if S_i satisfies E.

2. Computations with epsilon terms

From now on an arbitrary system E of closed critical formulas

$$(1) \qquad Cr_i : A[t] \to A[\epsilon x A]$$

is considered. The value of a closed ϵ-term $\epsilon x A$, i.e. the least x satisfying A, is not computable in general, but additional information can help: if $\epsilon x A$ is canonical and $A[N], \sim A[N-1], \ldots, \sim A[0]$ are known to be true, then $\epsilon x A{=}N$ provided the information is consistent. To be sure the value has been really computed, we will also require the information to be complete: the same kind of justification should be given for the subterms of $A[i]$ $(i \leq n)$ etc. First we define a weak derivability relation $\to$ meaning roughly that an object (term or formula) can be computed to a normal form (numeral, canonical term, truth-value) on the basis of accessible information.

Such information is encoded in a *sequent*, i.e. finite list X of formulas. Formulas which can be members of a sequent X (= occur in X) are closed formulas of the epsilon-calculus as defined in section 1 (also called *basic* formulas) as well as formulas of the form $?\epsilon x A, !\epsilon x A$ (*modal* formulas) with canonical $\epsilon x A$. $?\epsilon x A$ means that the term has trivial value 0 (roughly corresponding to undefined). $!\epsilon x A$ means that the term has numeric value N satisfying

$$(2) \qquad A[N], \sim A[N-1], \ldots, \sim A[0]$$

The list (2) is abbreviated by $\epsilon x A\# = N$.

The notation $!\epsilon x A{=}N$ means the sequent $!\epsilon x A, \epsilon x A\# = N$.

Unless stated otherwise we assume that the sequent X is *correct*: the formula $?\epsilon x A, !\epsilon x A$ can have at most one occurrence in X (as a member formula) and if $?\epsilon x A$ occurs, then $!\epsilon x A$ does not.

CALCULUS CX FOR COMPUTATIONAL VALIDITY

t, u are lists of terms, M, N are lists of numerals.

$$\frac{?\epsilon x A \text{ occurs in } X;}{\epsilon x A \to 0} \ (\to_0)$$

$$\frac{!\epsilon x A{=}N \text{ occurs in } X; \quad (\epsilon x A\# = N) \to \text{ true}}{\epsilon x A \to N} \ (\to_+)$$

$$\frac{t \to n; \quad A[N] \text{ is a primitive recursive true constant formula}}{A[t] \to \text{true}} \text{ (bool)}$$

$$\frac{t \to N; \quad \epsilon x A[x,t] \text{ occurs in } X, E;}{\epsilon x A[x,t] \to \epsilon x A[x,N]} \ (\epsilon \to) \qquad \begin{array}{l} \epsilon x A[x,N] \\ \text{is canonical} \end{array}$$

$$\frac{\epsilon x A \to \epsilon x B; \quad \epsilon x B \to N}{\epsilon x A \to N} \text{ (trans)}$$

$A \to$ false means $\sim A \to$ true. $CX : D \to v$ means derivability of $D \to v$ in CX. Reference to CX will be dropped sometimes. A sequent X is *computationally inconsistent* (c.i.) if CX: $A \to$ false for some formula A occurring in X. Otherwise it is computationally consistent (c.c.) provided it is correct. A numeric value $|t|, |A|$ of a term t or formula A relative to a sequent X is any numeral N or boolean value v such that CX: $t \to N$ or $A \to v$. The canonical value $\|t\|$ of an epsilon term t is any canonical term s such that $t \to s$, or t itself if t is canonical.

The *substitution* S_X *determined by* a given *sequent* X consists of numeric values of all canonical terms: $S_X(t) = |t|$. In other words, if $M = \epsilon x A[x,a]$ is an epsilon-matrix with free variables (=arguments) a, then $S_X(M)(N) = |\epsilon x A[x,N]|$. It is understood that $S_X(M)(N) = 0$ if $|\epsilon x A[x,N]|$ does not exist.

LEMMA 2.1. *If B is derivable in the calulus CX then one of the following conditions is satisfied:*

(a) $B = \epsilon x A \to N$ *with* $\epsilon x A$ *in* X, E *and* N *in* X;
(b) $B = \epsilon x A[x,t] \to \epsilon x A[x,N]$ *with* $\epsilon x A[x,t]$ *in* X, E, N *in* X *and canonical* $\epsilon x A[x,N]$;
(c) $B = A[t] \to$ *true where* $A[x]$ *is a primitive recursive formula with* t *in* X, E, *and* $CX : t \to N$ *for some numerals* N.

(Numeral 0 is assumed to occur in any sequent).

PROOF is by induction on the derivation in CX. Case (a) corresponds to the rules $\to_0, \to_+$, trans. Case (b) corresponds to the rule $\epsilon \to$. Case (c) corresponds to the rule bool. ⊣

If not explicitly stated otherwise we identify constant primitive recursive terms with their numeric values. We employ a somewhat non-standard notion of a *substitution instance* of a term. This will be any result of dropping some epsilon symbols and substituting numerals for all occurrences of corresponding variables and for some terms.

LEMMA 2.2. (a) *Derivability of $t \to u$ is decidable;*

(b) *derivability of* $A \to$ *true for any formula* A *is decidable;*
(c) *existence of* $\|t\|, |t|, |A|$ *is decidable.*

PROOF: (a) By Lemma 1 a derivation of $t \to u$ in the calculus CX contains only terms and formulas from X, E and their substitution instances by numerals in X. So there is only a finite number of derivations.
(b) follows from (a) by Lemma 2.1 (c): a formula can be reduced to truth in CX only if it is primitive recursive in a term t such that $CX : t \to M$ and t, M occur in X, E.
(c) For given t only u, N occuring in X, E (and their substitution instances by numerals in X) should be tested for derivability of $t \to u, t \to N$. To find $|A|$ one tests $A \to$ true and $\sim A \to$ true. ⊣

LEMMA 2.3. $|\epsilon x A| = N$ *implies the existence of a canonical* $\epsilon x B$ *such that* ($\epsilon x A = \epsilon x B$ *or* $\epsilon x A \to \epsilon x B$) *and* $\epsilon x B \to N$, *i.e.*

(3) *(*$?\epsilon x B$ *is in* X *and* $N = 0$*) or*

(4) $!\epsilon x B$ *is in* X *and* $A[N], \sim A[N-1], \ldots, \sim A[0]$ *are in* X *and reducible to truth in* CX.

PROOF is by induction on the length of the derivation of $\epsilon x A \to N$. If $\epsilon x A$ is not canonical, this derivation ends in the trans-rule, and over the right branch of this rule one of the rules $\to_0, \to_+$ is situated. The latter provides the data mentioned in (3),(4). ⊣

LEMMA 2.4. *A derivation of any relation* $e \to v$ *in* CX *contains only substitution instances of its subterms. In particular the rank of terms in the derivation does not exceed the maximal rank of terms in the derived relation.*

PROOF: The proof is by easy induction on the derivation. ⊣

Now we establish a version of the Church-Rosser theorem.

LEMMA 2.5. *If* X *is a correct c.c. sequent then* $|t|, \|t\|, |A|$ *are unique for any epsilon term* t *and formula* A.

PROOF: Apply induction on the sum of the lengths of derivations

$$\begin{aligned} &d : A \to v; \quad && d_1 : A \to v_1; \\ &d : t \to N; \quad && d_1 : t \to N_1; \\ &d : t \to \epsilon y B; \quad && d_1 : t \to \epsilon y B_1. \end{aligned}$$

Let L, L_1 be the last rule in d, d_1.

CASE 1. L is bool. Then L_1 is also bool. If v is different from v_1, then the premises of L and L_1 assign different values for some subterm t of A with shorter derivation, which contradicts the induction hypothesis. So $v = v_1$ as required.

CASE 2. L is one of $\to_0, \to_+$, trans with conclusion $t \to N$. Then L_1 is also one of these rules. Consider possible combinations L, L_1.
$(\to_0, \to_0)$ is as required.
$(\to_0, \to_+)$ is excluded since then X contains $?t, !t$ and so is not correct.
$(\to_0$, trans):

$$\frac{\epsilon xA \to \epsilon xB; \quad \epsilon xB \to N}{\epsilon xA \to N}\ (\text{trans}) \qquad \frac{?\epsilon xA \text{ in } X}{\epsilon xA \to M}\ (\to_0)$$

is impossible since ϵxA in $\to_0$ is canonical, and so the relation $\epsilon xA \to \epsilon xB$ cannot be derivable.
$(\to_+$,trans) is impossible for the same reason as $(\to_0$, trans).
$(\to_+, \to_+) : !\epsilon xA{=}N, !\epsilon xA{=}N_1$ occur in X and $\epsilon xA\# = N, \epsilon xA\# = N_1 \to$ true.

Assuming $N > N_1$ we see that $A[N-1], \sim A[N-1]$ are in X and $A[N-1] \to$ false since $\sim A[N-1] \to$ true. So X is c.i.
(trans, trans):

$$\frac{\epsilon xA \to \epsilon xB; \epsilon xB \to N}{\epsilon xA \to N} \qquad \frac{\epsilon xA \to \epsilon xB_1; \epsilon xB_1 \to N_1}{\epsilon xA \to N_1}$$

By induction hypothesis for $\|t\|$ one has $\epsilon xB = \epsilon xB_1$, so again by induction hypothesis $N = N_1$.

CASE 3. L is $\epsilon \to$. Then L_1 also is $\epsilon \to$.

$$\frac{t \to N; \quad t \to N_1; \quad \epsilon xA[x,t] \text{ is in } X, E;}{\epsilon xA[x,t] \to N, N_1}$$

Again $N = N_1$ by the induction hypothesis as required. ⊣

COROLLARY 2.6. *If $\epsilon xA[x,u] \to N$ for $\epsilon xA[x,a]$ of degree 1 and terms u different from numerals then $u \to M$ for a unique M and $\epsilon xA[x,M] \to N$.*

PROOF: The relation $\epsilon xA[x,u] \to N$ can be obtained only by the rule (trans) and both premises of that rule are uniquely determined by Lemma 2.5. The left premise is obtained by the rule $\to \epsilon$ as required. ⊣

Recall that the substitution S_X determined by a given sequent X consists of numeric values of all canonical terms: $S_X(t) = |t|$. Recall also the definition of the correct epsilon substitution from Hilbert-Bernays [1970]: substitution S is correct if for any matrix $M = \epsilon xA[x,a]$ with arguments a and any corresponding tuple N of numerals $S(M)(N) = 0$ or $S(M)(N) =$ the least K such that $S(A[K,N]) =$ true.

LEMMA 2.7. *The substitution S_X determined by a correct c.c. sequent X agrees with derivability in the calculus CX:*

$$\text{if } e \to v \text{ in } CX \text{ then } S(e) = S(v) \tag{5}$$

In particular S_X is uniquely determined and correct.

PROOF: Uniqueness of $S(M)(N)$ follows from the uniqueness of $|\epsilon x A[x, N]|$. (5) is proved by induction on the derivation. To establish correctness of S assume $\epsilon x A[x, N] \to K > 0$. Then by the lemma 2.3 and (5) one has $S(A[N]), S(\sim A[N-1]), \ldots, S(\sim A[0]) =$ true as required. $\dashv$

A closed term $\epsilon x A$ is *decided* by a sequent X if $|\epsilon x A|$ exists. A formula A is *supported* by X if $CX : A \to$ true. A sequent Y is *supported* by X if all formulas in Y are supported by X. A sequent X is *supported* if X is supported by X. X is *decided* if all terms in X, E are decided by X.

We say that a basic formula A has *rank* $r+$ where r is the maximal rank of epsilon terms in A. We put

$$r < r+ < r+1 \tag{6}$$

Formulas $?\epsilon x A, !\epsilon x A$ have rank $= rk(\epsilon x A)$. For a sequent X its part $X \leq r$ consists of members of X with ranks at most r. The sequent $X < r$ contains formulas of rank $< r$, including $(r-1)+$.

The following proposition expresses subformula properties of the calculus CX.

LEMMA 2.8. *Any formula of rank at most r supported by X is supported by $X \leq r$.*

PROOF: This is a reformulation of Lemma 2.4.

3. Gentzen-type system PA_ϵ. Statement of results.

Recall that a system E of closed critical formulas

$$Cr_i : A[t] \to A[\epsilon x A] \tag{1}$$

is assumed to be fixed.

Derivable objects of the Gentzen-type system PA_ϵ will be sequents X. A rough first approximation of the meaning of X is: any extension of X to a decided sequent either is inconsistent or determines a solving substitution. If E' is a closed arithmetic statement saying that the system E has a satisfying (solution) substitution, then X can be interpreted as

$X \to E'$, i.e. X implies the existence of a solution substitution for E. An alternative interpretation would be $X, \sim E' \to$ false (to stay in the negative fragment and make the derivation intuitionistic). Each sequent X in a derivation is decomposed into a *fixed part* X_f and a *provisional part* X_t:

$$X = X_f \cup X_t.$$

Components of epsilon substitution determined by formulas in the fixed part X_f should not be changed in a non-trivial way when the derivation is followed bottom-up; they can only be truncated to zero.

Formulas in the provisional part X_t can be changed during a H-step (for Hilbert) of a generating new epsilon-substitution. The notation $\{A\}$ means that A is in the t-part (i.e. provisional part X_t). The number $I(X)$ is the maximal i such that the i-th critical formula Cr_i is computationally true, i.e. CX: $Cr_i \to$ true.

Recall that a sequent X is decided if $|t|$ exists for all terms t in X, E, and that the sequent $X \leq r$ is obtained by deleting from X all formulas containing terms of rank $> r$ and all basic formulas containing terms of rank r.

Axioms different from the AxF(alse) below and conclusions of all rules should be c.c. The conclusion of any rule exept (cut) and (Fr) and any axiom exept AxF should be decided. In the axiom Axc and the rule H we write J-th critical formula as $Cr_J : B[t] \to B[\epsilon xB]$.

AXIOMS. A sequent X is an axiom in the following cases:

AxSol. $I(X) = I_{\max}$, i.e. $Cr_J \to$ true for all J

AxF. X is c.i.

Axc. $I(X) = J - 1 < I_{\max}, ?||\epsilon xB|| = ?\epsilon xA$ is in $X_f, (\epsilon xA\# = N) \to$ true for some N.

INFERENCE RULES.

$$\text{(Cut)} \quad \frac{?\epsilon xA, X; \ldots !\epsilon xA = N, X; \ldots}{X} \qquad \begin{array}{l}\text{all } N \\ ?\epsilon xA, !\epsilon x\ A = N \\ \text{are in the } f\text{-part}\end{array}$$

$$\text{(Fr)} \quad \frac{\{?\epsilon xA\}, X}{X} \qquad \text{The premise is c.c.}$$

$$\text{(H)} \qquad \frac{\{!\epsilon x\ A = N\}, X \leq r}{\{?\epsilon x A\}, X} \qquad \begin{array}{l} \text{where } r = rk(\epsilon x A), \\ I(X) = J - 1 < I_{\max} \\ \|\epsilon x B\| = \epsilon x A, \\ CX : (\epsilon x A\# = N) \rightarrow \text{ true} \end{array}$$

This concludes the description of PA_ϵ.

f-derivation is a derivation containing no rules (Fr), (H), i.e. only the rule cut and axioms.

The remaining part of the paper is devoted to proving the following theorems.

THEOREM 1. *The empty sequent is derivable in PA_ϵ by an f-derivation.*

THEOREM 2. *Every f-derivation of an empty sequent can be transformed into a cutfree derivation by standard cut-reduction transformations.*

THEOREM 3. *A cutfree derivation of the empty sequent is finite and linear:*

$$X_0, X_1, \ldots, X_n$$

It begins with axiom Sol (sequent X_0), ends with the empty sequent (X_n), and consists only of rules (Fr), (H). The corresponding sequence $S_n, S_{(n-1)}, \ldots, S_0$ of epsilon substitutions $S_i = S_{X_i}$ is a convergence protocol of the Hilbert's substitution method: S_n is the zero substitution and $S_{(i-1)}$ either coincides with S_i or is obtained from S_i by Hilbert's step. S_0 is the solution substitution.

4. Construction of the original derivation. Rule CutFr. The structure of derivations.

To derive the empty sequent 0 expressing solvability of our fixed set E of critical formulas we proceed bottom-up introducing cuts until all terms in E (and all necessary subterms) are decided. Then the resulting sequent either is computationally inconsistent (c.i.) or falsifies one of the critical formulas, or determines a solution substitution. In all these cases it is an axiom of our calculus PA_ϵ and so the whole tree is well-founded. To provide bounds for the length of its branches, i.e. for the height of the whole tree it is convenient to use computation from outside in addition to the computation from inside encoded in the calculus CX.

We introduce a new CALCULUS C_1 which differs from CX only in the formulation of the rule $(\rightarrow_+)$:

$$\frac{!\epsilon x A{=}N \text{ occurs in } X}{\epsilon x A \rightarrow N}\ (\rightarrow_+)$$

The requirement $(\epsilon xA\# = N) \to$ true is dropped compared with the calculus CX of section 2.

The notation $e \to_1 v$ means that $e \to v$ is derivable in C_1. The meaning of $|e|_1, \|t\|_1$, 1-decided etc. should be obvious. The next proposition shows that eventually decidability coincides with 1-decidability.

LEMMA 4.1. *If the sequent X is correct, c.c. and 1-decided then it is decided.*

PROOF: Let $r(e)$ be the maximum rank of subterms of the expression (term or formula) e. We apply induction on $r(e)$ to show that $|e|$ exists. The induction base $r(e) = 0$ is evident since e does not contain an epsilon symbol, i.e. is a primitive recursive term or formula.

The induction step for a formula A follows from the induction step for terms. The induction step for a term e is proved by induction on the degree $d(e)$.

INDUCTION BASE $d(e) = 1$.

Let $e = \epsilon xA$. Since e is 1-decided, we have $?e$ in X (and $e \to 0$) or $!e = N$, i.e $!\epsilon xA, A[N], \sim A[i]\,(i < N)$ in X. Since $\deg(e) = 1$, we have $r(A[j]) < rk(e) = r(e)$. By induction hypothesis $|A[j]|\,(j \leq N)$ exists, and since X is c.c., $|A[N]| =$ true, $|A[i]| =$ false $(i < N)$.

So the rule $(\to_+)$ of the calculus C_1 yields $(\epsilon xA\# = N) \to$ true as required.

INDUCTION STEP FOR DEGREE-INDUCTION.

One has $e = \epsilon xA[x, u]$ with $\deg(u) < \deg(e)$. Since $e \to_1 N$ is obtained by the rule $(\text{trans})_1$ of C_1, and the left premise of this rule should be obtained by $(\epsilon \to)_1$

$$(\text{trans})\ \frac{(\epsilon \to)\ \dfrac{u \to M}{\epsilon xA[x,u] \to \epsilon xA[x,M];} \quad \epsilon xA[x,M] \to N}{\epsilon xA[x,u] \to N}$$

we can use induction hypothesis for u and $\epsilon xA[x, M]$ to drop the subscript 1 in the rules. ⊣

A sequent is *well-formed* (wf) if it has the form

$$(4) \qquad ?\epsilon x_1 A_1, \ldots, ?\epsilon x_n A_n, !\epsilon y_1 B_1 = N_1, \ldots, !\epsilon y_m B_m = N_m$$

where for every $j \leq m$ all formulas in $!\epsilon y_j B_j = N_j$ are in the f-part, or all of them are in the t-part. In other words all formulas except $?\epsilon x_i A_i$ are separated into disjont clusters

$$!\epsilon y_j B_j = N_j.$$

NOTE. Any sequent occurring in a derivation is wf. This is established by bottom-up induction on the derivation.

Indeed the last empty sequent is obviously well-formed with $n = m = 0$.

Cut adds $?\epsilon xA$ to the left premise and the cluster $!\epsilon xA{=}N$ to the N-th right premise. Fr adds $?\epsilon xA$. CutFr (cf. next page) is a combination of Cut and Fr. Finally H adds a cluster $\{!\epsilon xA{=}N\}$ after deleting $?\epsilon xA$ and all formulas of rank $> r$. The latter operation deletes some clusters: if the rank s of ϵyB is greater than r, then the rank of formulas in $\epsilon yB\# = N$ is equal to $(s-1)+$ which is still greater than r. If $s \leq r$, then $!\epsilon yB$ is not deleted, and the same is true for $\epsilon yB\# = N$. From now on all sequents are assumed to be wf unless the opposite is explicitly stated.

Note that for any wf sequent X and any $!\epsilon xA$ in X the value $|\epsilon xA|_1$ exists. We assume as before that all considered sequents are wf.

LEMMA 4.2. *Let X be a correct sequent, L be any finite set of epsilon terms and r be the maximum rank of undecided terms in X, E, L. Then there is a deduction of X of height $< wr$ by the rule cut from some axioms and decided correct sequents X' containing X which decide all terms in X', E, L.*

PROOF: We shall construct a derivation from 1-decided sequents which is sufficient by the Lemma 4.1. We use induction on r and (inner) induction on the number of 1-undecided terms of rank r. Take one such term ϵxA of minimum degree and write down the list

$$\epsilon x_1 A_1, \ldots, \epsilon x_k A_k (= \epsilon xA) \tag{1}$$

of its ϵ-subterms including itself which are not 1-decided in order of increasing degree (i.e. lesser degree comes first). We apply the rule cut bottom-up to decide the term $\epsilon x_1 A_1$ to be written ϵzC. All subterms of ϵzC are 1-decided, otherwise it would not be the first in (1). So $\epsilon zC \to_1 \epsilon yB$ of degree 1 (or is of degree 1 itself).

CASE 1. Neither of $?\epsilon yB, !\epsilon yB$ is present in X. Then ϵyB is 1-decided in all premises of the rule

$$(\text{cut})\ \frac{?\epsilon yB, X; \ldots !\epsilon yB = N, X; \ldots}{X}$$

since $\epsilon yB \to 0$ in the leftmost premise, and $\epsilon yB \to N$ in the $(N+1)$-th premise. So ϵzC is also 1-decided as required.

CASE 2. One of $?\epsilon yB, !\epsilon yB$ is in X. In the first case $|\epsilon yB| = 0$, in the second case $|\epsilon yB|_1$ exists since X is wf. $\dashv$

PROOF OF THEOREM 1 (cf. section 3): First apply Lemma 4.2 to the empty sequent and the empty list of terms. Consider an arbitrary top sequent X of the resulting figure. We show how to extend X to an axiom if it is not one, i.e. if X is c.c. and $I(X) < I_{\max}$. Since X is decided we have $Cr_J \to$ false for $J = I(X) + 1$. Then writing

$$Cr_J : A[t] \to A[\epsilon x A] \tag{2}$$

we have $A[t] \to$ true, $A[\epsilon x A] \to$ false. Since all subterms in (2) are decided we have $t \to N$ for some N. Since $\epsilon x A$ is decided we have $?||\epsilon x A||$ in X and $|\epsilon x A| = 0$. Indeed, if $!||\epsilon x A||$ is in X and $|\epsilon x A| = M$ then $A[M] \to$ true by the Lemma 2.3 and so $A[\epsilon x A] \to$ true, which contradicts $A[\epsilon x A] \to$ false.

Now let L be the list of all ϵ-terms occuring in the formulas

$$A[N-1], \ldots, A[0] \tag{3}$$

Apply Lemma 4.2 to sequent X and list L. Each of the top sequents X' of the resulting deduction which is c.c. decides all formulas in (3). Let K be the least numeral such that $CX' : A[K] \to$ true. Such a K exists since $A[N] \to$ true. Since $Cr_J \to$ false we have $I(X') = I(X)$ and the sequent X' is an instance of the axiom Axc . ⊣

During cutelimination the cut rule will be gradually replaced by other rules, but for technical reasons (to be explained in the subsequent sections) some traces of eliminated cuts will be left in the derivation in the form of the rule CutFr below.

We introduce a new system PA_ϵ^+ by adding to PA_ϵ the new rule:

$$\text{(CutFr)}\ \frac{\{?\epsilon x A\}, X; \ldots !\epsilon x A{=}N, X; \ldots}{X}\quad \begin{array}{l}\text{The leftmost premise is c.c.}\\ !\epsilon x A{=}N \text{ are in the } f\text{-part}\end{array}$$

In other words the leftmost premise is exactly as in the rule (Fr), and remaining premises are exactly as the corresponding premises of Cut, i.e. are the same as in the omega-rule. We call the leftmost premise of a cut or CutFr the *major* premise of that rule, and remaining premises its *minor* premises.

THEOREM 4.3. *Let d be a derivation and X be a sequent in d.*

(a) *If $?\epsilon x A$ is in X then either of the following holds:*
(a1) *$?\epsilon x A$ is in X_f and traceable to the major premise of a cut, or*
(a2) *$?\epsilon x A$ is in X_t and traceable to the major premise of a Fr or CutFr.*

(b) *If* $!\epsilon xA$ *is in* X *then either of the following holds:*
(b1) $!\epsilon xA$ *is in* X_f *and traceable to a minor premise of a cut or CutFr;*
(b2) $!\epsilon xA$ *is in* X_t *and traceable to an H-rule.*
(c) *If a basic formula* A *is a member of* X *then either*
(c1) A *is in* X_f *and traceable to a side formula in a minor premise of a cut or CutFr, or*
(c2) A *is in* X_t *and traceable to a side formula of an H-rule.*

The proof is by bottom-up induction on the derivation. The induction base is trivial since the last sequent is empty. The induction step is proved by cases depending on the last rule. Only side formulas of this rule are to be considered. ⊣

Call the derivation *cutfree* if it contains neither cut nor CutFr.

COROLLARY 4.4. *Let* d *be a cutfree derivation of the empty sequent.*

(a) *only rules Fr and H are used in* d*;*
(b) $X = X_t$ *for every sequent* X *in* d*;*
(c) *any sequent* X *in* d *is supported and c.c.;*
(d) d *begins with the axiom Sol.*

PROOF: (a) By the definition of cutfree.
(b) By Lemma 4.3 any formula in X_f should be traceable to a cut or CutFr below X, so X_f is empty by (a).
(c) Use bottom-up induction on d. The empty sequent is obviously c.c. and supported. Rule (Fr) has c.c. proviso and its premise does not contain new basic formulas compared to the conclusion.

Consider the rule H

$$\frac{\{!\epsilon xA{=}N\}, X \leq r}{\{?\epsilon xA\}, X}$$

Assume that the conclusion $\{?\epsilon xA\}, X$ is a correct supported c.c. sequent. Then $!\epsilon xA$ does not occur in X, hence the premise $Y = \{!\epsilon xA{=}N\}, X \leq r$ is also correct. Suppose for contradiction that $CY : B \to$ false for some basic formula B in Y. Then $rk(B) < r$ since formulas of rank exactly r in Y are of the form $?\epsilon zC, !\epsilon zC$. By Lemma 2.8 we have $C(Y \leq r) : B \to$ false. The cluster $\{!\epsilon xA{=}N\}$ cannot be used in this derivation unless B is a formula in $\epsilon xA\# = N$. In that latter case we would have $C(X < r) : B \to$ false which contradicts the proviso of the H-rule: $CX : (\epsilon xA\# = N) \to$ true. All basic formulas in $X \leq r$ are supported as before (by Lemma 2.8) together with all basic formulas in $\epsilon xA\# = N$, as explained above.

This argument is our version of Ackermann's [1940] correctness argument (presented in Hilbert-Bernays [1970], section 2 of the Supplement

V) showing that the next substitution S' is correct if the previous substitution S was correct.
(d) The uppermost sequent should be an axiom. It cannot be the axiom False because of (c), and cannot be Axc since the f-part is empty. ⊣

LEMMA 4.5. *If X/X_1 is an inference according to the rule Fr then the corresponding epsilon substitutions coincide: $S_{X_1} = S_X$.*

If X/X' is an inference according to the rule (H) then the corresponding epsilon substitution S_{X_1} is a successor of S_X, i.e. is obtained from S_X by Hilbert's step : $S_{X_1} = (S_X)'$.

PROOF: By Corollary 4.4(c) all basic formulas in clusters $!\epsilon xA{=}N$ are supported, hence if such a cluster is present in a sequent X then the substitution S_X contains a component $\epsilon xA{=}N$, and consists exactly of such components plus zero substitutions for all remaining canonical terms. So the rule Fr only makes explicit existing zero components, and does not change the substitution. The rule H eliminates all components of rank $> r$ (by eliminating whole clusters) and adds one new component of rank r (in the form of supported cluster $!\epsilon xA{=}N$) exactly as required in Hilbert's substitution method. ⊣

PROOF OF THEOREM 3 (cf. section 3): Apply Corollary 4.4 and Lemma 4.5.

LEMMA 4.6. *If X is a sequent in a derivation of an empty sequent, then any basic formula in X_t is supported and any formula $!\epsilon xA$ is 1-decided.*

PROOF is by bottom-up induction on the derivation. New basic formulas in the provisional part X_t appear only in the rule (H), and there they are supported by Lemma 2.8. New formulas of the form $!\epsilon xA$ in X_t are 1-decided by the same argument. In X_f they appear in the minor premises of the rules (Fr) and (CutFr), and they are 1-decided by the very form of these premises. If some formula F looses its support or 1-support in the rule (H) since some of the supporting formula is deleted, then F is also deleted. ⊣

5. The weakening rule

The *weakening* (or *thinning*) *rule*

$$(1) \qquad \frac{X}{X, A}$$

in some form (postulated or proved admissible) is frequently used in the process of cutelimination. It is not in general admissible in our systems

with the ϵ-symbol because adding a new formula (say $?\epsilon xA$) to a correct sequent (say $!\epsilon xA = 1$) can make this sequent incorrect. The usual method of proving the admissibility of weakening (1) is to add formula A to the whole derivation of the premise X. Sometimes this transformation is called multiplication of the derivation by A. Even if correctness of sequents is preserved, some of our rules can be destroyed by this transformation for reasons similar to the unstability of the proviso for (generalized) eigenvariables of quantifier rules or (even more relevant) the restriction in the necessity-introduction rule of $S4$. If the derivation of X in the Gentzen-type formulation of the modal logic $S4$ contains this rule, one has to drop A altogether at the conclusion of this rule. Our treatment of weakening will be along these lines.

Recall that we consider wf sequents of the form

$$(*) \qquad ?\epsilon x_1 A_1, \ldots, ?\epsilon x_n A_n, !\epsilon y_1 B_1 = N_1, \ldots, !\epsilon y_m B_m = N_m$$

or $X?, X!$ for short.

The *product* $X * Y$ of two sequents will be defined by concatenating them and identifying the whole clusters.

DEFINITION. *Let X, Y be correct sequents, X be of the form $(*)$ and Y be of the the form*

$$X_1?, X_2!, Y_1?, Y_2!$$

where the parts of Y coinciding with the corresponding part of X are shown first, and the remaining parts are shown later. Then $X * Y$ is $X, Y_1?, Y_2!$ with the t-part preferred: if at least one of the identified formulas is in the t-part, the result is also placed in the t-part. If identified clusters $!\epsilon xA{=}N_1, !\epsilon xA{=}N_2$ have different N_1, N_2, then the maximum is preferred (in fact this case will never occur).

EXAMPLE. Assuming A_i to be different for different i we have

$$?\epsilon xA_1, ?\epsilon xA_2, !\epsilon xA_3 = 5 * ?\epsilon xA_1, !\epsilon xA_3 = 5, !\epsilon xA_4 = 1 =$$
$$?\epsilon xA_1, ?\epsilon xA_2, !\epsilon xA_3 = 5, !\epsilon xA_4 = 1$$

and

$$?\epsilon xA_1, ?\epsilon xA_2, \{!\epsilon xA_3 = 5\} * ?\epsilon xA_1, !\epsilon xA_3 = 5, !\epsilon xA_4 = 1 =$$
$$?\epsilon xA_1, ?\epsilon xA_2, \{!\epsilon xA_3 = 5\}, !\epsilon xA_4 = 1$$

Recall that $\{\}$ means being in the provisional part. A sequent X is *supported* (*t-supported*) if $CX : A \to$ true for every basic formula A in X (in X_t).

NOTE. If sequents X and Y are correct and do not conflict in a formula of the form $?\epsilon xA$ (i.e. one contains $?\epsilon xA$ and the second $!\epsilon xA$) then $X * Y$ is correct since such a conflict is the only possible reason of incorrectness.

LEMMA 5.1. (a) *If X, Y are c.c. supported sequents, and $X * Y$ is correct then $X * Y$ is also a supported c.c. sequent.*

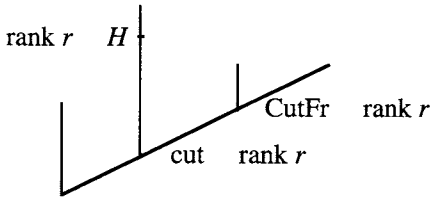
PROOF: The sequent $X * Y$ is supported since the X-part is supported by X and the Y-part is supported by Y. To prove computational consistency of $X * Y$ consider a derivation of $A \to$ false in $C(X * Y)$ for (say) A in X. Since X is c.c. and supported we have a derivation of $A \to$ true in CX. We apply induction on these derivations to prove that $A \to$ false for a formula A in X, Y, or $t \to N, N_1$ with different N, N_1, or $t \to t_1, t_2$ for different t_1, t_2 implies that $X * Y$ is incorrect. This induction is similar to the one used in the proof of the Church-Rosser property for CX. The induction step goes smoothly exept in the situation when t is canonical and the derivations of $t \to N, N_1$ are given. They should end in some combination of the rules $(\to_0), (\to_+)$. These rules cannot both be $(\to_0)$ since then $N = N_1 = 0$. If both of them are $(\to_+)$, use the induction assumption for the formula $A[\min(N, N_1)]$. Otherwise $X * Y$ contains $?\epsilon x A$ and $!\epsilon x A$ and so is incorrect. ⊣

LEMMA 5.2. *If X, Y are c.c. sequents, X is supported and $X * Y$ is correct but not c.c. then there is a basic formula A in Y not decided by Y and such that $C(X * Y) : A \to$ false.*

PROOF: Again apply induction on the pair of derivations of $A \to$ false in $C(X * Y)$ and $A \to$ true in CX or CY or a pair $t \to N, N_1$ with different N, N_1. ⊣

Cuts will be eliminated in the usual way beginning with maximum rank r. Eliminated cuts of rank r will be replaced by CutFr and H with the same main term, i.e. with the same rank. More precisely, a cut will be replaced by CutFr and then moved (permuted) up the derivation until one encounters Axc with main term $?\epsilon x A$ traceable to the main formula $?\epsilon x A$ of that CutFr. Then the Axc is replaced by the rule H , and the derivation of the corresponding right premise of the cut is placed over the H-rule. After all cuts of rank r are eliminated, these CutFr will be pruned to Fr. So finally cuts of rank r will be replaced by Fr of rank r. The cut to be eliminated will be one of the uppermost cuts of rank r. This motivates the following

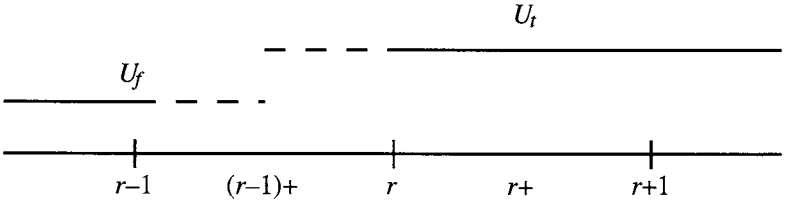
DEFINITION. *A derivation d is an r-*derivation *if it contains cuts of rank at most r, CutFr only of rank r, H only of rank $\geq r$, Fr only of rank $> r$, and no CutFr or H of rank r is below any cut of rank r.*



A sequent U is an r-sequent if all formulas in its provisional part U_t are of rank $> r-1$, and all formulas in its fixed part U_f have rank $\leq r$:

$$U_t \text{ is contained in } U(> r-1) \text{ and } U_f \text{ is contained in } U(\leq r).$$

So U_f consists of $U(\leq r-1)$ and some part of $U(\leq (r-1)+)$ and $U(=r)$, while U_t consists of $U(>r)$ and the remaining part of $U(\leq (r-1)+)$ and $U(=r)$:



Recall that basic formulas containing terms of rank $r-1$ have rank $(r-1)+ > (r-1)$ by the definition. This is motivated by the fact that all formulas in $\{!\epsilon xA{=}N\}$ for ϵxA of rank r should be treated together.

LEMMA 5.3. *Every sequent in an r-derivation is a t-supported r-sequent.*

PROOF is by the bottom-up induction on the derivation. Induction base (empty sequent) is obvious. For induction step consider cases according to the last rule applied.

RULE Fr. $?\epsilon xA$ is added to t-part and is of rank $> r$. Everything else in the t-part is supported in view of the induction assumption. The f-part is not changed at all.

RULE Cut. $?\epsilon xA$ is added to the f-part of the major (leftmost) premise. Its rank is $\leq r$, so the conditions for the r-sequent are satisfied and no new conditions for support have to be checked in this premise. In the remaining premises formulas $!\epsilon xA{=}N$ of rank $\leq r$ are added to f-part, so again conditions for the r-sequent are satisfied and no new conditions for support have to be checked.

RULE CutFr. This is the combination of the two previous cases.

RULE H. Here $\{!\epsilon xA{=}N\}$ with ϵxA of rank $s \geq (r-1)+ > r-1$ is added to the t-part and all formulas of rank $> s$ as well as $?\epsilon xA$ are deleted (from the conclusion). So the condition for the t-part in the definition of the r-sequent is preserved, and no new formulas are added to the f-part. The added formulas $\{\epsilon xA\# = N\}$ in the t-part are supported in view of the proviso in the rule H. The remaining formulas $X \leq r$ in the t-part of the premise are supported by the same derivations as for the conclusion: all of them have rank $\leq r$, and by Lemma 2.4 only formulas of rank $\leq r$ were used in the supporting derivation. But none of those formulas was deleted. ⊣

LEMMA 5.4. (a) *Let Z, U be c.c. t-supported r-sequents and $Z * U$ be correct. Then if $Z_f(\leq r) * U_f(\leq r)$ is c.c. then $Z * U$ is a c.c. t- supported r-sequent.*

(b) *If Z, U are correct c.c. t-supported r-sequents, Z coincides with $Z(\leq r)$ and contains $U(\leq r)$, then $Z{*}U$ is a correct c.c. t-supported r-sequent.*

PROOF: (a) $Z * U$ is a t-supported r-sequent since Z, U are such. We prove uniqueness of $|t|, ||t||, |A|$ by induction on the derivations in the C-calculus. If the rank is at most $r+$ then the derivations use only $Z(\leq r) * U(\leq r)$. These in turn use only $Z_f(\leq r) * U_f(\leq r)$ since (by the definition of r-sequent) the only formulas in the difference set are of the form $?\epsilon xA, !\epsilon xA{=}N$, with ϵxA of rank r, but the derivation of $\epsilon xA \to N$ uses only formulas of rank $< r$ which are in the f-part.

For the terms and formulas of rank $> r+$ we use induction on rank and the fact that $A \to$ true in CZ or CU for each basic formula A of rank $> r$ in $Z * U$. Indeed in this case A is in the t-part and Z, U are t-supported. (b) In this case $Z_f(\leq r) * U_f(\leq r)$ is $Z_f(\leq r)$ which is c.c. by the assumption. ⊣

LEMMA 5.5. *Let the conditions of Lemma 5.4(b) be satisfied and R be an inference rule with conclusion U. Then*

(a) if R is (Fr) with side formula $\{?\epsilon xA\}$ and $rk(?\epsilon xA) > r$ or

(b) R is (H) of rank $s > r$ and all terms in Z are decided,

then $Z * R$, i.e. the result of multiplying Z by all premises and the conclusion of R, is the same rule as R, and all premises are turned into correct c.c. t-supported r-sequents.

PROOF: (a) Consider

$$\frac{\{?\epsilon xA\}, U}{U}\ (R) \quad \text{and} \quad \frac{\{?\epsilon xA\}, Z * U}{Z * U}\ (Z * R)$$

Since $rk(\epsilon xA) > r$, and $rk(Z) \leq r$, neither of $?\epsilon xA, !\epsilon xA$ occurs in Z. The conditions of Lemma 5.4(b) are satisfied also by Z and $\{?\epsilon xA\}, U$, and moreover $Z * (\{?\epsilon xA\}, U) = ?\epsilon xA, (Z * U)$. So the premise is a correct c.c. t-supported r-sequent and $Z * R$ is an Fr-inference.
(b) Consider

$$\frac{\{!\epsilon xB, B[K], \sim B[i], i < K\}\,,\ X(\leq s)}{\{?\epsilon xB\}, X}\ (H)$$

$$\frac{\{!\epsilon xB, B[K], \sim B[i], i < K\}\,,\ Z * X(\leq s)}{?\epsilon xB, Z * X}\ (Z * H)$$

By the proviso for the rule (H) and the subformula property of the calculus C we see that the premise $\{!\epsilon xB = K\}, X(\leq s)$ is a t-supported r-sequent, and it is c.c. decided. Since Z is also t-supported, the conditions of Lemma 5.4(b) are satisfied for the premise, so the result of multiplication by Z is also c.c. decided, and since all terms in Z are decided, the proviso of H is satisfied. ⊣

DEFINITION. Let d be a deduction, that is a derivation from hypotheses, and Z be a correct wf sequent. Let us define a figure $Z * d$ by induction on d. [$Z * d$ will not be a deduction in general.] Roughly speaking $Z * d$ is the result of multiplying all sequents in d by Z, deleting everything which is above incorrect or computationally inconsistent (c.i.) sequents and pruning the branches over the premises of (cut) and (CutFr) which have become redundant. More precisely, some members of Z can be deleted in the process of going up the derivation when the rule H is encountered. So in fact $Z * d$ is defined in terms of $W * d'$ for W contained in Z and subderivations d' of d. We shall verify later that $Z * d$ is again a derivation in the cases we need. In the remaining (degenerate) cases the definition will be more or less arbitrary.

INDUCTION BASE. U is the last sequent in d. If Z*U is incorrect (degenerate case) or c.i. then $Z * d$ consists of the only sequent Z*U analysed as AxF if it is correct but c.i.

INDUCTION STEP. Let $W * d'$ be defined for all correct wf sequents W and immediate subderivations d' of the given derivation d. Let d end in a rule R with the conclusion U. If $Z * U$ is incorrect or c.i. then $Z * d$ consists of the only sequent $Z * U$. If $Z * U$ is correct and c.c. consider the following cases.

CASE 1. R is cut or (CutFr). If Z does not contain the cut formula, i.e. $?xA$ or $!\epsilon xA$, then multiply premises by Z and apply the rule R. This means that

$$d: \quad \frac{d_1 : U_1; \ldots d_n : U_n \ldots}{U}$$

is transformed into

$$Z * d: \quad \frac{Z * d_1 : Z * U_1; \ldots Z * d_n : Z * U_n \ldots}{Z * U}$$

If Z contains the cut formula, then prune the branch of R containing the complementary cut formula, delete the cut and proceed up the premise containing the considered cut formula:

$$d: \quad \frac{\begin{array}{cc} d_? & d_N \\ ?\epsilon xA, U; & !\epsilon x = N, U \end{array}}{U}$$

Derivation d is transformed into

$$?\epsilon xA, Z^{\sim} * d_? \; : ?\epsilon xA, Z^{\sim} * U$$

or

$$!\epsilon xA{=}N, Z * d_N \; : !\epsilon xA{=}N, Z^{\sim} * U$$

where $Z = ?\epsilon xA, Z^{\sim}$ or $Z = !\epsilon xA, Z^{\sim}$ with one exception. In a degenerate case when the rule is CutFr and $?\epsilon xA$ is in the fixed part Z_f, $Z * d$ is the figure

$$\frac{\{?\epsilon xA\}, Z \sim *U}{?\epsilon xA, Z \sim *U}$$

which is not an application of any rule.

CASE 2. R is (Fr). If Z does not contain the Fr-formula, i.e. $?\epsilon xA, !\epsilon xA$, then multiply by Z:

$$\frac{\{?\epsilon xA\}, U}{U}$$

is transformed into

$$Z * d: \quad \frac{\{?\epsilon xA\}, (Z * U)}{Z * U}$$

If Z contains $?\epsilon xA$ in the t-part, prune (Fr). If Z contains $!\epsilon xA$ (degenerate case) then $Z * d$ consists of the incorrect sequent $?\epsilon xA, !\epsilon xA, Z * U$.

In the degenerate case when Z_f contains $?\epsilon xA$ the derivation $Z * d$ consists of the only sequent $Z * U$.

CASE 3. R is (H). If Z does not contain $?\epsilon xA, !\epsilon xA$, multiply the premise by $Z(\leq r)$.

If Z contains $?\epsilon xA$, i.e. $Z = ?\epsilon xA, V$, multiply the premise by V.

Sequent Z cannot contain $!\epsilon xA$ since the conclusion U of the (H)-rule contains $?\epsilon xA$, and $Z * U$ is correct by the assumption. This concludes the definition.

We now begin to investigate sufficient conditions of the admissibility of the thinning rule $X/X, Y$.

Consider the structure of a derivation of the right premise of an uppermost rank r-cut in an r-derivation.

LEMMA 5.6. *Let X be a sequent in an r-derivation of the empty sequent 0 with no (CutFr) of rank r under X. Then*

$$Xf = X(\leq r); \qquad X_t = X(> r) \tag{7}$$

PROOF is by bottom-up induction. For the last sequent 0 all parts in (7) are empty. Consider possible inference rules.

CASE 1. Cut rule

$$\frac{?\epsilon xA, U; \qquad \ldots !\epsilon xA{=}N, U}{U}$$

Here $rk(\epsilon xA) \leq r$ and $?, !\epsilon xA{=}N$ is in the f-part, $rk(A[j]) \leq (r - 1)+ < r$, and so $(?, !\epsilon xA{=}N, U)_f = ?, !\epsilon xA{=}N, U_f = ?, !\epsilon xA{=}N, U(\leq r) = (?, !\epsilon xA{=}N, U)(\leq r)$, and $(?, !\epsilon xA{=}N, U)_t = U_t = U(> r) = (?, !\epsilon xA{=}N, U)(> r)$.

CASE 2. Rule (Fr).

$$\frac{\{?\epsilon xA\}, U}{U}\ (\mathrm{Fr})$$

By the definition of r-derivation $rk(\epsilon xA) > r$. Hence $(\{?\epsilon xA\}, U)_f = U_f = U(\leq r) = (\{?\epsilon xA\}, U)(\leq r)$, and $(\{?\epsilon xA\}, U)_t = ?\epsilon xA, U_t = ?\epsilon xA, U(> r) = (\{?\epsilon xA\}, U)(> r)$.

CASE 3. Rule (H).

$$\frac{\{!\epsilon xA{=}N\}, U}{\{?\epsilon xA\}, U}$$

By the induction hypothesis $rk(\epsilon xA) > r$, and so $rk(!\epsilon xA{=}N) \geq r+ > r$. We have $(\{!\epsilon xA{=}N\}, U)_f = U_f = U(\leq r) = (!\epsilon xA{=}N), U(\leq r)$, and $(\{!\epsilon xA{=}N\}, U)_t = !\epsilon xA{=}N, U_t = !\epsilon x = N, U(> r) = (\{!\epsilon xA{=}N\}, U)(> r)$.

CASE 4. Rule (CutFr). This case is impossible by the condition of the lemma since $rk(\text{CutFr}) = r$ in any r-derivation. ⊣

LEMMA 5.7. *Consider an r-derivation with a branch*

$$\begin{array}{lll} Axc & ?\epsilon yB, Z, V & \\ & \Big| & rk(\epsilon yB) = r, rk(Z) \leq r, \\ & \dfrac{?\epsilon yB, \ldots !\epsilon yB = N, X}{X}\ cut & rk(V) > r, (\epsilon yB = N) \rightarrow true \\ \Big| & & \\ d & & \\ \Big| & 0 & \end{array}$$

containing the left premise of an uppermost cut of rank r and corresponding axiom Axc . Let d be a deduction obtained by deleting from the given derivation the part above X. Then $!\epsilon yB = N, Z * d$ is a deduction of $!\epsilon yB = N, Z$ from $!\epsilon yB = N, Z * X$ by the rules (Fr), (H):

$$(Fr, H) \qquad \begin{array}{c} !\epsilon yB = N, Z * X \\ \Big| \\ Z, !\epsilon yB = N \end{array}$$

PROOF: We shall prove the following claims:

(a) Z contains $X(\leq r)$

(b) the segment from 0 to X contains neither (Fr) or H-rules of rank $\leq r$ nor (CutFr), and after multiplying by $Z, !\epsilon yB = N$ all cuts are pruned from this segment, but a branch contained in the segment is always chosen;

(c) all remaining rules, i.e. Fr and (H), are preserved, and all sequents $(!\epsilon yB = N, Z) * U$ in the segment are c.c. r-sequents.

Indeed, by the definition of r-derivation all (CutFr) have rank r, all H-rules have rank $\geq r$ and there are no (CutFr) or H-rules of rank r under any cut of rank r. So no formula of rank $\leq r$, in particular no side formula of a cut, is deleted in any H-rule (viewed bottom-up), and $Z(\leq r)$ inherits all these formulas from X, which proves (a). Claim (b) follows from the condition on the arrangement of cuts, H-rules and (CutFr) in an r-derivation. Note that rank$(H) \geq r$, rank (CutFr) $= r$, rank (Fr) $> r$ in an r-derivation. To prove claim (c) we verify conditions of Lemma 5.4. The relation $Z = Z(\leq r)$ is equivalent to $rk(Z) \leq r$. Since $?\epsilon yB, Z, V$ is decided, and $?\epsilon yB, Z$ is of rank $\leq r$, sequent Z is supported by the

subformula property of the calculus C. It is c.c. since $?\epsilon yB, Z, V$ is c.c. And it was just proved that Z is contained in $U(\leq r)$. Rules not listed in claims (a),(b) are (Fr) and (H) of rank $> r$. Their conclusions and premises are c.c. t-supported r-sequents. So by Lemma 5.5 multiplication by $Z(\leq r)$ preserves them. Claim (c) is established. ⊣

LEMMA 5.8. *Let d be an r-derivation of a sequent $?\epsilon xA, X$ with $?\epsilon xA$ of rank r in the f-part. Then the result of provisionalization of $?\epsilon xA$, i.e. moving $?\epsilon xA$ into the t-part is a deduction of a sequent*

$$\{?\epsilon xA\}, X$$

where all inference rules and axioms except Axc with main formula $\{?\epsilon xA\}$ are correct.

PROOF: By inspection of axioms and inference rules. ⊣

LEMMA 5.9. *Let d be a part of an R-derivation of 0 ending in an uppermost cut of rank r:*

$$\frac{d_? \quad ?t, X \quad \ldots !t = N, X \ldots \quad d_N}{\begin{array}{c} X \\ \ldots d- \end{array}} (c)$$

Then

(a) *$d_?, d_N$ contain only (Fr)-rules of rank$> r$, CutFrs of rank r, cuts of rank $< r$, H-rules of rank $\geq r$; any H-rule of rank r in $d_?, d_N$ is traceable to (CutFr) in $d_?, d_N$ (not in $d-$); any axiom Axc of rank r in d_N is traceable to a cut in $d-$; axiom Axc in $d_?$ can also be traceable to a cut in $d-$ and to the explicitly shown cut.*

(b) *$X_f = X(\leq r); X_t = X(> r)$*

(c1) *for any sequent V in $d_?$ one has $V_f = X_f \cup ?t \cup$ side formulas of cuts of rank $< r$ and of CutFrs of rank r (fixed premises) in $d_?$*

(c2) *for any axiom Axc : $?t, Z$ in $d_?$ one has*
$Z(\leq r) = X_f \cup$ side formulas of cuts of rank $< r$ in $d_? \cup$
fixed side formulas $!u = K$ of CutFrs of rank r in $d_? \cup$
(provisional) side formulas $\{!u = K\}$ of H-rules of rank r in $d_? \cup$
provisional side formulas $\{?u\}$ of CutFr of rank r.
In particular $(Z_f)_r = (X_f)_r \cup \{\{!u = K\} : rk(u) = r\}$ where V_r is $V \leq r$ intersected with $V(> (r-1))$.

(d) *for any sequent U in d_N one has*
$U_f = X_f \cup !t = N \cup$ side formulas of cuts of rank $< r$ and of CutFrs (fixed) of rank r in d_N.
$U_{f_r} = X_{f_r} \cup \{\{!u = K\} : rk(u) = r\}$.

This is summarized as follows:

$$\begin{array}{lll} Axc & ?t, Z & rk(\mathrm{H}) \geq r, rk(\mathrm{Cut}) < r, \\ & | & rk(\mathrm{CutFr}) = r, \\ d_? & \dfrac{?t, X \qquad !t = N, X}{X}\,(c) & rk(\mathrm{Fr}) > r, \\ & & Axc : ?u, w_1 \end{array}$$

$$d- \qquad \dfrac{\dfrac{?u, W \quad !u = M, W}{W}}{0}$$

PROOF: (a) The definition of an r-derivation implies the condition on the ranks of (Fr) (Cut) and (CutFr), and the bound $\geq r$ for ranks of remaining rules. The sharp bound $> r$ for CutFr and tracing of Axc follows since there are no cuts above (c). Part $d-$ below the cut (c) cannot contain (CutFr) by the definition of r-derivation which concludes the proof of (a). Parts (b)–(d) follow from (a). ⊣

THEOREM 5.10. *Under the conditions of Lemma 5.9 let Axc:* $?t, Z, V$ *with* $(?t, Z) = (?t, Z, V)(\leq r)$, $CZ : t\# = N \rightarrow$ *true be traceable to the main premise of cut (c). Then* $d_N * \{!t = N\}, Z$ *is an* r*-derivation.*

PROOF: In fact we prove by bottom-up induction on d_N the following lemma.

LEMMA 5.11. *Under the conditions of Lemma 5.9 let* U *be a sequent in* d_N *which is not pruned in the process of multiplication of* d_N *by* $\{!t = N\}, Z$, *and let* W *be the sequent to which* U *is multiplied in the process of computing* $d_N * (!t = N, Z)$. *Then*

(a) $U * W$ *is correct;*
(b) *if* $U * W$ *is the topmost sequent, it is an axiom;*
(c) *if some formulas in* U *and* W *are identified, and only one of them is in the* f*-part, it is the formula in* U*;*
(d) *every rule left in* $d_N * \{!t = N\}, Z$ *up to* $U * W$ *is correct.*

PROOF: Note that the sequent W for every U is of the form

$$(3) \quad W = \{!t = N\}Z';\ Z = Z' \cup \{\{?u_1\}, \ldots, \{?u_n\}\}, n \geq 0, rk(u_i) = r$$

i.e. Z' is obtained from Z by deleting of some formulas $\{?u\}$ with $rk(u) = r$. Indeed, only formulas of the form $\{?u\}$ can be dropped from W in the

process of the multiplication, so $rk(u) \leq r$ and since $\{!t = N\}, Z$ is an r-sequent and $\{?u\}$ is in Z, we have $rk(u) = r$. Since $\{!t = N\}, Z$ is a supported correct r-sequent of rank $\leq r$ and by the subformula property of the calculus C the sequent W is also a supported correct r-sequent of rank $\leq r$.

INDUCTION BASE: sequent $X * \{!t = N, Z\}$. From the fact that $\{!t = N\}, Z = W = W(\leq r)$ contains $X_f = X(\leq r)$ by Lemma 5.9 (c2), and the correctness of W it follows that

$$X * W = (X(\leq r), X(> r)) * W = W, X(> r)$$

is correct. This sequent is also a c.c. r-sequent by Lemma 5.4(b). (b) follows from $X(\leq r) = X_f$.

INDUCTION STEP. Let R be a rule in d_N, and (a)–(d) be verified up to conclusion U of R. If $U * W$ is c.i., then it is an axiom, we are done. Otherwise $U * W$ is c.c. and we consider cases depending of R.

CASE 1. U was an axiom in d. Then $U * W$ is an axiom of the same kind: since W is supported and U is supported, $W * U$ is supported, so all restrictions are satisfied for the new axiom.

CASE 2. R is Fr (of $rk > r$ by Lemma 5.9 (a))

$$\frac{\{?u\}, U}{U}\ (\mathrm{Fr}) \quad \frac{\{?u\}, U * W}{U * W}\ rk(u) > r$$

Since $\{?u\}$ is added to the common part of U and $U(> r)$, the sequent $\{?u\}, U * W$ is correct and $\{?u\}U$ is c.c. The sequent $U(\leq r) * W$ is a correct c.c. r-sequent by induction assumption, so $\{?u\}, U * W$ is a c.c. r-sequent by Lemma 5.4(a), and $R * W$ is again an instance of (Fr) of rank $> r$.

CASE 3. R is (Cut) (of rank $< r$ by Lemma 5.9(a))

$$\frac{\overset{e_0}{?u, U; \ldots} \quad \overset{e_N}{!u = N, U; \ldots}}{U}\ (R) \qquad rk(u) < r$$

CASE 3.1. Neither of $?u, !u$ is in W. Then all premises in

$$\frac{?u, U * W; \ldots !u = N, U * W \ldots}{U * W}\ (R * W)$$

are correct. They are r-sequents since $?u$ or $!u$ is added to the f-part and is of rank $< r$. So $R * W$ is a correct application of (Cut).

CASE 3.2. One of $?u, !u$ is in W. Then it is in W_f since $rk(u) < r$ and $?t, Z$ is an r-sequent. In particular u is different from t. So the cut is pruned into repetition

$$\frac{e_0 * W}{\dfrac{?u, U * W}{U * W, ?u}\ Rep} \qquad \text{or} \qquad \frac{e_M * W}{\dfrac{!u = M, U * W}{U * W, !u = M}}$$

We used the fact that W is wf and supported, so if $!u$ is in W, then $!u = M$ is in W for some M.

CASE 4. R is (CutFr) (of rank r by the definition of r-derivation):

$$\frac{\{?u\}, U; \ldots !u = N, U; \ldots}{U}\ (\text{CutFr})$$

CASE 4.1. Neither of $?u, !u$ is in W. Then $R * W$ is again (CutFr):

$$\frac{\{?u\}, U * W; \quad !u = N, U * W}{U * W}\ rk(u) = r$$

Indeed, the premises are correct since $?u, !u$ are not in W. They are r-sequents since only $\{?u\}$ is added to the t-part and $rk(u) = r$. The premise $\{?u\}, U * W$ is c.c. by Lemma 5.4(a), since $\{?u\}U$ is c.c.

CASE 4.2. $?u$ is in W. If $?u$ is in W_f then $?u$ is in X_{f^r} by Lemma 5.9(c2) and then $?u$ is in U_f, so the main premise of (CutFr) would be incorrect. So $?u$ is in W_t, i.e. $W = \{?u\}, W'$, and then (CutFr) is pruned into (Rep):

$$\frac{\{?u\}, U * W'}{U * W', \{?u\}}$$

CASE 4.3. $!u$ is in W. Since W is wf and supported, $!u = M$ for some M is in W, i.e. $W = W', !u = M$. If $!u$ is in W_f then CutFr is pruned exactly as a cut.

$$\frac{!u = M, U * W'}{U * W', !u = M}$$

If $!u$ is in W_t, then it is supported, so $(!u = M, U) * (\{!u = M\}, W) = \{!u = M\}, U * W$ and (CutFr) can be pruned as before.

CASE 5. R is the H-rule.

$$\frac{\{!u = M\}, U(\leq s)}{\{?u), U}\ rk(u) = s \geq r, \quad CU : u\# = M \rightarrow \text{true}$$

CASE 5.1. Neither of $?u, !u$ is in W. By the assumption the sequent $\{?u\}$, $U * W$ is a c.c. correct r-sequent. Since U supports $u = M$, that sequent also supports $u = M$. We have $(U * W)(\leq s) = U(\leq s) * W(\leq s) = (U \leq s) * W$, so $R * W$ is again the H-rule:

$$\frac{\{!u = M\}, U(\leq s) * W}{\{?u\}, U * W}$$

CASE 5.2. One of $?u, !u$ is in W. It cannot be $!u$, since otherwise $(?u, U) * W$ would be incorrect. So $W = \{?u\}, W$ where $?u$ should be in the t-part since otherwise $?u$ is in X_f which is contained in U_f. Now $(\{?u, U) * (\{?u\}, W) = \{?u\}, U * W$ and since $\{?u\}$ is pruned from W the figure $R * W$ is again the H-rule:

$$\begin{array}{lcl} \dfrac{\{!u = N\}, (U \leq s) * W}{\{?u\}, U * W} & = & (\{!u = N\}, (U \leq s)) * W \\ & = & (\{?u\}, U) * (\{?u\}, W) \end{array}$$

This concludes the proof of Lemma 5.11 and hence of Theorem 5.10. ⊣

6. Cutelimination

Now we can describe cutelimination transformation. Recall that in the standard Gentzen-type calculus a cut

$$(1) \qquad \begin{array}{ccc} & \sim A, Z, A & \\ & | & \\ d_? & \sim A, X \qquad A, Y & d_! \\ & X, Y & \end{array}$$

in a derivation where all occurrences of $\sim A$ are traceable to axioms (and not to main formulas of rules) is reduced to the following derivation:

$$\begin{array}{ll} d_! * Z & | \\ & Y, Z, A \\ d_? [\sim A := Y] & | \\ & Y, X \end{array}$$

obtained from the original derivation by first replacing $\sim A$ by Y and then deriving images Y, Z, A of former axioms $\sim A, Z, A$ by derivation $d_!$ multiplied by Z. The axioms of the form $\sim A, B, \sim B, Z$ are transformed into $Y, B, \sim B, Z$ and the second derivation is not needed.

In our ϵ-calculus the role of the axiom $\sim A, Z, A$ for the cut with main formulas $?s, !s$ is played by the axiom Axc: $?s, Z$ with $CZ : s\# = N \to$ true:

$$(2)\qquad \begin{array}{lc} & ?s, Z \\ & | \\ d_? & ?s, X \quad !s = N, X \qquad d_N \\ & \hline \\ & X \\ d- & | \\ & 0 \end{array}$$

It would be natural to transform such a derivation into the figure

$$(3)\qquad \begin{array}{lcl} d_N * Z & | & \\ & \{!s = N\}, Z, X & (H) \\ & \hline \\ & \{?s\}, Z, X & \\ d_? * X & | & \\ & \{?s\}, X & (Fr) \\ & \hline \\ & X & \\ d- & | & \\ & 0 & \end{array}$$

but there are two problems. First, the H-rule in the new figure will be incorrect in general since both Z and X can contain formulas of rank $> r$. Second, the thinning rule is not admissible in general, in particular the result of multiplying a derivation by a sequent is not necessarily a derivation.

These problems are solved as follows. We leave $d_?$ as it is, make $?s$ provisional, truncate Z to $Z(\leq r)$ over the H-rule (3) and reintroduce the missing formulas from X by using the part d- of the original derivation situated *below* X:

$$\begin{array}{lc} & X \\ d- & | \\ & 0 \end{array} \qquad\qquad \begin{array}{lc} & Z(\leq r) * X \\ Z(\leq r) * d- & | \\ & Z(\leq r) \end{array}$$

while the standard cut reduction uses only the part above X.

DEFINITION. *Let d be an r-derivation and (c) be a cut of rank r in d with no cuts of rank r above (c) (cf. (2)). Then $red(d, (c))$ is the figure:*

$$
\begin{array}{ll}
d_N * \{!s = N\}, Z(\leq r) & \\
 & \{!s = N\}, Z(\leq r) * X \\
 & \vdots \\
d - *\{?s = N\}, Z(\leq r) & \{!s = N\}, Z(\leq r) \\
 & \hline \\
 & \{?s\}, Z \\
 & \vdots \\
d_?\{\} & \{?s\}, X \quad !s = N, X \\
 & \hline \\
 & X \qquad (CutFr) \quad d_N \\
 & \vdots \\
d- & 0
\end{array}
$$

Note that in branches that do not end in the Axc with main formula traceable to the cut (c), no fancy transformations are needed: d_N is simply pruned.

The main syntactic result of this paper is the following

THEOREM 6.1. *red(d, c) is an r-derivation.*

PROOF: Consider the parts of red(d, c) bottom-up. $d-$ is left intact. $d_?\{\}$ is correct by Lemma 5.8. Next, part $d - *\{!s = N\}, Z(\leq r)$ is correct by Lemma 5.7. The part $d_N * \{!s = N\}, Z(\leq r)$ is correct by Theorem 5.10. It remains to verify the correctness of the rule (CutFr) introduced instead of cut (c). Indeed $\{?s\}$ has rank r, $?s, X$ is correct and $(?s, X)(\leq r) = ?s, X(\leq r)$ which is c.c. since $X(\leq r)$ and even X is c.c. So $\{?s\}, X$ is c.c. by Lemma 5.4(b) ⊣

We can now formulate

THEOREM 6.2. *Any derivation of the empty sequent 0 (of height less than ε_0) can be transformed into a cutfree deirvation (of height less than ε_0) by cut-reduction red and pruning.*

PROOF is standard: we use induction on the maximal cut rank r and induction on the height h in the induction step. The only non-trivial step left is the passage from an r-derivation to an $(r-1)$-derivation in the induction step of the induction on r. So suppose an r-derivation containing no cuts (only CutFr) of rank r is given. Prune all minor premiss of (CutFr) (and everything over them) and prove by bottom-up induction that the resulting figure is an $(r-1)$-derivation, i.e. contains only $(r-1)$-sequents. This is easily done by cases. ⊣

REFERENCES

W. ACKERMANN [1925], *Begründung des Tertium non datur mittels der Hilbertischen Theorie der Widerspruchsfreiheit*, Math. Ann. 93, 1–36.

W. ACKERMANN [1940], *Zur Widerspruchsfreiheit der Zahlentheorie*, Math. Ann. 117, 162–194.

G. GENTZEN [1939], *Neue Fassung des Widerspruchsfreiheitsbeweises für die reine Zahlentheorie*, Forsch. zur Logik u. z. Grundl. d. exacten Wiss., Hf. 4, Leipzig, Hirzel, 19–44.

W. GOLDFARB, T. SCANLON [1974], *The ω-consistency of Number theory via Herbrand Theorem*, J. Symbolic Logic, 39, 678–692.

D. HILBERT [1929], *Probleme der Grundlegung der Mathematik*, Math. Ann. 102, 1–9.

D. HILBERT, P. BERNAYS [1970], *Grundlagen der Mathematik, Bd. 2*, Springer.

G. KREISEL [1951], *On the Interpretation of Non-finitist Proofs I*, J. Symbolic Logic 16, 241–267.

G. KREISEL [1952], *On the Interpretation of Non-finitist Proofs II*, J. Symbolic Logic 17, 43–58.

G. MINTS [1982], *Simplified Consistency Proof for Arithmetic* (Russian), Proc. Estonian Acad. of Sci. Fiz.-Math. 31 N4, 376–382.

G. MINTS [1989], *Epsilon Substitution Method for the Theory of Hereditarily Finite Sets* (Russian), Proc. Estonian Acad. of Sci. Fiz.-Math. 1989 N2, 154–164.

J. VON NEUMANN [1927], *Zur Hilbertischen Beweistheorie*, Math. Zeitschrift 26, 1–46.

T. SCANLON [1973], *The consistency of Number Theory via Herbrand's Theorem*, J. Symbolic Logic 39, 25–98.

W. TAIT [1965], *Functionals defined by Transfinire Recursion*, J. symbolic Logic 30 N2, 155–174.

W. TAIT [1965a], *The Substitution method*, J. Symbolic Logic 30, N2, 175–192.

Logic, Methodology and Philosophy of Science IX
D. Prawitz, B. Skyrms and D. Westerståhl (Editors)

ADMISSIBLE PROOF THEORY AND BEYOND

MICHAEL RATHJEN

Universität Münster, Germany

1. Prologue

Ordinals made their entrance in proof theory through Gentzen's second consistency proof for Peano Arithmetic by transfinite induction up to ε_0, the latter being applied only to decidable predicates (cf. Gentzen [1938]). Gentzen's constructive use of ordinals as a method of analyzing formal theories has come to be a paradigm for much of proof theory from then on, particularly as exemplified in the work of Schütte, Takeuti and their schools.[1]

One of the strongest theories for which ordinal–theoretic bounds have been obtained is the impredicative subsystem of second order arithmetic based on Δ^1_2 comprehension plus bar induction. The latter result was achieved by employing the most advanced techniques in this area of research: cut elimination for infinitary calculi of ramified set theory with Π_2–reflection rules. This gathering of tools was entitled "Admissible Proof Theory" (cf. Pohlers [1982]), yet another appropriate title could have been "Proof Theory of Π_2–Reflection". Unfortunately, these methods are not strong enough for carrying through an ordinal analysis of Π^1_2 comprehension, let alone for second order arithmetic.

This article will survey the state of the art nowadays, in particular recent advance in proof theory beyond admissible proof theory, giving some prospects of success of obtaining an ordinal analysis of Π^1_2 comprehension.

Although a great deal of ordinally informative proof theory has been pursuing an extension of Hilbert's program, that is sought-for consistency proofs, I shall only indulge very little in this issue.[2] Even those who wish to detach themselves from consistency matters may benefit from ordinal analyses. Ordinal analysis has proved to be an important tool in reductive

[1] cf.Schütte [1977], Takeuti [1987], Pohlers [1987], Pohlers [1991].

[2] For details cf. Takeuti [1987] and also the papers Feferman [1988] and Sieg [1988] being written on the occasion of a special Symposium on Hilbert's Program.

proof theory and also for the determination of the provably total functions of various complexities of a variety of theories. Putting things into a broader perspective, a leit–motif for ordinal analysis could have been Kreisel's question:

> *What more than its truth have we recognized, when we have established a theorem in a formal theory?*

The article is divided into four parts. In Section 2 I roughly describe the role of ordinals and ordinal analysis in proof theory.

Section 3 will be concerned with the program of admissible proof theory as well as its achievements. Also, in a nutshell, the cut–elimination procedure for Kripke–Platek set theory is given.

After having witnessed a real ordinal analysis, the reader will be more prepared for a discussion of the many facets of ordinal analysis which will be the purpose of Section 4.

The final Section 5 deals with new cut–elimination procedures for reflections higher than Π_2. Cut–elimination for Π_n–reflection entails a proof–theoretic treatment of theories of nonmonotone inductive definitions. It is also touched upon the question of how far afield all this is from Π^1_2 comprehension.

2. Ordinal analysis

Let T be a theory the language of which is rich enough to contain formulas expressing well–foundedness properties. In addition, assume that T comprises primitive recursive arithmetic PRA and that T is faithful, i.e. whenever $T \vdash A$, then A is true. Under these conditions the *proof–theoretic ordinal* $|T|$ of T is often defined as follows:

$$|T| = \sup \{\alpha : \alpha \text{ provably recursive in } T\},$$

where α is said to be provably recursive in T if there is a recursive well–ordering $\prec$ with order–type α such that

$$T \vdash WO(\prec)$$

with $WO(\prec)$ expressing in the language of T that $\prec$ is a well–ordering.

The determination of $|T|$ is then called *ordinal analysis of* T.

The above definition of $|\ T\ |$ has the advantage of being mathematically precise, but as to the activity named 'ordinal analysis' it is left completely open what constitutes such an analysis and in what terms $|T|$ is to be given.

A nude set–theoretical ordinal is hardly ever of interest from the viewpoint of foundation. In proof theory attention focusses on structured ordinals which can be dealt with in a finitary manner. The paradigm is Gentzen's use of Cantor's representation of ordinals $< \varepsilon_0$. Every ordinal $0 < \alpha < \varepsilon_0$ has a unique representation of the form

$$\alpha = \omega^{\alpha_1} + \cdots + \omega^{\alpha_n}$$

with $\alpha_n \leq \cdots \leq \alpha_1 < \alpha$.

Therefore the ordinals $< \varepsilon_0$ can be represented by terms built up from a symbol for 0 and symbols for the function $+$ and $\lambda\xi.\omega^\xi$. We also gain finitary control on such (infinite) ordinals because of the following facts:

- For every expression E composed of the symbols $0, +, \omega$ it can be decided whether E represents an ordinal.
- Given two representations E_0, E_1 of ordinals α_0, α_1 respectively, we know how to compare α_0 and α_1 solely by means of the build–up of E_0 and E_1.

It is by now clear that $|\,T\,|$ is to be given in terms of a system of ordinal representation usually called ordinal notation system. Significant features that ordinal notation systems should have will be adressed in Section 4.

I have always found the description of ordinal analysis as a quest for proof–theoretic ordinals to be bad propaganda, above all, since it remains silent about the most interesting aspects of ordinal analysis and prejudices people against this enterprise. If experience has shown that the ordinal $|T|$ is intrinsically related to the proof power of T, it is rarely the sheer knowing of $|T|$ that lends itself to important information about T.[3] Most of the vital information springs from the proof itself. Turning attention to practice, an ordinal analysis of T provides, among others, the following results:

- A reduction of T to Heyting's Arithmetic, HA, plus a scheme of transfinite induction.
- A consistency proof of T.
- A classification of the provably recursive functions (on $\mathbb{N}$) of T.
- A classification of the provably hyperarithmetical functions of T.
- A classification of the provably Δ^1_2 functions of T.

[3]Actually, it has to be reckoned with theories where the proof–theoretic ordinal in the above sense doesn't reflect the proof–theoretic strength of the theory.

- A description of partial models of T, for instance models of all Π^1_2 or Π^1_3 theorems of T.

A discussion of these points will be more fruitful and lively after the reader has gained some experience with ordinal analysis, so we defer it to the next but one Section.

3. Admissible proof theory

Admissible proof theory arose out of the work of Jäger und Pohlers (cf. Pohlers [1982], Pohlers [1987]) who from a proof–theoretic stance started to investigate weak set theories featuring admissible sets. The direct proof–theoretical treatment of set theories is rather recent. Historically, the primary concern has been on subsystems of second order arithmetic and theories of iterated inductive definitions (cf. Buchholz et al. [1981]).

Admissible sets are the transitive models of a remarkable subsystem of ZF, known as Kripke–Platek set theory (hereinafter called KP). Admissible sets were a major source of interaction between model theory, recursion theory and set theory (cf. Barwise [1975]).

In this section I am going to sketch an ordinal analysis for KP. The motivation behind this is twofold. On the one hand, I would like to give some insight into admissible proof theory by presenting the basic ideas that underly its cut–elimination procedures. On the other hand, this will serve as a foil for a comparison with new cut–elimination procedures in Section 5 and also for the discussion in Section 4.

3.1. The system KP

Though considerably weaker than ZF, a great deal of set theory requires only the axioms of KP. The axioms of KP are:[4]

Extensionality:	$a = b \to [F(a) \leftrightarrow F(b)]$ for all formulas F.
Foundation:	$\exists x G(x) \to \exists x[G(x) \wedge (\forall y \in x)\neg G(y)]$
Pair:	$\exists x\ (x = \{a, b\})$.
Union:	$\exists x\ (x = \bigcup a)$.

[4]For technical convenience, $\in$ will be taken to be the only predicate symbol of the language of set theory. This does no harm, since equality can be defined by $a = b :\Leftrightarrow (\forall x \in a)(x \in b) \wedge (\forall x \in b)(x \in a)$, provided that we state extensionality in a slightly different form than usually.

Infinity:	$\exists x\,[x \neq \emptyset \wedge (\forall y \in x)(\exists z \in x)(y \in z)]$.[5]
Δ_0 *Separation:*	$\exists x\,(x = \{y \in a : F(y)\})$[6] for all Δ_0–formulas F in which x does not occur free.
Δ_0 *Collection:*	$(\forall x \in a)\exists y G(x,y) \rightarrow \exists z(\forall x \in a)(\exists y \in z)G(x,y)$ for all Δ_0–formulas G.

By a Δ_0 formula we mean a formula of set theory in which all the quantifiers appear restricted, that is have one of the forms $(\forall x \in b)$ or $(\exists x \in b)$.

KP arises from ZF by completely omitting the power set axiom and restricting separation and collection to absolute predicates (cf. Barwise [1975]), i.e. Δ_0 formulas. These alterations are suggested by the informal notion of 'predicative'. KP is an impredicative theory, notwithstanding. It is known from Howard [1968],[1981] and Jäger [1982] that KP proves the same arithmetic sentences as Feferman's system ID_1 of positive inductive definitions (cf. Feferman [1970]). Its proof–theoretic ordinal is the Howard–Bachmann ordinal $\theta\varepsilon_{\Omega+1}0$.

3.2. Infinitary calculi

Peano Arithmetic, PA, does not admit cut–elimination. However, it is well known that the infinitary calculus PA_ω which results from PA by replacing the induction scheme by the so–called ω–rule

$$\frac{\Gamma, A(\bar{n}) \text{ for all } n}{\Gamma, \forall x A(x)}$$

does admit cut–elimination.[7] An ordinal analysis for PA is then attained as follows:

- Each PA–proof can be unfolded into a PA_ω–proof of the same sequent.
- Each such PA_ω–proof can be transformed into a cut–free PA_ω–proof of the same sequent of length $< \varepsilon_0$.

In order to get a similar result for KP, we have to work a bit harder.

Experience has shown that the main obstacle for understanding ordinal analysis of impredicative theories is raised by its being intimately linked

[5] $x = \{y \in a : F(y)\}$ stands for the Δ_0–formula $(\forall y \in x)[y \in a \wedge F(y)] \wedge (\forall y \in a)[F(y) \rightarrow y \in x]$.

[6] This contrasts with Barwise [1975] where Infinity is not included in KP.

[7] $\bar{n}$ stands for the n^{th} numeral.

to specific systems of ordinal notations, even worse, to auxiliary deduction functions or relations needed in order for this method to work (cf. Pohlers [1981]).[8] Fortunately, Buchholz [1991] has presented a new approach which is distinguished by conceptual clarity and flexibility, and in particular by the fact that its basic concepts are in no way related to any system of ordinal notations. We are going to take up Buchholz's approach but in an even more relaxed atmosphere, thereby refraining from technical details as far as possible. Especially, we shall put forward that the collapsing of proof trees which is paramount in impredicative proof theory can be understood in terms of the usual Mostowski collapse familiar from set theory.

At the outset, we set up an infinitary calculus of ramified set theory which is modelled upon the constructible hierarchy.

For α an ordinal, L_α is the α^{th} level of Gödel's constructible hierarchy, i.e.

$$L_0 = \emptyset$$

$$L_\alpha = \bigcup_{\beta<\alpha} L_\beta \quad \text{if } \alpha \text{ is a limit ordinal}$$

$$L_\alpha = \{X : X \subseteq L_\beta \text{ and } X \text{ is definable over } \langle L_\beta, \in \rangle\} \quad \text{if } \alpha = \beta + 1.$$

Guided by the analogy with PA_ω, we would like to invent an infinitary rule which when added to KP enables us to eliminate cuts. However, as opposed to the natural numbers, it is not very clear how to bestow upon each element of the set–theoretic universe a name that reflects its generation; but within the confines of the constructible universe which is made from the ordinals it is pretty obvious how to name sets once we have given names to ordinals. Thus we are naturally led to the calculus RS we are going to introduce next.

3.2.1. Infinitary syntax

RS–terms and their levels are inductively defined as follows.

1. For every ordinal α, $\check{L}_\alpha$ is an RS–term of level α.
2. If $F(x, y_1, \cdots, y_n)$ is a formula of set theory with no free variables other than shown, and $s_1, \cdots s_n$ are RS–terms of levels $< \alpha$, then the formal expression
$$[x \in \check{L}_\alpha : F(x, s_1, \cdots s_n)^{\check{L}_\alpha}]$$
is an RS–term of level α.

[8] For instance, several years ago in a seminar at Münster devoted to the ordinal analysis of ν–fold iterated inductive definitions, more than half of the time was spent on developing collapsing functions with peculiar features, and those collapsing functions were not the ones that surfaced in the corresponding notation system.

We denote the level of an RS–term t by $|t|$. For a formula F, we denote by F^a the formula that is obtained by restricting any unbounded quantifier in F by a.

The interpretation i of an RS-term in L is, as was to expected,

- $\mathsf{i}(\check{L}_\alpha) = L_\alpha$
- $\mathsf{i}([x \in \check{L}_\alpha : F(x, s_1, \cdots s_n)^{\check{L}_\alpha}]) = \{x \in L_\alpha : L_\alpha \models F(x, \mathsf{i}(s_1), \cdots, \mathsf{i}(s_n))\}$
  $= \{x \in L_\alpha : L \models F(x, \mathsf{i}(s_1), \cdots, \mathsf{i}(s_n))^{L_\alpha}\}.$

An RS–formula is one that arises from a Δ_0 formula of set theory by replacing all its free variables with RS–terms. Let G be an RS–formula. By way of the interpretation i, validity of G in L, $L \models G$, is understood.

Abbreviations.

$k(G) = \{\alpha : \check{L}_\alpha$ occurs in $G\}$ (subterms included).

$|\, G \,| = sup\ k(G)$.

For RS–terms a, b with $|\, a \,| < |\, b \,|$, $\diamond$ a propositional junctor, and A an arbitrary RS–formula, we set

$$(a \check{\in} b) \diamond A = \begin{cases} B(a) \diamond A & \text{if } \ b \equiv [x \in \check{L}_\beta : B(x)] \\ A & \text{if } \ b \equiv \check{L}_\beta. \end{cases}$$

Obviously $(a \in b) \diamond A$ and $(a \check{\in} b) \diamond A$ have the same truth–value.

3.2.2. Infinitary rules

Next we introduce an infinitary sequent calculus, RS, that admits cut elimination.

$A, B, C, \ldots, F(t), G(t), \ldots$ range over RS–formulas. We denote by upper case Greek letters $\Gamma, \Delta, \Lambda, \ldots$ finite sets of RS–formulas. The intended meaning of $\Gamma = \{A_1, \cdots, A_n\}$ is the disjunction $A_1 \vee \cdots \vee A_n$. Γ, A stands for $\Gamma \cup \{A\}$.

The rules of RS are:

$$(\wedge) \quad \frac{\Gamma, A \quad \Gamma, A'}{\Gamma, A \wedge A'}$$

$$(\vee)\quad \frac{\Gamma, A_i}{\Gamma, A_0 \vee A_1} \text{ if } i \in \{0,1\}$$

$$(\forall)\quad \frac{\Gamma, s \check{\in} t \to F(s) \quad \text{for all } s \text{ such that } |s|<|t|}{\Gamma, (\forall x \in t) F(x)}$$

$$(\exists)\quad \frac{\Gamma, s \check{\in} t \wedge F(s)}{\Gamma, (\exists x \in t) F(x)} \text{ if } |s|<|t|$$

$$(\notin)\quad \frac{\Gamma, s \check{\in} t \to r \neq s \text{ for all } s \text{ such that } |s|<|t|}{\Gamma, r \notin t}$$

(where $r \neq s := \neg\,(r = s)$ and $r \notin t := \neg\,(r \in t)$)

$$(\in)\quad \frac{\Gamma, s \check{\in} t \ \wedge\ r = s}{\Gamma, r \in t} \text{ if } |s|<|t|$$

$$(Cut)\quad \frac{\Gamma, A \qquad \Gamma, \neg\, A}{\Gamma}.$$

As in Schwichtenberg [1977] we shall regard $\neg$ in front of a non–atomic formula as a defined operation:
$\neg\, A$ is defined to be the formula obtained from A by (i) putting a $\neg$ in front of any atomic formula, (ii) replacing $\wedge, \vee, (\forall x \in a), (\exists x \in a)$ by $\vee, \wedge$, $(\exists x \in a), (\forall x \in a)$, respectively, and (iii) dropping double negations.

Owing to the symmetry of the pairs of rules

$$(\wedge),\ (\vee)$$

$$(\forall),\ (\exists)$$

$$(\notin),\ (\in),$$

the usual cut–elimination procedure (cf. Schwichtenberg [1977]) applies to RS. But unequal to the situation for PA and PA_ω, RS does not allow of any nontrivial embedding of KP; the trivial one being provided by the fact that for any admissible ordinal κ, L_κ is a model of KP and the following completeness property of RS:

Theorem. (cf. Pohlers [1991], Theorem 3.2.6) *For each RS–formula G, if $L \models G$, then there is an RS proof of G.*

The only axioms of KP that shatter hopes of obtaining an informative embedding into RS are instances of Δ_0 collection. To remedy this, we simply add a new rule to RS which plainly entails Δ_0 collection. The reverse of the medal is that we need to be particular about permitting derivations in order to restore (partial) cut–elimination.

In the sequel, we fix an admissible ordinal Ω. Henceforth we will only be concerned with RS_Ω–formulas, i.e. RS–formulas of the form $F(s_1, \cdots, s_n)^{\check{L}_\Omega}$, where $s_1, \cdots, s_n$ are RS–terms of levels $< \Omega$ and $F(x_1, \cdots, x_n)$ is a formula of set theory. In case that $F(x_1, \cdots, x_n)$ contains no unbounded universal quantifiers, $F(x_1, \cdots, x_n)$ is said to be a Σ formula, and $F(s_1, \cdots, s_n)^{\check{L}_\Omega}$ will be called $\Sigma(\Omega)$ formula. Frequently we write A^α instead of $A^{\check{L}_\alpha}$. Occasionally, $(\exists x^\alpha)$ will be a shorthand for $(\exists x \in \check{L}_\alpha)$.

The already announced rule is

$$(\Sigma\text{-}Ref_\Omega) \quad \frac{\Gamma, A^\Omega}{\Gamma, (\exists z \in \check{L}_\Omega) A^z} \text{ if } A^\Omega \text{ is } \Sigma(\Omega).$$

The motivation behind this rule is that on the basis of the other axioms of KP, Δ_0 collection is equivalent to the scheme of Σ reflection, i.e.

$$B \to \exists z B^z$$

for every Σ formula B (cf. Barwise [1975]).

3.2.3. $\mathcal{H}$–controlled derivations

The concept of $\mathcal{H}$–controlled derivations stems from Buchholz [1991].

Let $P(ON) = \{X : X \text{ is a set of ordinals}\}$.

A class function

$$\mathcal{H} : P(ON) \to P(ON)$$

will be called operator if the following conditions are satisfied for $X, X' \in P(ON)$:

(H1) $0 \in \mathcal{H}(X)$. For $\alpha = \omega^{\alpha_1} + \cdots + \omega^{\alpha_n}$ with $\alpha_1 \geq \cdots \geq \alpha_n$, it holds $\alpha \in \mathcal{H}(X)$ if and only if $\alpha_1, ..., \alpha_n \in \mathcal{H}(X)$. (Especially, $\mathcal{H}(X)$ is closed with respect to $+$ and $\lambda\xi.\omega^\xi$, i.e., if $\alpha, \beta \in \mathcal{H}(X)$, then $\alpha + \beta, \omega^\alpha \in \mathcal{H}(X)$.)

(H2) $X \subseteq \mathcal{H}(X)$

(H3) $X' \subseteq \mathcal{H}(X) \Rightarrow \mathcal{H}(X') \subseteq \mathcal{H}(X)$.

Abbreviations. $\alpha \in \mathcal{H} := \alpha \in \mathcal{H}(\emptyset)$

$X \subseteq \mathcal{H} := X \subseteq \mathcal{H}(\emptyset)$

For an RS_Ω–term s, $\mathcal{H}[s]$ denotes the operator $(X \mapsto \mathcal{H}(k(s) \cup X))_{X \in P(ON)}$.

Let $\Omega \in \mathcal{H}$ and Γ be a finite set of RS_Ω–formulas. The relation $\mathcal{H} \mid\!\!\!\frac{\alpha}{}\!\!\!- \Gamma$ (*$\mathcal{H}$–controlled derivability*) is defined inductively by

$$\{\alpha\} \cup k(\Gamma) \subseteq \mathcal{H}$$

and the following rules

$$(\wedge) \qquad \frac{\mathcal{H} \mid\!\!\!\frac{\alpha_0}{}\!\!\!- \Lambda, A_0 \qquad \mathcal{H} \mid\!\!\!\frac{\alpha_1}{}\!\!\!- \Lambda, A_1}{\mathcal{H} \mid\!\!\!\frac{\alpha}{}\!\!\!- \Lambda, A_0 \wedge A_1} \qquad \alpha_0, \alpha_1 < \alpha$$

$$(\vee) \qquad \frac{\mathcal{H} \mid\!\!\!\frac{\alpha_0}{}\!\!\!- \Lambda, C}{\mathcal{H} \mid\!\!\!\frac{\alpha}{}\!\!\!- \Lambda, A \vee B} \text{ if } C \in \{A, B\} \qquad \alpha_0 < \alpha$$

$$(\forall) \qquad \frac{\cdots \mathcal{H}[s] \mid\!\!\!\frac{\alpha_s}{}\!\!\!- \Lambda, s \breve{\in} t \to F(s) \cdots (|s| < |t|)}{\mathcal{H} \mid\!\!\!\frac{\alpha}{}\!\!\!- \Lambda, (\forall x \in t) F(x)} \qquad \alpha_s < \alpha$$

$$(\exists) \qquad \frac{\mathcal{H} \mid\!\!\!\frac{\alpha_0}{}\!\!\!- \Lambda, s \breve{\in} t \wedge F(s)}{\mathcal{H} \mid\!\!\!\frac{\alpha}{}\!\!\!- \Lambda, (\exists x \in t) F(x)} \text{ if } |s| < |t| \qquad \alpha_0, |s| < \alpha,\ k(s) \subseteq \mathcal{H}$$

$$(\notin) \qquad \frac{\cdots \mathcal{H}[s] \mid\!\!\!\frac{\alpha_s}{}\!\!\!- \Lambda, s \breve{\in} t \to r \neq s \cdots (|s| < |t|)}{\mathcal{H} \mid\!\!\!\frac{\alpha}{}\!\!\!- \Lambda, r \notin t} \qquad \alpha_s < \alpha$$

$$(\in) \qquad \frac{\mathcal{H} \mid\!\!\!\frac{\alpha_0}{}\!\!\!- \Lambda, s \breve{\in} t \wedge r = s}{\mathcal{H} \mid\!\!\!\frac{\alpha}{}\!\!\!- \Lambda, r \in t} \text{ if } |s| < |t| \qquad \alpha_0, |s| < \alpha,\ k(s) \subseteq \mathcal{H}$$

$$(\Sigma\text{–}Ref_\Omega) \qquad \frac{\mathcal{H} \mid\!\!\!\frac{\alpha_0}{}\!\!\!- \Lambda, A^\Omega}{\mathcal{H} \mid\!\!\!\frac{\alpha}{}\!\!\!- \Lambda, (\exists z \in \breve{L}_\Omega) A^z} \text{ if } A^\Omega \in \Sigma(\Omega), \qquad \alpha_0 < \alpha$$

$$(Cut) \qquad \frac{\mathcal{H} \mid\!\!\!\frac{\alpha_0}{}\!\!\!- \Lambda, B \qquad \mathcal{H} \mid\!\!\!\frac{\alpha_0}{}\!\!\!- \Lambda, \neg B}{\mathcal{H} \mid\!\!\!\frac{\alpha}{}\!\!\!- \Lambda} \qquad \alpha_0 < \alpha.$$

Since we also want to keep control of the cuts of $\mathcal{H}$–controlled derivations, we assign a rank, $rk(A)$, to RS_Ω–formulas A. All we need to know is that $rk(A) = \omega \cdot |A| + n$ for some $n < \omega$, $rk(A) = rk(\neg A)$, and $rk((\exists z \in \breve{L}_\Omega) A^z) = \Omega$ if $A \in \Sigma(\Omega)$.

We write $\mathcal{H} \mid\!\!\!\frac{\alpha}{\rho}\!\!\!- \Gamma$ to express that there is an $\mathcal{H}$–controlled derivation of Γ such that $rk(B) < \rho$ holds for all cut formulas B in this derivation.

Having defined $\mathcal{H}$–controlled derivability, the notion of an $\mathcal{H}$–controlled derivation (or proof) is understood. To be more precise, an *$\mathcal{H}$–controlled derivation* is a well–founded tree the nodes of which are pairs $\langle \alpha, \Gamma \rangle$ resulting from its immediate successor nodes by one of the above rules.

We will use the notation $\Pi_{\mathcal{H}} \vert\!\frac{\alpha}{\varrho}\, \Gamma$ to indicate that $\Pi_{\mathcal{H}}$ is an $\mathcal{H}$–controlled derivation witnessing $\mathcal{H} \vert\!\frac{\alpha}{\varrho}\, \Gamma$.

3.3. Embedding KP

Let d be a KP–proof of a sentence F. Then there exists an integer n such that for every operator $\mathcal{H}$ with $\Omega \in \mathcal{H}$ we have an $\mathcal{H}$–controlled derivation

$$\Pi_{\mathcal{H},\Omega} \vert\!\frac{\Omega \cdot n}{\Omega + n}\, F^{\Omega}$$

(cf. Buchholz [1991]). Furthermore, it is to be noted that the construction of the $\mathcal{H}$–proof $\Pi_{\mathcal{H},\Omega}$ of F^{Ω} is uniform in $\mathcal{H}$ and Ω. This is reflected by the following facts: If $\mathcal{H}'$ majorizes $\mathcal{H}$, i.e. $\forall X(\mathcal{H}(X) \subseteq \mathcal{H}'(X))$, then $\Pi_{\mathcal{H},\Omega} = \Pi_{\mathcal{H}',\Omega}$. If $\Omega < \hat{\Omega} \in \mathcal{H}$, then $\Pi_{\mathcal{H},\Omega}$ and $\Pi_{\mathcal{H},\hat{\Omega}}$ are closely related to each other. $\Pi_{\mathcal{H},\Omega}$ can be obtained from $\Pi_{\mathcal{H},\hat{\Omega}}$ by the following pruning and substitution processes:

- Omit from each instance of a rule

$$\frac{\cdots \Delta, H(t) \cdots (\vert t \vert < \hat{\Omega})}{\Delta, H(x)}$$

  in $\Pi_{\mathcal{H},\hat{\Omega}}$ all premisses $\Delta, H(t)$ with $\hat{\Omega} \leq \vert t \vert$ as well as the subproofs of these premisses.

- Within the mutilated proof, each ordinal $\alpha > 0$ has Cantor normal form $\alpha = \hat{\Omega}^{k_1}\beta_1 + \cdots + \hat{\Omega}^{k_r}\beta_r$ where $k_1 > \cdots > k_r$ and $\beta_1, \cdots, \beta_r < \Omega$. Now replace α by $\Omega^{k_1}\beta_1 + \cdots + \Omega^{k_r}\beta_r$.

The use of a whole family of proofs is reminiscent of Girard's notion of β–proof (cf. Girard [1985]).

Indeed, there are more points of contact. Usually for a single $\mathcal{H}$, it will not be possible to transform an $\mathcal{H}$–proof into a cut–free $\mathcal{H}$–proof. To overcome this difficulty, we pass over to stronger and yet stronger operators during the cut–elimination procedure, but in a controlled manner, thereby working simultaneously on a whole family of proofs indexed by operators.

3.4. Cut–elimination

As already mentioned, $(\Sigma\text{–}Ref_{\Omega})$ is the only rule that spoils cut–elimination. Since an instance of $(\Sigma\text{–}Ref_{\Omega})$ always introduces a formula of rank Ω, we can at least remove all cuts of rank $> \Omega$. So we get

Cut-elimination I. *Let $n > 0$. Then:*

$$\mathcal{H}\left|\frac{\Omega\cdot n}{\Omega+n}\Gamma \;\Rightarrow\; \mathcal{H}\right|\frac{\Omega(n)}{\Omega+1}\Gamma,$$

where $\Omega(1) := \Omega$ and $\Omega(k+1) := \Omega^{\Omega(k)}$.[9]

A first step towards elimination of $(\Sigma\text{-}Ref_\Omega)$ is provided by the following

Bounding Theorem. *Let B^Ω be a $\Sigma(\Omega)$ formula. If $\alpha < \Omega$ and $\mathcal{H}\left|\frac{\alpha}{\rho}\Gamma, B^\Omega\right.$, then $\mathcal{H}\left|\frac{\alpha}{\rho}\Gamma, B^\alpha\right.$.*

This result is easily proved by induction on α. First let us focus on the case when the last inference is $(\Sigma\text{-}Ref_\Omega)$ with principal formula B^Ω. Then B^Ω is of the form $(\exists z \in \check{L}_\Omega)A^z$, and we have $\mathcal{H}\left|\frac{\alpha_0}{\Omega}\Gamma, A^\Omega\right.$ for some $\alpha_0 < \alpha$. By induction hypothesis we get

$$\mathcal{H}\left|\frac{\alpha_0}{\Omega}\Gamma, A^{\alpha_0}\right.,$$

which is the same as $\mathcal{H}\left|\frac{\alpha_0}{\Omega}\Gamma, \check{L}_{\alpha_0} \check{\in} \check{L}_\Omega \wedge A^{\check{L}_{\alpha_0}}\right.$, thus $\mathcal{H}\left|\frac{\alpha}{\Omega}\Gamma, (\exists z \in \check{L}_\Omega)A^z\right.$ follows by an inference $(\exists)$.

The key to an understanding of the Boundedness Theorem is provided by the case when the last inference is of the form

$$\frac{\mathcal{H}\left|\frac{\alpha_0}{\Omega}\Gamma, F(s)^\Omega, A\right.}{\mathcal{H}\left|\frac{\alpha}{\Omega}\Gamma, (\exists x \in \check{L}_\Omega)F(s)^\Omega\right.} \quad (\exists)$$

with $A \equiv (\exists x \in \check{L}_\Omega)F(x)$. Using the induction hypothesis we then get

$$\mathcal{H}\left|\frac{\alpha_0}{\Omega}\Gamma, F(s)^{\check{L}_\alpha}, A^\alpha\right..$$

The conditions imposed by $(\exists)$ ensure that $|s| < \alpha$, thus $\mathcal{H}\left|\frac{\alpha_0}{\Omega}\Gamma, A^\alpha\right.$ via an inference $(\exists)$.

The Boundedness Theorem also traces out the way for an elimination of $(\Sigma\text{-}Ref_\Omega)$ in the more general situation when $\mathcal{H}\left|\frac{\beta}{\Omega}\Gamma\right.$ with $\beta \geq \Omega$. However, we can no longer deal with arbitrary operators. In the sequel we shall restrict ourselves to operators $\mathcal{H}$ such that for each $\mathcal{H}$–controlled derivation $\Pi_\mathcal{H}$ without Ω–branchings the following "collapsing" properties are satisfied:

(C1) *The set $\{\, |s| < \Omega : s$ occurs in $\Pi_\mathcal{H}\}$ is bounded below Ω*

[9] This is the right place to explain why we demanded $\mathcal{H}(X)$ to be closed under $+$ and $\alpha \mapsto \omega^\alpha$: simply because these closure properties are needed for the above cut–elimination method.

(C2) *The Mostowski collapse of the set* $\{\alpha : \alpha \text{ occurs in } \Pi_{\mathcal{H}}\}$ *is less than* Ω.

Of course, requiring that $\Pi_{\mathcal{H}}$ has no Ω–branchings is a necessary condition for (C1) and (C2) to hold. But the reader might have a suspicion that the restrictions imposed by (C1) and (C2) give a too narrow class of operators in order for the cut–elimination to work. At the end of this Section we shall deliver a class of operators that fulfills (C1), (C2) and, in addition, is sufficiently rich for the purpose of cut–elimination.

Now let us fix an $\mathcal{H}$–controlled derivation $\Pi_{\mathcal{H}} \mid\!\!\frac{\alpha}{\Omega}\, \Gamma$ without Ω–branchings. This is for instance guaranteed if Γ is a set of $\Sigma(\Omega)$ formulas. On the other hand, if Γ entails a formula D which contains a quantifier $(\forall x \in \breve{L}_\Omega)$, then it can be shown that $\mathcal{H} \mid\!\!\frac{\alpha}{\Omega}\, \Gamma \setminus \{D\}$, i.e. D can be dropped from the derivation. Thus the exclusion of Ω–branchings is almost equivalent to Γ being a set of $\Sigma(\Omega)$ formulas.

Henceforth we assume that Γ is a set of $\Sigma(\Omega)$ formulas.

We are going to transform $\Pi_{\mathcal{H}}$ into a proof–tree without instances of $(\Sigma\text{–}Ref_\Omega)$. To this end, using (C1), pick $\beta < \Omega$ such that

$$\{\, |s| < \Omega \;:\; s \text{ occurs in } \Pi_{\mathcal{H}}\} \subseteq \beta.$$

By (C2), we then find an order preserving function

$$f : \beta \cup \{\alpha : \alpha \text{ occurs in } \Pi_{\mathcal{H}}\} \longrightarrow \gamma$$

onto some $\gamma < \Omega$.

Next let $\Pi^f_{\mathcal{H}}$ denote the tree that results from $\Pi_{\mathcal{H}}$ by replacing every node $\langle \xi, \Gamma\rangle$ in $\Pi_{\mathcal{H}}$ by $\langle f(\xi), \Gamma\rangle$. If we now define $\mathcal{H}_\beta$ via the equation

$$\mathcal{H}_\beta(X) = \mathcal{H}(X \cup (\beta + 1)),$$

we may expect that $\Pi^f_{\mathcal{H}} \mid\!\!\frac{f(\alpha)}{\Omega}\, \Gamma$ is an $\mathcal{H}_\beta$–controlled derivation. Indeed, this is readily verified. Since $f(\alpha) < \Omega$, we can employ the technique of the Bounding Theorem to get rid of all instances of $(\Sigma\text{–}Ref_\Omega)$ in $\Pi^f_{\mathcal{H}}$. We just have to replace the transitions in $\Pi^f_{\mathcal{H}}$ that are under the command of $(\Sigma\text{–}Ref_\Omega)$ by suitable instances of $(\exists)$. So we come up with a derivation $\Pi_{\mathcal{H}[\beta]} \mid\!\!\frac{f(\alpha)}{\Omega}\, \Gamma$ that no longer contains $(\Sigma\text{–}Ref_\Omega)$.

After having devised ways and means to remove $(\Sigma\text{–}Ref_\Omega)$ from derivations

$$\tilde{\Pi}_{\mathcal{H}} \mid\!\!\frac{\alpha}{\Omega}\, \Gamma$$

with $\Gamma \subseteq \Sigma(\Omega)$, we may now attack the problem of removing cuts of rank Ω from derivations $\Pi_{\mathcal{H}} \mid\!\!\frac{\alpha}{\Omega+1}\, \Gamma$.

The reason why the usual cut–elimination method fails for cuts with rank Ω is that it is too limited to treat a cut in the following context:

$$\dfrac{\dfrac{\Pi^0_{\mathcal{H}} \vdash^{\xi_0}_{\Omega} \Gamma, A^{\Omega}}{\Pi^1_{\mathcal{H}} \vdash^{\xi}_{\Omega} \Gamma, (\exists z \in \breve{L}_{\Omega}) A^z}(\Sigma\text{–}Ref_{\Omega}) \qquad \dfrac{\cdots \Pi_{\mathcal{H}[s]} \vdash^{\xi_s}_{\Omega} \Gamma, \neg A^s \cdots (|\, s \,| < \Omega)}{\Pi^2_{\mathcal{H}} \vdash^{\xi}_{\Omega} \Gamma, (\forall z \in \breve{L}_{\Omega}) \neg A^z}(\forall)}{\Pi_{\mathcal{H}} \vdash^{\alpha}_{\Omega+1} \Gamma}(Cut)$$

In this situation we are apt to apply the above introduced collapsing technique to $\Pi^0_{\mathcal{H}}$. Thus from $\Pi^0_{\mathcal{H}}$ we can extract a $\beta < \Omega$ and a function f such that $\mathcal{H}_\beta \vdash^{f(\xi_0)}_{\Omega} \Gamma, A^{f(\xi_0)}$. Next we single out the $f(\xi_0)^{th}$ premiss of the last inference of $\Pi^2_{\mathcal{H}}$, that is

$$\Pi_{\mathcal{H}[f(\xi_0)]} \vdash^{\xi_{f(\xi_0)}}_{\Omega} \Gamma, \neg A^{f(\xi_0)},$$

and, as $rk(A^{f(\xi_0)}) < \Omega$, a cut yields

$$\mathcal{H}_\beta \vdash^{\delta}_{\Omega} \Gamma$$

for some δ.

In order to get rid of all cuts of rank Ω in an arbitrary derivation $\tilde{\Pi}_{\mathcal{H}} \vdash^{\alpha}_{\Omega+1} \Gamma$, one has to repeat the foregoing process at worst α many times.

3.5. The functions θ_α

Yet another point is that we want to extract bounds from proofs of Σ formulas in KP. Therefore we have to take account of the quantitative aspects of "collapsing". Specifically, the "seize" of the operator after reducing the cut–rank from $\Omega + 1$ to Ω has to be related (via a functional dependence) to the "seize" of the input operator.

Through the above construction of $\mathcal{H}_\beta$ from $\mathcal{H}$, one is quite naturally led to processes lying behind the construction of the Feferman–Aczel functions θ_α (cf. Schütte [1977]).

The functions $\theta_\alpha : \Omega \to \Omega$ are inductively generated as follows:[10]

Let

$$C(\alpha, \beta) = \left\{ \begin{array}{l} \textit{closure of } \{0, \Omega\} \cup \beta \\ \textit{under } +,\ \xi \mapsto \omega^\xi,\ (\xi, \zeta \mapsto \theta_\xi(\zeta))_{\xi<\alpha, \zeta<\Omega} \end{array} \right.$$

[10] On the basis of the assumption $\Omega = \aleph_1$ (cf. Schütte [1977]) it is easily verified that $\theta_\alpha(\xi) < \Omega$ holds for $\xi < \Omega$ because of the countability of the set $C(\alpha, \xi)$. If, instead, Ω is merely supposed to be an admissible $> \omega$, it is by no means trivial to show that $\theta_\alpha(\xi) < \Omega$ (cf. Rathjen [1991b],[1991c]).

and

$$\theta_\alpha(\eta) = \eta^{th} \textit{ ordinal } \delta \textit{ such that } \delta \notin C(\alpha, \delta).$$

So this is a recursion with regard to α. If we now define operators $\mathcal{H}^\alpha$ by

$$\mathcal{H}^\alpha(X) = \bigcap \{C(\gamma,\beta) : X \subseteq C(\gamma,\beta) \wedge \alpha < \gamma\},$$

then the family of operators

$$(\mathcal{H}^\alpha)_{\alpha < \varepsilon_{\Omega+1}},$$

where $\varepsilon_{\Omega+1} = \sup\limits_{n<\omega} \Omega(n)$, is sufficient for all our purposes. However, it takes some efforts to show that the operators $\mathcal{H}^\alpha$ $(\alpha < \varepsilon_{\Omega+1})$ meet the requirements (C1) and (C2). Moreover, the technical details of the cut–elimination procedure via the familiy $(\mathcal{H}^\alpha)_{\alpha<\varepsilon_{\Omega+1}}$ are very delicate and fiddly; but we shall be satisfied by having pointed out the key ideas.

3.6. Π_2–reflection

As yet we have been dealing merely with Σ–reflection. One could argue that by doing so we covered Π_2–reflection as well since Π_2–reflection is a consequence of Σ–reflection, at least for structures of the form L_α (cf. Barwise [1975]). On the other hand if instead of (Σ–Ref_Ω) we incorporated the rule (Π_2–Ref_Ω) in the infinitary calculus, cut–elimination could be handled in almost the same spirit. By (Π_2–Ref_Ω) is meant the rule

$$\frac{\Gamma, \forall x^\Omega \exists y^\Omega F(x,y)}{\Gamma, \exists z\ [Tran(z) \wedge z \neq \emptyset \wedge (\forall x \in z)(\exists y \in z) F(x,y)]}$$

where $Tran(z)$ says that z is transitive and F ranges over the $\Delta_0(\Omega)$–formulas.

At first glance it might be surprising that the collapsing technique of 3.4 also renders (Π_2–Ref_Ω) accessible since, as a rule, a derivation with instances of (Π_2–Ref_Ω) will have Ω–branchings whilst the collapsing technique is evidently constrained to derivation without such branchings. To overcome this difficulty, one employs an asymmetrical interpretation of the quantifiers. To explain this, let $\Pi_\mathcal{H} \frac{}{\Omega} \Lambda$ be a derivation, possibly containing instances of (Π_2–Ref_Ω), and suppose that Λ is a set of RS_Ω–formulas of utmost complexity $\Pi_2(\Omega)$. Now proceed as follows:

- Pick $\gamma < \Omega$, and remove from each rule

$$\frac{\cdots \Delta, H(t) \cdots (|t| < \Omega)}{\Delta, \forall x^\Omega H(x)}\ (\forall)$$

in $\Pi_{\mathcal{H}}$ all the premisses $\Delta, H(t)$ with $\gamma \leq |\ t\ |$ as well as their subproofs. In the remaining tree replace every quantifier $\forall x^{\Omega}$ by $\forall x^{\gamma}$. For a suitably chosen operator (uniformly in γ) this will give a new proof without Ω–branchings to which the collapsing technique of 3.4 can thus be applied. After collapsing employ the Bounding Theorem to the collapsed derivation in order to extract a bound $g(\gamma) < \Omega$ for the existential quantifiers $(\exists x^{\Omega})$.

- Compute the function g' which enumerates the fixed points of g.
- Construct a new operator $\mathcal{H}'$ from $\mathcal{H}$ which is closed under g', i.e. $\eta \in \mathcal{H}'(X)$ entails $g'(\eta) \in \mathcal{H}'(X)$.
- By combining all the previous steps one receives an $\mathcal{H}'$–controlled derivation $\Pi_{\mathcal{H}'} \models_{\Omega} \Lambda$ without any instances of $(\Pi_2\text{–}Ref_{\Omega})$.

The final result reads as follows:

Theorem. *If $KP \vdash \forall x \exists y F(x,y)$ with F a Σ formula, then there exists an n such that for all $\xi < \Omega$,*

$$(\forall x \in L_{\xi})(\exists y \in L_{\theta_{\Omega(n)}(\xi+1)})F(x,y).$$

3.7. More admissibles

The cut–elimination procedure we have seen operating so well on KP can be adapted to extensions of the form

$$KP + \text{`there are many admissibles'}.$$

A prominent example for such a theory is Jäger's system KPi which, in addition to KP, has an inaccessibility axiom saying that for every set x there is an admissible set y containing it, i.e. $x \in y$.

It turned out that KPi is of the same proof–theoretic strength as the subsystem of second order arithmetic, $(\Delta^1_2 - CA) + BI$. The latter system consists of arithmetic plus

$$(\Delta^1_2 - CA): \qquad \forall n[F(n) \leftrightarrow G(n)] \rightarrow \exists X \forall n[n \in X \leftrightarrow F(n)]$$

$$\text{for all } F \in \Pi^1_2,\ G \in \Sigma^1_2,$$

$$BI: \qquad WO(<_X) \wedge \forall n[\forall m <_X n\ H(m) \rightarrow H(n)] \rightarrow \forall n H(n),$$

where $m <_X n := 3^m \cdot 5^n \in X$.

However, adjusting the methods which have been fruitfully employed to KP to KPi, is easier said than done. When ascending from KP to KPi, the ordinal notation systems as well as the cut–elimination procedures get more and more complicated. Notwithstanding that, the key idea pervades.

Finally, I shall briefly report on the theory KPM which is somewhat on the verge of admissible proof theory. KPM is designed to axiomatize essential features of a recursively Mahlo universe of sets, i.e. a universe that is a model of KPi and the scheme

$$(M) \quad \forall x \exists y H(x,y) \to \exists z[Ad(z) \wedge (\forall x \in z)(\exists y \in z)H(x,y)]$$

for all Δ_0–formulas $H(a,b)$, where $Ad(z)$ signifies that z is an admissible set.

It is easily verified that L_α is a model of KPM if and only if α is a recursively Mahlo ordinal (cf. Hinman [1978]).[11]

An ordinal analysis for KPM was published in Rathjen [1991] and has also been obtained independently by Arai [1989].

Roughly speaking, the central scheme of KPM falls under the heading "Π_2–reflection with constraints". The main stumbling block for an analysis of KPM was the invention of a suitable ordinal notation system. Till that time the recipes for creating ordinal notation systems had been based on ideas of Veblen and Bachmann. But these ideas only enabled one to engender collapsing functions which take as their values ordinals that, even when looked at from within the notation system, have cofinality ω, thus are highly singular ordinals. To be more precise, from the viewpoint of a notation system N the regularity of an ordinal $\kappa \in N$ is manifested by its being equipped with a collapsing function $\psi_\kappa : N \to N \cap \kappa$. Yet, in the approaches we have been just alluding to, the image of ψ_κ would never contain an ordinal π that is anew equipped with a collapsing function ψ_π, whereas the ordinal analysis of KPM requires a collapsing function always having this property. Eventually, such collapsing were developed in Rathjen [1990].

Not to leave any stone unturned, a characterization of KPM in terms of subsystems of second order arithmetic may be found in Rathjen [1991d]. It turns out that KPM proves the same sentences of second order arithmetic as $(\Delta^1_2 - CA) + BI$ augmented by an axiom schema expressing that every true Π^1_3 sentence (possibly including parameters) is already satisfied in a β–model of $(\Delta^1_2 - CA)$.

[11] An admissible ordinal α is said to be recursively Mahlo if for every total function $f : \alpha \longrightarrow \alpha$ that is Σ–definable in L_α there exists some $0 < \beta < \alpha$ such that $(\forall \xi < \beta)(f(\xi) < \beta)$.

4. Aspects of ordinal analysis

This Section is reserved to the discussion of consequences of ordinal analysis which were exhibited at the end of Section 2.

To explain these points, let $(D, \prec, \cdots)$ be an ordinal notation system where D stands for a set of terms and $\prec$ denotes their ordering relation.[12] Moreover, let T be a theory which has been analyzed by way of $(D, \prec, \cdots)$, resulting in $\mid T \mid = \mid \prec \mid$.

4.1. Consistency

By $PRWO(\prec)$ we mean the Π^0_2–sentence of arithmetic expressing that $\prec$ is primitive recursively well–ordered, i.e. for every primitive recursive function p a strictly $\prec$–descending chain $p(0) \prec p(1) \prec \cdots$ must terminate after finitely many steps.

Then a consistency proof of T can be carried out in PRA extended by $PRWO(\prec)$.

PRA is distinguished here since it is widely agreed that this system does not go beyond finitary reasoning in Hilbert's sense.

However, $PRA + PRWO(\prec)$ proves a much stronger consistency property, namely the 1–consistency of T, signifying that any Σ^0_1 sentence which is provable in T is also true.

As to PA, the result $PRA + PRWO(\varepsilon_0) \vdash Con(PA)$ can be easily drawn from Gentzen's 1938 paper. There he assigned ordinal notations $ord(d) < \varepsilon_0$ to PA–derivations d and gave a primitive recursive reduction procedure R such that, for any derivation d of an inconsistency, $R(d)$ is also a derivation of an inconsistency and, in addition, $ord(R(d)) < ord(d)$.

Later on, the ordinal ε_0 was reobtained as the ordinal of PA by use of derivations in infinitary logic with ω–rule, especially through Schütte's work. In the infinitary setting ordinals make a canonical appearance as a measure of the lengths of proof trees as well as of their cut–ranks. One is naturally led to ask whether Gentzen's result can also be achieved by employing cut–elimination for infinitary logic. This can be answered in the affirmative. It has turned out that primitive recursive proof–trees suffice and that the syntactical transformations employed in the course of cut–elimination can be represented by primitive recursive functions on the codes (cf. Schwichtenberg [1977]). Thus the use of infinitary derivations in metamathematics is much in keeping with Gentzen's extension of the finite standpoint since the only principle for dealing with them that transcends finitistic means is a de-

[12] "..." is supposed to indicate that such a notation system usually conveys a much richer structure.

scending chain principle to show that certain 'concrete' (primitive recursive) processes terminate.

4.2. Reduction

$\prec$ will arise as the union of initial segments $\prec_n$ ($n \in \mathbb{N}$) such that, for any $n \in \mathbb{N}$, T proves $\prec_n$ being well–ordered.

Let $PA_{<|T|}$ stand for Peano Arithmetic endowed with the scheme of transfinite induction for all the orderings $\prec_n$. Then T is conservative over $PA_{<|T|}$ with respect to all arithmetic sentences or, equivalently, T is conservative over the intuitionistic system $HA_{<|T|}$ with respect to all arithmetic sentences modulo $\neg\neg$ translations.

Just to mention two applications of such reductions:

By an ordinal analysis of the theories ID_ν formalizing ν–fold iterated inductive definitions, Pohlers and Buchholz (cf. Pohlers [1981]) showed that these theories were reducible to their intuitionistic counterparts ID^i_ν.

Another famous example is provided by the reduction of Δ^1_2 comprehension plus bar induction to Feferman's constructive theory T_0 of functions and classes. T_0 is based on intuitionistic logic and is a suitable framework for Bishop style constructive mathematics. In 1977, Feferman (cf. Feferman and Sieg [1981]) had shown that T_0 is interpretable in $(\Delta^1_2 - CA) + BI$. The ordinal analysis of the latter system is due to joint work of Jäger and Pohlers [1982]. Jäger [1983] then showed that the well–ordering proof for any ordinal $<|(\Delta^1_2 - CA) + BI|$ can be carried out in T_0; thereby completing the reduction.

4.3. A classification of the provably recursive functions

The *$\prec$–descent recursive functions*, $DCR(\prec)$, constitute the smallest class of recursive functions that has all the closure properties of the primitive recursive functions and, in addition, is closed with respect to the scheme:

> If g and h are in the class, and there is some natural number k such that $h(x,y) \prec k$ holds for all $x, y \in \mathbb{N}$, then so is
>
> $$f(m) = g(\mu n.[h(n,m) \preceq h(n+1,m)], m),$$
>
> where μn indicates the least n in the ordering of the integers.

The reason for introducing the class $DCR(\prec)$ is (as was to be expected) that this class coincides with the provably recursive functions of T.

The concept of descent recursive function is for instance discussed in Smith [1985].

5. Beyond admissible proof theory

The strength of Π^1_2 comprehension is greatly bigger than that of Δ^1_2 comprehension. In particular, there is no way to describe this comprehension in terms of admissibility.

As to the set–theoretic side, Π^1_2 comprehension corresponds to Σ separation, i.e. the set of axioms

$$\exists z(z = \{x \in a : F(x)\})$$

for all Σ formulas F in which z does not occur free.

The precise relationship reads as follows:

5.1 Theorem. *$KP + \Sigma$ separation and $(\Pi^1_2 - CA) + BI$ prove the same sentences of second order arithmetic.*[13]

The ordinals κ such that $L_\kappa \models KP + \Sigma$ *separation* are familiar from ordinal recursion theory.

5.2 Definition. *An admissible ordinal κ is said to be nonprojectible if there is no total κ–recursive function mapping κ one–one into some $\beta < \kappa$, where a function $F : L_\kappa \to L_\kappa$ is called κ–recursive if it is Σ definable in L_κ.*

The key to the 'largeness' properties of nonprojectible ordinals is:

5.3 Theorem. *For any nonprojectible ordinal κ, L_κ is a limit of Σ_1–elementary substructures*[14]*, i.e. for every $\beta < \kappa$ there exists a $\beta < \rho < \kappa$ such that L_ρ is a Σ_1–elementary substructure of L_κ, written $L_\rho \prec_1 L_\kappa$.*

Such ordinals satisfying $L_\rho \prec_1 L_\kappa$ have strong reflecting properties. For instance, if $L_\rho \models F$ for some set–theoretic sentence F (possibly containing parameters from L_ρ), then there exists a $\gamma < \rho$ such that $L_\gamma \models F$. This is because $L_\rho \models F$ implies $L_\kappa \models \exists \gamma F^{L_\gamma}$, hence $L_\rho \models \exists \gamma F^{L_\gamma}$ using $L_\rho \prec_1 L_\kappa$.

The last result makes it clear that an ordinal analysis of Π^1_2 comprehension would necessarily involve a proof–theoretic treatment of reflections beyond those surfacing in admissible proof theory. Here one encounters two difficulties.

1. Significantly stronger notation systems are required. The problem is (as always in this area) to develop a constructive object, i.e. a notation system, that shares "enough" properties with a (recursively) large ordinal. So far

[13] Warning: It is crucial to this result that Infinity is among the axioms of KP.

[14] L_ρ is said to be a Σ_1–elementary substructure of L_κ if every Σ_1–sentence with parameters from L_ρ that holds in L_κ also holds in L_ρ.

definition procedures based on ideas of Veblen and Bachmann have been paramount, but it seems that this approach is constrained to admissible proof theory. So some new ideas will be needed.

2. New cut–elimination procedures have to be invented. Of course, this task cannot be completely separated from the previous one since the ideas giving rise to a notation system should lend themselves to a cut–elimination procedure.

Recently we have been able to get hold on Π_n–reflection for arbitrary n.

5.4 Definition. *A set–theoretic formula is said to be* Π_n *(respectively* Σ_n*) if it consists of a string of* n *alternating quantifiers beginning with an universal one (respectively existential one), followed by a* Δ_0 *formula. By* Π_n*–reflection we mean the scheme*

$$F \to \exists z[Tran(z) \wedge z \neq \emptyset \wedge F^z]$$

where F *is* Π_n*, and* $Tran(z)$ *expresses that* z *is a transitive set.*

$\alpha > 0$ *is said to be* Π_n*–reflecting if* $L_\alpha \models \Pi_n$*–reflection.*

Σ_n*–reflection and* Σ_n*–reflecting ordinals are defined analogously.*

Π_n–reflecting ordinals have interesting points of contact with non–monotone inductive definitions.

5.5 Definition. *A function* Γ *from the power set of* $\mathbb{N}$ *into itself is called an operator on* $\mathbb{N}$*.* Γ *determines a transfinite sequence* $\langle \Gamma^\xi : \xi \in ON \rangle$ *of subsets of* $\mathbb{N}$ *defined by*

$$\Gamma^\lambda = \Gamma^{<\lambda} \cup \Gamma(\Gamma^{<\lambda}),$$

where $\Gamma^{<\lambda} = \bigcup_{\xi<\lambda} \Gamma^\xi$.

The closure ordinal $|\Gamma|$ *of* Γ *is the least ordinal* ρ *such that* $\Gamma^{\rho+1} = \Gamma^\rho$.

Γ *is said to be* Π^0_k *in case there is an arithmetic* Π^0_k *formula* $F(U, u)$ *with free second order variable* U *such that for* $X \subseteq \mathbb{N}$,

$$\Gamma(X) = \{n \in \mathbb{N} : F(X, n)\}.$$

Let $|\Pi^0_k| := sup\{|\Gamma| : \Gamma \text{ is } \Pi^0_k\}$.

By work of Aczel and Richter [1974] we have the following characterization.

5.6 Theorem.

$$|\Pi^0_k| = \textit{first } \Pi_{k+1}\textit{–reflecting ordinal}\,.$$

Several notions of recursively large ordinals are modelled upon notions of large cardinals. This is especially true of notions like "recursively inaccessible ordinal" and "recursively Mahlo ordinal". It turns out that the least Π_3–reflecting ordinal is greater than the least recursively Mahlo ordinal, indeed much greater than any iteration of "Mahloness" into the transfinite from below.

5.7 Definition. *Assume that κ is recursively Mahlo. κ is called recursively α–Mahlo if for every κ–recursive function $f : \kappa \to \kappa$ there is an ordinal $\beta < \kappa$ closed under f such that β is recursively γ–Mahlo for any $\gamma < \alpha$.*

κ is recursively hyper–Mahlo if κ is recursively κ–Mahlo.

As a matter of fact, there are 'many' recursively hyper–Mahlo ordinals below the first Π_3–reflecting ordinal. Aczel and Richter [1974] have convincingly argued that Π_3–reflecting ordinals are the recursive analogue of weakly compact cardinals also known as Π^1_1–indescribable cardinals. The same considerations justify the view that Π_{n+2}–reflecting ordinals provide the recursive analogue for the Π^1_n–indescribable cardinals for all $n > 0$.

Next we shall glimpse at an ordinal notation system which in some respect internalizes the first Π_3–reflecting ordinal. Rather than exhibiting such a notation system, it is more appropriate to give a model for the peculiar functions the notation system is made from. Such a model can be provided on the basis of a weakly compact cardinal.

So let us indulge in a little science fiction and fix a weakly compact cardinal κ.

5.8 Definition. *Let*

$$V = \bigcup_{\alpha \in ON} V_\alpha$$

be the cumulative hierarchy of sets, i.e.

$$V_0 = \emptyset, \quad V_{\alpha+1} = \{X : X \subset V_\alpha\}, \quad V_\lambda = \bigcup_{\xi<\lambda} V_\xi \text{ for limit ordinals } \lambda.$$

A cardinal κ is weakly compact *if whenever $U \subseteq V_\kappa$ and $A(P)$ is a Π^1_1 formula of set theory with P a class variable such that $\langle V_\kappa, \in\rangle \models A(U)$, then for some $\alpha < \kappa$:*

$$\langle V_\alpha, \in\rangle \models A(U \cap V_\alpha).$$

For κ a regular cardinal, a subset $S \subseteq \kappa$ is stationary in κ *if $S \cap C \neq \emptyset$ holds for every set $C \subseteq \kappa$ that is closed and unbounded in κ.*

5.9 Definition. *Let κ be a weakly compact cardinal. By recursion on α we define sets $B(\alpha,\beta)$, M^α and the function Ξ_κ as follows:*

$$B(\alpha,\beta) = \left\{ \begin{array}{l} \textit{closure of } \beta \cup \{0,\kappa\} \\ \textit{under } +, \lambda\xi.\omega^\xi \textit{ and } (\xi \mapsto \Xi_\kappa(\xi))_{\xi<\alpha} \end{array} \right.$$

$$M^\alpha = \{\pi < \kappa : B(\alpha,\pi) \cap \kappa = \pi \wedge \forall \xi \in B(\alpha,\pi) \cap \alpha [\pi \cap M^\xi \textit{ stationary in } \pi]\}$$

$$\Xi_\kappa(\alpha) = \textit{least element of } M^\alpha.$$

The hypothesis that κ be weakly compact will be needed to ensure that $M^\alpha \neq \emptyset$ and thus to show that $\Xi_\kappa(\alpha)$ is defined.

In a second step, for every $\pi \in M^\alpha$ and $\xi \in B(\alpha,\pi) \cap \alpha$, one defines collapsing functions

$$\Theta_\pi^\xi : ON \longrightarrow \pi \cap M^\xi.$$

With the aid of (symbols for) the functions and constants $\Xi_\kappa, \Theta_\pi^\xi, +, \omega, \kappa, 0$, and special constraints needed to ensure uniqueness of notations, it is then possible to construct a primitive recursive system of ordinal notations $N(\kappa)$ which reflects some properties of the rather large cardinal κ.

Akin to RS_Ω one can invent an infinitary calculus RS_κ, which in addition has the following rules:

$$(\Pi_3\text{–}Ref_\kappa)\ \frac{\Gamma, A^\kappa}{\Gamma, (\exists z \in \check{L}_\kappa)[Tran(z) \wedge z \neq \emptyset \wedge A^z]}$$

for every $\Pi_3(\kappa)$ formula A and

$$(\Pi_2\text{–}Ref_\pi^\xi)\ \frac{\Gamma, B}{\Gamma, (\exists z \in \check{L}_\pi)(z \in M^\xi \wedge B^z)}$$

for every $\Pi_2(\pi)$–formula B, where $\pi \in M^\alpha$, $\xi < \alpha$, $\xi \in B(\alpha,\pi)$.

The rules $(\Pi_2 - Ref_\pi^\xi)$ are not needed for an embedding of $KP + \Pi_3$–reflection into RS_κ. They are only required for carrying through the cut–elimination procedure. Usually, removing one instance of $(\Pi_3\text{–}Ref_\kappa)$ in a derivation can be done only at the expense of introducing a bunch of new $(\Pi_2\text{–}Ref_\pi^\xi)$ rules. This discriminates the cut–elimination for RS_κ sharply from that for RS_Ω, where instances of the impredicative rule $(\Sigma\text{–}Ref_\Omega)$ are replaced by instances of the predicative rule $(\exists)$.

Cut–elimination for RS_κ can be achieved by using the $\mathcal{H}$–controlled RS_κ–derivations, with $\mathcal{H}$ ranging over the operators

$$\mathcal{H}_\gamma(X) = \bigcap\{B(\alpha,\beta) : X \subseteq B(\alpha,\beta) \wedge \gamma < \alpha \wedge \beta < \kappa\}$$

where $\gamma \in N(\kappa)$.

For $\Gamma = \{A_1, \cdots, A_n\}$ we set $\Gamma^\pi := \{A_1^\pi, \cdots, A_n^\pi\}$.
The key to the elimination of $(\Pi_3 - Ref_\kappa)$ is the following theorem.

5.10 Theorem. *If Γ is a set of $\Pi_3(\kappa)$ formulas and $\mathcal{H}_\gamma \vdash^\alpha_{\kappa+1} \Gamma$, then, for every $\pi \in M^{f(\alpha,\gamma)}$,*

$$\mathcal{H}_{f(\alpha,\gamma)}[\pi] \vdash^{\Xi_\kappa(f(\alpha,\gamma)+\pi)}_{\Xi_\kappa(f(\alpha,\gamma)+\pi)} \Gamma^\pi ,$$

where f is a function that depends only on Γ.

It is not by accident that in Theorem 5.9 a single derivation is 'collapsed' into a family of derivations indexed by a stationary subset of κ. The elimination of $(\Pi_3$–$Ref_\kappa)$ requires such a "stationary collapsing" technique.

Unfortunately, we will not be able to go any further into details. The interested reader is referred to Rathjen [1991e].

At the end we hasten to assure that this is not the first of an infinite series of new cut–elimination procedures. Π_3–reflection just served as a paradigm. Stationary collapsing is applicable to all of the theories $KP + \Pi_n$–reflection.

To close, we raise the question of how far afield from Π^1_2 comprehension all this is. The idea is to approach Π^1_2 comprehension by stronger and yet stronger reflection principles in an autonomous manner. I conjecture that the large cardinal analogue for a suitable notation system resides below the first Ramsey cardinal, and, moreover, is compatible with $V = L$.

References

ACZEL P. and RICHTER W.H. [1974], *Inductive definitions and reflecting properties of admissible ordinals*, in: J.E.Fenstad and P.G. Hinman, eds., Generalized recursion theory, North Holland, Amsterdam, 301–381.

ARAI T. [1989], *Proof theory for reflecting ordinals II: recursively Mahlo ordinals*, handwritten notes.

BARWISE J. [1975], *Admissible sets and structures*, Springer Verlag, Berlin, Heidelberg, New York.

BUCHHOLZ W. [1991], *A simplified version of local predicativity*, to appear in: Leeds Proof Theory 1990, Cambridge University Press.

BUCHHOLZ W., FEFERMAN S., POHLERS W., SIEG W. [1981], *Iterated inductive definitions and subsystems of analysis*, Springer Verlag, Berlin, Heidelberg, New York.

FEFERMAN S. [1970], *Formal theories for transfinite iterations of generalized inductive definitions and some subsystems of analysis*, in: J. Myhill, A. Kino, R.F. Vesley, eds., North Holland, Amsterdam.

FEFERMAN S. and SIEG W. [1981], *Proof theoretic equivalence between classical and constructive theories for analysis*, in: Buchholz, Feferman, Pohlers, Sieg, [1981], 78–142.

FEFERMAN S. [1988], *Hilbert's program relativized*, Journal of Symbolic Logic 53, 364–384.

GENTZEN G. [1938], *Neue Fassung des Widerspruchsfreiheitsbeweises für die reine Zahlentheorie*, Forschungen zur Logik und zur Grundlegung der exakten Wissenschaften, Neue Folge 4, 19–44.

GIRARD J.-Y. [1985], *Introduction to Π^1_2–logic*, Synthese 62, 191–216.

HINMAN P.G. [1978], *Recursion-theoretic hierarchies*, Springer Verlag, Berlin, Heidelberg, New York.

HOWARD W.A. [1968], *Functional interpretation of bar induction by bar recursion*, Comp. Math. 20, 107–124.

HOWARD W.A. [1981], *Ordinal analysis of bar recursion of type zero*, Comp. Math. 42, 105–119.

JÄGER G. [1982], *Zur Beweistheorie der Kripke-Platek Mengenlehre über den natürlichen Zahlen*, Archiv für Mathematische Logik und Grundlagenforschung 22, 121–139.

JÄGER G. [1983], *A well-ordering proof for Feferman's theory T_0*, Archiv für Mathematische Logik und Grundlagenforschung 23, 65–77.

JÄGER G. and POHLERS W. [1982], *Eine beweistheoretische Untersuchung von $(\Delta^1_2 - CA) + BI$ und artverwandter Systeme*, Sitzungsberichte der Bayerischen Akademie der Wissenschaften, Mathematisch–Naturwissenschaftliche Klasse.

POHLERS W. [1981], *Proof-theoretical analysis of ID_ν by the method of local predicativity*, in: Buchholz, Feferman, Pohlers, Sieg, [1981], 261–357.

POHLERS W. [1982], *Admissibility in proof theory*, in: L.J. Cohen, J. Los, H. Pfeiffer, K.-P. Podewski, eds., Logic, Methodology and Philosophy of Science VI, North Holland, Amsterdam, 123–139.

POHLERS W. [1987], *Contributions of the Schütte school in Munich to proof theory*, in: Takeuti [1987], 406–431.

POHLERS W. [1989], *Proof theory: an introduction*, Springer Verlag, Berlin.

POHLERS W. [1991], *Proof theory and ordinal analysis*, Arch. Math. Logic 30, 311–376.

RATHJEN M. [1990], *Ordinal notations based on a weakly Mahlo cardinal*, Arch. Math. Logic 29, 249–263.

RATHJEN M. [1991a], *Proof-theoretic analysis of KPM*, Arch. Math. Logic 30, 377–403.

RATHJEN M. [1991b], *Fragments of Kripke-Platek set theory with infinity*, to appear in: Leeds Proof theory 90, Cambridge University Press.

RATHJEN M. [1991c], *How to develope proof-theoretic ordinal functions on the basis of admissible ordinals*, to appear in: Zeitschrift für Mathematische Logik und Grundlagen der Mathematik.

RATHJEN M. [1991d], *An interpretation of KPM in second order arithmetic and a characterization of 1-section (superjump)*, preprint, Universität Münster.

RATHJEN M. [1991e], *Proof-theoretic analysis of theories of non-monotone induction*, in preparation.

SCHÜTTE K. [1977], *Proof Theory*, Springer Verlag, Berlin, Heidelberg, New York.

SCHWICHTENBERG H. [1977] *Proof Theory: some applications of cut-elimination*, in: J. Barwise, ed., Handbook of mathematical logic, North Holland, Amsterdam, 867–895.

SIEG W. [1988], *Hilbert's program sixty years later*, Journal of Symbolic Logic 53, 338–348.

SMITH R.L. [1985], *The consistency strengths of some finite forms of the Higman and Kruskal theorems*, in: L.A. Harrington, M.D. Morley, A. Scedrov, S.G. Simpson eds., Harvey Friedman's research on the foundations of mathematics, North Holland, Amsterdam, 119–136.

TAKEUTI G. [1987], *Proof theory*, second edition, North Holland, Amsterdam.

Logic, Methodology and Philosophy of Science IX
D. Prawitz, B. Skyrms and D. Westerståhl (Editors)

ON THE REDUCIBILITY ORDER BETWEEN BOREL EQUIVALENCE RELATIONS

ALAIN LOUVEAU

Equipe d'Analyse, Université Paris VI, Paris, France

Introduction

An equivalence relation E on a set X is a *Borel equivalence relation* if both X and E are Borel, in some Polish space (which can always be taken to be the space $\mathbf{2}^\omega$), and its square, respectively. We denote by $BOREQ$ the class of all Borel equivalence relations.

We say that (X, E) is *reducible* to (Y, F) if there is a Borel function $f : X \to Y$ such that

$$\forall x \in X \, \forall y \in X \, \big(xEy \leftrightarrow f(x)Ff(y)\big).$$

This defines a quasi-ordering $\leq$ on $BOREQ$, with associated equivalence $\equiv$. For more information about this quasi-ordering, see the paper of Kechris [6].

In [2], H. Friedman and L. Stanley prove:

Fact 1. $(BOREQ, \leq)$ has no maximum element.

In fact, they introduce a "jump" operator, a version of which is defined as follows: To each E, associate E^+ on X^ω, defined by

$$(x_n)E^+(y_n) \leftrightarrow \forall n \exists m \, (x_n E y_m) \wedge \forall n \exists m \, (x_m E y_n).$$

Fact 1 then follows from:

THEOREM (Friedman-Stanley). *For all E in $BOREQ$ with at least two classes, $E < E^+$.*

The original proof of this theorem used the deep results of H. Friedman on Borel diagonalizations, and in particular was not elementary (i.e. in second-order arithmetics). In [3], Harrington gives an elementary proof of this theorem. For each countable ordinal ξ, let $E(\xi)$ be the ξth iterated jump, using the above operator $+$, of $(\mathbf{2}^\omega, =)$. Harrington proves:

THEOREM (Harrington). *The family* $E(\xi)$, $\xi \in \omega_1$ *is unbounded in* $BOREQ$. *In fact for each* ξ, *and each* Σ^0_ξ *equivalence relation* E, $E(\xi+1) \not\leq E$.

This result easily implies the Friedman-Stanley theorem, hence Fact 1, but also

Fact 2. $BOREQ \cap \Sigma^0_\xi$ is not cofinal in $BOREQ$.

Harrington's proof is elementary, but uses a delicate forcing argument. The aim of this paper is to give a different, and much simpler, proof of Facts 1 and 2, based on a different "jump" operator, for which we will prove analogs of the two theorems above. This proof also brings in an interesting invariant of the reducibility equivalence relation $\equiv$, the *potential Wadge class* of a Borel equivalence relation.

I would like to thank R. Sami and J. Saint Raymond for the discussions we had on the subject.

1. Potential Wadge classes

DEFINITION 1. Let Γ be a Wadge class, and X a Borel set. A subset A of X^2 is *potentially of class* Γ, written $A \in pot\Gamma$, if for some finer Polish topology τ on X, A is in Γ in $(X,\tau)^2$.

One can define the potential Wadge class of a Borel set $A \subseteq X^2$, $pot\Gamma(A)$, as the least Γ such that $A \in pot\Gamma$ (this is clearly well-defined). Now if $A \subseteq X^2$ and $B \subseteq Y^2$ are such that there exists a Borel function $f : X \to Y$ with $(x,y) \in A \leftrightarrow (f(x),f(y)) \in B$, then $pot\Gamma(A) \subseteq pot\Gamma(B)$. We will apply this remark to Borel equivalence relations.

Note that the notion of potential Wadge classes is non trivial: For each non self dual Borel Wadge class Γ, with dual class $\check{\Gamma}$, there is in $(\mathbf{2}^\omega)^2$ a set in Γ which is not in $pot\check{\Gamma}$, namely any Γ-universal set. To see this, note that any two Polish topologies, with one finer than the other, coincide on a dense G_δ set, hence on a perfect set, which contains a set in $\Gamma \backslash \check{\Gamma}$.

It is usually hard to compute the exact potential Wadge class of a Borel equivalence relation. However, we will be able to do it in enough particular cases.

Let $\mathcal{F}$ be a filter on ω. We define the relation $2^{\mathcal{F}}$ on $\mathbf{2}^\omega$ by

$$\alpha \; 2^{\mathcal{F}} \; \beta \leftrightarrow \{n : \alpha(n) = \beta(n)\} \in \mathcal{F}.$$

THEOREM 2. *Let* Γ *be a Wadge class closed under intersections. If* $2^{\mathcal{F}}$ *is in* $pot\Gamma$, *then* $\mathcal{F}$ *is in* Γ.

PROOF: Let τ be the finer Polish topology on $\mathbf{2}^\omega$ for which $2^{\mathcal{F}}$ is in Γ, and $H \subseteq \mathbf{2}^\omega$ be a dense G_δ set on which the two topologies coincide. We

claim that there is a partition of ω into two sets A_0, A_1, and two sets B_0, B_1, with for $i = 0, 1$ $B_i \subseteq A_i$, such that for $i = 0$ or 1, if $A \subseteq \omega$ satisfies $A \cap A_i = B_i$, then $A \in H$. This claim will finish the proof, for one has, for $A \subseteq \omega$:

$$A \in \mathcal{F} \leftrightarrow (A \cap A_0) \cup B_1 \; 2^{\mathcal{F}} B_1 \; \wedge \; (A \cap A_1) \cup B_0 \; 2^{\mathcal{F}} B_0$$

(here and below, we identify a subset of ω with its characteristic function). And as $2^{\mathcal{F}}$ is in Γ on H^2, this gives a Γ definition of $\mathcal{F}$, as desired.

To prove the claim, note first that for any dense open set G in $\mathbf{2}^\omega$, and any k, there is an $l > k$ and a subset S of $[k, l[$ such that any $A \subseteq \omega$ with $A \cap [k, l[= S$ is in G: Enumerate all subsets of $[0, k[$ as $(S_n)_{n<2^k}$, and build inductively k_n and $T_n \subseteq [k_n, k_{n+1}[$, starting with $k_0 = k$, so that for each $n < 2^k$, if $A \cap [k_i, k_{i+1}[= T_i$ for all $i \leq n$ and $A \cap [0, k[= S_n$, then $A \in G$, using the density of G. Then $l = k_{2^k}$ and $S = \cup_n T_n$ work. Applying the subclaim successively to a decreasing sequence $(G_n)_{n\in\omega}$ of dense open sets with intersection H gives a sequence k_n with $k_0 = 0$, and sets $S_n \subseteq [k_n, k_{n+1}[$ such that if $A \cap [k_n, k_{n+1}[= S_n$, $A \in G_n$. Then $A_i = \bigcup_n [k_{2n+i}, k_{2n+i+1}[$ and $B_i = \bigcup_n S_{2n+i}$, for $i = 0, 1$, satisfy the claim. ⊣

Remark. The claim used in the previous proof is a folklore result. It can be used e.g. to show that a free Borel filter $\mathcal{F}$ on ω is meager, or that there exists a finite-to-one function $\varphi : \omega \to \omega$ with $\varphi(\mathcal{F}) = \mathcal{N}$. More interestingly, W. Just uses it in [5] to prove that there are in $(BOREQ, \leq)$ antichains of arbitrary finite cardinality.

By the previous result, the computation of the potential Wadge class of $2^{\mathcal{F}}$ is reduced to the computation of the Wadge class of $\mathcal{F}$. We do not know exactly which Wadge classes are Wadge classes of filters on ω (Easily, Δ^0_1, Π^0_1 and Σ^0_2 are such classes, and by a Baire category argument, any Π^0_2 filter is Π^0_1. Calbrix [1] has exhibited filters of Wadge class Π^0_ξ and Σ^0_ξ for all $\xi > 2$). Nevertheless, it is easy to check that Borel filters have Wadge classes unbounded in Δ^1_1. In fact if we let $\mathcal{N}$ be the Fréchet filter, $\mathcal{N} = \{A \subseteq \omega : A \text{ is cofinite}\}$, and if we define its iterates $(\mathcal{N}_\xi)_{\xi\in\omega_1}$ by induction by

$$\mathcal{N}_1 = \mathcal{N}$$

$$A \in \mathcal{N}_{\xi+1} \leftrightarrow \{n : \{m : \varphi(n, m) \in A\} \in \mathcal{N}_\xi\} \in \mathcal{N}$$

where φ is a bijection between ω and ω^2, and for limit λ

$$A \in \mathcal{N}_\lambda \leftrightarrow \{n : \{m : \varphi(n, m) \in A\} \in \mathcal{N}_{\psi(n)}\} \in \mathcal{N}$$

where ψ is a bijection between ω and λ, then one easily checks that all Borel sets are obtained from the clopen sets by the operation of liminf along one of these iterates. So their Wadge classes are unbounded, and by the theorem above, we get:

COROLLARY 3.

(a) *The sequence* $(2^{\mathcal{N}_\xi})_{\xi\in\omega_1}$ *is unbounded in BOREQ.*

(b) *Given any countable* η, *there is a* ξ *such that for any* Σ^0_η *equivalence relation* E, $2^{\mathcal{N}_\xi} \not\leq E$ *(In fact, by the exact computations of Calbrix [1], one can take* $\xi = \eta$ *).*

Remarks. 1. Theorem 2 has another nice consequence. In [4], Harrington-Kechris-Louveau prove that any Borel equivalence relation either is smooth, i.e. reducible to $(\mathbf{2}^\omega, =)$, or else reduces $2^{\mathcal{N}}$. From this, they infer, using a measure theoretic argument, that every G_δ equivalence relation is smooth. This can also be derived from Theorem 2, by noting that otherwise $2^{\mathcal{N}}$ would be potentially Π^0_2, hence $\mathcal{N}$ would be Π^0_2 in $\mathbf{2}^\omega$, a clear contradiction.

2. In the proofs above, the only property used of the reducing function was the Baire Property, so that our arguments would apply, using the appropriate level of determinacy, to more general notions of reducibility, up to reducibility by arbitrary functions in the context of AD, as noticed by A.S. Kechris.

2. A jump operator in *BOREQ*

DEFINITION 4. Let $\mathcal{F}$ be a Borel filter on ω, and E a Borel equivalence relation on some Borel X. We define the relation $E^{\mathcal{F}}$ on X^ω by

$$(x_n)E^{\mathcal{F}}(y_n) \leftrightarrow \{n : x_n E y_n\} \in \mathcal{F}$$

Note that with this notation, $2^{\mathcal{F}}$ is just $(2, =)^{\mathcal{F}}$.

THEOREM 5. *The operator* $E \mapsto E^{\mathcal{N}}$ *is a jump operator in BOREQ: For every Borel* E *with at least two classes,* $E < E^{\mathcal{N}}$.

PROOF: Clearly, $E \leq E^{\mathcal{N}}$, by sending any $x \in X$ to the constant sequence (x). Assume that $E^{\mathcal{N}} \leq E$. We claim that for every $\xi < \omega_1$, $E^{\mathcal{N}_\xi} \leq E$. This is proved by induction on ξ. Suppose first $\xi = \eta + 1$ is successor, and let f be a Borel reduction of $E^{\mathcal{N}_\eta}$ to E. Define $F : X^\omega \to X^\omega$ by:

$$F((x_k)) = (f((x_{\varphi(n,m)})_m))_n.$$

One gets

$$\begin{aligned}(x_k)E^{\mathcal{N}_\xi}(y_k) &\leftrightarrow \{k : x_k E y_k\} \in \mathcal{N}_\xi \\ &\leftrightarrow \{n : \{m : x_{\varphi(n,m)} E y_{\varphi(n,m)}\} \in \mathcal{N}_\eta\} \in \mathcal{N} \\ &\leftrightarrow ((x_{\varphi(n,m)})_m)_n (E^{\mathcal{N}_\eta})^{\mathcal{N}}((y_{\varphi(n,m)})_m)_n \\ &\leftrightarrow F((x_k))E^{\mathcal{N}}F((y_k))\end{aligned}$$

So $E^{\mathcal{N}_\xi} \leq E^{\mathcal{N}} \leq E$, as desired.

The proof for limit ξ is similar: Let for each $\eta < \xi$ f_η reduce $E^{\mathcal{N}_\eta}$ to E, and set

$$F((x_k)_k) = (f_{\psi(m)}((x_{\varphi(n,m)})_m))_n.$$

By a computation similar to the one above, one checks that F reduces $E^{\mathcal{N}_\xi}$ to $E^{\mathcal{N}}$, and as above we get the claim.

Suppose now that $(2, =) \leq E$. Then easily for any $\mathcal{F}$, one gets $2^{\mathcal{F}} \leq E^{\mathcal{F}}$, hence by the previous claim, for all $\xi < \omega_1$, we get $2^{\mathcal{N}_\xi} \leq E$, contradicting Theorem 2. ⊣

REFERENCES

[1] J. CALBRIX, *Classes de Baire et espaces d'applications continues*, Note aux C. R. Acad. SC. Paris, 301, 1985, 759–762.

[2] H. FRIEDMAN, L. STANLEY, *A Borel reducibility theory for classes of countable structures*, J. Symb. Logic 54 (1989), 894–914.

[3] L. HARRINGTON, *On the complexity of Borel equivalence relations*, abstract, International Workshop on Set Theory, Marseille-Luminy, 1990.

[4] L. HARRINGTON, A. S. KECHRIS, A. LOUVEAU, *A Glimm-Effros dichotomy for Borel equivalence relations*, Journal of the A.M.S.4(3),1990,903–928.

[5] W. JUST, *More mutually irreducible ideals*, preprint, 1990.

[6] A. S. KECHRIS, *The structure of Borel equivalence relations in Polish spaces*, to appear in the Proceedings of the Workshop on Set Theory and the Continuum, MSRI, Berkeley 1989.

Logic, Methodology and Philosophy of Science IX
D. Prawitz, B. Skyrms and D. Westerståhl (Editors)

THE CORE MODEL UP TO A WOODIN CARDINAL

WILLIAM MITCHELL

Dept. of Mathematics, Univ. of Florida, Gainsville, Fl 32611, USA

Inner models, and in particular core models, have made important contributions to the theory of measurable cardinals and of other large cardinal properties of similar consistency strength. Until recently, however, very little was known about the inner model theory of cardinals beyond measurable cardinals. Part of the reason for this weakness was our lack of understanding of the potential of large cardinal properties: it was generally believed that the smallest important cardinal larger than a measurable cardinal was a supercompact cardinal, and even supercompact cardinals were believed to be relatively weak — far weaker, for example, than the axiom of determinacy for $\mathbf{\Sigma}^1_2$ sets of reals. This made inner model theory a hard nut to crack: very little is known even today about inner model theory for supercompact cardinals. Work of Foreman, Magidor, and Shelah [6] started to reverse these views, showing that supercompactness was stronger than previously believed, and following this Woodin built on their ideas to pinpoint a property, now known as a Woodin cardinal, which is far weaker than supercompactness but appears to be at least as interesting. Further work has confirmed the importance of Woodin cardinals, most notably through the discovery by Martin and Steel [8] and by Woodin [10] that the existence of a Woodin cardinal is equiconsistent with the axiom of determinacy for $\mathbf{\Sigma}^1_2$ formulas, and that the existence of infinitely many Woodin cardinals is equiconsistent with the full axiom of determinacy.

This work has led to major advances in core model theory. On the one hand it has given an additional impetus to the study of cardinals small enough that a core model theory is practical at the present time, while on the other hand it has provided new tools together with a new understanding of some of the older tools. The purpose of this paper is to give an exposition of the current state of core model theory, and in particular of the core model theory for large cardinals up to a Woodin cardinal. The core model which we will be describing is due to J. Steel

[19]; other major contributors to the inner model theory leading up to this have been S. Baldwin, R. Jensen, D. A. Martin, W. Mitchell and H. Woodin.

This paper is not intended to be technical discussion of the core model. It does not assume a knowledge of core model theory for measurable cardinals, and will not attempt to give more than a superficial knowledge of the new core model theory. It will attempt to give some idea of what has been accomplished, of the main difficulties that have been surmounted, and of some of the gaps and limitations in the current theory.

§ 1. What is a core model?

What we will call "the true core model" is a model, to be denoted by $\mathbf{K}$, which contains all of the large cardinal structure existing in the universe, but which is, at the same time, as much as possible like the constructible sets L. The existence of this model is speculative, but $\mathbf{K}$ is known to exist under appropriate assumptions restricting the size of large cardinals existing in the universe. The results discussed in this paper imply that it exists under the assumption that there is no Woodin cardinal together with a further technical assumption.

For the core model program to succeed it is not enough that this "true" core model exist, but it is also necessary that it be recognizable when it is found. With this in mind we will begin by describing some examples of core models for smaller cardinals, looking at their characteristics as a guide to what to expect as we move up to larger cardinals.

EXAMPLES OF THE CORE MODEL. The simplest example of a core model is simply the class L of constructible sets. This model does contain the large cardinal structure of the universe, provided that this structure is not too large: for example any cardinal which is inaccessible, Mahlo, or weakly compact in V has the same property in L. Measurable cardinals, on the other hand, cannot exist in L. A large cardinal property which is weaker than measurability but still cannot hold in L is the existence of a nontrivial embedding $i\colon L \to L$, which is necessarily not a member of L. This property is equivalent to the existence of a class of indiscernibles for L, and also to the existence of a particular subset, known as $0^{\#}$, of ω. The existence of $0^{\#}$ is a critical dividing line: it is inconsistent with L, but every smaller large cardinal is not only consistent with L but reflects to L just like inaccessibility and weak compactness. Thus we say that L is the core model up to $0^{\#}$, that is, $\mathbf{K} = L$ provided that $0^{\#}$ does not exist.

The second example of a core model to be studied was $L[\mu]$, the class of sets constructible from a measure μ. It is easy to check that $L[\mu]$ is the minimal model in which μ is a measure, but Kunen showed in [7] that in addition $\mu \cap L[\mu]$ is the only measure in $L[\mu]$, and that the model $L[\mu]$ does not depend on the choice of the measure μ, but only on the cardinal κ where μ lives. This work, together with work of Silver ([17], [18]) led to the recognition that $L[\mu]$ has the properties which now lead us to call it a core model: we say that $\mathbf{K} = L[\mu]$ provided that $L[\mu]$ exists (that is, that $\exists \mu, \kappa\ L[\mu] \models$ "μ is a measure on κ"), that κ is as small as possible, and that $0^\dagger$ does not exist. The nonexistence of $0^\dagger$, by analogy with $0^\#$, asserts that there is no nontrivial embedding from $L[\mu]$ to $L[\mu]$. Many of the techniques which are used in more general core models are taken from this basic work on $L[\mu]$.

There is a large gap between L and $L[\mu]$. This gap ranges from $0^\#$ up to a measurable cardinal, so that these models are of little help in understanding important intermediate notions such as Ramsey cardinals. In addition, $L[\mu]$ is frequently not very useful even in dealing with conditions which are as strong as a measurable cardinal. The problem is that if we want to show that some property $\mathbf{P}$ implies the consistency of a measurable cardinal, then it may not be of much use to have a model which can only be constructed after we have our hands on a measure μ. What is needed is a model which exists without any preconditions but which will give us a model with a measurable cardinal if there is such a model. The birth of true core model theory came with the construction by Dodd and Jensen ([2], [3], [4], [5]) of a model satisfying these conditions. The idea of the Dodd-Jensen core model K is that even if there are no measurable cardinals there may exist approximations to measurable cardinals. These approximations, called *mice*, are models $M = L_\alpha[U]$ such that

(1) $L_\alpha[U] \models U$ is a measurable cardinal on some ordinal $\kappa < \alpha$.
(2) All iterated ultrapowers of $L_\alpha[U]$ by U are well founded.
(3) $L_{\alpha+1}[U] \models |\alpha| = \rho$ for some $\rho < \kappa$.

The Dodd-Jensen core model K is equal to $L[\mathcal{M}]$, where $\mathcal{M}$ is the class of all mice. The measure U in a mouse $L_\alpha[U]$ is partial by necessity; clause (3) guarantees that U cannot be extended to a measure on any set extending $\mathcal{P}(\kappa) \cap L_\alpha[U]$. These partial measures can, however, be used in connection with cardinal properties smaller than measurable cardinals. Suppose, for example, that κ is a Ramsey cardinal in V. This is equivalent to the existence of a certain kind of ultrafilter on each field of subsets of κ of cardinality at most κ. This ultrafilter can be used to construct the ultrafilter U which is needed to define a mouse and the mice constructed

in this way can then be used to show that κ is a Ramsey cardinal in $\mathbf{K}$ (see [13]).

The simplest example of the Dodd-Jensen core model K occurs when $0^{\#}$ exists, but no class of indiscernibles exists for $L[0^{\#}]$. In this case K is equal to $L[0^{\#}]$. To see why this is true, recall that $0^{\#}$ holds if and only if there is an embedding $i\colon L \to L$ which is not the identity. If κ is the critical point of i then set $U = \{x \subset \kappa : \kappa \in i(x)\}$, so that U is the ultrafilter on $\mathcal{P}(\kappa) \cap L$ which is generated by the embedding i. Then $L[0^{\#}] = L[U]$. The ultrafilter U is not a measure on κ in $L[U]$; in fact if i is chosen so that its critical point κ is as small as possible then $L_{\kappa+1}[U]$ is a mouse.

An important fact about K is that the measures in the mice can be used to define a canonical well ordering of the class $\mathcal{M}$ of mice. To decide which of two mice $L_\alpha[U]$ and $L_{\alpha'}[U']$ is smaller, we take iterated ultrapowers

$$i : L_\alpha[U] \to L_\beta[U^*] \qquad \text{and} \qquad i' : L_{\alpha'}[U'] \to L_{\beta'}[U^*]$$

of each of them until the measures U^* in the two resulting structures agree. The lengths β and β' of the iterated ultrapowers then determine the order of the the mice. This well ordering of the mice, to which will return later in this paper, yields a L-like structure on the Dodd-Jensen core model K.

The model $L[\mu]$ has been generalized ([12], [15]) to an inner model $L[\mathcal{U}]$ for a sequence $\mathcal{U}$ of measures, and to the associated core model $K[\mathcal{U}]$. A key concept here is the notion of a *coherent sequence* of measures, which allows for models with measures concentrating on measurable cardinals. In this paper we will not discuss further the concepts of coherence or general sequences of measures, though the reader should be aware that they are important ingredients in the core model we will be studying.

CRITERIA FOR A CORE MODEL. Suppose that we are given a model: how do we recognize it as the core model? There is only a limited sense in which we can give an answer to this question — as we consider larger and larger cardinals the core models will look less and less like L, and a good part of the interest in the investigation of large cardinals, including core model theory, is in the discovery of these necessary differences. One possible way around this question is to simply assert that we will recognize the core model when we see it. This answer is not entirely frivolous: part of the strength of core model theory so far has been that the core models defined have been clearly and unambiguously recognizable as such. Nevertheless we can describe some properties of the known core models which we can

expect to continue to hold, at least for core models at the levels which we are currently considering.

The most basic characteristic of L which we would like to be preserved in future core models is its method of construction: it is built around the ordinals, which form its basic skeleton, and it is built up "from below" along this skeleton by pieces L_α which are defined by recursion along the ordinals using some very simple closure properties. One consequence of this construction is that the logical complexity of L comes entirely from that of the ordinals. Another consequence of this construction is the absoluteness of L: if M is any well founded model of set theory then $L^{(M)}$, the model L as constructed inside M, is equal to L_α where α is the order type of the ordinals of M. For the Dodd-Jensen core model K described above we have a similar situation, but in this case the central skeleton is given not by the well ordering of the ordinals but instead by the more complicated well ordering of the class $\mathcal{M}$ of mice. As with L the complexity of K comes entirely from the complexity of the well ordering of its skeleton. The well ordering of the reals in L, for example, is Δ^1_2. The reason for this is that if $\phi(E)$ is the formula asserting that (ω, E) is isomorphic to some countable initial segment $(L_\alpha, \in)$ then the only nonarithmetical part of ϕ is the assertion that the relation E is well founded, which is a Π^1_1 condition. The assertion that (ω, E) be isomorphic to some mouse $L_\alpha[U]$ involves the assertion that every iterated ultrapower of (ω, E) by its measure is well founded. This is a Π^1_2 condition, and hence the well ordering of the reals of K is Δ^1_3.

An important consequence of the method of construction of L is the *condensation principle:* if α is an ordinal and $M \cong H \prec_1 L_\alpha$ then $M = L_{\alpha'}$ for some ordinal $\alpha' \leq \alpha$. The condensation principle leads immediately to the proof that GCH and $\diamondsuit$ hold in L and is the basic fact behind the fine structure theory of L, including such applications as $\square$ and morasses. This principle is obscured in the normal description of $L[\mu]$, and is also obscured (though to a lesser degree) in the Dodd-Jensen core model K, but it is as critical to the theory of each of these models as it is to that of L. One of the recent advances in inner model theory [11] is a way of defining core models (including K and $L[\mu]$) so that condensation is (almost) literally true instead of being hidden in the machinery.

A second basic characteristic of the core model is *rigidity.* We mentioned that an elementary embedding $i : L \to L$ is the least large cardinal inconsistent with L, and similarly the existence of an nontrivial embedding $i : K \to M$ of the Dodd-Jensen core model implies that there is a model $L[\mu]$ with a measurable cardinal. This is one of two basic ideas for using the core model to prove the existence of inner models with large

cardinals. In general a basic condition for recognizing a model $\mathbf{K}$ as the correct core model up to some large cardinal property $\mathbf{P}(\kappa)$ is that the existence of a nontrivial embedding $i : \mathbf{K} \to \mathbf{K}$ is equivalent to the existence of a model with a cardinal κ satisfying $\mathbf{P}(\kappa)$. In the well established core models we have the stronger property that there is no nontrivial embedding $i : \mathbf{K} \to M$ for any well founded class M, but it is still open whether this stronger rigidity property will hold past a strong cardinal.

A third basic characteristic of the core model is the covering lemma, but it is difficult to use this as a criterion for a core model since it is not clear what we can hope for in an abstract covering lemma. Jensen's original covering lemma [1] for L asserted that if $0^{\#}$ does not exist — that is, if L is the true core model — then any set x of ordinals is contained in some set $y \in L$ such that $|y| \leq |x| + \aleph_1$. This same covering lemma holds for the Dodd-Jensen core model K, provided that there is no model with a measurable cardinal, and this is the second of the two basic ideas for using the core model to prove the existence of a model with a measurable cardinal. For the model $L[\mu]$ it becomes necessary to admit an exception: Prikry forcing over $L[\mu]$ will yield a ω-sequence C of indiscernibles over $L[\mu]$ which is cofinal in the measurable cardinal κ of $L[\mu]$, and this set C is not contained in any member of the core model $L[\mu]$ of cardinality smaller than κ. If there are longer sequences of measures this exception expands: the set C of indiscernibles for the single measure μ becomes a system of indiscernibles for the measures in the sequence, and each set x to be covered begins to require a different system of indiscernibles. There is, however, one corollary of the covering lemma which holds in all of these models. This corollary, known as the *weak covering lemma*, asserts that if λ is any singular strong limit cardinal then $\lambda^+ = \lambda^{+(\mathbf{K})}$. The weak covering lemma remains true up to a Woodin cardinal, but at that point either it or the immutability of $\mathbf{K}$ will also have to be sacrificed, as another exception appears: we will later discuss the *stationary tower forcing* at a Woodin cardinal κ which, among other things, will collapse λ^+ for an unbounded set of singular cardinals λ below κ.

While the statement of the covering lemma has been changing so dramatically, however, one thing has stayed relatively constant: the basic proof of the covering lemma. Suprisingly, the increasing awkwardness necessary to incorporate the exceptions such as Prikry sequences into the statement of the covering lemma is not reflected in the proof, which remains essentially the same except for some fairly difficult but straightforward adaptations to the complications of the larger models. The awkwardness comes from interpreting what it is that the proof actually proves. It is still open whether there will be a similarly straightforward adaption

to handle the latest exception, stationary tower forcing, but there are indications of how the proof may eventually take account of this case as well.

We will mention one final characteristic of core models: *correctness* (or absoluteness). If Γ is a class of formulas over the reals then we say that a model M is Γ-*correct* if for all formulas ϕ in Γ and all reals $r \in M$ we have $M \models \phi(r) \iff V \models \phi(r)$. The concept of correctness may not at first appear to be an extrapolation from L, but recall the statement of Shoenfield's Σ^1_2-absoluteness theorem [16]: Any model M containing L_{ω_1} is Σ^1_2-correct. The theorem is usually stated with ω_1 in place of L_{ω_1}, but as long as M is a model of a fragment of ZF the statements are equivalent. It has turned out that for stronger correctness theorems it is not enough to use some ordinal larger than ω_1, but instead it is necessary to look at the more complex well ordering of the mice. We have the following conjecture:

CONJECTURE *Assume that the sharp $a^\#$ of a exists for every real a (and probably something more), and let K^* be the core model for a Woodin cardinal (and slightly more) if there is an inner model with such a cardinal, and let K^* be the true core model* $\mathbf{K}$ *if there is no such model. Then any model M of set theory which contains an iterated ultrapower of K^* is Σ^1_3-correct.*

This is known to be true under the additional assumption that there is no inner model with a strong cardinal (or with more than a few strong cardinals), but the point is that the correctness theorem should be true, for this fixed model K^* which cannot contain more than one Woodin cardinal (plus a bit more), no matter what additional cardinal structure may exist in the universe.

It appears that correctness may be not only a pleasant application of core model theory, but also a prerequisite to the further development of core model theory. We will mention later how the iterability of a model — the property that every iterated ultrapower of the model is well founded — becomes much more complex as we get into larger cardinals. A major jump occurs at a Woodin cardinal, as is to be expected from the theorem of Martin and Steel that slightly more than a Woodin cardinal implies the axiom of determinacy for Σ^1_2 formulas and hence that there can be no Δ^1_3 well ordering of the reals. One scenario suggests that the Σ^1_2-correctness given by the conjecture will be necessary for a full theory of iterability for models with up to two Woodin cardinals. An alternate scenario, due to Woodin, suggests that the necessary correctness will come instead from descriptive set theorical considerations. This reflects the increasingly strong interconnections between the core model and descriptive set theory.

§ 2. Why is a Woodin cardinal interesting (and what is it)?

Woodin cardinals, like most large cardinal properties above measurable cardinals, are defined in terms of the existence of elementary embeddings $i\colon V \to M$ where M is a well founded model which is, in some sense, large. A cardinal κ is *measurable* provided that there is such an embedding $i\colon V \to M$ with κ equal to the critical point of i, the least cardinal α such that $i(\alpha) > \alpha$. Strong cardinals increase the constraints on M: κ is *λ-strong* if there is such an embedding with critical point κ such that $V_\lambda \subset M$ and κ is *strong* if κ is λ-strong for every λ. The definition of a Woodin cardinal κ uses embeddings with critical point smaller than κ:

DEFINITION. *A cardinal κ is a Woodin cardinal if for any function $f\colon \kappa \to \kappa$ there is an embedding $i\colon V \to M$ with critical point $\alpha < \kappa$ such that $f\text{“}\alpha \subset \alpha$ and $V_{i(f)(\alpha)} \subset M$.*

A cardinal is *λ-supercompact*, for $\lambda > \kappa$, if there is an embedding $i\colon V \to M$ with critical point κ such ${}^{\lambda}M \subset M$. For κ to be a Woodin cardinal is much weaker than for it to be even κ^+-supercompact.

The definition of a Woodin cardinal has been included to satisfy the reader's curiosity, but it was probably more puzzling than satisfying. As the title of this section suggests, we will rely on the consequences of a Woodin Cardinal rather than the definition to justify the claim that the extension of the core model to a Woodin cardinal is an important step in the progress of set theory.

The most important of these consequences involve the axiom of determinacy. Martin and Steel have shown [8] that a Woodin cardinal, plus slightly more, implies that the axiom of determinacy holds for $\mathbf{\Sigma}^1_2$ sets of reals, while Woodin has shown [10] that $\mathbf{\Sigma}^1_2$ determinacy implies that there is an inner model with a Woodin cardinal. Later, Woodin showed that the full axiom of determinacy is equivalent to the existence of a model with infinitely many Woodin cardinals.

Since the axiom of determinacy for $\mathbf{\Sigma}^1_2$ sets implies that every Σ^1_3 set of reals is Lebesgue measurable it follows that no model containing more than a Woodin cardinal can have a Δ^1_3 well ordering of the reals. Thus the minimal inner model for a Woodin cardinal is the the largest inner model having a Δ^1_3 well ordering of the reals, and in this respect it is just one step past L, which is the largest inner model which has a Δ^1_2 well ordering of the reals.

The second reason for the importance of a Woodin cardinal is the *stationary tower forcing* which was briefly referred to earlier. A condition of the stationary tower forcing on a cardinal κ is a pair (S, X), where $X \in V_\kappa$ and S is a stationary subset, in the appropriate sense, of X. A condition

(S', X') is stronger than (S, X) if $X' \supset X$ and $S' \cap X \subset S$. Thus a generic object G for the stationary tower forcing is a tower of ultrafilters on the power sets of the members X of V_κ, and it is possible to use this tower to form a generalized ultrapower $i^G \colon V \to \mathrm{ult}(V, G)$. It can be shown that if κ is a Woodin cardinal then $V_\kappa^{(\mathrm{ult}(V,G))} = V_\kappa^{(V[G])}$, and in particular κ is still a cardinal in $V[G]$.

Now suppose that $\lambda < \kappa$ is a cardinal and $\nu \leq \lambda$ is a regular cardinal, and consider the condition (S_ν, H_{λ^+}), where H_{λ^+} is the set of sets hereditarily of cardinality less than λ^+ and

$$S_\nu = \{M \prec H_{\lambda^+} : \lambda \subset M \text{ and } \mathrm{cf}(\mathrm{ordinals}(M)) = \nu\}.$$

This condition will force that λ^+ is the critical point of i^G, so that λ^+ is collapsed in $\mathrm{ult}(V, G)$ and hence in $V[G]$. On the other hand λ remains a cardinal in $\mathrm{ult}(V, G)$, and since $V_\kappa^{(\mathrm{ult}(V,G))} = V_\kappa^{(V[G])}$ it remains a cardinal in $V[G]$ as well. Furthermore $\lambda^{+(V)}$ has cofinality ν in $\mathrm{ult}(V, G)$ and hence in $V[G]$. Notice that no assumptions were imposed on the cardinal λ: it could be choosen to be $\aleph_0$, or $\aleph_\omega$, or a measurable cardinal. Thus we can collapse the successor of a singular cardinal while giving it any cofinality we wish, or we can collapse the successor of a measurable cardinal while keeping the cardinal measurable. Neither of these is possible below a Woodin cardinal.

There is one further, rather startling, consequence of a Woodin cardinal which should be mentioned. Woodin (unpublished) has shown that if $L[\mathcal{E}]$ is the minimal model of a Woodin cardinal κ then there is a notion P of forcing in $L[\mathcal{E}]$, having the κ-chain condition, such that if X is any set whatsoever, taken from the universe V, then there is an iterated ultrapower $i \colon L[\mathcal{E}] \to L[\mathcal{E}']$ such that X is $i(P)$-generic over $L[\mathcal{E}']$. Of course i will not in general be a member of $L[\mathcal{E}]$ (and in fact slightly more than a model of a Woodin cardinal is needed to prove that this iteration can take place) but the point is that X could be any set whatsoever: for example, X could code a class model containing a supercompact or a huge cardinal, or even the sharp of a minimal model for a Woodin cardinal.

§ 3. How is the model constructed?

The definition of a Woodin cardinal, or even a strong cardinal, requires the existence of embeddings $i \colon V \to M$ with $V_\lambda \subset M$ for ordinals λ larger than the critical point κ of i, and hence we will need to have a way to represent such embeddings. A measurable cardinal κ is traditionally given by a κ-complete, normal ultrafilter U on κ. The canonical embedding

$i^U: V \to M = \text{ult}(V, U)$ using such an ultrafilter has the property that

$$M = \{i^U(f)(\kappa) : f \in V \cap {}^{\kappa}V\}. \tag{1}$$

The embeddings for cardinals stronger than measurable cardinals are given by *extenders*. A (κ, λ)-extender E, with $\kappa + 1 \leq \lambda$, has an embedding $i^E: V \to M = \text{ult}(V, E)$ with the property that

$$M = \{i^E(f)(\mathbf{a}) : f \in V \wedge \exists n < \omega\, (f: [\kappa]^n \to V \wedge \mathbf{a} \in [\lambda]^n)\}, \tag{2}$$

which is essentially formula (1) with the ordinal κ replaced by the interval $[\kappa, \lambda)$. This may look unwieldy, but it is made workable by realizing that, just as the embedding i^U defines the ultrafilter U by setting $U = \{x : \kappa \in i^U(x)\}$, an embedding i^E with property (2) defines a sequence of ultrafilters $E_{\mathbf{a}} = \{x \subset \kappa^n : \mathbf{a} \in i^E(x)\}$. The extender E itself is defined to be the sequence $(E_{\mathbf{a}} : \mathbf{a} \in [\lambda]^{<\omega})$ of all these ultrafilters. Notice in particular that a measure on κ may be regarded as a $(\kappa, \kappa + 1)$-extender.

There are three basic problems connected with the movement from ultrafilters to more general extenders, all of which arise from the problem of *overlapping extenders*. The models we are looking at have the form $L[\mathcal{E}]$ where each member $\mathcal{E}_\nu$ of the sequence $\mathcal{E}$ is a $(\kappa_\nu, \lambda_\nu)$-extender for some pair $(\kappa_\nu, \lambda_\nu)$ of ordinals. Now the length λ_ν of the extenders is almost a nondecreasing sequence, but once we pass a strong cardinal there are essentially no restraints on the behavior of the critical points κ_ν of the extenders. Thus there are ordinals α such that $\kappa_\nu < \alpha$, and hence $i^{\mathcal{E}_\nu}(\alpha) \neq \alpha$, for cofinally many extenders $\mathcal{E}_\nu$ in the sequence $\mathcal{E}$. This contrasts with the behavior of a model $L[\mathcal{U}]$, where $\mathcal{U}$ is a sequence of measures. Here each $\mathcal{U}_\nu$ is a measure on a cardinal κ_ν and the sequence κ_ν is nondecreasing, so that if $\nu > \alpha^{++}$ then $\kappa_\nu > \alpha$ and hence α is not moved by $i^{\mathcal{U}_\nu}$.

The first problem in its most serious form is connected with the fact that we may have two extenders E_0 and E_1 which overlap in the sense that E_0 is a (κ_0, λ_0)-extender and E_1 is a (κ_1, λ_1)-extender, with $\kappa_0 \leq \kappa_1 < \lambda_0$. To see why this is a problem, let us look briefly at iterated ultrapowers as they come up in the theory of inner models for sequences of measures. A typical two stage iteration would look like

$$i : M_0 \xrightarrow{i_0} M_1 = \text{ult}(M_0, U_0) \xrightarrow{i_1} M_2 = \text{ult}(M_1, U_1) \tag{3}$$

where $\kappa_0 < \kappa_1$ and U_i is an ultrafilter on κ_i in M_i for $i = 0, 1$. By property (1) above, if $x \in \mathcal{P}(\kappa_0) \cap M_0$ then $x \in U_0 \iff \kappa_0 \in i_0(x)$, and since $\kappa_1 > \kappa_0$ we have $i_1(\kappa_0) = \kappa_0$ and hence $x \in U_0 \iff \kappa_0 \in i(x)$.

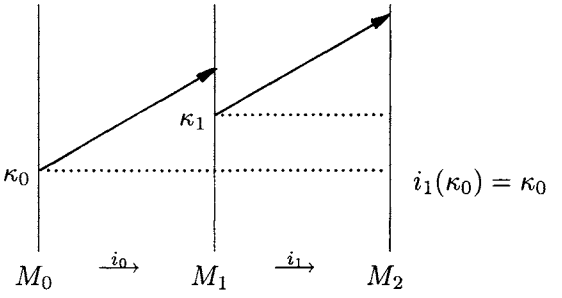


Figure 1

Since κ_0 is never moved after the first embedding i_0, we can keep track of U_0 by keeping track of the ordinal κ_0. This situation is illustrated in figure 1.

Now consider the same situation with the measures U_0 and U_1 replaced by overlapping extenders E_0 and E_1, so that the embedding (3) becomes

$$(4) \qquad i : M_0 \xrightarrow{i_0} M_1 = \mathrm{ult}(M_0, E_0) \xrightarrow{i_0} M_2 = \mathrm{ult}(M_1, E_1).$$

Now we want to use property (2) to keep track of E_0, but this will require keeping track of the whole interval $[\kappa_0, \lambda_0)$ instead of just the ordinal κ_0. Unfortunately $\kappa_1 < \lambda_0$, and as a consequence i_1 is not the identity on this interval, as is illustrated by figure 2 where the generators of E_0, indicated by the thicker line, are broken up by the embedding i_1.

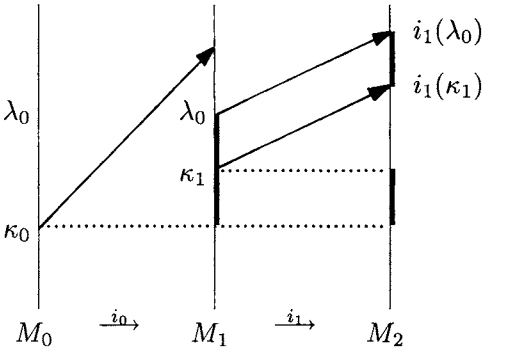
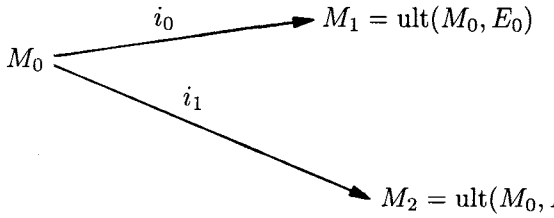


Figure 2

The theory breaks down because of the difficulty of keeping track of these moving generators. Fortunately there is a way out, namely *iteration*

trees. The iteration (4) is restructured to become a tree



where the extender E_1 is applied to M_0 even though it is a member of, and would thus naturally be applied to, M_1. In a more general iterations of length α this leads to a tree T with nodes $(M_\nu : \nu < \alpha)$, where there is an embedding from $M_{\nu'}$ to M_ν just in case M_ν is below $M_{\nu'}$ in the tree T. The tree T may branch any number of times at any node, and may have arbitrarily large height.

Now suppose that α is a limit ordinal and the iteration tree T has been defined with nodes $(M_\nu : \nu < \alpha)$. In the case where there only ultrafilters this tree never splits, so that the tree is a well ordering and the limit M_α of the iterated ultrapower is just the direct limit of the sets M_ν for $\nu < \alpha$. In this case the only question is whether M_α is well founded. With a tree iteration, there is more of a problem. In the first place, there may or may not even be a cofinal branch — the tree could conceivably consist of a root node and infinitely many branches each of length one. If it does have cofinal branches then we can take the direct limit M_b along any cofinal branch b of T, and the models M_b may or may not be well founded. In order for M to be iterable we will need to have an *iteration strategy*, that is, a function σ on iteration trees such that $\sigma(T)$ is always a cofinal branch b of T such that M_b is well founded, provided that T is a tree that was formed by following σ at all previous limit points.

If there is no model with a Woodin cardinal then this iteration strategy has a simple description: A theorem of Martin and Steel [9] asserts that if there are no Woodin cardinals then no tree has more than one cofinal branch with a well founded limit, so that the only possible iteration strategy is to always pick the unique well founded branch and the condition for iterability is that every iteration tree have at least one well founded branch. This would seem to be a Π^1_3 condition, but Martin and Steel also show that if T is a countable tree with no well founded branch then there is a witness to this ill foundedness which is continuous in T, so that the condition is Π^1_2. Recall that the well ordering of the reals of L is Δ^1_2 because of the fact that "E is well founded" is Π^1_1, while the ordering of the reals in the Dodd-Jensen core model is Δ^1_3 because the assertion that every countable iterated ultrapower of a countable model $M = (\omega, E)$ is

well founded is Π^1_2. Thus the well ordering of the reals of the core model is still Δ^1_3 up through a Woodin cardinal.

As soon as we pass a Woodin cardinal we will have trees with more than one well founded branch b, but Martin and Steel prove that below two Woodin cardinals there is at most one of these branches such that the direct limit M_b has the additional property that every iteration tree starting anew from M_b has a well founded branch. The iteration strategy then calls for choosing the branch with this stronger property, and the iterability condition on M is that there always is such a branch. This is a Π^1_3 condition, and thus the well ordering of the reals is now Δ^1_4.

There remains the problem of showing that there exists some sequence $\mathcal{E}$ such that every iteration tree on $L[\mathcal{E}]$ has a well founded branch. With the unbranching trees which come from using only ultrafilters this problem of *iterability* has a simple solution: If all of the ultrafilters in $\mathcal{E}$ are countably complete then every iterated ultrapower will be well founded. Steel's solution for iteration trees — which so far is only a partial solution — is given by *background embeddings*. For each (κ, λ)-extender E in the inner model M we require that there be a background embedding $i^* : V \to M^*$ in the real world which extends E in the sense that $E_{\mathbf{a}} = \{x \in \mathcal{P}(\kappa^n) \cap M : \mathbf{a} \in i^*(x)\}$ for each $\mathbf{a} \in [\lambda]^n$, but which is stronger than the embedding i^E. The appropriate notion of "stronger" has changed with the development of the theory. When Steel first introduced background extenders he required that the embedding i^* map a rank into a rank, so that much more than huge cardinals were needed in order to get an inner model with a Woodin cardinal. Martin and Steel then reduced this requirement [9] so that an inner model with a Woodin cardinal could be constructed using only a Woodin cardinal in the universe, and finally Steel [19] further weakened the requirements on the background extenders so that they can be used, with restrictions, to construct a core model.

Background extenders, while useful, present a problem: it is not presently known how to get the extenders without assuming some sort of extra large cardinal structure in the universe. Steel's original notes assumed that there was a complete ultrafilter on the definable subsets of the class of ordinals. This condition can apparently be weakened to a strong form of ineffability, but it is not know whether it can be eliminated. The difficulty that this causes in applications is illustrated by a theorem of Steel asserting that if there is a ω_2-saturated ideal on ω_1 together with a measurable cardinal then there is an inner model with a Woodin cardinal. The measurable cardinal is needed in order to provide the extra large cardinal structure needed for the core model. The conclusion probably follows from a saturated ideal alone, but no proof is known.

The other two problems arising from the movement to extenders come from the fact that the sequence $\mathcal{E}$ of extenders may have cofinally many extenders $\mathcal{E}_\nu$ with critical point below some fixed ordinal α. Recall that a mouse in the Dodd-Jensen core model is a set $\mathfrak{m} = L_\alpha[U]$, where $\mathfrak{m} \models$ "U is a measure on κ" and $L_{\alpha+1}[U] \models |\alpha| = \rho < \kappa$. Thus there is a subset $x_\mathfrak{m}$ of ρ which is definable in, but not a member of, $L_\alpha[U]$. To see the critical role this set $x_\mathfrak{m}$ plays in the theory, we recall the definition of the well ordering of the mice. Suppose that $\mathfrak{m} = L_\alpha[U]$ and $\mathfrak{m}' = L_{\alpha'}[U']$ are two mice, each of which adds a new subset of an ordinal $\rho < \min(\kappa, \kappa')$. To see which is larger we form iterated ultrapowers

$$L_\alpha[U] \xrightarrow{i} L_\nu[U^*]$$
$$L_{\alpha'}[U'] \xrightarrow{i'} L_{\nu'}[U^*]$$

so that $i(U) = i'(U') = U^*$ agree. Since $\rho < \min(\kappa, \kappa')$ both of the embeddings i and i' are the identity on ρ and hence the sets $x_\mathfrak{m}$ and $x_{\mathfrak{m}'}$ are definable in, but not members of, $L_\nu[U^*]$ and $L_{\nu'}[U^*]$ respectively. Suppose that $\nu < \nu'$. Then $x_\mathfrak{m} \in L_{\nu+1}[U^*] \subset L_{\nu'}[U^*]$ and hence $x_\mathfrak{m} \in \mathfrak{m}' = L_{\alpha'}[U]$ so that $\mathfrak{m}'$ is larger than $\mathfrak{m}$ in the well ordering of the mice. If the mice $\mathfrak{m}$ and μ' contain extenders then this no longer works, since the mice $\mathfrak{m}$ and $\mathfrak{m}'$ may have cofinally many extenders $\mathcal{E}_\nu$ with critical point smaller than ρ, and hence the embeddings i and i' need not be the identity on ρ. There is a way around this: we build into the theory of iteration trees the condition that whenever this situation arises we will apply the offending extender to the core model $\mathbf{K}$ rather than to the current iterate of the mouse $\mathfrak{m}$. The weak covering lemma referred to earlier, asserting that $\lambda^+ = \lambda^{+(\mathbf{K})}$ for singular cardinals λ, can then be used to substitute for the use of the set $x_\mathfrak{m}$ above, and in fact the argument shows that this situation never occurs on the unique well founded branch of the final tree. Unfortunately this approach is circular — we need to be able to compare mice long before we can prove the covering lemma. Steel avoided this problem by again using the extra large cardinal structure in the universe to show that the weak covering lemma holds on a stationary class even without the covering lemma. This allows the core model theory to go through, eventually proving a form of the covering lemma which implies that the weak covering lemma is true everywhere. Again, this trick comes at a cost: it only works with the assumption that we have this extra large cardinal structure.

The final problem is more subtle. In order to see where the problem arises we will consider the core model for two measures, which has the form $L[U_1, U_2, \mathcal{M}]$ where U_1 and U_2 are the two measures and $\mathcal{M}$ is the class of mice. The measures U_1 and U_2 are chosen in sequence, so that

nothing is known about U_2 at the time U_1 is chosen. Thus some care is required in picking U_1 to ensure that it will remain iterable after U_2 is chosen. There is no such problem with the mice. To see why this is so it is enough to recall the Dodd-Jensen mice below one measure, which have the form $\mathfrak{m} = L_\alpha[W]$ where W is a measure on a cardinal κ in $L_\alpha[W]$ but $|\alpha| < \kappa$ in $L_{\alpha+1}[W]$. There is no need to worry about the well foundedness of ultrapowers of the core model by W, since W is not a measure even in $L_{\alpha+1}[W]$.

With extenders this situation can change. An extender which first appears in a mouse $\mathfrak{m}$ may have critical point smaller than the projectum ρ of $\mathfrak{m}$, so that it may survive to be an extender in the final core model. Thus the same care has to be taken in adding mice to the core model as is necessary in adding extenders.

This section has concentrated on problems arising in dealing with larger embeddings, and seems to suggest that what was a difficult construction in the original core models has become exponentially more complicated. It is true that the models have become more complicated, but we conclude this section on a more positive note, outlining one development which has lead to a substantial simplifications even in the older cases where there were no extenders, only measures.

The original construction of the core model for sequences of measures [14] involved three different sorts of models. First there were ordinary inner models for sequences of measures, which had the form $L[\mathcal{U}]$ where $\mathcal{U}$ is a sequence of measures. Next were the mice, which had the form $L_\alpha[\mathcal{U} \restriction \rho, \mathcal{W}]$ where $\mathcal{W}$ was a sequence of measures peculiar to the mouse, and finally there was the core model itself, which had the form $L[\mathcal{U}, \mathcal{M}]$ where $\mathcal{M}$ was the class of mice. Some ten years ago I proposed, in an attempt at a solution of the first two problems described above, a first approximation to the concept of iteration trees together with the suggestion that all three models should have the same form. In particular a mouse should look like the core model, having the form $\mathfrak{m} = L_\alpha[\mathcal{U} \restriction \alpha, \mathcal{M} \restriction \alpha]$ where $\mathcal{M} \restriction \alpha$ is the set of all mice smaller than $\mathfrak{m}$. I dropped work on this approach, partly because I was unable to solve the iterability problem but also because this approach seemed likely to become excessively complicated. Recently Stu Baldwin mentioned to me that he had been thinking of a approach which, on examination, turned out to be essentially the same as mine. He remarked that this approach made things much simpler, so I went back and looked at it again and I found that he was right: this approach not only makes a good inner model theory for extenders possible, but it makes the core model theory much simpler even for well understood models such as $L[\mu]$ and the Dodd-Jensen core model.

The proper approach turns out to be to code the class $\mathcal{M}$ of mice by the extenders appearing in the mice. Thus the core model $\mathbf{K}$ has the form $L[\mathcal{E}]$ where $\mathcal{E}$ is a sequence containing both full extenders in $L[\mathcal{E}]$ and partial extenders which are extenders only in mice from $L[\mathcal{E}]$. The actual condition on the sequence $\mathcal{E}$ is that each member $\mathcal{E}_\nu$ of the sequence is an extender only on sets in $L[\mathcal{E} \restriction \nu]$. The extender $\mathcal{E}_\nu$ may or may not survive to be an extender in $L[\mathcal{E}]$; this depends on whether there is a subset x of the critical point κ_ν of $\mathcal{E}_\nu$ which is in $L[\mathcal{E}]$ but is not in $L[\mathcal{E} \restriction \nu]$. The mice in this approach turn out to be simply the initial segments $L_\alpha[\mathcal{E} \restriction \alpha]$ of the full core model $L[\mathcal{E}]$. Fine structure can be defined in this model word for word the same (except for the need for a Σ_0 code) as it is defined in L; and condensation holds exactly[1] as in L, at least for substructures of $L_\alpha(\mathcal{E})$ of the form $\mathcal{H}^{L_\alpha(\mathcal{E})}(\rho \cup p)$ where ρ is an ordinal, p is finite, and $\mathcal{H}(\rho \cup p)$ is the Skolem hull.

§ 4. What is to be done?

I will conclude this paper with a list of three conjectures, some of which are repeated from earlier in this paper. These conjectures will illustrate the depths and gaps of our current understanding of the core model. In general, the depths become shallower and the gaps deeper as the size of the cardinals increase: We have a very good understanding of the core model up to a strong cardinal, and with the results outlined in this paper we have a good understanding, with serious gaps, up to a Woodin cardinal. Steel has looked at the theory for larger cardinals: there are problems at the levels of a Woodin cardinal and of infinitely many Woodin cardinals, and there are major difficulties at the level of a measurable limit of Woodin cardinals. Beyond this point very little is known, and much of what is known consists of results of Woodin and others which indicate that some of the properties which we expect to hold of a core model will eventually have to fail. Studies of inner models for these larger cardinals will provide a fertile field for research, but most of the questions listed below concentrate on the better understood area discussed in this paper.

CONJECTURE 1. *Assume that the sharp $a^\#$ exists for every real a (and possibly something more), and let K^* be either the core model for a Woodin cardinal (and slightly more), or the true core model $\mathbf{K}$ if there is no model with this much large cardinal structure. Then any model M of set theory which contains an iterated ultrapower of K^* is Σ^1_3-correct.*

[1] Actually there is one remaining technicality, although Sy Friedman has suggested a way in which this technicality may be avoided.

Recall that this is currently known to be true up to a strong cardinal and slightly beyond. A solution to this problem would both materially advance our understanding of the core model below a Woodin cardinal and provide a valuable tool for the understanding of the core model above a Woodin cardinal.

CONJECTURE 2. *If there is a ω_2-saturated, countably complete ideal on ω_1 then there is an inner model with a Woodin cardinal.*

It will be recalled that Steel has proved this with the additional hypothesis of a measurable cardinal. The problem here is to eliminate the need for the extra large cardinal structure in the construction of the core model, perhaps by giving a proof of the covering lemma which does not presuppose the weak covering lemma and does not require background extenders stronger than given by the proof of the covering lemma itself. It is possible that the extra structure is in fact essential to a full core model theory, in which case the behavior in the important special cases in which there is no such structure should be very interesting.

CONJECTURE 3. *Suppose that λ is a singular cardinal and $\lambda^+ \neq \lambda^{+(\mathbf{K})}$. Then there is a model M of set theory with a Woodin cardinal and a set G such that G is generic over M for some variant of the stationary tower forcing and $\lambda^+ = \lambda^{+(M[G])}$.*

The question here is how stationary tower forcing, the next exception to the covering lemma, is to be handled. It is already known (at least with the extra large cardinal structure required for the core model) that a Woodin cardinal is required to collapse the successor of a singular cardinal. It should be noted that there is an additional question hidden here: exactly what happens if there is inner model with a Woodin cardinal, but nothing more. For technical reasons it appears likely that there may not be a core model, at least in the traditional sense, in this situation.

I will conclude with one more conjecture, which is related to conjecture (3) but goes far beyond the known core models. The second condition was suggested by Woodin.

CONJECTURE 4. *Suppose that M is a model of set theory, $M[G]$ is set generic over M, and λ is a singular cardinal of $M[G]$ such that $\lambda^{+(M)}$ is collapsed in $M[G]$ and either (i) λ is the only such cardinal or (ii) $\lambda^{+(M[G])}$ is accessible in M. Then there is a model with a κ^+-supercompact cardinal κ.*

The results outlined in this paper show that large cardinals, which once seemed an amorphous contrast to the order of the constructible sets, can

in fact be gathered into a similarly rigid and powerful structure, and the problems listed in this section show that there is yet much to be done. I expect this area to continue to be an exciting and fruitful field in the foreseeable future.

REFERENCES

1. K. DEVLIN and R. JENSEN, *Marginalia to a Theorem of Silver*, ISILC Logic Conference (Kiel 1974), Lecture Notes in Mathematics 499, Springer-Verlag, Berlin and New York, 1975, pp. 115–142.
2. A. DODD, *The Core Model*, London Math. Soc. Lecture Notes 61, Cambridge University Press, Cambridge, 1982.
3. A. DODD and R. JENSEN, *The Core Model*, Annals of Mathematical Logic 20 (1981), 43–75.
4. A. DODD and R. JENSEN, *The Covering Lemma for K*, Annals of Mathematical Logic 22 (1982), 1–30.
5. A. DODD and R. JENSEN, *The Covering Lemma for $L[U]$*, Annals of Mathematical Logic 22 (1982), 127–135.
6. M. FOREMAN, M. MAGIDOR and S. SHELAH, *Martin's Maximum, Saturated Ideals, and Non-Regular Ultrafilters, I*, Annals of Mathematics (2nd Series) 127 (1988), 1–47; *Correction to Martin's Maximum, Saturated Ideals, and Non-Regular Ultrafilters, I*, Annals of Mathematics (2nd Series) 127 (1988), 521–545.
7. K. KUNEN, *Some Applications of Iterated Ultrapowers in Set Theory*, Ann. Math Logic 1 (1970), 179–227.
8. D. A. MARTIN and J. STEEL, *A Proof of Projective Determinacy*, Jour. of the AMS 2 (1989), 71–125.
9. D. A. MARTIN and J. STEEL, *Iteration Trees*, submitted to the Journal of the American Mathematical Society.
10. A. MATHIAS, R. SOLOVAY and H. WOODIN, *The Consistency Strength of the Axiom of Determinacy*, (in preparation).
11. W. MITCHELL and J. STEEL, *Fine Structure and Iteration Trees*, submitted to Assoc. for Symbolic Logic Lecture Notes in Logic (1990).
12. W. MITCHELL, *Sets Constructible from Sequences of Ultrafilters*, Jour. of Symbolic Logic 39 (1974), 57–66.
13. W. MITCHELL, *Ramsey Cardinals and Constructibility*, Journal of Symbolic Logic 44 (1979), 260–266.
14. W. MITCHELL, *The Core Model for Sequences of Measures, I*, Math. Proc. of the Cambridge Philosophical Society 95 (1984), 41–58.
15. W. MITCHELL, *Sets Constructible from Sequences of Measures: Revisited*, Jour. of Symbolic Logic 48 (1983), 600–609.

16. J. R. SHOENFIELD, *The Problem of Predicativity*, Essays on the Foundations of Mathematics (Y. Bar-Hillel, E. I. J Poznanski, M. O. Rabin and A. Robinson, eds.), The Magnes Press, Jerusalem, 1961, pp. 132–142.

17. J. SILVER, *The Consistency of GCH with the Existence of a Measurable Cardinal*, Axiomatic Set Theory, part 1, Proc. of Symposia in Pure Mathematics, vol. 13, AMS, Providence, 1971, pp. 391–395.

18. J. SILVER, *Measurable Cardinals and* $\mathbf{\Delta}^1_3$ *Well-orderings*, Ann. of Math 94 (1971), 414–446.

19. J. STEEL, *The Core Model Iteration Problem*, (preprint) (1990).

Logic, Methodology and Philosophy of Science IX
D. Prawitz, B. Skyrms and D. Westerståhl (Editors)

LATTICE EMBEDDINGS INTO THE R.E. DEGREES PRESERVING 1

KLAUS AMBOS-SPIES, STEFFEN LEMPP, MANUEL LERMAN

Math. Inst., Universität Heidelberg, D-6900 Heidelberg, Germany
Dept. of Math., University of Wisconsin, Madison, WI 53706, USA
Dept. of Math., University of Connecticut, Storrs, CT 06269, USA

1. Introduction

The characterization of the finite lattices embeddable into the recursively enumerable (r.e.) degrees (possibly with various additional restrictions, such as preserving the least and/or greatest element) is important to recursion theorists for two reasons: On the one hand, it gives insight into the (very complicated) structure of the r.e. degrees. On the other hand, it constitutes a crucial step in determining the decidability of the universal-existential theory of the partial ordering of the r.e. degrees and of the existential theory of the r.e. degrees in the language of lattices (where meet is a ternary relation), possibly with constant symbols for the least and/or greatest element.

Unfortunately, even though substantial progress has been made, the full characterization of the lattices embeddable into the r.e. degrees remains open. Work by Lachlan, Lerman, Thomason, Yates, and others [6,13,14] led to a proof of the embeddability of all countable distributive lattices into the r.e. degrees, while Lachlan [7] showed the embeddability of the two nondistributive five-element lattices, M_5 and N_5. Hopes that all finite lattices might embed into the r.e. degrees were dashed by Lachlan and Soare [9], who exhibited the counterexample S_8. The latest word on lattice embeddings into the r.e. degrees is Ambos-Spies and Lerman [3,4], who isolate sufficient conditions (for both embeddability and nonembeddability). It is not known whether these conditions are complementary.

This research was partially supported by the Mathematical Sciences Research Institute (where all three authors stayed in the spring of 1990), the Deutsche Forschungsgemeinschaft (for the first author), and the last two authors' NSF grants DMS-8901529 and DMS-8900349.

All known lattice embeddings into the r.e. degrees preserve the least element, 0. Preserving the greatest element, 1, turned out to be quite a bit harder. Lachlan [8], and independently Shoenfield and Soare [10], showed the embeddability of the diamond lattice preserving 1, and Ambos-Spies [1] extended this proof to all countable distributive and some nondistributive lattices (the latter all generalizations of N_5).

Here, we show the embeddability of M_5 into the r.e. degrees preserving 1, which is harder since the usual proof for embedding requires infinitary traces. We also reprove the embeddability of N_5 published in Ambos-Spies's thesis [1] but not elsewhere.

Our notation is standard and generally follows Soare [12] with two exceptions. Here, the *use* of a computation $\Phi^X(y)$ is the largest number *actually* used in the computation and is denoted by $\varphi(y)$ (and similarly for other Greek letters). If the oracle is given as the join of two sets then we assume the use function to give the use separately for each set of the join, thus $\Phi^{(X\oplus Y)\upharpoonright(\varphi(z)+1)}(z)$ is the same as $\Phi^{X\upharpoonright(\varphi(z)+1)\oplus Y\upharpoonright(\varphi(z)+1)}(z)$.

2. The theorems

We consider embeddings of two lattices, M_5 and N_5. Both have five elements (including the least element, 0, and the greatest element, 1) and are nondistributive. M_5 is a modular lattice and contains three pairwise incomparable elements while N_5 is a nonmodular lattice and contains two comparable elements both of which are incomparable to a third element. An embedding of a lattice into the r.e. degrees is said to *preserve the greatest element*, 1, if the image of 1 under the embedding is the complete r.e. degree $\mathbf{0}'$.

The purpose of this paper is to give the proofs of the following two theorems:

THEOREM 1. *The modular nondistributive five-element lattice, M_5, can be embedded into the r.e. degrees preserving the greatest element.*

THEOREM 2. *The nonmodular nondistributive five-element lattice, N_5, can be embedded into the r.e. degrees preserving the greatest element.*

(In [1], Ambos-Spies also shows the embeddability of several other lattices (similar to N_5) preserving 1.)

The proofs of the two above theorems are fairly unrelated. We begin with the first and more complicated proof.

3. The requirements and the intuition for M_5

We need to construct three incomplete r.e. sets A_0, A_1, and A_2 and an r.e. set B such that any two of the degrees of A_0, A_1, and A_2 join to $\mathbf{0}'$ and meet to the degree of B. We thus also build partial recursive (p.r.) functionals Γ_0, Γ_1, and Γ_2 and infinitely many p.r. functionals Δ and p.r. functions Λ (of which we suppress the indices), and we ensure the following requirements:

$$\begin{aligned}
\mathcal{S}_i &: B \leq_{\mathrm{T}} A_i \text{ (for } i < 3),\\
\mathcal{P}_i &: \Gamma_i^{A_j \oplus A_k} = K \text{ (where } \{i,j,k\} = \{0,1,2\} \text{ and } j < k),\\
\mathcal{M}_{j,k}^{\Phi} &: \Phi^{A_j} = \Phi^{A_k} \text{ total} \implies \exists \Delta (\Delta^B = \Phi^{A_j})\\
&\quad \text{(for } j < k < 3, \text{ all p.r. functionals } \Phi), \text{ and}\\
\mathcal{N}_i^{\Psi} &: \Psi^{A_i} = K \implies \exists \Lambda (\Lambda = K) \text{ (for } i < 3, \text{ all p.r. functionals } \Psi).
\end{aligned}$$

(Here K is the complete r.e. set of the halting problem. Notice that we assume Posner's trick (see Soare [12]) for the $\mathcal{M}$-requirements, so we can assume the same p.r. functional Φ for both A_j and A_k.)

The global requirements $\mathcal{S}_i$ are easily met by putting all numbers entering B also into all A_i so as to ensure $B = A_i \cap R_i$ for recursive sets R_i.

The global requirements $\mathcal{P}_i$ are met by ensuring that the functionals $\Gamma_i^{A_j \oplus A_k}$ are total and correctly compute K. (The hard part here will be totality.)

For the local requirements $\mathcal{M}_{j,k}^{\Phi}$, we use Fejer's strategy [5]. Whenever $\Delta^B(x)$ is defined but equal neither to $\Phi^{A_j}(x)$ nor to $\Phi^{A_k}(x)$ then that strategy puts a number $y \leq \delta(x)$ into B to allow the correction of $\Delta^B(x)$.

For the local requirements $\mathcal{N}_i^{\Psi}$, the problem in meeting $\Psi^{A_0} \neq K$ (setting $i = 0$ to simplify notation) is that protecting computations $\Psi^{A_0}(n)$ for the Sacks preservation strategy conflicts with higher-priority $\mathcal{P}_j$- (and $\mathcal{M}_{j,k}^{\Phi}$-) requirements putting numbers into A_0 (either directly or via B). The usual way to resolve this conflict with $\mathcal{P}_j$ is to fix a number y_0 (independent of n) and to "lift" uses $\gamma_j(y_0) \leq \psi(n)$ by enumerating $\gamma_j(y_0)$ into A_k (for $k \neq 0$) so that $\gamma_j(y_0) > \psi(n)$ can be achieved without having injured $\Psi^{A_0}(n)$. But in order to "lift" all three $\gamma_j(y_0)$ (for $j = 0,1,2$), we need to put numbers into at least two sets, namely A_1 and A_2 (since $\Psi^{A_0}(n)$ must not be injured). If we put numbers into A_1 and A_2 simultaneously, this may injure a higher-priority $\mathcal{M}_{1,2}^{\Phi}$-requirement and cause it to destroy $\Psi^{A_0}(n)$ through the correction process. So we have to put a number into A_1 first, wait for Φ^{A_1} to recover, and then put a number

into A_2. While we wait for Φ^{A_1} to recover, $\Psi^{A_0}(n)$ is still unprotected and thus may be destroyed before we can put a number into A_2. If this pattern repeats infinitely often then $\Psi^{A_0}(n)$ is undefined but also $\gamma_0(y_0)$ and $\gamma_2(y_0)$ enter A_1 infinitely often, so $\Gamma_0^{A_1 \oplus A_2}(y_0)$ and $\Gamma_2^{A_0 \oplus A_1}(y_0)$ are undefined, injuring our highest-priority requirement.

We use a trick first used by Ambos-Spies, Lachlan, and Soare in their refutation of the existence of a minimal cupping pair of r.e. degrees [2]. It consists in not using $y = y_0$ at first but some $y = y_1 > y_0$, and then repeating the procedure for $y = y_1 - 1, y_1 - 2, \ldots, y_0$. We will be able to show that once we have reached $y < y_1$, only a $K \upharpoonright (y+1)$-change can cause the destruction of $\Psi^{A_0}(n)$.

The full strategy σ for an $\mathcal{N}_0^{\Psi}$-requirement thus proceeds intuitively as follows (for a fixed number y_0):

(1) Fix y_1 "big", set $n = 0$.
(2) Wait for $\Psi^{A_0} \upharpoonright (n+1) = K \upharpoonright (n+1)$.
(3) For $y = y_1, y_1 - 1, \ldots, y_0 + 1, y_0$, proceed as follows:

(a) Put $\gamma_0(y)$ and $\gamma_2(y)$ into A_1; if $y < y_1$ then also put $\gamma_1(y+1)$ into A_0. Wait for all higher-priority $\mathcal{M}$-strategies to recover.
(b) Put $\gamma_0(y)$ and $\gamma_1(y)$ into A_2, and put $\gamma_2(y)$ into A_0. If $y > y_0$ then wait for all higher-priority $\mathcal{M}$-strategies to recover.

(4) Define $\Lambda_\sigma(n) = K(n)$, increment n by $+1$, and go to 2.

Our strategy assumes that $K \upharpoonright y_0$ will no longer change; so whenever $K \upharpoonright y_0$ does change, we "reset" σ (thus discarding Λ_σ) and start again at (1) with the same y_0. (This constitutes only finite injury to σ.) Furthermore, while σ is in (3) it may be injured by higher-priority $\mathcal{M}_{i,j}^{\Phi}$-strategies τ with $\tau^\frown\langle 0\rangle \subseteq \sigma$ (i.e. τ's of which σ assumes the infinite outcome). Before performing (3)(b), σ will check if A_0 or A_2 have changed (on an initial segment to be specified later). Before performing (3)(a) (for $y < y_1$), σ will check if A_0 or A_1 have changed (again on an initial segment to be specified later). If so (in either case), σ will destroy $\Psi^{A_0}(n)$ (by putting some $\delta(y) \leq \psi(n)$ into B and thus also into A_0), increment y_1 by $+1$, and go back to (2).

The possible outcomes of the $\mathcal{N}_0^{\Psi}$-strategy σ (neglecting the finite injury by $K \upharpoonright y_0$) are thus as follows:

(A) σ eventually waits at (2) forever. Then clearly $\Psi^{A_0} \neq K$.
(B) Λ_σ is total. Then we will be able to show $\Lambda_\sigma = K$, a recursive computation for the nonrecursive set K. Thus this outcome cannot actually occur.
(C) Otherwise. Then n must come to a limit, n_0, say; y_1 is incremented infinitely often; and $\Psi^{A_0}(n_0)$ must be destroyed infinitely often (we

call this the "infinite outcome for n_0"). We will be able to show in this case that for each y', eventually $\Psi^{A_0}(n_0)$ will always be destroyed before we put $\gamma_j(y')$ into any A_k (for $j,k \in \{0,1,2\}$), thus allowing each Γ_j to be total. In order to show this, we will use the fact that when $y_1 > y'$ and $n = n_0$ then $\Psi^{A_0}(n_0)$ can only be destroyed if either $y \geq y'$ or $K \upharpoonright (y'+1)$ changes (where the latter, of course, can occur at most finitely often for each y'). The hard part will be to show that no higher-priority $\mathcal{M}$-strategy τ with $\tau \hat{} \langle 0 \rangle \subseteq \sigma$ (i.e. of which σ assumes the infinite outcome) will injure σ infinitely often while $y < y'$. Here we will use the fact that A_0 "holds one side" for τ if τ is an $\mathcal{M}^{\Phi}_{0,1}$- or $\mathcal{M}^{\Phi}_{0,2}$-strategy, and that A_j "holds one side" for τ if τ is an $\mathcal{M}^{\Phi}_{1,2}$-strategy where $j = 2$ between (3)(a) and (3)(b) and $j = 1$ between (3)(b) and (3)(a). (For this, we use a variant of the concept of "configurations" from Slaman's proof of the density of the branching degrees [11].)

We are now ready to describe the full construction.

4. The construction for M_5

Our tree of strategies is the full binary tree $T = 2^{<\omega}$ with the ordering on T induced by the ordering on 2. The requirements $\mathcal{S}_i$ and $\mathcal{P}_i$ are global and will not be put on the tree. We effectively ω-order the $\mathcal{M}^{\Phi}_{j,k}$- and $\mathcal{N}^{\Psi}_i$-requirements as $\{\mathcal{M}_n\}_{n\in\omega}$ and $\{\mathcal{N}_n\}_{n\in\omega}$, respectively. A node $\rho \in T$ works on $\mathcal{M}_n$ if $|\rho| = 2n$ is even, and on $\mathcal{N}_n$ if $|\rho| = 2n+1$ is odd (we call ρ an $\mathcal{M}_n$- or $\mathcal{N}_n$-*strategy*, respectively). We identify 0 with the infinite outcome and 1 with the finite outcome of a strategy ρ.

Each $\mathcal{M}_n$- (or $\mathcal{N}_n$-) strategy $\rho \in T$ builds a p.r. functional Δ_ρ (or a p.r. function Λ_ρ, respectively) to satisfy its requirement. (We will frequently suppress the index on Δ and Λ.)

A strategy $\rho \in T$ is *initialized* by making all its parameters undefined and its functional undefined on all arguments. A strategy $\rho \in T$ is *reset* by initializing it, except that if ρ is an $\mathcal{N}$-strategy then ρ's parameter $y_{0,\rho}$ remains defined. A parameter is defined *big* by setting it to a number greater than any number mentioned thus far in the construction.

We now describe each stage of the construction.

At stage 0, we initialize all strategies and let all $\Gamma_i^{A_j \oplus A_k}$ be undefined on all arguments.

A stage $s+1$ consists of substages $t \leq s+1$ with some additional action before the first and after the last substage. At each substage t, a strategy $\rho \in T$ of length t is "eligible to act" and either "ends the stage" or determines the strategy $\rho' \supset \rho$ eligible to act at substage $t+1$.

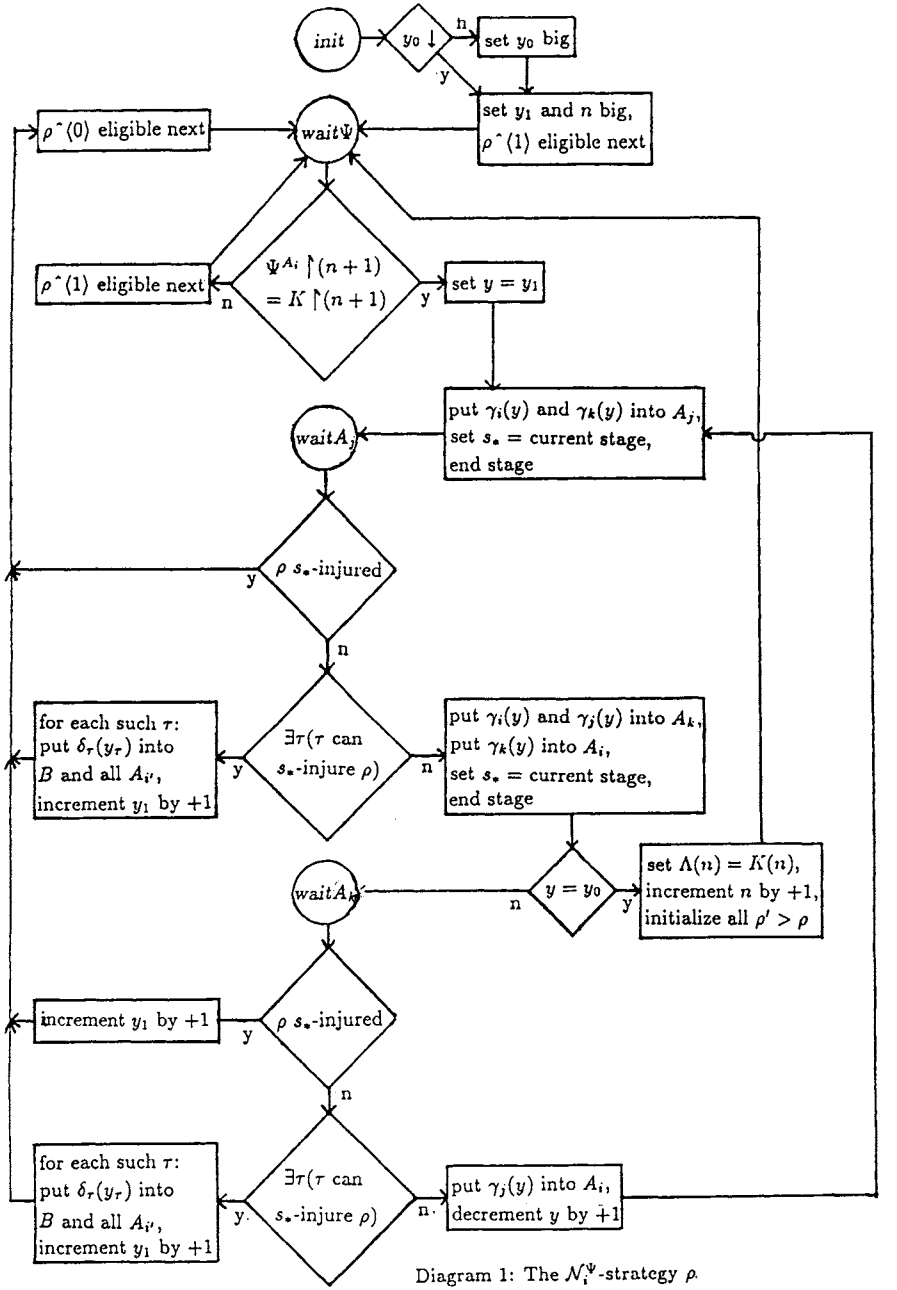


Diagram 1: The $\mathcal{N}_i^\Psi$-strategy ρ.

Diagram

Before the first substage of stage $s+1$, we determine the (least) $n \in K_{s+1} - K_s$ (if any). If n exists then, for all i, j, and k such that $\Gamma_i^{A_j \oplus A_K}(n)$ is currently defined, put $\gamma_i(n)$ into B, and thus into $A_{i'}$ for all $i' < 3$, and reset all strategies $\rho \in T$ for which there is an $\mathcal{N}$-strategy $\sigma \leq \rho$ whose parameter $y_{0,\sigma}$ is defined and $> n$. Now proceed to substage 0 of stage $s+1$, at which the strategy $\emptyset$ is eligible to act.

At a substage t of stage $s+1$, suppose ρ is eligible to act. We distinguish cases depending on whether ρ is an $\mathcal{M}$- or an $\mathcal{N}$-strategy.

If ρ is an $\mathcal{M}_{j,k}^{\Phi}$-strategy we first check if there is a (least) x_0 such that $\Delta_\rho^B(x_0)$ is defined but equals neither $\Phi^{A_j}(x_0)$ nor $\Phi^{A_k}(x_0)$. If so then pick $x_1 \leq x_0$ minimal such that

$$(1) \qquad \forall x(x_1 \leq x < x_0 \Longrightarrow \forall i \in \{j,k\}(\Phi^{A_i}(x) \downarrow = \Delta_\rho^B(x) \rightarrow \varphi^{A_i}(x) \geq \delta_\rho(x+1))).$$

(I.e. correcting $\Delta_\rho^B(x_0)$ by putting $\delta_\rho(x_0)$ into B and thus all A_i's would trigger a cascade of corrections ending with the correction of $\Delta_\rho^B(x_1)$.) Then put $\delta_\rho(x_1)$ into B and all A_i's.

Next check if the *length of agreement*

$$\ell(\rho) = \max\{x \mid \forall x' < x(\Phi^{A_j}(x') \downarrow = \Phi^{A_k}(x') \downarrow)\}$$

is now greater than at any previous stage at which ρ was eligible to act. If so then for each $x < \ell(\rho)$ (for which now $\Delta_\rho^B(x)$ is undefined) set $\Delta_\rho^B(x) = \Phi^{A_j}(x)$ with the previous use (if $\Delta_\rho^B(x)$ was defined before and no $\delta_\rho(y)$ (for $y \leq x$) has entered B since the last definition of $\Delta_\rho^B(x)$) or with big use (otherwise), and end the substage by letting $\rho^\frown\langle 0\rangle$ be *eligible to act next*. Otherwise, end the substage by letting $\rho^\frown\langle 1\rangle$ be *eligible to act next*.

Now assume an $\mathcal{N}_i^{\Psi}$-strategy ρ is eligible to act at substage t. We describe its action using the flow chart in Diagram 1. After each initialization, ρ starts in state *init*, and at each substage at which it is eligible to act, it proceeds from one state (denoted by a circle) to the next, following the arrows and along the way executing the instructions (in rectangular boxes) and deciding the truth of statements (in diamonds, following the y-arrow iff the statement is true). The parameters defined in the flow chart intuitively have the following meaning: n is the argument at which we currently attempt to define Λ; y is the number for which the $\gamma_{i'}(y)$'s are currently lifted by strategy ρ to a large number; y_0 and y_1 are the current lower and upper bounds for y; i, j, and k are indices of the sets $A_{i'}$ where $\{i,j,k\} = \{0,1,2\}$, i is determined by $\mathcal{N}_i^{\Psi}$, and $j < k$; and s_*

is the latest stage at which markers were lifted (this parameter is needed to measure (potential) injury).

Given an $\mathcal{M}^{\Phi}_{j',k'}$-strategy τ with $\tau\hat{}\langle 0\rangle \subseteq \rho$ (i.e. of which ρ guesses the infinite outcome and by which ρ could be injured), we define (for $l = j'$ or k')

$$m_l^\tau = \mu z \geq \max\{\varphi^{A_l}(y) \mid \delta_\tau(y) \leq \psi(n)\}\forall y(\delta_\tau(y)\downarrow \leq z \\ \Longrightarrow \varphi^{A_l}(y)\downarrow \leq z))\,.$$

(Note that we allow $m_l^\tau = \infty$ if $\Phi^{A_l}(y)\uparrow$ for some y with $\Delta^B_\tau(y)\downarrow$.) We say τ *can s_*-injure* ρ at stage $s+1$ if (a) ρ was in *wait*A_j at the beginning of stage $s+1$, $j' = j$, $k' = k$, and some number $\leq m_k^\tau[s_*]$ has entered A_k since stage s_* (note that A_k was supposed to "hold one side" for τ); or (b) ρ was in *wait*A_k at the beginning of stage $s+1$, $j' = j$, $k' = k$, and some number $\leq m_j^\tau[s_*]$ has entered A_j since stage s_* (note that A_j was supposed to "hold one side" for τ); or (c) $j' \neq j$ or $k' \neq k$, and some number $\leq m_i^\tau[s_*]$ has entered A_i since stage s_* (note that A_i was supposed to "hold one side" for τ). If (a), (b), or (c) applies then we define

$$y_\tau = \max\{y \mid \delta_\tau(y)[s_*] \leq \psi(n)[s_*]\}\,.$$

We say *ρ has been s_*-injured* if some number $\leq \psi(n)[s_*]$ has entered A_i since stage s_* (this takes care of miscellaneous injury). Note here that we may assume

$$\forall n(n \leq \psi(n))\,. \tag{2}$$

We end the stage if Diagram 1 specifies so or if $s \leq t$, otherwise we go to substage $t+1$.

At the end of stage $s+1$, i.e. after the last substage, we define $\Gamma_i^{A_j \oplus A_k}(n)$ (for each $\Gamma_i^{A_j \oplus A_k}$ and each $n \leq s$ such that $\Gamma_i^{A_j \oplus A_k}(n)$ is now undefined) with the previous use (if $\Gamma_i^{A_j \oplus A_k}(n)$ was defined before and no $\gamma_i(n')$ (for $n' \leq n$) has entered A_j or A_k since the last definition of $\Gamma_i^{A_j \oplus A_k}(n)$) or with big use (otherwise). Furthermore, we initialize all strategies $>$ the strategy last eligible to act, and proceed to the next stage.

This ends the description of the construction.

5. The verification for M_5

Our first two lemmas are easy:

LEMMA 1 ($\mathcal{S}_i$-SATISFACTION LEMMA). *$B \leq_T A_i$ for $i = 0, 1$, and 2.*

PROOF: Fix any number x. By the construction, $B \subseteq A_i$. So assume $x \in A_i$, say $x \in A_{i,s}$ for some stage s. But then $x \in B$ iff $x \in B_s$ by the construction. This establishes the claim. ■

LEMMA 2 (RESETTING LEMMA). *If $\rho \in T$ is initialized at most finitely often then it is reset at most finitely often.*

PROOF: Since ρ is initialized only finitely often, the same holds for any $\rho' \leq \rho$, and thus $y_{0,\rho}$ comes to a limit. Furthermore, $y_{0,\rho'} \leq y_{0,\rho}$ at any stage at which $y_{0,\rho'}$ is defined for any $\rho' \leq \rho$. Thus ρ is never reset after $y_{0,\rho}$ and $K \upharpoonright \lim_s y_{0,\rho,s}$ settles down. ■

We define the *true path* $f \in [T]$ by induction as follows: Let $\rho = f \upharpoonright n$. Then $f(n) = 0$ if $\rho \hat{} \langle 0 \rangle$ is eligible to act infinitely often, and $f(n) = 1$ otherwise.

We now turn to the $\mathcal{M}$-requirements:

LEMMA 3 ($\mathcal{M}^{\Phi}_{j,k}$-SATISFACTION LEMMA). *If an $\mathcal{M}^{\Phi}_{j,k}$-strategy $\tau \subset f$ is eligible to act infinitely often and is initialized at most finitely often then it satisfies its requirement.*

PROOF: Suppose $\Phi^{A_j} = \Phi^{A_k}$ are both total. By Lemma 2, τ is reset at most finitely often, so Δ_τ is never discarded after some (least) stage s_0, say. By the first part of τ's action in the construction, $\Delta^B_\tau(x) \downarrow \neq \Phi^{A_j}(x)$ is impossible for any x.

Thus we only have to show $\Delta^B_\tau(x) \downarrow$ for all x. Suppose this fails for some (least) x_0, and $\Delta^B_\tau \upharpoonright x_0$ as well as $\Phi^{A_j}(x_0)$ and $\Phi^{A_k}(x_0)$ are defined by correct computations after some (least) stage $s_1 \geq s_0$. Since $\lim_s \ell_s(\tau) = \infty$, we have $\rho \hat{} \langle 0 \rangle \subset f$ and thus $\Delta^B_\tau(x_0)[s]$ must be defined at infinitely many stages s. Since $\Delta^B_\tau(x_0) \uparrow$, we have $\lim_s \delta_{\tau,s}(x_0) = \infty$. By the way $\delta_{\tau,s}(x_0)$ is defined, it can only be increased by the action of τ or some $\mathcal{N}$-strategy $\sigma \supseteq \tau \hat{} \langle 0 \rangle$. Once $\delta_{\tau,s}(x_0) \geq \varphi^{A_j}(x_0)$, $\varphi^{A_k}(x_0)$ and $s > s_1$, τ will not increase $\delta_{\tau,s}(x_0)$ by our assumption on s_1 and by (1). There are only finitely many $\mathcal{N}$-strategies $\sigma \supseteq \tau \hat{} \langle 0 \rangle$ that ever set their $s_* \leq s_1$. Let $s_2 \geq s_1$ be the least stage such that each such σ will either never put $\delta_\tau(y)$ (for $y \leq x_0$) into B after stage s_2 or has already set its $s_* \geq s_1$. Suppose some $\mathcal{N}$-strategy $\sigma \supseteq \tau \hat{} \langle 0 \rangle$ causes $\delta_\tau(x_0)$ to increase by putting $\delta_\tau(y_\tau)$ into B (for $y_\tau \leq x_0$) at a stage $s > s_2$. By our assumption on s_1 and the minimality of x_0, we have $y_\tau = x_0$. Then $A_j \upharpoonright (m^\tau_j[s_*] + 1)$ or $A_k \upharpoonright (m^\tau_k[s_*] + 1)$ must have changed between stage s_* and s; without loss

of generality assume the former has changed. Since $s_* \geq s_1$ and by our assumption on s_1, $m_j^\tau[s_*] > \varphi^{A_j}(x_0)$. But by the definition of y_τ $(= x_0)$, we have $\delta_\tau(x_0+1)[s_*] > \psi(n)[s_*]$, and thus, by $\delta_\tau(x_0)[s_*] \geq \varphi^{A_j}(x_0)$, we have $m_j^\tau[s_*] \leq \varphi^{A_j}(x_0)$, a contradiction. ■

We now prove a very technical lemma, which constitutes the inductive step in the proofs of the satisfaction of both the $\mathcal{N}$- and the $\mathcal{P}$-requirements:

LEMMA 4 (CONFIGURATION LEMMA). *Let $\{i,j,k\} = \{0,1,2\}$ with $j < k$. Let $\sigma \subset f$ be an $\mathcal{N}_i^\Psi$-strategy, and suppose that σ is not initialized or reset after some (least) stage s_0. If σ reaches state waitA_k with parameter y at a stage $s_1 > s_0$, then $\Psi^{A_i}(n)$ will not be destroyed after stage s_1 unless $K \upharpoonright y$ changes. If σ reaches waitΨ, having defined $\Lambda(n)$ for some n at a stage $s_1 > s_0$, then $\Psi^{A_i}(n)$ will not be destroyed after stage s_1.*

PROOF: Let $\{i,j,k\} = \{0,1,2\}$ with $j < k$. Let

$$\mathcal{T}_1 = \{\tau \ \ \mathcal{M}_{j,k}^\Phi\text{-strategy} \mid \tau^\frown\langle 0\rangle \subseteq \sigma \wedge \Phi \ \text{ p.r. functional}\}, \text{ and}$$
$$\mathcal{T}_2 = \{\tau \ \ \mathcal{M}_{j',k'}^\Phi\text{-strategy} \mid \tau^\frown\langle 0\rangle \subseteq \sigma \wedge \Phi \ \text{ p.r. functional} \wedge \ (j \neq j' \text{ or } k \neq k')\}.$$

(These are the $\mathcal{M}$-strategies "dangerous" to σ.) We will first note that by reverse induction on $y \in [y_0, y_1]$ the following hold:

(3) If σ is in waitA_j with $y = y_1$ then

(3a) $\gamma_i(y_1), \gamma_k(y_1) > m_k^\tau[s_*], \psi(n)[s_*]$ for all $\tau \in \mathcal{T}_1$, and
(3b) $\gamma_i(y_1), \gamma_k(y_1) > m_i^\tau[s_*], \psi(n)[s_*]$ for all $\tau \in \mathcal{T}_2$.

(4) If σ is in waitA_k with $y = y_1$ then

(4a) $\gamma_i(y_1), \gamma_j(y_1), \gamma_k(y_1) > m_j^\tau[s_*], \psi(n)[s_*]$ for all $\tau \in \mathcal{T}_1$, and
(4b) $\gamma_i(y_1), \gamma_j(y_1), \gamma_k(y_1) > m_i^\tau[s_*], \psi(n)[s_*]$ for all $\tau \in \mathcal{T}_2$.

(5) If σ is in waitA_j with $y < y_1$ then

(5a) $\gamma_i(y), \gamma_j(y+1), \gamma_k(y) > m_k^\tau[s_*], \psi(n)[s_*]$ for all $\tau \in \mathcal{T}_1$, and
(5b) $\gamma_i(y), \gamma_j(y+1), \gamma_k(y) > m_i^\tau[s_*], \psi(n)[s_*]$ for all $\tau \in \mathcal{T}_2$.

(6) If σ is in waitA_k with $y < y_1$ then

(6a) $\gamma_i(y), \gamma_j(y), \gamma_k(y) > m_j^\tau[s_*], \psi(n)[s_*]$ for all $\tau \in \mathcal{T}_1$, and
(6b) $\gamma_i(y), \gamma_j(y), \gamma_k(y) > m_i^\tau[s_*], \psi(n)[s_*]$ for all $\tau \in \mathcal{T}_2$.

(7) If σ reaches waitΨ while defining $\Lambda(n)$ for some n at a stage $s > s_0$ then (6a)–(6b) hold for this n and s until σ reaches a new state.

It is not hard to verify (3)–(7), keeping in mind the definition of the Γ-uses, the actions of σ, and the fact that the right-hand sides are defined and finite since all the $\tau \in \mathcal{T}_1 \cup \mathcal{T}_2$ have outcome 0 at stage s_* (i.e. greater length of agreement than at all previous τ-stages).

Note that σ ends the stage when it reaches any state other than *wait*Ψ and initializes all $\rho > \sigma$ when it defines Λ on some n. By this feature, the definition of the m_l^τ's, our assumption on s_0, and (3)–(7) we have after stage s_0:

(8) After σ reaches *wait*A_k for some y and n, we have

(8a) $\gamma_i(y), \gamma_j(y), \gamma_k(y) > m_j^\tau, \psi(n)$ for all $\tau \in \mathcal{T}_1$, or
(8b) $\gamma_i(y), \gamma_j(y), \gamma_k(y) > m_k^\tau, \psi(n)$ for all $\tau \in \mathcal{T}_1$; and always
(8c) $\gamma_i(y), \gamma_j(y), \gamma_k(y) > m_j^\tau, \psi(n)$ for all $\tau \in \mathcal{T}_2$

unless $K \upharpoonright y$ changes later.

(9) After σ reaches *wait*Ψ having defined $\Lambda(n)$, we have (8a)–(8c) for $y = y_0$.

By (8a)–(8c) and (9a)–(9c), we have established the lemma since the $\tau \in \mathcal{T}_1 \cup \mathcal{T}_2$ are the only strategies able to destroy $\Psi^{A_i}(n)$ but are prevented from doing so by m_j^τ or m_k^τ for $\tau \in \mathcal{T}_1$, and by m_i^τ for $\tau \in \mathcal{T}_2$, respectively. ■

The satisfaction of the $\mathcal{N}$-requirements now follows easily:

LEMMA 5 ($\mathcal{N}_i^\Psi$-SATISFACTION LEMMA). *If an $\mathcal{N}_i^\Psi$-strategy $\sigma \subset f$ is eligible to act infinitely often and is initialized at most finitely often then it defines $\Lambda_\sigma = K \upharpoonright \operatorname{dom} \Lambda_\sigma$ correctly and satisfies its requirement.*

PROOF: Suppose σ is not initialized or reset after some (least) stage s', using Lemma 2. Then, after stage s', σ will pick a big n (call it n_0) and will try to define $\Lambda_\sigma(n)$ for all $n \geq n_0$. Once $\Lambda(n)$ is defined, the corresponding $\Psi^{A_i}(n)$ cannot be destroyed by Lemma 4 and our assumption on s', establishing the first half of our claim.

Now suppose $\Psi^{A_i} = K$, and fix $s_n \geq s'$ (for $n \geq n_0$) such that $\Psi^{A_i} \upharpoonright (n+1)$ is never destroyed after stage s_n and such that $K_{s_n} \upharpoonright (n+1) = K \upharpoonright (n+1)$. Then, after stage s_n, σ will enter states *wait*A_j and *wait*A_k with this n and, since $\Psi^{A_i}(n)$ is no longer destroyed, return to state *wait*Ψ only after having defined $\Lambda_\sigma(n)$. Thus $\Lambda_\sigma(n) = K(n)$ for cofinitely many n, establishing the satisfaction of the $\mathcal{N}_i^\Psi$-requirement. ■

LEMMA 6 (INITIALIZATION/ELIGIBILITY LEMMA). *Each $\rho \subset f$ is eligible to act infinitely often and is initialized at most finitely often. Thus all $\mathcal{M}_{j,k}^\Phi$- and $\mathcal{N}_i^\Psi$-requirements are satisfied.*

PROOF: Since K is not recursive the domain of Λ_σ for any $\mathcal{N}$-strategy $\sigma \subset f$ must be finite. Thus no $\mathcal{N}$-strategy $\sigma \subset f$ will initialize $\sigma^\frown\langle 0\rangle$

infinitely often. Furthermore, σ can be in states $waitA_j$ and $waitA_k$ at most finitely often before returning to state $wait\Psi$, and it will return almost always via injury. Thus σ will not end the stage at infinitely many of the stages at which it is eligible to act. This establishes the first half of the lemma. The rest follows by Lemmas 2, 3, and 5. ■

LEMMA 7 ($\mathcal{P}_i$-SATISFACTION LEMMA). *If* $\{i,j,k\} = \{0,1,2\}$ *and* $j < k$ *then* $\Gamma_i^{A_j \oplus A_k} = K$.

PROOF: By the construction it is impossible to have $\Gamma_i^{A_j \oplus A_k}(z) \downarrow \neq K(z)$ for any z. It thus suffices to show that $\Gamma_i^{A_j \oplus A_k}$ is total. So assume $\Gamma_i^{A_j \oplus A_k}(z)$ is undefined for some (least) z. By the construction, $\Gamma_i^{A_j \oplus A_k}(z)$ is defined infinitely often, and by the assumption on z, $\Gamma_i^{A_j \oplus A_k}(z)$ is destroyed infinitely often by some $\mathcal{N}$-strategy σ. By the way y_0 is picked, only finitely many $\mathcal{N}$-strategies σ can destroy $\Gamma_i^{A_j \oplus A_k}(z)$, so say σ_0 is the $<$-least of them destroying $\Gamma_i^{A_j \oplus A_k}(z)$ infinitely often. Then necessarily $\sigma_0 \subset f$ (by initialization). By Lemmas 6 and 2, σ_0 is not initialized or reset after some (least) stage s_0. Thus $\lim_s y_{1,s} = \infty$, say $y_{1,s} > z+1$ for all $s \geq s_1$ (for some $s_1 > s_0$). Then σ_0 must reach $waitA_k$ with $y = z+1$ infinitely often, and, by the first sentence of the proof of Lemma 6, almost always with the same n. So $\Psi^{A_i}(n)$ is defined infinitely often after stage s_0 when σ_0 reaches $waitA_k$ with $y = z+1$ but later destroyed. By Lemma 4, $K \upharpoonright y$ must change every time, a contradiction. ■

This concludes the proof of Theorem 1. We now turn to the proof of Theorem 2.

6. The requirements and the strategies for N_5

We have to construct r.e. sets A_0, A_1, A_2, and B such that for $\mathbf{a}_i = \deg(A_i)$ $(i = 0,1,2)$ and $\mathbf{b} = \deg(B)$,

$$\begin{aligned} \mathbf{a}_0 \cup \mathbf{a}_2 &= \mathbf{0}', \\ \mathbf{a}_1 \cap \mathbf{a}_2 &= \mathbf{b}, \\ \mathbf{b} < \mathbf{a}_0 &< \mathbf{a}_1, \text{ and} \\ \mathbf{b} &< \mathbf{a}_2. \end{aligned}$$

We ensure this by the following requirements:

$\mathcal{S} : B \leq_T A_0 \leq_T A_1$ and $B \leq_T A_2$,

$\mathcal{P} : K \leq_T A_0 \oplus A_2$,

$\mathcal{M}^\Phi : \Phi^{A_1} = \Phi^{A_2}$ total $\implies \exists \Delta(\Delta^B = \Phi^{A_1})$ (for all p.r. functionals Φ),

$\mathcal{D}_i^\Psi : A_i \neq \Psi^B$ (for $i = 0, 2$ and all p.r. functionals Ψ), and

$\mathcal{D}_1^\Psi : A_1 \neq \Psi^{A_0}$ (for all p.r. functionals Ψ).

The global requirement $\mathcal{S}$ is met by direct coding, i.e. whenever a number enters B (A_0, resp.) it also enters A_0, A_1, and A_2 (A_1, respectively).

To satisfy the global requirement $\mathcal{P}$ we construct a functional Γ which computes the complete set K from $A_0 \oplus A_2$. We will define Γ implicitly by a marker function $\gamma(x)$ which may be viewed as the use function of Γ. The xth position of γ at the end of stage s will be denoted by $\gamma(x)[s]$. The marker obeys the following rules (for any numbers x, y, s, t):

(γ^0) $\quad x \neq y \implies \gamma(x)[s] \neq \gamma(y)[t]$,

(γ^1) $\quad \gamma(x)[s] \neq \gamma(x)[s+1] \implies \gamma(x)[s] < \gamma(x)[s+1]$
and, for some $i \in \{0, 2\}$, $\gamma(x)[s] \in A_i[s+1]$,

(γ^2) $\quad \lim_s \gamma(x)[s]$ exists,

(γ^3) $\quad \gamma(x)[s] \notin (A_0 \cup A_2)[s]$, and

(γ^4) $\quad x \in K_{s+1} - K_s \implies \gamma(x)[s] \in (A_0 \cup A_2)[s+1]$.

Then, by (γ^1) and (γ^2), $\gamma^*(x) := \lim_s \gamma(x)[s]$ exists and $\gamma^*(x) = \sup_s \gamma(x)[s]$. Moreover, by (γ^1) and (γ^3), $\gamma^* \leq_T A_0 \oplus A_2$. Finally, by (γ^3) and (γ^4),

$$\gamma^*(x) = \gamma(x)[s] \implies K_s(x) = K(x).$$

So to compute $K(x)$ from $A_0 \oplus A_2$, $\Gamma^{A_0 \oplus A_2}$ computes the first stage s such that $\gamma^*(x) = \gamma(x)[s]$ and checks whether x has entered K by the end of this stage. If so, $x \in K$; otherwise $x \notin K$.

For the local requirements $\mathcal{M}^\Phi$, as in the preceding proof, we use Fejer's strategy [5]: Whenever $\Delta^B(x)$ is defined but equal neither to $\Phi^{A_1}(x)$ nor to $\Phi^{A_2}(x)$ then the strategy puts a number $y \leq \delta(x)$ into B to allow the correction of $\Delta^B(x)$.

For the local requirements $\mathcal{D}_i^\Psi$ for $i = 0$ ($i = 2$, respectively) we basically use the Friedberg-Muchnik strategy: The strategy has a follower x. If $\Psi^B(x) = 0$ at some stage then it puts x into A_i and tries to preserve the

computation $\Psi^B(x)$. The latter conflicts with the $\mathcal{M}$-strategies which are the only ones which put numbers into B. To prevent that an infinitary higher-priority strategy $\mathcal{M}^\Phi$ destroys $\Psi^B(x)$, the $\mathcal{D}_i^\Psi$-strategy attacks only at Φ-expansionary stages and tries to protect the computations $\Phi^{A_2}(y) = \Delta^B(y)$ (or $\Phi^{A_1}(y) = \Delta^B(y)$, respectively). This is achieved by initializing lower-priority $\mathcal{D}$- and $\mathcal{M}$-strategies and by lifting markers $\gamma(z) \leq \varphi(y)$ by enumerating them into A_i.

The strategy for the local requirements $\mathcal{D}_1^\Psi$ is similar but slightly more complicated. Again the strategy has a follower x and waits for $\Psi^{A_0}(x) = 0$. Then it wants to put x into A_1 and hold $A_0 \upharpoonright (\psi(x) + 1)$ to ensure $A_1(x) \neq \Psi^{A_0}(x)$. Now, to lift a marker $\gamma(z)$, however, we have to put $\gamma(z)$ into A_2, since putting $\gamma(z)$ into A_0 might destroy $\Psi^{A_0}(x)$. So if we put x into A_1 at the same time, for some $\mathcal{M}^\Phi$ and y as above we might destroy both sides of an agreement

$$\Phi^{A_1}(y) = \Delta^B(y) = \Phi^{A_2}(y),$$

thereby causing $\mathcal{M}^\Phi$ to put a number $u \leq \delta(y)$ into B and therefore into A_0 (by $\mathcal{P}$), which might destroy $\Psi^{A_0}(x)$. This problem is overcome by doing the attack in two stages.

At the first expansionary stage we lift markers $\gamma(z)$ via A_2 to protect $\Psi^{A_0}(x)$ (and hold A_1 to prevent $\mathcal{M}^\Phi$ from acting). Then, at the next expansionary stage, we put $\gamma(z)$ into A_0 and x into A_1, thereby diagonalizing (and now hold A_2 to prevent $\mathcal{M}^\Phi$ from acting).

7. The construction for N_5

We define the tree of strategies to be

$$T = \{x \in 2^{<\omega} \mid \forall n(\alpha(2n+1)\downarrow \implies \alpha(2n+1) = 1)\}.$$

Let $\{\mathcal{M}_n\}_{n\in\omega}$ and $\{\mathcal{D}_n\}_{n\in\omega}$ be effective listings of the $\mathcal{M}^\Phi$- and $\mathcal{D}_i^\Psi$-requirements, respectively. As before, node σ works on $\mathcal{M}_n$ if $|\sigma| = 2n$ is even and on $\mathcal{D}_n$ if $|\sigma| = 2n+1$ is odd; and we call σ an $\mathcal{M}_n$- or $\mathcal{D}_n$-strategy, respectively. (Since the $\mathcal{D}_n$-strategies are finitary we have put only their finitary outcome 1 on the tree T.) Every $\mathcal{M}_n$-strategy σ builds a functional Δ_σ to satisfy $\mathcal{M}_n$. Initializing a strategy σ is defined as in the previous construction. We let $\mathrm{In}(\sigma)[s]$ be the greatest stage $t \leq s$ at which σ is initialized.

For n with $\mathcal{M}_n = \mathcal{M}^\Phi$ we let

$$\ell(n)[s] = \max\{x : \forall y < x(\Phi^{A_1}(y)[s]\downarrow = \Phi^{A_2}(y)[s]\downarrow)\}.$$

Here we adopt the convention that if $\Phi^{A_i}(y)[s] \downarrow \neq \Phi^{A_i}(y)[s+1]$ then $\Phi^{A_i}(y)[s+1] \uparrow$ ("hat-trick", see Soare [12]).

Based on the length function ℓ, σ-stages and σ-expansionary stages are defined as usual by induction on $|\sigma|$: Any stage is a $\emptyset$-stage, and stage s is $\emptyset$-expansionary if $s = 0$, or $s > 0$ and $\forall t < s\ (\ell(0)[t] < \ell(0)[s])$. For σ with $|\sigma| > 0$, s is σ-expansionary if $|\sigma| = 2n$ is even, s is a σ-stage, and

$$\ell(n)[s] > \max\{\ell(n)[t] : t < s \wedge t \text{ is a } \sigma\text{-stage}\}\,.$$

Finally s is a $\sigma^\frown\langle i\rangle$-stage if $|\sigma| < s$ and either $i = 0$ and s is σ-expansionary, or $i = 1$ and s is a σ-stage but not a σ-expansionary stage.

The unique string σ of length s such that s is a σ-stage will be denoted by $\alpha[s]$.

In the following description of the stages of the construction, a number y is called big if y is bigger than all numbers mentioned in the construction up to this point (with the exception of the values of the marker function γ).

Stage 0: Initialize all strategies σ. Let $\gamma(x)[0] = 2\langle x, 0\rangle$.

Stage $s+1$: The stage consists of 6 steps.

Step 1: Initialize all strategies σ with $\alpha[s] < \sigma$.

Step 2 ($\mathcal{D}$-Strategies): For any $\sigma \subseteq \alpha[s]$ with $|\sigma| = 2n+1$ odd and $\mathcal{D}_n = \mathcal{D}_i^\Psi$, σ requires attention if either σ has no follower or, for the follower x, $A_i(x)[s] = \Psi^B(x)[s] = 0$ (if $i \in \{0,2\}$) or $A_1(x)[s] = \Psi^{A_0}(x)[s] = 0$ (if $i = 1$).

Fix the least σ which requires attention, say $|\sigma| = 2n+1$, $\mathcal{D}_n = \mathcal{D}_i^\Psi$. (If no σ requires attention, Step 2 is vacuous.) Say that σ acts. Initialize all strategies σ' with $\sigma < \sigma'$. If σ has no follower, let x be the least big odd number and appoint x as a σ-follower. If σ has a follower, say x, then distinguish the following 3 cases.

Case 1: $i \in \{0,2\}$. Then put x into A_i. Moreover, for any $y \geq \mathrm{In}(\sigma)[s]$, put $\gamma(y)[s]$ into A_i and let $\gamma(y)[s+1] = 2\langle y, s+1\rangle$.

Case 2: $i = 1$ and x is not yet confirmed. Then, for any $y \geq \mathrm{In}(\sigma)[s]$, put $\gamma(y)[s]$ into A_2, and let $\gamma(y)[s+1] = 2\langle y, s+1\rangle$. Say that x is *confirmed.*

Case 3: $i = 1$ and x is confirmed. Then put x into A_1. Moreover, for any $y \geq \mathrm{In}(\sigma)[s]$, put $\gamma(y)[s]$ into A_0 and let $\gamma(y)[s+1] = 2\langle y, s+1\rangle$. Moreover, for any $y \geq In(\sigma)[s]$, put $\gamma(y)[s]$ into A_0 and let $\gamma(y)[s+1] = 2\langle y, , \rangle s+1$.

Step 3 ($\mathcal{P}$-Strategy): For any x such that $x \in K_{s+1} - K_s$ and $\gamma(x)[s+1]$ has not been redefined in Step 2, put $\gamma(x)$ into A_0, let $\gamma(x)[s+1] = 2\langle x, s+1\rangle$, and, for any σ such that $|\sigma|$ is odd and $x < \mathrm{In}(\sigma)[s]$, cancel the σ-follower (if there is any).

If not stated otherwise above, $\gamma(y)[s+1] = \gamma(y)[s]$.

Step 4 ($\mathcal{M}$-Strategies; correction): For any σ such that σ has not been initialized in the previous steps and $|\sigma|$ is even, say $|\sigma| = 2n$ and $\mathcal{M}_n = \mathcal{M}^\Phi$, and for any number y do the following: If $\Phi^{A_1}(y)[s] \neq \Delta_\sigma^B(y)[s] \downarrow \neq \Phi^{A_2}(y)[s]$ then put $\delta_\sigma(y)[s]$ into B, let $\delta_\sigma(y)[s+1] \uparrow$ and $\Delta_\sigma^B(y)[s+1] \uparrow$, and initialize all σ' with $\sigma^\frown\langle 0\rangle <_L \sigma'$.

Step 5 ($\mathcal{M}$-Strategies; extension): For any σ such that $\sigma^\frown\langle 0\rangle \subseteq \alpha[s]$, σ has not been initialized in the previous steps and such that $|\sigma|$ is even, say $|\sigma| = 2n$ and $\mathcal{M}_n = \mathcal{M}^\Phi$, and for any number y do the following: If $y < \ell(n)[s]$ and $\Delta_\sigma^B(y) \uparrow$ then let $\Delta_\sigma^B(y)[s+1] = \Phi^{A_1}(y)[s] = \Phi^{A_2}(y)[s] \downarrow$ and let $\delta_\sigma(y)[s+1]$ be the previous use (if $\Delta_\sigma^B(y)$ has been defined before and no $\delta_\sigma(y')$ for $y' \leq y$ has entered B since the last definition of $\Delta_\sigma^B(y)$) or the least big odd number (otherwise).

Step 6 ($\mathcal{S}$-Strategy): Put any number which has entered B (A_0) in one of the previous steps also into A_0, A_1, and A_2 (A_1, respectively).

This completes the description of the construction.

8. The verification for N_5

LEMMA 1 ($\mathcal{S}$-LEMMA). *$B \leq_T A_0 \leq_T A_1$ and $B \leq_T A_2$.*

PROOF: Any number x which enters any set under construction at stage $s+1$ has not entered any other set under construction at any previous stage. So the claim is immediate by Step 6 of stage $s+1$. ■

The *true path* $f \in [T]$ is defined to be the leftmost path through T such that for any n, $f \upharpoonright n \subseteq \alpha[s]$ for infinitely many s.

We say x is a permanent σ-follower if x is σ-follower from some stage on.

LEMMA 2 (INITIALIZATION LEMMA). *Let $\sigma \subset f$.*

(a) *σ is initialized only finitely often.*
(b) *If $|\sigma|$ is odd then σ acts only finitely often and has a permanent follower.*

PROOF: We proceed by induction on $|\sigma|$.

Fix s_0 such that $\sigma \leq \alpha[s]$ for all $s \geq s_0$ and such that, by inductive hypothesis, no σ' with $\sigma' \subset \sigma$ acts after stage s_0. Then σ will not be initialized in Step 1 or 2 of any stage $s > s_0$. Moreover, no $\delta_\tau(y)$ for $\tau^\frown\langle 0\rangle <_L \sigma$ will be appointed after stage s_0 (in Step 5), whence, there will be a stage $s_1 > s_0$ such that σ will not be initialized in Step 4 of any stage $s \geq s_1$ and hence will not be initialized after stage s_1 at all.

Now, if $|\sigma|$ is odd, fix $s_2 > s_1$ such that $K_{s_2} \upharpoonright (s_1+1) = K \upharpoonright (s_1+1)$. Then no follower of σ will be cancelled in Step 3 of any stage $s \geq s_2$, whence any σ-follower existing after stage s_2 is permanent. Moreover, if s_3 is the least σ-stage $> s_2$ then either there is a σ-follower at the end of stage s_3 or a σ-follower is appointed at stage s_3. So σ will act at most once (if $i \in \{0,2\}$) or twice (if $i = 1$) after stage $s_3 + 1$. ■

LEMMA 3 ($\mathcal{P}$-LEMMA). $K \leq_T A_0 \oplus A_1$.

PROOF: By the discussion of the $\mathcal{P}$-strategy preceding the construction it suffices to show that the function $\gamma(x)[s]$ satisfies conditions (γ^0)–(γ^4). For (γ^0)–(γ^1) and (γ^3)–(γ^4) this is immediate by the construction. For a proof that $\lim_s \gamma(x)[s]$ exists fix x. By Lemma 2, choose stages s_1 and s_0 such that $s_1 > s_0 > x$, $\alpha[s_0] \subseteq \alpha[s_1] \subset f$, and no σ with $\sigma \leq \alpha[s_0]$ acts after stage s_1. Since any σ with $\alpha[s_0] < \sigma$ is initialized in Step 1 of stage $s_0 + 1 > x$ and since only such σ will act after stage s_1, $\gamma(x)[s]$ will not be redefined in Step 2 of any stage $s + 1 > s_1$. So the value of $\gamma(x)[s]$ will change at most once after stage s_1, namely if x enters K after that stage. ■

LEMMA 4 (Δ-CORRECTNESS LEMMA). *Let $\mathcal{M}^\Phi = \mathcal{M}_n$, $|\sigma| = 2n$, and $\sigma\hat{}\langle 0\rangle \subset f$. If s is a $\sigma\hat{}\langle 0\rangle$-stage and $\Delta_\sigma^B(y)[s] \downarrow$ then $\Delta_\sigma^B(y)[s] = \Phi^{A_1}(y)[s]$.*

PROOF: For a contradiction assume that $\Delta_\sigma^B(y)[s] \neq \Phi^{A_1}(y)[s]$. Let t be the greatest stage $< s$ such that $\Delta_\sigma^B(y)[t] \uparrow$. Then t is a σ-expansionary stage and

$$\Delta_\sigma^B(y)[s] = \Delta_\sigma^B(y)[t+1] = \Phi^{A_1}(y)[t] = \Phi^{A_2}(y)[t]\,.$$

Since s is σ-expansionary, too, we must have $\Delta_\sigma^B(y)[s] \neq \Phi^{A_1}(y)[s] \downarrow$ and $\Delta_\sigma^B(y)[s] \neq \Phi^{A_2}(y)[s] \downarrow$. So, by the "hat-trick", there must be a stage v such that $t < v < s$ (whence $\Delta_\sigma^B(y)[t+1] = \Delta_\sigma^B(y)[v] = \Delta_\sigma^B(y)[s]$), and $\Delta_\sigma^B(y)[v] \neq \Phi^{A_1}(y)[v]$ and $\Delta_\sigma^B(y)[v] \neq \Phi^{A_2}(y)[v]$ (where one of the right-hand side computations is undefined). So, by Step 4 of the construction, $\Delta_\sigma^B(y)[v+1] \uparrow$ contrary to the choice of t. ■

LEMMA 5 ($\mathcal{M}$-LEMMA). *Each $\mathcal{M}^\Phi$ is met.*

PROOF: Without loss of generality, we may assume that $\Phi^{A_1} = \Phi^{A_2}$ is total. Pick n and σ such that $\mathcal{M}^\Phi = \mathcal{M}_n$, $|\sigma| = 2n$, and $\sigma \subset f$. Then, by assumption, $\lim_s \ell(n)[s] = \infty$. So there are infinitely many σ-expansionary stages, whence $\sigma\hat{}\langle 0\rangle \subset f$. Moreover, by Lemma 2, there is a stage after which σ is never initialized. It easily follows from Step 5 in the construction that Δ_σ^B is total and, with Lemma 4, that $\Delta_\sigma^B = \Phi^{A_1}$. ■

LEMMA 6 ($\mathcal{D}$-LEMMA). *Each $\mathcal{D}_i^\Psi$ is met.*

PROOF: We give the proof for $i = 1$. (The other cases are similar and somewhat simpler.) Fix n and σ such that $\mathcal{D}_n = \mathcal{D}_i^\Psi$, $|\sigma| = 2n + 1$ and $\sigma \subset f$. By Lemma 2 there is a stage s_0 such that at stage $s_0 + 1$ a σ-follower x is appointed which will never be cancelled. We will show that $A_1(x) \neq \Psi^{A_0}(x)$. We distinguish two cases.

Case 1. There is a σ-stage $s > s_0$ such that

$$\Psi^{A_0}(x)[s] \downarrow = 0\,.$$

Then let s_1 be the least such stage. By the choice of s_0, σ acts at stage $s_1 + 1$ and x becomes confirmed. Now let s_2 be the least σ-stage $> s_1$.

We claim that

$$(*) \qquad B[s_1] \upharpoonright s_1 = B[s_2] \upharpoonright s_1 \text{ and } A_i[s_1] \upharpoonright s_1 = A_i[s_2] \upharpoonright s_1$$

for $i = 0, 1$, whence in particular

$$\Psi^{A_0}(x)[s_2] = \Psi^{A_0}(x)[s_1] = 0$$

via the same computation. For a proof of $(*)$ we note that all strategies $\sigma' > \sigma$ are initialized whence such strategies cannot injure $(*)$. Moreover, since, by choice of s_0, σ is not initialized after this stage, no $\mathcal{M}$-strategy σ' with $\sigma' \hat{} \langle 0 \rangle <_L \sigma$ will put a number into B after stage s_0 and no $\mathcal{D}$-strategy σ' with $\sigma' < \sigma$ will put a number into any set A_j $(j = 0, 1, 2)$. Since σ itself does not destroy $(*)$ (at stage $s_1 + 1$ it puts numbers into A_2 only and it does not act before stage $s_2 + 1$ again), this leaves only the $\mathcal{P}$-strategy and $\mathcal{M}$-strategies τ with $\tau \hat{} \langle 0 \rangle \subseteq \sigma$. Now, by action of σ at stage $s + 1$, $\gamma(y)[s_1 + 1] > s_1$ for all y with $y \geq \mathrm{In}(\sigma)[s_1]$. So if the $\mathcal{P}$-strategy injures $(*)$, then it enumerates some $\gamma(y)[s]$ with $y < \mathrm{In}(\sigma)[s_1] \leq \mathrm{In}(\sigma)[s]$ into some A_i, which will result in cancellation of the follower x contrary to the choice of x. Finally, consider an $\mathcal{M}$-strategy τ with $\tau \hat{} \langle 0 \rangle \subseteq \sigma$. Then s_1 is τ-expansionary and, by Lemma 4,

$$\forall y (\Delta_\tau^B(y)[s_1] \downarrow \implies \Delta_\tau^B(y)[s_1] = \Phi^{A_1}(y)[s_1] \downarrow = \Phi^{A_2}(y)[s_1])\,.$$

Now τ will injure $(*)$ at a stage $s + 1 > s_1$ only if, for such a number y, $\Phi^{A_1}(y)[s_1] \neq \Phi^{A_1}(y)[s]$ and $\Phi^{A_2}(y)[s_1] \neq \Phi^{A_2}(y)[s]$, i.e. if some other strategy has injured $A_1[s_1] \upharpoonright s_1 = A_1[s_2] \upharpoonright s_1$ *before*. As we have shown, however, this will not happen.

Now at stage $s_2 + 1$, σ becomes active again and puts x into A_1. To show that $\Psi^{A_0}(x) = \Psi^{A_0}(x)[s_2] = 0$ (whence $\mathcal{D}_n$ is met) it suffices to show

$$(**) \qquad B[s_2] \upharpoonright s_2 = B \upharpoonright s_2 \text{ and } A_2[s_2] \upharpoonright s_2 = A_2 \upharpoonright s_2 .$$

This is shown as $(*)$. We only have to note that σ acts at stage $s_2 + 1$ for the last time and that it does not put any numbers into B or A_2 at this stage. (Also note that the $\gamma(y)[s_2]$ which σ enumerates into A_0 have been lifted at stage s_1 already whence they cannot injure the computation $\Psi^{A_0}(x)$.)

Case 2. Otherwise. Then $\Psi^{A_0}(x) \neq 0$ and x never enters A_1. So $A_1 \neq \Psi^{A_0}$ whence $\mathcal{D}_1^{\Psi}$ is met. ■

This completes the proof of Theorem 2.

REFERENCES

1. K. AMBOS-SPIES, "On the Structure of the Recursively Enumerable Degrees", Doctoral Dissertation, University of Munich, 1980.
2. K. AMBOS-SPIES, A. H. LACHLAN, R. I. SOARE, *The continuity of cupping to* $\mathbf{0}'$, Ann. Pure Appl. Logic (to appear).
3. K. AMBOS-SPIES, M. LERMAN, *Lattice embeddings into the recursively enumerable degrees*, J. Symbolic Logic 51 (1986), 257–272.
4. K. AMBOS-SPIES, M. LERMAN, *Lattice embeddings into the recursively enumerable degrees, II*, J. Symbolic Logic 54 (1989), 735–760.
5. P. A. FEJER, *The density of the nonbranching degrees*, Ann. Pure Appl. Logic 24 (1983), 113–130.
6. A. H. LACHLAN, *Lower bounds for pairs of recursively enumerable degrees*, Proc. London Math. Soc. 16 (1966), 537–569.
7. A. H. LACHLAN, *Embedding nondistributive lattices in the recursively enumerable degrees*, in "Conference in Mathematical Logic, London, 1970", Lecture Notes in Mathematics No. 255, SpringerVerlag, Berlin, Heidelberg, New York, 1972, pp. 149–177.
8. A. H. LACHLAN, *Decomposition of recursively enumerable degrees*, Proc. Amer. Math. Soc. 79 (1980), 629–634.
9. A. H. LACHLAN, R. I. SOARE, *Not every finite lattice is embeddable in the recursively enumerable degrees*, Adv. in Math. 37 (1980), 74–82.
10. J. R. SHOENFIELD, R. I. SOARE, *The generalized diamond theorem*, abstract # 219, Recursive Function Theory Newsletter 19 (1978).
11. T. A. SLAMAN, *The density of infima in the recursively enumerable degrees*, Ann. Pure Appl. Logic 52 (1991), 155–179.
12. R. I. SOARE, "Recursively Enumerable Sets and Degrees", Perspectives in Mathematical Logic, Omega Series, Springer-Verlag, Berlin, Heidelberg, New York, London, Paris, Tokyo, (1987).
13. S. K. THOMASON, *Sublattices of the recursively enumerable degrees*, Z. Math. Logik Grundlag. Math. 17 (1971), 272–280.
14. C. E. M. YATES, *A minimal pair of recursively enumerable degrees*, J. Symbolic Logic 31 (1966), 159–168.

Logic, Methodology and Philosophy of Science IX
D. Prawitz, B. Skyrms and D. Westerståhl (Editors)

CONTRIBUTIONS TO THE HISTORY OF VARIATIONS OF WEAK DENSITY IN THE n-R.E. DEGREES

MARAT M. ARSLANOV

Department of Mathematics, University of Kazan

A set $A \subseteq \omega$ is called n-r.e. if $n = 1$ and A is recursively enumerable (r.e) or $n > 1$ and there is an r.e. set A_1 and an n-r.e. set $A_2 \subseteq A_1$ such that $A = A_1 - A_2$. A Turing degree is called a n-r.e. degree if it contains a n-r.e. set; it is called properly n-r.e. if it is n-r.e. but not $n-1$-r.e. Clearly a set A is n-r.e. if and only if there is a recursive function f such that for all $x \lim_s f(s,x) = A(x), f(0,x) = 0$ and

$$\text{card}\{s : f(s+1,x) \neq f(s,x)\} \leq n. \tag{1}$$

In the obvious way, a set $A \subseteq \omega$ is called ω-r.e. iff it satisfies the same definition where (1) is replaced by

$$\text{card}\{s : f(s+1,x) \neq f(s,x)\} \leq h(x) \tag{1'}$$

for some recursive h. The reader should note that if a set A satisfies (1') for some recursive h then it satisfies (1') for any recursive unbounded function g (see [1]).

The existence of properly α-r.e. degrees was first proved for $1 < \alpha < \omega$ by Cooper [2] and for $\alpha = \omega$ by Epstein [5] and Lachlan (1968, unpublished), who showed that there is an ω-r.e. minimal degree, and that every nonrecursive n-r.e. degree for $1 < n < \omega$ bounds a nonrecursive r.e. degree, respectively.

During the past decade, an intensive study of the structure of n-r.e. (and more particularly d-r.e.=2-r.e.) degrees was initiated. Interest in the n-r.e. degrees stems from their affinity with the r.e. degrees, although a number of recent papers have sought several essential differences between these structures. Probably the most fundamental result in this direction is the Cooper-Harrington-Lachlan-Lempp-Soare Nondensity Theorem, which states that the partial orderings of n-r.e. degrees for any $n > 1$ are not dense.

THEOREM 1 (Cooper, Harrington, Lachlan, Lempp, Soare [3]) *There exists a d-r.e. degree* $\underset{\sim}{d} < \underset{\sim}{0}'$ *with no n-r.e. (for all* $n \geq 1$*) such that* $\underset{\sim}{d} < \underset{\sim}{a} < \underset{\sim}{0}'$.

Some authors studied variations of the next "weak" density problem for n-r.e.degrees. Let $n \geq 1$ and $I(n)$ denote the set of integers k such that for any two n-r.e.degrees $\underset{\sim}{a} < \underset{\sim}{b}$ there is a properly k-r.e.degree $\underset{\sim}{c}$ such that $\underset{\sim}{a} < \underset{\sim}{c} < \underset{\sim}{b}$. The question is whether $k \in I(n)$ for several k, n. The next two results assert that for any $n > m \geq 1$, $m \notin I(n)$ and $2 \in I(1)$, respectively.

THEOREM 2. (Hay and Lerman [6]) *For all* n, m *such that* $n > m \geq 1$, *there exist n-r.e.degrees* $\underset{\sim}{a} < \underset{\sim}{b}$ *such that there is no m-r.e.degree between them.*

THEOREM 3 (Cooper, Lempp, Watson [4]) *Given r.e. degrees* $\underset{\sim}{a} < \underset{\sim}{b}$ *there exists a properly d-r.e. degree* $\underset{\sim}{c}$ *such that* $\underset{\sim}{a} < \underset{\sim}{c} < \underset{\sim}{b}$.

Further, the Nondensity Theorem for n-r.e. degrees asserts that for any $n \geq 2$ and $k \geq 1$, $k \notin I(n)$. Therefore, the only open question in this area is whether $n \in I(1)$ for all $n > 2$.

Our notation generally follows Soare [8].

1. The theorems

In this paper we study the next related question. Let $\underset{\sim}{a}$ and $\underset{\sim}{b}$ be d-r.e. degrees such that $\underset{\sim}{a} < \underset{\sim}{b}$ and either $\underset{\sim}{a}$ or $\underset{\sim}{b}$ is r.e. In which cases do r.e. degrees $\underset{\sim}{c}$ exist such that $\underset{\sim}{a} < \underset{\sim}{c} < \underset{\sim}{b}$?

Before presenting our results, we survey the related results known to us. We first consider the case when $\underset{\sim}{a}$ is r.e.

THEOREM 4. (Lachlan, unpublished) *For any n-r.e.* $\underset{\sim}{b} > 0$ *there is a r.e. degree* $\underset{\sim}{c}$ *such that* $\underset{\sim}{0} < \underset{\sim}{c} < \underset{\sim}{b}$.

THEOREM 5. (Kaddach, unpublished) *There are r.e.* $\underset{\sim}{a}$ *and d-r.e.* $\underset{\sim}{b}$ *such that* $\underset{\sim}{a} < \underset{\sim}{b}$ *and there is no r.e. degree* $\underset{\sim}{c}$ *between them.*

Assume now that $\underset{\sim}{a}$ is d-r.e. and $\underset{\sim}{b}$ is r.e., $\underset{\sim}{a} < \underset{\sim}{b}$.

LEMMA 1. (Cooper, Harrington, Lachlan, Lempp, Soare [3]) *There exists a d-r.e. degree* $\underset{\sim}{a} < \underset{\sim}{0}'$ *with no n-r.e.* $\underset{\sim}{c}$ *such that* $\underset{\sim}{a} < \underset{\sim}{c} < \underset{\sim}{0}'$ *for any* $n \leq \omega$.

This result compares with theorem 7 below. For the case $\underset{\sim}{c}$ r.e. it was first obtained by Hay and Shore (unpublished).

The next theorem also follows from known results.

THEOREM 6. *For any high r.e. degree* $\underset{\sim}{b} < \underset{\sim}{0}'$ *there exists a d-r.e. degree* $\underset{\sim}{a} < \underset{\sim}{b}$ *with no r.e. degree* $\underset{\sim}{c}$ *such that* $\underset{\sim}{a} < \underset{\sim}{c} < \underset{\sim}{b}$.

PROOF: Let $\underset{\sim}{b} < \underset{\sim}{0}'$ be an r.e. degree such that $\underset{\sim}{b}' = \underset{\sim}{0}''$. Harrington proved (see [7]) that every high r.e. degree has the anticupping property: there is a nonzero r.e. degree $\underset{\sim}{d} < \underset{\sim}{b}$ such that for no r.e. $\underset{\sim}{c} < \underset{\sim}{b}$ does $\underset{\sim}{b} = \underset{\sim}{d} \cup \underset{\sim}{c}$. Cooper, Lempp and Watson [4] proved that there is a *d*-r.e. degree $\underset{\sim}{a} < \underset{\sim}{b}$ such that $\underset{\sim}{a} \cup \underset{\sim}{d} = \underset{\sim}{b}$. Obviously, there is no r.e. degree $\underset{\sim}{c}$ between $\underset{\sim}{a}$ and $\underset{\sim}{b}$.

Studying the last question we add to this list of known results the next two theorems.

THEOREM 7. *Given r.e. sets* A, P *and a d-r.e. set* $B = B_1 - B_2$ *such that* $A <_T B, B|_T P$ *and* $B_2 \leq_T P$, *there is an* ω*-r.e. set* X *such that* $A <_T X <_T B$.

We have noted above that we can choose the sets A and B so that in this theorem X cannot be r.e. A very interesting open question is the following: is it possible to generalize the theorem so that the condition "$X\omega$-r.e." is replaced by "Xn-r.e." for some $n, 1 < n < \omega$?

THEOREM 8. *Given an r.e. set* $A <_T \emptyset'$, $A' \equiv_T \emptyset''$, *there is a d-r.e. set* B *such that* $A <_T B$ *and there is no r.e. set* C *such that* $A <_T C <_T B$.

Theorems 6 and 8 imply the following interesting

COROLLARY. *For any high r.e. degree* $\underset{\sim}{a} < \underset{\sim}{0}'$ *there are d-r.e. degrees* $\underset{\sim}{b}$ *and* $\underset{\sim}{c}$ *such that* $\underset{\sim}{b} < \underset{\sim}{a} < \underset{\sim}{c}$ *and* $\underset{\sim}{a}$ *is the single r.e. degree between* $\underset{\sim}{b}$ *and* $\underset{\sim}{c}$. *(Of course, it means that any high r.e. degree* $\underset{\sim}{a}$ *is definable in* $\mathcal{R}$ *(the structure of the r.e. degrees) using two parameters from* D_2 *(the set of d-r.e. degrees) by the formula* $\Phi(x, b, c) = b < x < c$.)

2. The proof of Theorem 7

We first note that given an r.e. set A and a *d*-r.e. set B such that $A <_T B <_T \emptyset'$, it is easy to prove the existence of an r.e. set P such that $A <_T P$

and $B|_T P$. Indeed, it suffices to construct such a set $P = A \oplus Q$ meeting for all e the requirements:

$$S_e : B \neq \Phi_e^{A \oplus Q};$$
$$T_e : Q \neq \Phi_e^A.$$

The overall requirements will ensure that either $A <_T P <_T B$ and the theorem is proved with the r.e. set P instead of X, or $A <_T P$ and $P|_T B$.

Let us give a recursive enumeration $\{K_s\}_{s \in \omega}$ of the creative set K. To meet T_e we compute $\Phi_e^A(\langle e, x\rangle)$ for any $x \in K_s$. If for some $t \geq s$ and $\phi \subset \omega$ $\Phi_{e,s}^{A_s \restriction \phi}(\langle e, x\rangle) \downarrow = 0$ then $\langle e, x\rangle \in Q_t$, otherwise $\langle e, x\rangle \notin Q$. Obviously, either there is an x such that $\Phi_e^A(\langle e, x\rangle) \neq Q(\langle e, x\rangle)$, or for any x we have $x \in K_s \,\&\, \Phi_{e,s}^{A_s \restriction \phi}(\langle e, x\rangle) \downarrow$ for the least $s \to A_s \restriction \phi \neq A \restriction \phi$. Thus, we have $x \in K \leftrightarrow x \in K_{s'}$, where $s' = \mu s\{\exists \phi \subset \omega \left(\Phi_{e,s}^{A_s \restriction \phi}(\langle e, x\rangle) \downarrow \,\&\, As = A \restriction \phi\right)\}$ and K is hence recursive in A.

The strategy for meeting S_e (cycle k) is next.

1. Fix a length of agreement $x > \langle e, k\rangle$.
2. Wait for a stage s at which
$$B_s \restriction x = \Phi_{e,s}^{(A_s \oplus Q_s) \restriction \phi} \restriction x \text{ for some } \phi.$$
3. Start cycle $k+1$, to run simultaneously. Set
$$B_s(k) = \Delta_e^{A_s \restriction \phi}(k)$$
and restrain $Q \restriction \phi$ from strategies of lower priority.
4. Wait for $B(k)$ to change at some stage $t > s$.
(If $A \restriction \phi$ changes between stages s and t, then go back to 2.)
5. Cancel cycles greater than k. Wait for a stage $u > t$ at which a new agreement $B_u \restriction x = \Phi_{e,u}^{(A_u \oplus Q_u) \restriction \phi'} \restriction x$ is achieved for some $\phi' \supseteq \phi$. (Obviously, this happens only if $A_u \restriction \phi \neq A_s \restriction \phi$).
6. Set $B(k) = B_u(k) = \Delta_e^{A_u \restriction \phi}(k)$. Cancel the restraint of this cycle and start cycle $k+1$.

There are three possible outcomes of this strategy.

A. Some cycle waits forever at 2 or 5. It clearly means that $B \neq \Phi_e^{A \oplus Q}$ and we meet S_e.

B. There are infinitely many cycles, and every cycle finishes in 4 or 6. It obviously means that $B \leq_T A$ which contradicts the condition. Therefore, S_e cannot have this outcome.

C. There are infinitely many cycles, and $A \restriction \phi$ changes at any time between stages s and t. It means that $\Phi_e^{A \oplus Q}$ is partial and we again meet S_e.

It is easy to see that S_e as well as $T_e, e \in \omega$, causes finite injury to lower priority requirements. So, the whole construction is a finite injury argument on the priority orderings $S_0, T_0, \ldots, S_e, T_e, \ldots$.

Therefore, let P be an r.e. set such that $P|_T B$ and $\{P_s\}_{s\in\omega}$ be a recursive enumeration of P. We construct an ω-r.e. set $M \leq_T B$ meeting for all $e \in \omega$ the requirements

$$U_e : M \neq \Phi_e^A;$$
$$V_e : B \neq \Phi_e^{A\oplus M}.$$

Then $X = A \oplus M$ will be the desired ω-r.e. set.

The priority ranking of requirements is $U_0, V_0, \ldots, U_e, V_e, \ldots$.

To ensure that $M \leq_T B$ we use the permitting argument. The strategy for V_e proceeds as follows (cycle k).

1. Choose an $x > \langle e, k\rangle$.
2. Wait for a stage s at which $B_s \upharpoonright x = \Phi_{e,s}^{(A_s\oplus M_s)\upharpoonright\phi} \upharpoonright x$ for some ϕ.
3. Start cycle $k+1$, to run simultaneously. Set
$$B_s(k) = \Gamma_e^{A_s\upharpoonright\phi}(k).$$
Restrain $M \upharpoonright \phi$ from other strategies with priority S_e.
4. Wait for $B(k)$ to change at some stage $t > s$.
5. Cancel cycles $> k$. Wait for a stage $u \geq t$ at which a new agreement
$$B_u \upharpoonright x = \Phi_{e,u}^{(A_u\oplus M_u)\upharpoonright\phi'} \upharpoonright x$$
is achieved for some ϕ'. (Notice that this can happen only if $A_s \upharpoonright \phi \neq A_u \upharpoonright \phi$.)
6. Cancel the restraint of $M \upharpoonright \phi$.
7. If $B_u(k) = 0$ (thus $B(k) = B_u(k)$) then set $B(k) = \Gamma_e^{A_u\upharpoonright\phi'}(k)$. Start cycle $k+1$.
8. If $B_u(k) = 1$ then restrain $M \upharpoonright \phi'$, set $B_u(k) = \Gamma_e^{A\upharpoonright\phi'}(k)$ and start cycle $k+1$, to run simultaneously.
9. Wait for $B(k)$ to change at some stage $v > u$.
10. Cancel cycles $> k$. Wait for a stage $w \geq v$ at which a new agreement
$$B_w \upharpoonright x = \Phi_{e,w}^{(A_w\oplus M_w)\upharpoonright\phi''} \upharpoonright x$$
is achieved for some ϕ''. (Again, this can happen only if $A_v \upharpoonright \phi' \neq A_w \upharpoonright \phi'$.)
11. Cancel the restraint of $M \upharpoonright \phi'$ and set $B(k) = B_w(k) = \Gamma_e^{A\upharpoonright\phi'}(k)$. Start cycle $k+1$.

There are three possible outcomes:

A. Some cycle waits forever at 2, 5 or 10. Then clearly $B \neq \Phi_e^{A \oplus M}$ and we meet V_e.

B. There are infinitely many cycles, and every cycle finishes in 4, 9 or 11. Then $B \leq_T A$, which contradicts the condition. Therefore V_e can not have this outcome.

C. There are infinitely many cycles, and $A \upharpoonright \phi$ or $A \upharpoonright \phi'$ changes at any time between stages s and t or between stages u and v, respectively. Obviously, this means that Φ_e^A is not total and we again meet V_e.

INJURY LEMMA for V_e. *Any V_e-strategy causes finite injury to lower priority requirements.*

PROOF: If V_e has outcome (A) then each of its finitely many cycles causes finite injury to lower priority requirements. If V_e has outcome (C) then each of its cycles creates restraints which later are cancelled. Therefore, V_e again causes finite injury to lower priority requirements.

Strategy for U_e (cycle k).

1. Wait for a stage s at which $k \in P_s$. Start cycle $k+1$ to run simultaneously.
2. Wait for a stage $t \geq s$ at which $\Phi_{e,t}^{A_t}(\langle e,k \rangle) \downarrow$.

3a. If $\Phi_{e,t}^{A_t}(\langle e,k \rangle) = 1$ then restrain $\langle e,k \rangle$ from X and close this cycle.

3b. If $\Phi_{e,t}^{A_t}(\langle e,k \rangle) = 0$, then

4. Wait for $B \upharpoonright \langle e,k \rangle$ to change at some stage $u \geq s$.
5. Put $\langle e,k \rangle$ into X if it is not restrained by some V_i of higher priority. (If $\langle e,k \rangle$ is restrained by some $V_i, i < e$, then

$5\frac{1}{2}$. Wait for a stage $u' > u$ at which the restraint of V_i is lifted and then go to 5.)

6. Wait for a stage $v > u$ at which again $B_v \upharpoonright \langle e,k \rangle = B_{t'} \upharpoonright \langle e,k \rangle$ for some $t' \leq t$.
7. Remove $\langle e,k \rangle$ from X and go to 4 with $s = v$ if $\langle e,k \rangle$ is not restrained by some $V_i, i < e$. (If $\langle e,k \rangle$ is restrained by $V_i, i < e$, then

$7\frac{1}{2}$. Wait for a stage $v' > v$ at which the restraint of V_i is lifted and then go to 7).

LEMMA 2. *$X \leq_T B$ and X is ω-r.e.*

PROOF: Obviously, by the construction we have $X \leq_T B$ and X is ω-r.e.: there exist not more than $2\langle e,k\rangle + 2$ stages s such that $B_s \upharpoonright \langle e,k\rangle \neq B_{s+1} \upharpoonright \langle e,k\rangle$. Therefore, for any k, $X(\langle e,k\rangle)$ changes its value not more than $2\langle e,k\rangle + 2$ times.

LEMMA 3. *$X \leq_T A$.*

PROOF: Assume that $X = \Phi_e^A$ for some e and choose the least k such that any $x \geq k$ is never injured. Then $\exists^\infty x\{x \in P_s \,\&\, \Phi_{e,s}^{A_s \upharpoonright \phi} \downarrow$ for some least $s \in \omega$ and $\phi \subset \omega \,\&\, (A_s \upharpoonright \phi = A \upharpoonright \phi)\}$. (Obviously otherwise we have $P \leq_T A$). If $\forall x \geq k\{[x \in P_s \,\&\, \Phi_{e,s}^{A_s \upharpoonright \phi}(\langle e,x\rangle) \downarrow = 0$ for some least s and $\phi \,\&\, (A_s \upharpoonright \phi = A \upharpoonright \phi)] \to B_s \upharpoonright \langle e,x\rangle = B \upharpoonright \langle e,x\rangle\}$, then obviously we have $B \leq_T A \oplus P \equiv_T P$. Therefore, for some $x \geq k$ we have $x \in P_s \,\&\, \Psi_{e,s}^{A_s \upharpoonright \phi}(\langle e,x\rangle) \downarrow$ for the least $s, \phi \to A_s \upharpoonright \phi = A \upharpoonright \phi$ and $\Phi_{e,s}^{A_s \upharpoonright \phi}(\langle e,x\rangle) = 0 \to B_s \upharpoonright \langle e,x\rangle \neq B \upharpoonright (\langle e,x\rangle)$. This means that at the cycle x of the U_e-strategy we meet U_e.

LEMMA 4. *$B \not\leq_T A \oplus X$.*

PROOF: At the end of the description of the V_e-strategy we have seen that if the strategy for U_e is finite injury then for any e we meet V_e.

The above lemmas establish the theorem.

3. The proof of Theorem 8

Let A be an r.e. set, $A' \equiv_T \emptyset''$ and $\emptyset' \not\leq_T A$. We construct a d-r.e. set B meeting for all $e \in \omega$ the requirements

$$T_e : B \neq \Phi_e^A;$$
$$S_e : W_e = \Phi_e^{A \oplus B} \quad \to \quad W_e = \Gamma_e^A.$$

Then $A \oplus B$ will be the desired set.

In satisfying S_e we shall construct a functional Γ_e. The priority ranking of the requirements is $T_0, S_0, T_1, S_1, \ldots, T_e, S_e, \ldots$.

The strategy for T_e is the same as that for T_e of Theorem 7. Let us now consider the requirement S_e. We may assume that A is e-dominant, namely, the computation function c_A defined by

$$c_A(x) = \mu s\{A_s \upharpoonright x = A \upharpoonright x\}$$

dominates every total recursive function f, i.e.,

$$\exists x_0 \forall x \geq x_0\{c_A(x) > f(x)\}.$$

The strategy for S_e (cycle k) proceeds as follows:

1. Choose an $x > \langle e, k \rangle$.
2. Wait for a stage s at which

(x) $$W_{e,s} \upharpoonright x = \Phi_{e,s}^{(A_s \oplus B_s)\upharpoonright \phi} \upharpoonright x \text{ for some } \phi.$$

3. Start cycle $k+1$, to run simultaneously. Set

$$W_{e,s}(k) = \Gamma_e^{A_s \upharpoonright \phi_1}(k),$$

where $\phi_1 \geq$ any ϕ used to stage s of the construction.

4. Wait for $W_e(k)$ to change at some stage $t > s$.
(Note: If $A \upharpoonright \phi$ or $B \upharpoonright \phi$ changes between stages s and t, go back to 2.)
5. Restrain $B \upharpoonright \phi$ from other strategies of lower priority from now on, cancel cycles greater than k, and
6. Wait for a stage $u > t$ at which a new agreement

(xx) $$W_{e,u} \upharpoonright x = \Phi_{e,u}^{(A_u \oplus B_u)\upharpoonright \phi'} \upharpoonright x$$

is achieved for some ϕ'.

7a. If (xx) is achieved because of $A_u \upharpoonright \phi_1 \neq A_t \upharpoonright \phi_1$ (clearly we have $W_{e,t} \upharpoonright x = \Phi_{e,t}^{(A_t \oplus B_t)\upharpoonright \phi} \upharpoonright x$, see Note at 4.) then set

$$W_{e,u}(k) = W_e(k) = \Gamma_e^{A_u \upharpoonright \phi_1}(k),$$

cancel the restraint of $B \upharpoonright \phi$ and start cycle $k+1$.

7b. Suppose now that $A_u \upharpoonright \phi_1 = A_t \upharpoonright \phi_1$. Therefore we have $B_t \upharpoonright \phi \neq B_u \upharpoonright \phi$. There are the following possible cases.

$7b_1$. Some b is removed from B at stage $t', t < t' < u$. This means (see $7b_3$ below) that either some requirement $S_i, i < e$, of higher priority is satisfied at some stage k' because $W_i(k') \neq \Phi_e^{(A \oplus B)\upharpoonright \psi'}(k')$, or later $A \upharpoonright \psi'$ will be changed. By construction we restrain $B \upharpoonright \psi'$ in the S_i-strategy at some stage $< t'$ hence $\phi_1 > \phi'$ (see 3.). Therefore, in this case we must wait for $A \upharpoonright \psi' (\subseteq A \upharpoonright \psi_1)$ to change at some stage $v > u$ and then go to 8.

$7b_2$. $B_u \supseteq B_t$, and $\exists b \leq \phi_1 (b \in B_u - B_t$ and b is restrained by some S_i-requirement with restraint $B \upharpoonright \phi')$. In this case for all such b

we wait for $A \upharpoonright \psi'$ to change at some stage $v > u$ and then go to $7b_3$.

$7b_3$. $B_u \supseteq B_t$, and $\forall b$ ($b \in B_u - B_t \rightarrow b$ is not restrained by requirements of higher priority). Remove all these b from B and wait for $A \upharpoonright \phi'$ to change at some stage $v > u$.

8. Set $W_e(k) = W_{e,v}(k) = \Gamma_e^{A \upharpoonright \psi_1}(k)$. Cancel restraints of S_e and start cycle $k+1$.

The possible outcomes for S_e.

A. Some cycle waits forever at 2. or 6. Clearly, then $W_e \neq \Phi_e^{A \oplus B}$ and we meet S_e.

B. There are infinitely many cycles, and every cycle (beginning for some cycle k) finishes in 4., 7a. or 8. Then clearly $W_e \leq_T A$ and we again meet S_e.

C. There are infinitely many cycles which finish in $7b_1$. This means that any of their $A \upharpoonright \psi'$ is not changed after stage u. Hence, some requirement $S_i, i < e$, is satisfied at stage u. But there are finitely many requirements of higher priority, therefore we cannot have this outcome.

D. There are infinitely many cycles which finish in $7b_2$. To prove that we again cannot have this outcome we define by induction a total recursive function h_e which contradicts that A is e-dominant.

Suppose x_0 is the greatest integer such that $h_e(x)$ is defined for any $x \leq x_0$. Define $h_e(x_0 + 1) = u$ (the stage where a new agreement is achieved, see point 6. of the description of the S_e-strategy). Clearly if h_e is not total then we do not come to 6. from 4. at some cycle k. Therefore, $W_e \neq \Phi_e^{A \oplus B}$ and we meet S_e.

Let $b \leq \psi_1$ be an integer which is enumerated in B at some stage $t' > t$ (see point 4. of the description of the S_e-strategy) and after that at point 3. of cycle k' of some S_i-strategy, i.e., creates a $B \upharpoonright \phi^*$-restraint with $\phi^* \geq b$. (Otherwise, if it creates a $B \upharpoonright \phi^*$-restraint at stage $t' < t$, b may be removed from B without problems.) To force $A \upharpoonright \phi^*$ to change we want to have $x_0 + 1 \leq \phi^*$ and we slightly modify the definition of ϕ_1 in 3.: ϕ_1 must be greater than $\tilde{x}$, where $\tilde{x} = \max\{x : h_e(x)$ is defined at some stage $< s$ and for some $e\}$.

Now it is obvious that the existence of infinitely many cycles which finish in $7b_2$ means that h_e is total and c_A does not majorize h_e.

Therefore, there are only two possible outcomes (A) and (B) for the S_e-strategy. If the strategy S_e has outcome (A), then it causes finite injury to lower priority requirements. If it has outcome (B), then each of its cycles

creates restraints which later are cancelled. Therefore, S_e again causes finite injury to lower priority requirements.

REFERENCES

1. M. M. ARSLANOV, *On the structure of degrees below* $0'$, Recursion Theory Week (Oberwolfach, 1989), Lecture Notes in Math., 1432, Springer, Berlin, 1990, 23–32.
2. S. B. COOPER, *Degrees of Unsolvability*, Ph. D. Thesis, Leicester University, Leicester, 1971.
3. S. B. COOPER, L. HARRINGTON, A. H. LACHLAN, S. LEMPP, R. I. SOARE, *The d-r.e. degrees are not dense*, to appear.
4. S. B. COOPER, S. LEMPP, P. WATSON, *Weak density and cupping in the d-r.e. degrees*, Israel J. Math. 67, 1989, 137–152.
5. R. L. EPSTEIN, *Degrees of Unsolvability: Structure and Theory*, Lecture Notes in Math., 759, Springer, Berlin, 1979.
6. L. HAY, M. LERMAN, *On the degrees of Boolean combinations of r.e. sets*, Recursive Function Theory Newsletter, 1976.
7. D. MILLER, *High recursively enumerable degrees and the anticupping property*, Logic Year, 1979–80, Lecture Notes in Math., 859, Springer, Berlin, 1981, 230–245.
8. R. I. SOARE, *Recursively Enumerable Sets and Degrees*, Springer, Berlin, 1987.

Logic, Methodology and Philosophy of Science IX
D. Prawitz, B. Skyrms and D. Westerståhl (Editors)

RIGIDITY AND DEFINABILITY IN THE NONCOMPUTABLE UNIVERSE[1]

S. BARRY COOPER

University of Leeds, Leeds LS2 9JT, England

§ 1. The noncomputable universe

It followed from Gödel [1931], [1934] that most functions are not effectively computable and most interesting mathematical theories are undecidable. This led to an awareness of, and a growing interest in, a *noncomputable universe* intimately connected with the world of everyday mathematics. This noncomputability is of a fundamental nature, and does not arise from mere practical limitations such as those on capacity of memory or duration of computational processes. An important aim of recursion theory is to investigate the context of interesting mathematical objects (for example, Gödel's undecidable theories) within the noncomputable universe, and Kleene and Post [1954] proposed the *degrees of unsolvability* (or *Turing degrees*) as an appropriate theoretical framework, or fine structure theory, within which to do this.

The subsequent development of *local* degree theory (concerned with that part of the noncomputable universe already manifest in the work of Gödel) was largely based on an autonomy of interest and motivation through which evolved elegant techniques and striking results, while its general impact amongst mathematicians and computer scientists was limited by its seeming preoccupation with pathology and technique of ever more prohibitive complexity. The need to match this complexity with an understanding of the wider significance of the theory has meant a more recent emphasis (already apparent in Lerman [1980]) on the *global* theory of the Turing degrees (focused largely on a number of basic questions raised by Rogers [1967] and by Kleene and Post [1954]).

[1]The author received support from SERC Research Grants nos. GR/F 42003 and GR/H 02165 during the preparation of this article.

There are of course a number of notions on which to base a useful fine structure theory, of which two are especially important.

DEFINITION 1.1 (Turing [1939], Kleene [1943], Post [1943] etc.). *Let A, B be sets of numbers. Then*
(1) (Many-one reducibility) $A \leq_m B \Leftrightarrow A = f^{-1}(B)$, some recursive f.
(2) (Turing reducibility) $A \leq_T B \Leftrightarrow \exists$ an oracle Turing machine T which computes χ_A using an oracle for B.

Many-one reductions are historically significant as the recursion theoretic analogues of natural translations between formal theories (see for example Davis [1958]), while Turing reducibility is thought to include, essentially, all possible notions of effective computability relative to oracles.

§ 2. Basic structure theory

We use standard notation and terminology (see for example Soare [1987]).

For instance, corresponding to the ith Turing machine, Φ_i denotes the ith partial recursive (p.r.) functional $2^\omega \to 2^\omega$, so that $A \leq_T B$ if and only if $A = \Phi_i^B$ for some $i \in \omega$.

DEFINITION 2.1 (Post [1944], Kleene-Post [1954]). *A, B are Turing equivalent ($A \equiv_T B$) if and only if $A \leq_T B$ and $B \leq_T A$. The degree of unsolvability or Turing degree of A is defined by*

$$deg(A) = \{X \in 2^\omega \mid A \equiv_T X\}.$$

We write $\leq$ for the partial ordering on $\mathcal{D}$, the set of all degrees, $\mathbf{0}$ for the least degree, consisting of all recursive sets of numbers, and $\boldsymbol{\mathcal{D}}$ for the structure $\langle \mathcal{D}, \leq \rangle$.

To notate the analagous degree theoretic notions derived from many-one reducibility in place of Turing reducibility it is usual to append a subscript m. For instance $\boldsymbol{\mathcal{D}}_m$ denotes the structure of all m-degrees with the ordering induced by $\leq_m$.

The most important nonrecursive degree $\mathbf{0}'$ is that containing the (coded) undecidable axiomatic theories of Gödel, as well as many other natural mathematical objects. When relativised to an arbitrary set A of degree $\mathbf{a}$, it gives rise to a *jump operator* taking $\mathbf{a}$ to a strictly higher degree $\mathbf{a}'$, defined as the largest degree containing sets which are effectively enumerable (or *recursively enumerable* , written r.e.) using oracle A. Post's Theorem [1944] showed a close relationship between the quantifier forms of most naturally occuring sets of numbers and the ascending sequence $\mathbf{0} < \mathbf{0}' < \mathbf{0}'' < \ldots < \mathbf{0}^{(n+1)} = (\mathbf{0}^{(n)})' < \ldots$

($\mathbf{0}$ being the degree of the recursive sets). Feferman [1957] and Shoenfield [1958] showed the r.e. degrees to be exactly those degrees containing (coded) recursively axiomatisable first-order theories.

DEFINITION 2.2. *Let $W_i^A = dom\,\Phi_i^A$ denote the ith recursively enumerable in A (A-r.e.) set ($W_i = W_i^\phi$ being the ith r.e. set). Then the jump ($n+1$th jump) of a set A is defined by $A' = A^{(1)} = \{x \mid x \in W_x^A\}$ ($A^{(n+1)} = (A^{(n)})'$).*

The jump operator on degrees is defined by $\mathbf{a}' = deg(A')$, $A \in \mathbf{a}$, where $\mathbf{a} < \mathbf{a}'$, and $\mathbf{a}'$ is the least upper bound of the degrees of sets r.e. in $A \in \mathbf{a}$. We also write $\mathbf{a}^{(n+1)} = deg(A^{(n+1)}) = (\mathbf{a}^{(n)})'$. We define the standard ω-jump of $\mathbf{a}$ by $\mathbf{a}^{(\omega)} = deg(\oplus_{n\in\omega} A^{(n)})$, $A \in \mathbf{a}$.

We write $\mathcal{D}'$ for the structure $\langle \mathcal{D}, \leq, ' \rangle$, and $\mathcal{R}$ for the structure of the r.e. degrees.

We assume that we have standard recursive sequences $\{\Phi_{i,s}\}_{s\geq 0}$, $\{W_{i,s}^A\}_{s\geq 0}$ of finite approximations to the p.r. functionals and r.e. sets, respectively. We denote by $\mathcal{A}[s]$ the corresponding approximation to an expression $\mathcal{A}$ at a stage s. A superscript s may also be used to convert a particular set, function or relation to its sth-stage approximation.

§ 3. Questions

Given an oracle A, an examination of its context within the noncomputable universe can be approached via two general and related questions:

I. *Which relations on $\mathcal{D}$ are (first-order)* definable *in $\mathcal{D}$ or $\mathcal{D}'$?*
II. *Which relations on $\mathcal{D}$ are unchanged under all automorphisms of $\mathcal{D}$, i.e., are* order-theoretic*?*

In particular, we have:

3.1. (Kleene-Post [1954]) *Are $\mathbf{0}'$ and the jump operator definable in $\mathcal{D}$?*
3.2. (Rogers [1967]) *Are $\mathbf{0}'$ and the jump operator order-theoretic?*
3.3. (Slaman-Woodin [1986]) *Is $\mathcal{R}$ definable in $\mathcal{D}$?*
3.4. (Rogers [1967]) *Is the relation of "r.e. in" definable in $\mathcal{D}$/order theoretic?*

Since the notion of a relation being order-theoretic is not language based it is more widely applicable than that of definability, but it may be trivial without an answer to the fundamental question:

III. *Do there exist any non-trivial automorphisms of $\mathcal{D}$ or $\mathcal{D}'$; i.e., Is $\mathcal{D}$ or $\mathcal{D}'$* rigid*?*

Related to the question of rigidity are those of *homogeneity*:

3.5. (Yates [1970]) *For which* $\mathbf{a}, \mathbf{b} \in \mathcal{D}$ *is* $\boldsymbol{\mathcal{D}}(\geq \mathbf{a}) \equiv \boldsymbol{\mathcal{D}}(\geq \mathbf{b})$ *or* $\boldsymbol{\mathcal{D}}'(\geq \mathbf{a}) \equiv \boldsymbol{\mathcal{D}}'(\geq \mathbf{b})$*?*

and of *strong homogeneity*:

3.6. (Rogers [1967]) *For which* $\mathbf{a}, \mathbf{b} \in \mathcal{D}$ *is* $\boldsymbol{\mathcal{D}}(\geq \mathbf{a}) \cong \boldsymbol{\mathcal{D}}(\geq \mathbf{b})$ *or* $\boldsymbol{\mathcal{D}}'(\geq \mathbf{a}) \cong \boldsymbol{\mathcal{D}}'(\geq \mathbf{b})$*?*

Of course, analagous versions of the above questions exist for the other known degree structures, in particular for the many-one degrees $\boldsymbol{\mathcal{D}}_m$.

§ 4. Rigidity and the many-one degrees

In contrast to the Turing degrees, initial segments of $\boldsymbol{\mathcal{D}}_m$ can be constructed by successive extensions. The elements of the technique for doing this appears in the simplest such construction:

PROPOSITION 4.1. (Lachlan [1972]) *Every m-degree* $\mathbf{a}$ *has a strong minimal cover (that is, a degree* $\mathbf{b}$ *such that* $\boldsymbol{\mathcal{D}}_m(\leq \mathbf{a}) = \boldsymbol{\mathcal{D}}_m(< \mathbf{b})$*).*

This leads to a striking characterisation of the structure of the many-one degrees. $\boldsymbol{\mathcal{D}}_m$ is fixed by a few known basic properties and the special feature allowing extensions of ideals of $\boldsymbol{\mathcal{D}}_m$ isomorphic to any 'reasonable' ideal. Call a partially ordered set L an *m-ideal* if and only if

1. L is a distributive uppersemilattice with least element.
2. L has the countable predecessor property.

Then

THEOREM 4.2. (Ershov [1975], Paliutin [1975]) $\boldsymbol{\mathcal{D}}_m$ *can be characterised (up to isomorphism) as the only m-ideal with power* $2^{\aleph_0}$ *satisfying:*

($\maltese$) *Any ideal* I *of* $\boldsymbol{\mathcal{D}}_m$ *can be extended to one isomorphic to any m-ideal with power* $< 2^{\aleph_0}$ *which contains an ideal isomorphic to* I*.*

PROOF: Given any two m-ideals L_1, L_2 with power $2^{\aleph_0}$ satisfying ($\maltese$), we can use ($\maltese$) as the basis for a back-and-forth argument which builds an isomorphism between L_1 and L_2.

Theorem 4.2 gives immediate answers for $\boldsymbol{\mathcal{D}}_m$ to all the global questions of the last section.

COROLLARY 4.3 (STRONG HOMOGENEITY). *Any two upper cones of* $\boldsymbol{\mathcal{D}}_m$ *are isomorphic.*

PROOF: Straightforward relativisation shows that any $\boldsymbol{\mathcal{D}}_m(\geq \mathbf{a})$ is an m-ideal with power $2^{\aleph_0}$ satisfying ($\maltese$), so is isomorphic to $\boldsymbol{\mathcal{D}}_m$ by Theorem 4.2.

COROLLARY 4.4 (DEFINABILITY). $\mathbf{0}_m$ *is the only order-theoretic m-degree, so is the only definable m-degree.*

PROOF: Given $\mathbf{a} > \mathbf{0}_m$, find $\mathbf{b} \neq \mathbf{a}$ with $\mathcal{D}_m(\leq \mathbf{a}) \cong \mathcal{D}_m(\leq \mathbf{b})$. Extend this isomorphism to an automorphism of $\mathcal{D}_m$ along the lines of the proof of Theorem 4.2, where $\mathbf{a}$ is not a fixed point.

In fact (since there are $2^{\aleph_0}$ choices of $\mathbf{b}$):

Every order-theoretic/definable set of m-degrees $\neq\{\mathbf{0}_m\}$ *has power of the continuum.*

Finally:

THEOREM 4.5 (AUTOMORPHISMS). (Shore) *There are* $2^{2^{\aleph_0}}$ *automorphisms of* $\mathcal{D}_m$.

PROOF: Note that if $\mathbf{a} = \mathbf{0}_m$ the back-and-forth argument underlying the proof of Strong Homogeneity produces an isomorphism $\mathcal{D}_m \to \mathcal{D}_m$ in $2^{\aleph_0}$ steps. By exercising choice at successor stages in extending the partial isomorphisms build a tree of height $2^{\aleph_0}$ of automorphisms of $\mathcal{D}_m$.

A more detailed discussion of the global characteristics of $\mathcal{D}_m$ can be found in Odifreddi [1989].

§ 5. Rigidity and the Turing degrees

The fact that many-one reductions are unable to use to the full the individual information content of a given oracle is reflected in the fact that $\mathcal{D}_m$ is as far from being rigid as it can be. Nothing is nontrivially definable in $\mathcal{D}_m$, strong homogeneity holds, and there are many automorphisms. The extra subtlety possible with Turing reductions gives a very different, more differentiated kind of degree structure. Firstly, many degrees and classes of degrees are definable in $\mathcal{D}$.

THEOREM 5.1 (DEFINABILITY). *The following are first-order definable in* $\mathcal{D}$, *and so are order-theoretic:*

(a) $\mathbf{0}'$ *(and so* $\mathbf{0}^{(2)}$, $\mathbf{0}^{(3)}$, ... *),*
(b) the jump operator,
(c) *all the* jump classes $High_n$, Low_n, $n > 0$ *(where* $\mathbf{a} \in High_n \Leftrightarrow \mathbf{a} \leq \mathbf{0}' \,\&\, \mathbf{a}^{(n)} = \mathbf{0}^{(n+1)}$ *and* $\mathbf{a} \in Low_n \Leftrightarrow \mathbf{a} \leq \mathbf{0}' \,\&\, \mathbf{a}^{(n)} = \mathbf{0}^{(n)}$*),*
(d) *(Jockusch and Shore [1984])* $\mathcal{A}$ = *the* set of all arithmetical degrees *(*$\mathbf{a} \in \mathcal{A} \Leftrightarrow \mathbf{a} \leq \mathbf{0}^{(n)}$, *some* n*),*
(e) $\mathcal{R}$,[2]

[2]Slaman and Woodin [1986] used codings in the degrees to show $\mathcal{R}$ definable in $\mathcal{D}(\leq \mathbf{0}')$ using a finite number of parameters.

(f) *the relation* 'REA' *(i.e., 'recursively enumerable in and above'), and*
(g) *the relation* 'r.e. in'.[3]

PROOF: For (b) relativise (a) to get a definition of $\mathbf{a}$ in $\mathcal{D}(\geq \mathbf{a})$. (c) follows immediately from (b). Since (Jockusch and Soare [1970]) the set of arithmetical degrees is definable in $\mathcal{D}'$ (using Spector's exact pair theorem) as the smallest jump ideal of $\mathcal{D}'$, we can also get (d) from (b). Standard relativisation of (e) gives (f). We return to (a), (e) and (g) later.

For information on the historical background to Theorem 5.1 see Shore [1981], Lerman [1983], Odifreddi [1989] or Cooper [ta1]. One important consequence of Theorem 5.1 is that results concerning definability in $\mathcal{D}'$ also hold in $\mathcal{D}$.

COROLLARY 5.2. *Any relation on* $\mathcal{D}(\geq \mathbf{0}^{(3)})$ *which is definable in second-order arithmetic is definable in* $\mathcal{D}$.

PROOF: Shore [1982] showed it for $\mathcal{D}'$. Result follows by the definability of the jump in $\mathcal{D}$.

However, the definability of the jump does not help us in defining jump classes in the degrees below $\mathbf{0}'$. Shore [1988] showed that $High_n$ and Low_n are definable in $\mathcal{D}(\leq \mathbf{0}')$ for $n \geq 3$ using coding methods, but the question for $n < 3$ is still open. Also, defining $\mathcal{R}$ leaves us a long way from answering the questions of Slaman/Harrington concerning the existence of a definable/nondefinable (respectively) r.e. $\mathbf{a} \neq \mathbf{0}$ or $\mathbf{0}'$.

COROLLARY 5.3. *(a)* (STRONG HOMOGENEITY) *If* $\mathcal{D}(\geq \mathbf{a}) \cong \mathcal{D}(\geq \mathbf{b})$ *then* $\mathbf{a}^{(3)} = \mathbf{b}^{(3)}$.

(b) (HOMOGENEITY) If $\mathcal{D} \equiv \mathcal{D}(\geq \mathbf{a})$ then $\mathbf{a}^{(3)} = \mathbf{0}^{(3)}$.

PROOF: (a) follows from Richter's [1979] result for the theory of $\mathcal{D}'$.

(b) follows from Shore's [1981] result for $\mathcal{D}'$.

(Both proved using degree theoretic codings.)

Parts (a) and (b) of Corollary 5.3 already suggest contrasting situations with regard to homogeneity and strong homogeneity. In fact Jockusch (private communication) has extended the methods of Jockusch [1980] (see also Odifreddi [1989], p.546) for proving elementary equivalence of lower cones to obtain results for upper cones:

[3]This extension of 5.1(f) originated with an observation of C. G. Jockusch (August 1991), which we gratefully acknowledge.

PROPOSITION 5.4. *Assuming Projective Determinacy (PD), there is a comeager set of degrees which are bases of elementarily equivalent cones.*

PROOF: If φ is a sentence in the first-order language of partial orderings, let $P_\varphi(A)$ be the second-order predicate which says that φ is true in the cone of degrees above $\deg(A)$. Since the family of sets satisfying $P_\varphi(A)$ is projective, PD implies that it has the Baire property. Let $\mathcal{A}_\varphi$ be this set if it is comeager, and its complement otherwise. Forming the (countable) intersection of all such sets $\mathcal{A}_\varphi$, we get a comeager set $\cap\mathcal{A}_\varphi$ such that the truth-value of any such φ in $\boldsymbol{\mathcal{D}}(\geq \deg(A))$ is independent of $A \in \cap\mathcal{A}_\varphi$.

Martin [1968] previously showed (again using determinacy) that there is a $\mathbf{d}$ such that all cones having base $\geq \mathbf{d}$ are elementarily equivalent (see also Shore [1982]).

Proposition 5.4 suggests a number of questions. Since rigidity of $\boldsymbol{\mathcal{D}}$ may have little direct relevance for the context of everyday mathematics within the noncomputable universe, one would prefer a language-based formulation of rigidity — despite the fact that only countably many degrees can be definable. Noting that $\boldsymbol{\mathcal{D}}$ is rigid if and only if

$$(\forall \mathbf{a}, \mathbf{b} \in \mathcal{D})[\langle \boldsymbol{\mathcal{D}}, \mathbf{a}\rangle \cong \langle \boldsymbol{\mathcal{D}}, \mathbf{b}\rangle \Rightarrow \mathbf{a} = \mathbf{b}],$$

we say:

DEFINITION 5.5. $\boldsymbol{\mathcal{D}}$ *is* first-order rigid *if and only if*

$$(\forall \mathbf{a}, \mathbf{b} \in \mathcal{D})[\langle \boldsymbol{\mathcal{D}}, \mathbf{a}\rangle \equiv \langle \boldsymbol{\mathcal{D}}, \mathbf{b}\rangle \Rightarrow \mathbf{a} = \mathbf{b}].$$

Equally as interesting a question as that of rigidity is:

QUESTION 5.6. *Is* $\boldsymbol{\mathcal{D}}$ *first-order rigid?*[4]

More generally:

DEFINITION 5.7. $\mathbf{a} \in \mathcal{D}$ *is* first-order characterisable *in* $\boldsymbol{\mathcal{D}}$ *if and only if*

$$(\forall \mathbf{b} \in \mathcal{D})[\langle \boldsymbol{\mathcal{D}}, \mathbf{a}\rangle \equiv \langle \boldsymbol{\mathcal{D}}, \mathbf{b}\rangle \Rightarrow \mathbf{a} = \mathbf{b}].$$

That is, if and only if

$$(\forall \mathbf{b} \in \mathcal{D})(\exists \text{ a sentence } \varphi)[\langle \boldsymbol{\mathcal{D}}, \mathbf{a}\rangle \models \varphi \,\&\, \langle \boldsymbol{\mathcal{D}}, \mathbf{b}\rangle \models \neg\varphi].$$

Then:

[4] Jokusch (private communication) can extend the methods of the proof of Proposition 5.4 to show (assuming PD) that $\boldsymbol{\mathcal{D}}$ is *not* first-order rigid.

QUESTION 5.8. *Which degrees are first-order characterisable?*

There remains the main question left open by Corollary 5.3 as to whether there exist *any* distinct $\mathbf{a}, \mathbf{b}$ which are bases of *isomorphic* cones.

Using the theorem of Epstein [1979] and Richter [1979] on automorphisms of $\mathcal{D}'$, one can use the definability of the jump to show that all automorphisms of $\mathcal{D}$ are the identity above $\mathbf{0}^{(3)}$. But Slaman and Woodin (see Slaman [ta]) have announced the following, proved directly via their simplified coding techniques based on the Spector exact pair construction:

THEOREM 5.9 (Slaman and Woodin [ta]) (AUTOMORPHISMS).

(a) Any automorphism of $\mathcal{D}$ is the identity above $\mathbf{0}''$.

(b) Any automorphism of $\mathcal{D}$ is represented by an arithmetically definable function on reals – so there are only countably many such automorphisms.

Even if rigidity fails, one would hope to see $\mathbf{0}''$ eventually replaced by $\mathbf{0}'$ in Theorem 5.9.

§ 6. Automorphism bases

The question of rigidity of a degree structure can sometimes be reduced to that for a familiar substructure.

DEFINITION 6.1 (Lerman [1977]). *$\mathcal{A} \subset \mathcal{D}$ is an* automorphism base *for $\mathcal{D}$ if and only if any automorphism $\psi : \mathcal{D} \to \mathcal{D}$ is completely determined by $\psi \restriction \mathcal{A}$.*

For instance:

THEOREM 6.2. *The following are automorphism bases for $\mathcal{D}$:*

(a) *(Jockusch and Posner [1981]) Any comeager set $\mathcal{A} \subseteq \mathcal{D}$,*

(b) *(Jockusch and Posner [1981]) $\mathcal{M}$ = the set of all minimal degrees, and*

(c) *(Slaman and Woodin [ta]) $\mathcal{R}$.*

PROOF: (a) Jockusch and Posner show that $\mathcal{D}$ is actually *generated* by any given comeager $\mathcal{A}$ — if $\mathbf{d} \in \mathcal{D}$ then $\mathbf{d} = (\mathbf{a}_1 \cup \mathbf{a}_2) \cap (\mathbf{a}_3 \cup \mathbf{a}_4)$, some $\mathbf{a}_i$'s $\in \mathcal{A}$.

(b) Show that $\mathcal{M}$ generates $\mathcal{D}$ using the same idea.

(c) Exploit the rigidity of the standard model of second-order arithmetic by showing that there is a finite set F of r.e. degrees such that (in a sense made precise by Slaman and Woodin) $\mathcal{D}$ is *biinterpretable with* second-order arithmetic in the parameters from F. So $\mathcal{D}$ is biinterpretable with second-order arithmetic using r.e. parameters.

The Slaman-Woodin result has important consequences concerning the question of rigidity of $\mathcal{D}$. From part (c) of Theorem 6.2 we get that if $\mathcal{A}$ is any set of degrees which generates the recursively enumerable degrees, then $\mathcal{A}$ is an automorphism base for $\mathcal{D}$. For instance, since Posner (see Jockusch and Posner [1981]) has shown that $\mathcal{M}(\leq \mathbf{0}')$ generates $\mathcal{D}(\leq \mathbf{0}')$, the set of minimal degrees below $\mathbf{0}'$ forms an automorphism base for $\mathcal{D}$.

Further:

COROLLARY 6.3. *Let $\mathcal{A}$ be a definable set of degrees which generates the recursively enumerable degrees. Then rigidity of $\mathcal{A}$ implies rigidity of $\mathcal{D}$.*

Combining this with the definability results of Theorem 5.1, we get:

COROLLARY 6.4. *If the Turing degrees at any level of the arithmetical hierarchy are rigid, then so is $\mathcal{D}$.*

In particular, nontrivial automorphisms of $\mathcal{D}$ can only exist if there are nontrivial automorphisms of $\mathcal{R}$ and $\mathcal{D}(\leq \mathbf{0}')$.

Hence, in this context, proving rigidity of the noncomputable universe reduces to the local problem of proving $\mathcal{R}$ or $\mathcal{D}(\leq \mathbf{0}')$ to be rigid. The converse, of course, depends on an affirmative answer to the question of Slaman as to whether all automorphisms of $\mathcal{R}$ are extendable to automorphisms of $\mathcal{D}$.

§ 7. The definability of "recursively enumerable in"

We have seen above that the definability of the relation of "r.e. in" is central to the reduction of the rigidity problem to questions of local degree theory. The definition of "r.e. in $\mathbf{c}$" is a natural one depending on an adaptation of known splitting properties for r.e. degrees. The proof that this definition does not include any $\mathbf{d}$ not r.e. in the given $\mathbf{c}$ necessarily falls into two parts, each part depending on a new nonsplitting theorem. Given $\mathbf{d}$ not r.e. in $\mathbf{c}$, one needs to construct a context for $\mathbf{d}$ within which adapted splitting fails.

The first part, for $\mathbf{d} \not\leq \mathbf{c}'$, combines a nonsplitting theorem for d-r.e. degrees (which is independent of $\mathbf{d}$) with the Jockusch-Shore [1983], [1984] pseudo-jump machinery. An immediate corollary is the definability of $\mathbf{c}'$ in $\langle \mathcal{D}, \mathbf{c} \rangle$.

The second part, for $\mathbf{d} < \mathbf{c}'$, requires the nonsplitting context for $\mathbf{d}$ to be constructed directly from $\mathbf{d}$, and below $\mathbf{c}'$. This gives a proof of the definability of "r.e. in $\mathbf{c}$" in $\mathcal{D}(\leq \mathbf{c}')$, and hence in $\langle \mathcal{D}, \mathbf{c} \rangle$ by the first part of the proof.

DEFINITION 7.1. *Given* $\mathbf{a}, \mathbf{b}, \mathbf{d}$, *we say* $\mathbf{d}$ *is* unsplittable over $\mathbf{a}$ avoiding $\mathbf{b}$ *if and only if* $\mathbf{a}, \mathbf{b} \leq \mathbf{d}$, $\mathbf{b} \not\leq \mathbf{a}$, *and for all* $\mathbf{d}_0, \mathbf{d}_1 < \mathbf{d}$, *if* $\mathbf{a} < \mathbf{d}_0, \mathbf{d}_1$ *then either* $\mathbf{b} \leq \mathbf{d}_0$ *or* $\mathbf{d}_1$, *or* $\mathbf{d} \neq \mathbf{d}_0 \cup \mathbf{d}_1$.

$\mathbf{d}$ *is* relatively unsplittable *if and only if* $\mathbf{d}$ *is unsplittable over* $\mathbf{a}$ *avoiding* $\mathbf{b}$, *some* $\mathbf{a}, \mathbf{b}$.

It is important to notice that, by the relativised Sacks Splitting Theorem (see Soare [1987], p.124), there is no relatively unsplittable r.e. degree.

THEOREM 7.2. *Given* $\mathbf{c}$, *a degree* $\mathbf{d}$ *is r.e. in* $\mathbf{c}$ *if and only if*

$$(\forall \mathbf{a}, \mathbf{b} > \mathbf{c})[\mathbf{a} \cup \mathbf{d} \textit{ is not unsplittable over } \mathbf{a} \textit{ avoiding } \mathbf{b}].$$

PROOF SKETCH: (1) Assume that $\mathbf{d}$ is r.e. in $\mathbf{c}$, $\mathbf{a}, \mathbf{b} > \mathbf{c}$, $\mathbf{b} \leq \mathbf{a} \cup \mathbf{d}$ and $\mathbf{b} \not\leq \mathbf{a}$. Then $\mathbf{a} < \mathbf{a} \cup \mathbf{d}$ and $\mathbf{a} \cup \mathbf{d}$ is r.e. in $\mathbf{a}$ since $\mathbf{d}$ is r.e. in $\mathbf{c} < \mathbf{a}$. If $\mathbf{d} \leq \mathbf{a}$, then $\mathbf{a} \cup \mathbf{d} = \mathbf{a}$ is trivially not unsplittable over $\mathbf{a}$ avoiding $\mathbf{b}$. Otherwise $\mathbf{a} < \mathbf{a} \cup \mathbf{d}$, so by the relativised Sacks Splitting Theorem (p.124 of Soare [1987]) there exist $\mathbf{a}$-REA degrees $\mathbf{d}_0, \mathbf{d}_1$ such that $\mathbf{a} \cup \mathbf{d} = \mathbf{d}_0 \cup \mathbf{d}_1$ and $\mathbf{b} \not\leq \mathbf{d}_0$ or $\mathbf{d}_1$. Again, this means $\mathbf{a} \cup \mathbf{d}$ is not unsplittable over $\mathbf{a}$ avoiding $\mathbf{b}$, as required.

(2) Let $\mathbf{d}$ be a degree which is not r.e. in $\mathbf{c}$. We need to show that

$$(\exists \mathbf{a}, \mathbf{b} > \mathbf{c})[\mathbf{a} \cup \mathbf{d} \text{ is unsplittable over } \mathbf{a} \text{ avoiding } \mathbf{b}].$$

The case $\mathbf{d} \not\leq \mathbf{c}'$ was described in Cooper [ta1]. We describe in more detail what happens when $\mathbf{d} \leq \mathbf{c}'$ and $\mathbf{d} \not\leq \mathbf{c}$.

Let $D \in \mathbf{d}$. We construct sets $A, B \in \Delta_2^C$ satisfying the requirements:

$$P_k: \quad B \neq \Theta_k^A \vee (\exists A^* \in \Pi_1^C)(A^* \equiv_T D),$$

$$Q_k: \quad D = \Psi_k(\Phi_k^{A,D}, \widehat{\Phi}_k^{A,D}) \Rightarrow B = \Gamma_k^{\Phi_k^{A,D},A} \vee B = \Lambda_k^{\widehat{\Phi}_k^{A,D},A},$$

$k \geq 0$, where $(\Theta_k, \Psi_k, \Phi_k, \widehat{\Phi}_k)$ is a standard list of all quadruples of p.r. functionals and Γ_k, Λ_k are C-partial recursive functionals to be constructed. The fact that $B \leq_T A \oplus D$ will follow from the satisfaction of the Q-requirements. We assume a standard coding of C into A and B to give $C \leq_T A$ and $\leq_T B$.

We consider just two requirements P ($= P_{k'}$, say) and Q ($= Q_k$, say) in relation to each other, Q being of higher priority than P. We follow the convention of writing $\theta_k, \varphi_k, \widehat{\varphi}_k, \psi_k, \gamma_k, \lambda_k$ etc for the respective standard use functions of $\Theta_k, \Phi_k, \widehat{\Phi}_k, \Psi_k, \Gamma_k, \Lambda_k$ etc.

The naive P-strategy: *Monitor* the growth of the length $\ell(B, \Theta^A)$ of the initial segment of agreement of B and Θ^A. As $\ell(B, \Theta^A)$ grows

larger, progress a modelling of B on D in $B \upharpoonright \ell(B, \Theta^A)$ and make any subsequent change in A below the A-use of $\Theta^A \upharpoonright \ell(B, \Theta^A)$ an *extraction* from A dependent on a change in $B \upharpoonright \ell(B, \Theta^A)$. The result in the case $\operatorname{Lim\,inf}_s \ell(B, \Theta^A)$ is unbounded will be that $D \leq_T B$ (because of the modelling) $= \Theta^A \leq_T A \in \Pi_1^C$, and $A \leq_T B$ (because of the restriction on A-changes) $\leq_T D$ (due to the modelling again), giving $D \equiv_T A \in \Pi_1^C$, this contradiction indicating a pseudo-outcome.

The naive Q-strategy: *First* try to implement the Γ-strategy: Previous to P requiring us to make a $B(x)$-change (due to a $D(x)$-change), try to prepare:
(a) A situation such that $\gamma(x) > \theta(x)$, so we can rectify the equation $B(x) = \Gamma(\Phi^{A,D}, A)(x)$ following the $B(x)$-change with an A-change bigger than $\theta(x)$.

This may entail:
(b) Getting $\gamma(x) >$ some $\psi(y)$, and then hoping to get a $\Phi^{A,D} \upharpoonright \gamma(x)$-change through a $D \upharpoonright y$-change forcing a $\Phi^{A,D} \upharpoonright \psi(y)$-change via the equation $D = \Psi_k(\Phi_k^{A,D}, \widehat{\Phi}_k^{A,D})$.

If it looks like we always get a $\widehat{\Phi}^{A,D} \upharpoonright \psi(y)$-change in (b), start to implement the Λ-strategy.

We consider in detail some of the problems involved in reconciling the strategies for P and Q.

Further discussion: Roughly speaking, our strategy for P and Q together is as follows. If $\ell(D, \Psi(\Phi^{A,D}, \widehat{\Phi}^{A,D}))$ (the standard length of agreement function at stage $s+1$) grows large, we follow the naive Q-strategy in initially implementing the Γ-strategy for making $B \leq_T \Phi^{A,D} \oplus A$. P may initiate B-changes intent on making B look like D below the $\ell(B, \Theta^A)$ level. This may conflict with the Γ-strategy in that changing $B(x)$, say, to agree with $D(x)$ at stage $s+1$ may not be accompanied by the establishment of a new $\Phi^{A,D} \upharpoonright \gamma(x)$. This will require the B-change to be signalled through a positive change in $A \upharpoonright \gamma(x) \subseteq A \upharpoonright \theta(x)$, in opposition to that part of the P-strategy aimed at limiting A-changes below the A-use of $\Theta^A \upharpoonright \ell(B, \Theta^A)$ to extractions from A.

According to the naive Q-strategy, our first approach to a resolution of this conflict will be to anticipate such a $D(x)$-change by trying to make $\gamma(x) > \theta(x)$. Then this will free $\gamma(x)$ for unrestricted use in indicating a $B(x)$-change via Γ, and any positive A-change resulting on the $D(x)$-change will be above the relevant use of $\Theta^A(x)$ (the impact of unrestricted $\gamma(x)$-changes on $\Theta^A(x')$, $x' > x$, can be discounted because of an inductive relationship between $B(x) = \Theta^A(x)$ and such values $\Theta^A(x')$). But in general we can only do this by injuring the existing use of $\Theta^A(x)$, in the

hope that our new larger $\gamma(x)$ will be greater than $\theta(x)$ when this becomes defined again (that is, by moving Γ-markers). This process (essentially Harrington's "capricious destruction") may be repeated using A-changes on larger and larger numbers. We will be able to recognise the P-strategy to be working up to the x level when we see that all the potentially damaging Γ-markers have been cleared from the use of $\Theta^A \upharpoonright x+1$.

There are various possible outcomes to this. We may succeed in obtaining a suitable relatively small use for Θ^A, resulting in a successful outcome to the P-strategy up to the x level. On the other hand, infinite repetition of this process at the x level will lead to $\Theta^A(x) \uparrow$ (P satisfied again), but (without further analysis) we will also end up with $\Gamma^{\Phi^{A,D},A}$ not total so that the Γ-strategy fails. We need to pursue a further possibility for avoiding this in order for such an outcome to provide the conditions for replacing the failed Γ-strategy with a successful Λ-strategy for Q. Since we are only concerned about Q if Φ^D is a total function, we may allow some $D \upharpoonright y$-change, with $\psi(y) < \gamma(x)$, to initiate an attempt to move $\gamma(x)$, but defer making any A-changes needed for such a move until at least $\Phi^{A,D} \upharpoonright \gamma(x)$ has become redefined. This leaves open the possibility that we may get a completely new $\Phi^{A,D} \upharpoonright \gamma(x)$ (that is, not containing as an initial segment any previously defined $\Phi^{A,D} \upharpoonright \gamma(x)$) which can be used to clear Γ-markers from the use of $\Theta^A(x)$ without the need for any A-changes to be made.

But then, assuming the $D \upharpoonright y$-change has been timed to coincide with $\Theta^A \upharpoonright x \downarrow (= B \upharpoonright x)$, we avoid disturbing $A \upharpoonright \theta(x)$ while clearing all Γ-markers $\geq \gamma(x)$ to positions above $\theta(x)$. Hence we get $\gamma(x') > \theta(x)$ for $x' \geq x$ following the above actions, so satisfying P up to the x level, and in the process leaving the Γ-strategy intact.

We can assist this outcome by using A to increase the likelihood of a new $\Phi^{A,D} \upharpoonright \gamma(x)$ being produced which can be used to clear the Γ-markers. $D \not\leq_T C$ provides infinitely many changes in beginnings of D. We try to ensure the occurrence of a $D \upharpoonright y$-change with $\psi(y) \downarrow \leq \gamma(x)$ and $\lambda(x)$, such a change being termed a *level x D-agitation* , with *D-agitator* y, so that the $D \upharpoonright y$-change will at least produce some sort of change in either $\Phi^{A,D} \upharpoonright \gamma(x)$ or $\widehat{\Phi}^{A,D} \upharpoonright \lambda(x)$. The provision of the right conditions for a level x D-agitation is attempted via a process similar to Harrington's "honestification" whereby if $\gamma(x)$ or $\lambda(x) < \psi(y)$, some designated y, we make an $A \upharpoonright w$ change, say, with $w \leq \min\{\gamma(x), \lambda(x)\}$, redefining $\gamma(x), \lambda(x) \geq \psi(y)$ when $\psi(y)$ is next defined. In fact honestification is extended by making $w \leq \min\{\gamma(x), \lambda(x)\}$ for all such $\gamma(x), \lambda(x)$ defined since the last occurrence of honestification, thereby ensuring that previously defined $\Phi^{A,D} \upharpoonright \gamma(x)$ or $\widehat{\Phi}^{A,D} \upharpoonright \lambda(x)$ will not return in tan-

dem with corresponding $A \upharpoonright \gamma(x)$ or $A \upharpoonright \lambda(x)$ to prevent us $\Phi^{A,D}$- or $\widehat{\Phi}^{A,D}$-permitting a Γ- or Λ-marker clearance following the $D \upharpoonright y$-change.

There is a problem here (apart from that of not knowing whether we get a $\Phi^{A,D}$- or $\widehat{\Phi}^{A,D}$-change following such a $D \upharpoonright y$-change), in that honestification will very likely also involve an $A \upharpoonright \theta(x)$ change, so that a suitable $D \upharpoonright y$-change can only be recognised following redefinition of $\theta(x)$, by which time the effects of honestification may have worn off, demanding renewed honestification. If this repetition developes into an infinite outcome, we get $\Gamma^{\Phi^{A,D},A}, \Lambda^{\widehat{\Phi}^{A,D},A}$ not total. But Q is then satisfied since we must have $\min\{\gamma(x), \lambda(x)\} < \psi(y)$ infinitely often so that $\Psi(\Phi^{A,D}, \widehat{\Phi}^{A,D})(y) \uparrow$ also. And $\theta(x) \uparrow$ infinitely often, so P is satisfied through $\Theta^A(x) \uparrow$. Unlike the construction for the definability of the jump, we no longer have control of D (while on the positive side A-changes no longer need D-permitting), so we must also wait for the D-agitations to be provided. This means we must work with an upwardly shifting choice of y, relying on $D \notin \Pi^C_1$ for the corresponding possible infinitary outcome being a pseudo-outcome.

Even then, honestification as described will still not be sufficient to supply the ideal conditions for exploiting the $D \upharpoonright y$-change. This is because of a new complication resulting from the possibility of returns to strings $\Phi^{A,D} \upharpoonright \gamma(x)$ (following the $D \upharpoonright y$-change) which appeared since the last occurrence of honestification for (P, Q). So before recognising a $D \upharpoonright y$-change as a usable D-agitation we further ask that $\Phi^{A,D} \upharpoonright \psi(y)$, $\widehat{\Phi}^{A,D} \upharpoonright \psi(y)$ are unchanged at all stages since the previous occurrence of honestification, and if this condition is not met, we again honestify (even if $\psi(y) \leq \gamma(x), \lambda(x)$). If continued honestification is required we still get $\Psi(\Phi^{A,D}, \widehat{\Phi}^{A,D})(y) \uparrow$, and Q is again satisfied (along with P since $\Theta^A(x) \uparrow$).

So far, there is nothing in the above discussion to indicate that A need not be r.e. in C over the segment $A \upharpoonright \theta(x)$. It is only when we consider the interactions of more than one P-requirement with a Q-requirement of higher priority that we find that we may need to extract some numbers from A. We briefly look at what happens.

Say we have a P', of priority intermediate between that of Q and P, and that we get a $D \upharpoonright y$-change leading to a suitable new $\Phi^{A,D} \upharpoonright \gamma(x)$ which we restrain in order to be able to preserve $B(x) = D(x)$, with $A \leq_T B$ up to the $\theta(x)$ level, while maintaining the Γ-strategy. It may happen at a later stage that we act on some y' through P', resulting in a loss of the new $\Phi^{A,D} \upharpoonright \gamma(x)$ (replaced by a new $\widehat{\Phi}^{A,D} \upharpoonright \lambda(x)$, presumably). It may not be possible now to rectify Γ by a suitable positive A-change, as

this may conflict with the actions for P' (for instance). We then have no alternative but to reverse any consequent $B(x)$-change and any moving of markers for x beyond $\theta(x)$, and to extract the relevant marker from A in the case that the new $\Phi^{A,D} \upharpoonright \gamma(x)$ had been used to permit a $B(x)$-change via Γ. (There is still the possibility of the new $\Phi^{A,D} \upharpoonright \gamma(x)$ which permitted $B(x) = D(x)$ via Γ reasserting itself at a later stage, but in defining the corresponding $\gamma(x)$ we will have been able to have regard for higher priority P' to the extent that we can avoid having to redefine $B(x)$ by making a suitable $A \upharpoonright \gamma(x)$ change.) It remains to follow through the consequences of infinitely many occurrences of post-honestification level x D-agitations for (P, Q) which produce no appropriate new strings of the form $\Phi^{A,D} \upharpoonright \gamma(x)$. In this case we utilise the fact that following each such D-agitation we get a new $\widehat{\Phi}^{A,D} \upharpoonright \lambda(x)$ to satisfy P, Q through the Λ-strategy. In fact, this outcome for (P, Q) gives us a successful Λ-strategy for each (P', Q) with P' $(= P_{k''}$ say) of lower priority than P, so we will not assume that the infinite set of y's we act on necessarily relates to P'.

We now assume that the Λ-strategy has its own set of followers $z \geq 0$, disjoint from any other set of followers, and its own procedure for D-agitation (called $\widehat{D}$-agitation). We $\widehat{D}$-agitate with the pre-knowledge that we get infinitely many $A \upharpoonright \gamma(x)$ changes through capricious destruction, and infinitely many usable $\widehat{\Phi}^{A,D} \upharpoonright \lambda(y)$ changes (or, more relevant, no usable $\Phi^{A,D} \upharpoonright \gamma(y)$ changes). This means we only bother to act in the interests of $B(z) \neq \Theta^A_{k''}(z)$ if $\theta_{k''}(z) < \gamma(x)$. Since $\gamma(x)$ goes to infinity, this will still provide sufficient space in which to satisfy $P_{k''}$.

In order to use a $\widehat{D}$-agitator $\widehat{y}$ we also need to obtain $\psi(\widehat{y}) \leq \psi(y)$ and $\psi(\widehat{y}) \leq \lambda(z)$. The failure of y to successfully D-agitate for (P, Q) will mean that $\widehat{y}$ must $\widehat{D}$-agitate in the interests of obtaining a usable $\widehat{\Phi}^{A,D} \upharpoonright \lambda(z)$ change to clear markers from the A-use of $\Theta^A_{k''}(z)$ via Λ without the need to injure $\Theta^A_{k''}(z) \neq D(z)$ with an $A \upharpoonright \lambda(z)$ change. This requires its own honestification, which we can time to coincide with the honestification for (P, Q). Again, the honestification takes the stronger form described previously.

As for the construction for the definability of the jump, one needs a tree of outcomes on which to reconcile the strategies for different pairs (P', Q'), relative to which (in this case) we can determine the final outcomes along the true path recursively in $C^{(3)}$. It is worth mentioning here some of the special complications arising from the fact that D is not r.e. in C but is only Δ^C_2. An inevitable consequence is a certain amount of 'Δ^C_2-noise' in the construction. This means that the questions appearing in the basic module for the earlier construction, which acted as gateways through which the module irrevocably passed, now appear as *tests* . The

results of these tests can change leading to a revoking of the license to proceed, and demanding a reversion to another phase of the module.

The extra unpredictability of D- and hence B-changes does not in itself cause too many problems for A. With the help of redefinitions of γ, λ A can cope with the corresponding demands of the Γ- and Λ-strategies. And to some extent the added technical complexity merely acts to mask the eventual, but still inevitable situation, the strategy covering this eventuality working with ultimate outcomes in place of recognisable events. There are slightly more problems with the consequent lack of control over $\Phi^{A,D}$ and $\widehat{\Phi}^{A,D}$. We relied above, in certain situations, on $\Phi^{A,D}$- or $\widehat{\Phi}^{A,D}$-changes enabling us to avoid certain sorts of A-changes in the interests of the A^*-strategy for the P-requirements. When these changes are in doubt, we will have to fall back on the undesired type of A-changes. However, temporary vacillations in $\Phi^{A,D}$ or $\widehat{\Phi}^{A,D}$ can be matched by a corresponding flexibility in the A-changes; and where the ultimate outcomes for $\Phi^{A,D}$- or $\widehat{\Phi}^{A,D}$-changes are assured we will be able to maintain our aims in regard to the A^*-strategies. On the other hand, in reconciling the demands of different P-requirements, the A^*-strategies may demand negative A-changes where the immediate need may seem to be new positive A-changes. The key factor here is, of course, that the success of the A^*-strategy for P lies in producing $B \neq \Theta^A$, not in an infinitary outcome of $D \equiv_T A^* \in \Pi^C_1$.

We now give a more formal description of the strategies for (P, Q).

The basic module for P confronted with one higher priority Q

(All statements in the description below are assumed to relate to stage $s+1$ of the construction.)

Let

$$\ell(D, \Psi(\Phi^{A,D}, \widehat{\Phi}^{A,D})) = \mu z\,[D(z) \neq \Psi(\Phi^{A,D}, \widehat{\Phi}^{A,D}; z)] \text{ and}$$
$$\ell(B, \Theta^A) = \mu z\,[B(z) \neq \Theta^A(z)] \text{ (at stage } s+1).$$

We say that a D- or $\widehat{D}$-agitator y for x (say) associated with (P, Q) is *realised* if (P, Q) is in possession of a suitable $D \upharpoonright y$-change for (respectively) D- or $\widehat{D}$-agitation. Otherwise y is *unrealised*. Let $z =$ the largest unrealised D- or $\widehat{D}$-agitator for any x associated with (P, Q) (at stage $s+1$), if such a z exists, and $= s$ otherwise. We have an overall constraint on the construction relative to (P, Q) that if $\ell(D, \psi(\Phi^{A,D}, \widehat{\Phi}^{A,D})) > z$ then we must define $\Gamma^{\Phi^{A,D},A} \upharpoonright x$ or $\Lambda^{\widehat{\Phi}^{A,D},A} \upharpoonright x$.

And if y, z are unrealised D-, $\widehat{D}$-agitators respectively for x (at stage $s+1$) we ask that whenever we redefine $\Gamma^{\Phi^{A,D},A} \upharpoonright x$ or $\Lambda^{\widehat{\Phi}^{A,D},A} \upharpoonright x$

we choose $\gamma(x), \lambda(x)$ so that $\gamma(x) \geq \psi(y)$ or $\lambda(x) \geq \psi(z)$ respectively. Also, whenever we redefine $\Gamma^{\Phi^{A,D},A}(w)$ or $\Lambda^{\widehat{\Phi}^{A,D},A}(w)$, $w \geq 0$, we define $\Gamma^{\Phi^{A,D},A}(w) = B(w)$ or $\Lambda^{\widehat{\Phi}^{A,D},A}(w) = B(w)$, respectively.

Whenever we redefine values of $\Gamma^{\Phi^{A,D},A}$ or of $\Lambda^{\widehat{\Phi}^{A,D},A}$ in such a way that $\Gamma^{\Phi^{A,D},A} \simeq B$ (that is, they agree on all values on which both are defined) or $\Lambda^{\widehat{\Phi}^{A,D},A} \simeq B$, we say that we *rectify* Γ or Λ, respectively.

We assume that (P, Q) has available an infinite set η of potential D-agitators and a set ζ of potential markers $\gamma(x)$ for numbers $x \geq 0$. At stage $s+1$ η, ζ consist of those members of η, ζ respectively which have not yet been used in the construction.

The basic module consists of the following phases together with the above overall constraints.

1. We Γ-*select* $x =$ the least number not previously Γ-selected for (P, Q).
2. We *select* the least $y \in \eta$, $y > x$, as a *D-agitator* for x, and *initiate* a *y-cycle* with phase 3 below.
3. We *wait* for $\ell(D, \Psi(\Phi^{A,D}, \widehat{\Phi}^{A,D}))$ to grow bigger than y and *define* the least $\in \zeta$ to be the Γ-*marker* $\gamma(x)$ for x.
4. And we *test* for $\ell(B, \Theta^A) > x$.
   (a) A positive test allows us to *restrain* $A_- \upharpoonright \theta(x)$, where A_- is the set of numbers in $\overline{A}$ (at stage $s+1$) and to *continue* with the cycle from 5 onwards so long as $A \upharpoonright \theta(x)$ is unchanged. An $A \upharpoonright \theta(x)$-change at any subsequent stage dictates a return to 4(b).
   (b) Following the emergence of a negative test result we *return* to 4 for a retest, and *cancel* any existing restraints dependent on an earlier positive test.
5. We *test* for $\gamma(x) > \theta(x)$.
   (a) If $\gamma(x) > \theta(x)$ we *proceed* to 6.
   (b) (Honestification and capricious destruction combined.)
   If $\gamma(x) \leq \theta(x)$ we *enumerate* $\gamma(x)$ into A, and *proceed* through phases 3, 4 and then 7.
6. In the absence of a return to 4 we *restrain* $A \upharpoonright \theta(x)$ and get:
   Outcome: P is *satisfied* up to the x level, and x ceases to interfere with Q.
7. We now *test* if $\gamma(x) \geq \psi(y)$ and if $\Phi^{A,D} \upharpoonright \psi(y)$, $\widehat{\Phi}^{A,D} \upharpoonright \psi(y)$ are unchanged at all stages since they became redefined following the latest application of phase 5(b) (at stage $u+1$ say).
   (a) At each stage at which the test is positive we can *proceed* through 8 and beyond.
   (b) While a failed test at any stage dictates a *return* to 3.

8. We *wait* for $D \upharpoonright y \neq D^{u'} \upharpoonright y$, each u' with $u+1 \leq u' < s+1$.
   (a) If we achieve such a D-change we *go to* 9.
   (b) During the wait phase, we *return* once to 2 to *choose* a new D-agitator (y' say) and *initiate* a new y'-cycle.
9. We *wait* for $\ell(D, \Psi^{\Phi^{A,D}, \widehat{\Phi}^{A,D}}) > y$, and then:
10. *Test* if $\Gamma^{\Phi^{A,D},A}(x) \uparrow$.
    (a) If the test is positive, we *redefine* $\gamma(x) > \theta(x)$ (that is, we *move the marker* $\gamma(x)$), and *restrain* $A \upharpoonright \theta(x)$.
    Outcome: P is *satisfied* up to the x level, and following a $B(x)$-change Γ can be *rectified* without an $A \upharpoonright \theta(x)$-change being required.
    (b) During negative testing we *return* to 2, replacing any marker previously moved through 10(a).

In the case of infinitely many returns to 2 on behalf of x and (P, Q), we need to describe the Λ-strategy. This is an auxiliary strategy that synchronises its activities with phases 2, 5, 7 and 8 of the Γ-strategy. As mentioned before, it can relate to (P', Q) even if $P' \neq P$.

$\widehat{2}$. (Simultaneous with 2.) (a) We Λ-*select* the least number x' not previously Λ-selected for (P', Q) and then (b) *select* the least $\widehat{y} \in \widehat{\eta}$, $\widehat{y} > x'$, as a $\widehat{D}$-*agitator* for x' (with (P', Q)).
$\widehat{7}$. (Simultaneous with 7.) We *test* if $\lambda(x') \geq \psi(\widehat{y})$.
   (a) At each stage at which the tests in 7 and $\widehat{7}$ are positive we can *proceed* through $\widehat{8}$ and beyond.
   (b) While a failed test at any stage dictates a *return* to 3 as already described.
$\widehat{8}$. We *wait* for $D \upharpoonright \widehat{y} \neq D^{u'} \upharpoonright \widehat{y}$, each u' with $u+1 \leq u' < s+1$.
   (a) If we achieve such a D-change we *go to* $\widehat{9}$.
   (b) (Simultaneous with 8(b).) During the wait phase, we *return* once to $\widehat{2}$(b) (working with the existing Λ-selected x') to *choose* a new $\widehat{D}$-agitator ($\widehat{y}'$ say).
$\widehat{9}$. We *wait* for $\ell(D, \Psi^{\Phi^{A,D}, \widehat{\Phi}^{A,D}}) > \widehat{y}$, and then:
$\widehat{10}$. *Test* if $\Lambda^{\Phi^{A,D},A}(x') \uparrow$ and $\gamma(x') > \theta'(x')$.
   (a) During a positive test, we *redefine* $\lambda(x') > \theta'(x')$ and *restrain* $A \upharpoonright \theta'(x')$, while initiating a new cycle of the Λ-strategy for (P', Q) with a return to $\widehat{2}$.
   Outcome: P' is *satisfied* up to the x' level, and following a $B(x')$-change Λ can be *rectified* without an $A \upharpoonright \theta'(x')$-change being required.
   (b) During negative testing we *return* to $\widehat{2}$(b).

Summary of outcomes of the Γ- and Λ-strategies for (P, Q), (P', Q)

The finite outcomes:

$\boxed{w_1}$: The strategy halts at 3. Then $D \neq \Psi(\Phi^{A,D}, \widehat{\Phi}^{A,D})$ and Q is satisfied and ceases to interfere with P.

$\boxed{t_1}$: The strategy proceeds infinitely often through phase 4(b). Then $B \neq \Theta^A$ and P is satisfied. Any interference with Q is transitory.

$\boxed{s_1}$: The strategy for making $D \equiv_T A^*$ r.e. in C is defended by the placement of $\gamma(x)$ and the imposition of an A_--restraint in 4(a). P is satisfied up to level x and ceases to interfere with Q up to that level, due to phase 6 applying. $\boxed{s_1}$ will only apply with respect to finitely many followers x of (P, Q), so represents a genuine finite outcome.

$\boxed{w_1'}$: The strategy halts at 9. Outcome as for $\boxed{w_1}$.

$\boxed{s_2}$: Strategy halts at 10(a). The strategy for making $D \equiv_T A^*$ r.e. in C is defended by the moving of $\gamma(x)$ and the imposition of an A_--restraint in 4(a). P is satisfied up to level x and ceases to interfere with Q up to that level, while maintaining $\Gamma^{\Phi^{A,D},A} = B$ via a $\Phi^{A,D}$-change.

$\boxed{\widehat{w}_1}$: The strategy halts at $\widehat{9}$. Outcome as for $\boxed{w_1}$ and $\boxed{w_1'}$.

$\boxed{\widehat{s}_2}$: Strategy halts at $\widehat{10}$(a). P' is satisfied up to level x' by the moving of $\lambda(x')$, and ceases to interfere with Q up to that level. $\Lambda^{\Phi^{A,D},A} = B$ is maintained via a $\widehat{\Phi}^{A,D}$-change.

The infinitary outcomes:

$\boxed{i_1}$: The strategy passes through phase 7(b) infinitely often.

Since we infinitely often pass through phases 3 and 4, $\gamma(x)$ goes to infinity. Since we never halt at 6, $\theta(x) \geq \gamma(x)$ infinitely often, so $\Theta^A(x) \uparrow$ and P is satisfied.

Since we go through 7(b) infinitely often, either $\psi(y) > \gamma(x)$ infinitely often, or $\Phi^{A,D} \uparrow \psi(y)$ or $\widehat{\Phi}^{A,D} \uparrow \psi(y)$ changes infinitely often, so in either case $\Psi^{\Phi^{A,D},\widehat{\Phi}^{A,D}}(y) \uparrow$ (possibly with $\psi(y)$ bounded but $\Phi^{A,D}(u)$ or $\widehat{\Phi}^{A,D}(u) \uparrow$, some $u \leq \psi(y)$), giving Q also satisfied.

$\boxed{i_2}$: We wait at phase 8 for D-changes relative to infinitely many D-agitators y for x with (P, Q), leading to infinitely many returns via 8(b) to 2. An application of Miller and Martin's Lemma (see Soare [1987]. p.223) relativised to C removes this as a possibility since D is not r.e. in C ($\boxed{i_2}$ is a pseudo-outcome).

$\boxed{i_3}$: The strategy passes through phase 10(b) infinitely often. *Outcome*: We implement the Λ-strategy, P' is satisfied as in $\boxed{i_1}$.

$\boxed{\widehat{i_3}}$: The strategy passes through $\widehat{10}$(b) infinitely often.

As for $\boxed{i_1}$ we get P' satisfied through $\Theta'^{A}(x') \uparrow$, while the Λ-strategy for (P', Q) is maintained. This is because, by the conditions of 7(a)/8(a) and $\widehat{7}$(a)/$\widehat{8}$(a) we must arrive at $\widehat{10}$(a) with either $\Lambda^{\widehat{\Phi}^{A,D},A}(x') \uparrow$ or $\Gamma^{\Phi^{A,D},A}(x) \uparrow$. Since 10(a) does not apply, we must have $\Lambda^{\widehat{\Phi}^{A,D},A}(x') \uparrow$, so we get to move $\lambda(x')$ before returning to $2/\widehat{2}$.

Combining the strategies for more than one pair P, Q

We look at some immediate consequences of combining the P, Q-strategies, giving separate consideration to what happens when the number of P- or Q-requirements alone is increased.

(a) Infinitely many P-requirements below one Q-requirement

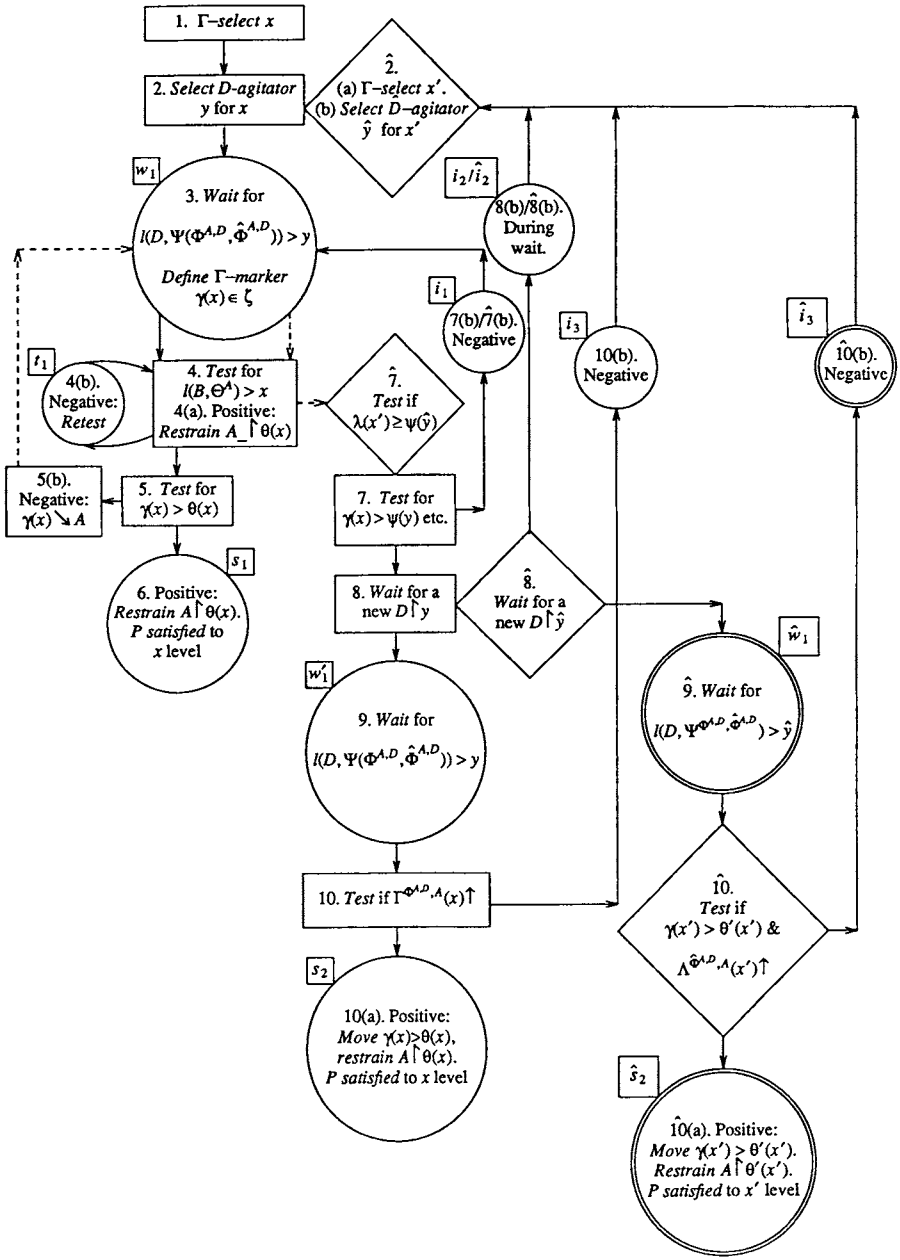
We need only look at the situation in which for some P above Q the outcome for P, Q is infinitary.

If $\boxed{i_1}$ is the outcome for P, Q (with P minimal), each P' of priority between Q and P has a finite outcome. The main consequence for the requirements P' below P is that the strategy for P', Q can ignore the Γ- and Λ-strategies for Q, and at stage $s+1$ pursue the satisfaction of P' within the upper boundary determined by the top of the column of markers for P, Q already in A at stage $s+1$.

The outcome $\boxed{i_3}$ for P, Q is similar in effect in that we need to pursue outcomes for P', Q, P' below P, within the upper boundary determined by the already utilised markers for P, Q. Then $\boxed{i_3}$ for P, Q presents such P', Q with the possibility of a successful Λ-strategy with outcome $\boxed{\widehat{s_2}}$ or $\boxed{\widehat{i_2}}$.

We also note that it is at the level of one Q-requirement above just two P-requirements that we make full use of the ability to extract numbers from A, resulting in A being Δ_2^C. Say we have P, P' below Q, P' of higher priority than P, and that we are pursuing outcome $\boxed{s_2}$ for P and P' within the context of a successful Γ-strategy for Q. So we can assume that during the construction we pass through phases 7(a), 8(a), 9, 10and 10(a) of the basic module in relation to outcome $\boxed{s_2}$ for P, Q. That is, we get a suitable $D \upharpoonright y$-change, some D-agitator y, achieving a suitable new $\Phi^{A,D} \upharpoonright \gamma(x)$, arriving at $A \upharpoonright \theta(x)$ cleared of markers for x and

The following diagram relates the strategies to the outcomes:



a possible rectification of Γ in the interests of the Γ-strategy for Q. It maylater happen that we follow a similar process for P', Q, acting on some D-agitator y' through 7(a)/8(a), resulting in a loss of the new $\Phi^{A,D} \upharpoonright \gamma(x)$ (replaced by a new $\widehat{\Phi}^{A,D} \upharpoonright \lambda(x)$, presumably). It may not be possible now to rectify Γ by a suitable A-change, as this may conflict with the actions for P' (or for some other P-requirement of higher priority than that of P). We then have no alternative but to allow $\Gamma\left(\Phi^{A,D}, A; x\right)$ to find its value relative to existing axioms for Γ, while using an at most finite number of $B(x)$ changes in maintaining $B = \Gamma\left(\Phi^{A,D}, A; x\right)$. Reinstatement of $\gamma(x) \notin A$ following the ultimately unsuccessful D-agitation will now be necessary in the interests of ensuring $\Gamma\left(\Psi^{A,D}, A\right) \in \Delta_2^C$.

(b) One P-requirement below two Q-requirements

Say we have Q above Q', with both Q and Q' of higher priority than P. An infinitary outcome to the strategy for P, Q or P, Q' results in $\Theta^A(x) \uparrow$ for some follower x of P, so any problems for P arise in this simple case in the situation in which both Q and Q' either maintain viable Γ-strategies or switch to a Λ-strategy due to some *other* P requirement. This means that (discarding the simple outcomes $\boxed{w_1}$, $\boxed{t_1}$, $\boxed{w_1'}$ and $\boxed{\widehat{w}_1}$) we aim to satisfy P through a combination of outcomes $\boxed{s_1}$, $\boxed{s_2}$ and $\boxed{\widehat{s}_2}$ for subrequirements P, Q, P, Q'.

If we follow the Γ-strategy for P, Q or P, Q' (say P, Q) and fail to get outcome $\boxed{s_1}$ on P, Q we have no need to pursue such an outcome on P, Q'. This is because phase 5 of the basic module is only necessary for us to be sure that $\theta(x)$ is unbounded if outcomes $\boxed{i_1}$ or $\boxed{i_2}$ apply to P with some higher priority Q-requirement, and for this the repeated failure of $\boxed{s_1}$ on P, Q would be sufficient. In any case, we cannot expect to combine outcomes $\boxed{s_2}$ or $\boxed{\widehat{s}_2}$ for P, Q with outcome $\boxed{s_1}$ for P, Q' since the former require A-changes of unpredictable magnitude to counteract changes in $\Phi^{A,D}$- or $\widehat{\Phi}^{A,D}$-permissions via Γ or Λ, respectively, and these may conflict with the A-restraints required for the latter. So, failing $\gamma(x) > \theta(x)$ while following the Γ-strategy for P, Q, we either work with the one follower x and attack P, Q and P, Q' simultaneously through step 5(b) of the basic module, aiming to satisfy P through outcome $\boxed{s_2}$ for both subrequirements, or the Λ-strategy will be appropriate to P, Q'.

We can use the tree of outcomes to help us work with the right combination. The main problem is that of asynchronicity of occurrences of $\Gamma(\Phi^{A,D}, A; x) \uparrow$ and $\Gamma'(\Phi'^{A,D}, A; x) \uparrow$ or $\Lambda'(\widehat{\Phi}'^{A,D}, A; x) \uparrow$. This is because the completion of phase 9 or $\widehat{9}$ for the Γ- or Λ-strategy for P, Q, say, has to include steps to ensure that $\Gamma(\Phi^{A,D}, A; x) \downarrow$ or $\Lambda(\widehat{\Phi}^{A,D}, A; x) \downarrow$, respec-

tively, without waiting for P, Q' to complete 9 or $\widehat{9}$. This means we can only anticipate a successful outcome 10(a) to an implementation of 8(a) at stage $s+1$, say, by also setting up a restraint for $\Theta^A(x)$ at stage $s+1$, and then, if 10(b) is the actual result cancelling the restraint for $\Theta^A(x)$ and retaining follower x in any further capricious destruction through phase 5(b), but necessarily choosing a new follower x_1, say, for the purposes of further pursuit of outcome $\boxed{s_2}$ through 7(a)/8(a) etc. The situation is similar but simpler with regard to the Λ-strategy. Although we must define $\Lambda'(\widehat{\Phi}'^{A,D}, A; x')$, say, at $\widehat{8}$(a), and set up a (possibly temporary) A-restraint, we can do this in such a way that a failure of $\widehat{\Phi}'^{A,D}$-permission means we must initialise the Λ-strategy.

There are some new features arising from combining the strategies for pairs P, Q where both multiple P- and Q-requirements are involved. For instance, the temporary A-restraints consequent on $\boxed{i_3}$ define a lower boundary on the usage of markers which present extra timing difficulties. See Cooper [ta2] for further discussion and a formal proof.

The tree of outcomes

Our tree will reflect the fact that restraints, followers and agitators are chiefly manipulated in the interests of the P-requirements, while the management of the Γ- and Λ-strategies is in direct response to activity on lower priority P-requirements. Since each P-requirement P_i is considered in relation to its higher priority Q-requirements, the strategy for each pair having consequences for the requirements below P_i, it is convenient to consider P_i via a corresponding block of *subrequirements* , or *P, Q-requirements* , namely $(P_i, Q_0), (P_i, Q_1), \ldots, (P_i, Q_{i-1})$. The priority ordering of the P-, Q- and P, Q-requirements is the obvious one, (P_i, Q_j) having priority relative to $R_{i'} \neq P_i$ given by the priority of P_i relative to $R_{i'}$, P_i itself having priority greater than each of its subrequirements, and within the block of subrequirements for P_i by $(P_i, Q_j) \geq (P_i, Q_{j'})$ if and only if $j \leq j'$.

During the construction, requirements and their individual ingredients are only considered in relation to nodes σ on a tree T which correspond to either immediate or more long-term assessments of the true outcome for the higher priority requirements (that is, those assigned to the levels of T below $\ell h(\sigma)$). If $\mathcal{A}_i$ is an expression being considered at node $\sigma \in T$ we will usually write $\mathcal{A}_\sigma$ for $\mathcal{A}_i$. We write R_σ for the requirement located at node $\sigma \in T$, calling σ a *P-, Q- or P, Q-node* according as R_σ is a P-, Q- or P, Q-requirement, respectively. $(P, Q)_\sigma$ denotes the P, Q-requirement situated at P, Q-node σ, (P_σ, Q_τ) denotes *some* subrequirement of the full P-requirement P_σ. In referring to P_σ, Q_σ or $(P, Q)_\sigma$ below we presuppose

R_σ to be (respectively) a P-, Q- or P,Q-requirement.

The P,Q-nodes lay out a framework of possible outcomes for the higher priority requirements, within which a full P-requirement actively pursues its strategy in relation to the higher priority Q-requirements.

For the purposes of satisfying P_σ, the activity on $(P_{\sigma'}, Q_{\tau'})$ of higher priority limits the segment of numbers over which the strategy for P_σ can usefully operate. The restraints consequent on $(P_{\sigma'}, Q_{\tau'})$, and any finitary use of markers, present a lower boundary to activity on P_σ. The use of infinitely many markers in relation to $(P_{\sigma'}, Q_{\tau'})$ will require P_σ to wait until activity on agitators for $(P_{\sigma'}, Q_{\tau'})$ has moved above the region in which we seek to satisfy P_σ, so presenting an upper boundary to activity on P_σ. Meanwhile the main feature of the activity on Q_τ of higher priority than P_σ is the ultimate choice of Γ- or Λ-strategy for Q_τ, although this depends on what happens on lower priority P-requirements. Outcome $\boxed{i_1}$ is not of special interest to P_σ except in so far as knowing whether there is an infinite column of markers for a higher priority $(P_{\sigma'}, Q_\tau)$ used in relation to this outcome. On the other hand, a switch from Γ- to Λ-strategy for Q_τ may be due to some $(P_{\sigma'}, Q_\tau)$ of lower priority than that of P_σ, so cannot be deduced from that part of the tree above the node for P_σ.

§ 8. Definability in $\mathcal{R}$ and $\mathcal{D}(\leq \mathbf{0}')$

As noted above, if $\mathcal{R}$ or $\mathcal{D}(\leq \mathbf{0}')$ is rigid then so is $\mathcal{D}$, so particular interest attaches to the question of definability in these degree classes. The more we can define in $\mathcal{R}$ or $\mathcal{D}(\leq \mathbf{0}')$, the less likely are non-trivial automorphisms to exist.

The main definable classes below $\mathbf{0}'$ are $\mathcal{R}$ and $High_n$, Low_n for $n \geq 3$. Classes and relations we would like to define in $\mathcal{R}/\mathcal{D}(\leq \mathbf{0}')$ include:

(1) $High_n$, Low_n for $n < 3$.
(2) *Jump equivalence* : $\mathbf{x}' = \mathbf{y}'$.
(3) Individual *atomic jump classes* : $\mathbf{c}^{-1} = \{\mathbf{x} | \mathbf{x}' = \mathbf{c}\}$, $\mathbf{c}$ $\mathbf{0}'$-REA.
(4) Individual *r.e. degrees* $\in (\mathbf{0}, \mathbf{0}')$.
(5) $\mathcal{R}(\leq \mathbf{a})$, *given* $\mathbf{a}$ *r.e.* $> \mathbf{0}$. *The set of n-REA degrees* $\leq \mathbf{0}'$, $n \geq 2$.
(6) $\mathcal{R}_n$ = *the set of n-r.e. degrees*, $n > 1$.

A first step in defining *High* / Low_2 is provided by Shore and Slaman [ta2] who succeeded in separating *High* and Low_2 in the r.e. degrees:

THEOREM 8.1. *There is a definable* $\mathcal{C} \subset \mathcal{R}$ *such that* High $\subset \mathcal{C}$ *and* $\text{Low}_2 \subset \mathcal{R} - \mathcal{C}$.

PROOF: Take

$$\mathcal{C} = \{\mathbf{x} | \exists \mathbf{z} < \mathbf{y} \leq \mathbf{x} \forall \mathbf{u}, \mathbf{v} [\mathbf{u} \cup \mathbf{v} = \mathbf{y} \Rightarrow \mathbf{u} \cup \mathbf{z} = \mathbf{y} \vee \mathbf{v} \cup \mathbf{z} = \mathbf{y}]\}.$$

Apply the Harrington Low_2 Splitting Theorem (see Shore and Slaman [ta1]), and push the Lachlan Non-Splitting Theorem [1975] below any given high r.e. degree.

There are a number of other known structural features of the High/Low_2 degrees which indicate a definability result below $\mathbf{0}'$:

(1) (Cooper [1974]) Any $\mathbf{h} \in \mathcal{R}$ bounds a minimal r.e. pair.
(2) (Posner [1977], Cooper [1972]) Any $\mathbf{h} \in \mathcal{D}(\leq \mathbf{0}')$ is the join of a pair of minimal degrees, whereas (Jockusch-Posner [1978]) all minimal degrees are Low_2.
(3) (Harrington, Cooper, Yates) There is a relatively non-cuppable degree below any $\mathbf{h} \in \mathcal{R}$.

There is a related basic question:

QUESTION 8.2. *Can we (first-order) distinguish* $\mathcal{D}(\leq \mathbf{0}')$ *from* $\mathcal{D}(\leq \mathbf{a})$ *for each* $\mathbf{a} < \mathbf{0}'$*?*

A (relativisable) positive answer to this question would enable us to define the low degrees: A degree $\mathbf{a}$ would be low if and only if $\mathcal{D}[\mathbf{a}, \mathbf{0}']$ looked like $\mathcal{D}[\mathbf{a}, \mathbf{a}']$ should do.

It is not easy to see what sort of property is required to answer Question 8.2, although pointers are provided by results of Soare and Stob [1982], Shore [1982a] and Harrington (the degrees below any $\mathbf{a} < \mathbf{0}'$ do not realise all jumps in the $\mathbf{0}'$-REA degrees), and Cooper and Epstein [1987] (who found a finite injury construction giving a first-order distinction between $\mathcal{D}(\leq \mathbf{0}')$ and $\mathcal{D}(\leq \mathbf{a})$, some r.e. $\mathbf{a} > \mathbf{0}$).

§ 9. Rigidity and definability for other degree structures

Nerode and Shore [1980a] and Slaman and Woodin [ta] have extended a number of their global results for the Turing degrees to other degree structures such as the 1-, m-, tt-, btt-, wtt-, arithmetic, hyperarithmetic and e-degrees. But for some degree structures there are very strong results concerning rigidity and definability and for others much less is known. A comprehensive review of information can be found in Odifreddi [1989] and [ta]. We mention one or two of the more striking results, and related questions.

THEOREM 9.1 (Slaman and Woodin [ta]). *The* hyperdegrees *are rigid.*

THEOREM 9.2 (Slaman and Woodin [ta]. *Any automorphism of the* arithmetic degrees $\mathcal{D}_a$ *is the identity above* $\mathbf{0}_a^{(\omega)}$.

Shore (see Odifreddi [ta]) has a number of global results for the theory of $\mathcal{D}'_a$, and in particular shows that every automorphism of $\mathcal{D}'_a$ is the identity above $\mathbf{0}'_a$.

QUESTION 9.3. *Is* $\mathbf{0}'_a$ *definable in* $\mathcal{D}_a$*?*

For the *1-degrees* there is no characterisation of $\mathcal{D}_1$ to parallel that for the m-degrees. But:

THEOREM 9.4 (Nerode and Shore [1980a]). *If* $\mathcal{D}_1{}' \equiv \mathcal{D}'_1\,(\geq \mathbf{b})$ *(or* $\mathcal{D}'_m \equiv \mathcal{D}'_m\,(\geq \mathbf{b})$*) then* $\mathbf{b}$ *is arithmetical.*

Non-homogeneity for the *truth-table degrees* follows from results of Mohrherr [1984] ($\mathcal{D}_{tt}\,(\geq \mathbf{0}'_{tt})$ is dense) and Martin (any hyperimmune-free minimal Turing degree is also a minimal tt-degree) that $\mathcal{D}_{tt} \not\equiv \mathcal{D}_{tt}\,(\geq \mathbf{0}'_{tt})$.

Slaman and Woodin [ta] have some partial results for the *enumeration degrees* , although most global questions are open. Homogeneity is known to fail by results (see Cooper [1990]) of Gutteridge (there does not exist a minimal e-degree) and Cooper (there exists a minimal cover in $\mathcal{D}_e$).

QUESTION 9.5. *Do the Turing degrees form an automorphism base for the enumeration degrees (under the natural embedding)?*

QUESTION 9.6. *Are the Turing degrees definable in the enumeration degrees?*

REFERENCES

S. B. COOPER [1972], *Degrees of unsolvability complementary between recursively enumerable degrees, Part I*, Ann. Math. Logic 4, 31–73.

S. B. COOPER [1974], *Minimal pairs and high recursively enumerable degrees*, J. Symbolic Logic 39, 655–660.

S. B. COOPER [1990], *Enumeration reducibility, nondeterministic computations and relative computability of partial functions*, in "Recursion Theory Week, Oberwolfach 1989," (eds. K. Ambos-Spies, G. Müller, G. E. Sacks), Springer-Verlag, Berlin, Heidelberg, New York, pp. 57–110.

S. B. COOPER [ta1], *Definability and global degree theory*, to appear in the proceedings of Logic Colloquium '90, Helsinki.

S. B. COOPER [ta2], *The recursively enumerable degrees are absolutely definable*, to appear.

S. B. COOPER and R. L. EPSTEIN [1987], *Complementing below recursively enumerable degrees*, Ann. of Pure and Applied Logic 34, 15–32.

M. DAVIS [1958], "Computability and Unsolvability," McGraw-Hill, New York.

R. L. EPSTEIN [1979], "Degrees of Unsolvability. Structure and Theory," Lecture Notes in Mathematics No. 759, Springer-Verlag, New York.

Y. L. ERSHOV [1975], *The upper semilattice of numerations of a finite set*, Alg. Log. 14, 258–284 (Russian); 14 (1975), 159–175 (English translation).

S. FEFERMAN [1957], *Degrees of unsolvability associated with classes of formalized theories*, J. Symbolic Logic 22, 161–175.

K. GÖDEL [1931], *Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I*, Monatsh. Math. Phys. 38, 173–198.

K. GÖDEL [1934], *On undecidable propositions of formal mathematical systems*, mimeographed notes, in "The Undecidable. Basic Papers on Undecidable Propositions, Unsolvable Problems, and Computable Functions," (M. Davis, ed.), Raven Press, New York, 1965, pp. 39–71.

C. G. JOCKUSCH, Jr. [1980], *Degrees of generic sets*, Lond. Math. Soc. Lect. Notes 45, 110–139.

C. G. JOCKUSCH, Jr. and D. POSNER [1978], *Double jumps of minimal degrees*, J. Symbolic Logic 43, 715–724.

C. G. JOCKUSCH, Jr. and D. POSNER [1981], *Automorphism bases for degrees of unsolvability*, Israel J. Math. 40, 150–164.

C. G. JOCKUSCH, Jr. and R. A. SHORE [1983], *Pseudo jump operators I: The R.E. case*, Trans. Amer. Math. Soc. 275, 599–609.

C. G. JOCKUSCH, Jr. and R. A. SHORE [1984], *Pseudo jump operators II: Transfinite iterations, hierarchies, and minimal covers*, J. Symbolic Logic 49, 1205–1236.

C. G. JOCKUSCH, Jr. and R. I. SOARE [1970], *Minimal covers and arithmetical sets*, Proc. Amer. Math. Soc. 25, 856–859.

S. C. KLEENE [1943], *Recursive predicates and quantifiers*, Trans. Amer. Math. Soc. 53, 41–73.

S. C. KLEENE and E. L. POST [1954], *The upper semi-lattice of degrees of recursive unsolvability*, Ann. Math. (2) 59, 379–407.

A. H. LACHLAN [1972], *Recursively enumerable many-one degrees*, Alg. Log. 11, 326–358 (Russian); 11 (1972), 186–202 (English translation).

A. H. LACHLAN [1975], *A recursively enumerable degree which will not split over all lesser ones*, Ann. Math. Logic 9, 307–365.

M. LERMAN [1977], *Automorphism bases for the semilattice of recursively enumerable degrees*, A-251, Abstract no.77T-E10, Notices Amer. Math. Soc. 24.

M. LERMAN [1980], *The degrees of unsolvability: Some recent results*, in "Recursion Theory: Its Generalisations and Applications," (eds. F. R. Drake and S. S. Wainer), London Math. Soc. Lecture Notes Series No. 45, Cambridge Univ. Press, Cambridge, U.K., pp. 140–157.

M. LERMAN [1983], "Degrees of Unsolvability," Perspectives in Mathematical Logic, Omega Series, Springer-Verlag, Berlin, Heidelberg, London, New York, Tokyo.

D. A. MARTIN [1968], *The axiom of determinateness and reduction principles in the analytical hierarchy*, Bull. Amer. Math. Soc. 74, 687–689.

J. MOHRHERR [1984], *Density of a final segment of the truth-table degrees*, Pacific J. Math. 115, 409–419.

A. NERODE and R. A. SHORE [1980], *Second order logic and first order theories of reducibility orderings*, in "The Kleene Symposium," (J. Barwise et al., eds.), North-Holland, Amsterdam, pp. 181-200.

A. NERODE and R. A. SHORE [1980a], *Reducibility orderings: theories, definability and automorphisms*, Ann. Math. Logic 18, 61–89.

P. ODIFREDDI [1989], "Classical Recursion Theory," North-Holland, Amsterdam, New York, Oxford.

P. ODIFREDDI [ta], "Classical Recursion Theory, Vol. II," North-Holland, Amsterdam, New York, Oxford (to appear).

E. PALIUTIN [1975], *Addendum to the paper of Ershov [1975]*, Alg. Log. 14, 284–287 (Russian); 14 (1975) pp. 176–178 (English translation).

D. POSNER [1977], "High Degrees," Ph.D. Dissertation, University of California, Berkeley.

E. L. POST [1943], *Formal reductions of the general combinatorial decision problem*, Amer. J. Math. 65, 197–215.

E. L. POST [1944], *Recursively enumerable sets of positive integers and their decision problems*, Bull. Amer. Math. Soc. 50, 284–316.

L. J. RICHTER [1979], *On automorphisms of the degrees that preserve jumps*, Israel J. Math. 32, 27–31.

H. ROGERS, Jr. [1967], "Theory of recursive functions and effective computability," McGraw-Hill, New York.

J. R. SHOENFIELD, *Degrees of formal systems 23*, J. Symbolic Logic, 389–392.

R. A. SHORE [1981], *The degrees of unsolvability: global results*, in "Logic Year 1979-80: University of Connecticut," (eds. M. Lerman et al.), Lecture Notes in Mathematics No. 859, Springer-Verlag, Berlin, Heidelberg, New York, pp. 283–301.

R. A. SHORE [1982], *On homogeneity and definability in the first order theory of the Turing degrees*, J. Symbolic Logic 47, 8–16.

R. A. SHORE [1982a], *Finitely generated codings and the degrees r.e. in a degree* $\mathbf{d}$, Proc. Amer. Math. Soc. 84, 256–263.

R. A. SHORE [1988], *Defining jump classes in the degrees below* $\mathbf{0}'$, Proc. Amer.Math. Soc. 104, 287–292.

R. A. SHORE and T. A. SLAMAN [ta1], *Working below a low$_2$ recursively enumerable degree*, to appear.

R. A. SHORE and T. A. SLAMAN [ta2], *Working below a high recursively enumerable degree*, to appear.

T. A. SLAMAN [ta], *Degree structures*, to appear.

T. A. SLAMAN and W. H. WOODIN [1986], *Definability in the Turing degrees*, Illinois J. Math. 30, 320–334.

T. A. SLAMAN and W. H. WOODIN [ta], *Definability in degree structures*, to appear.

R. I. SOARE [1987], "Recursively enumerable sets and degrees," Springer-Verlag, Berlin, Heidelberg, London, New York.

R. I. SOARE and M. STOB, *Relative recursive enumerability*, in "Proceedings of the Herbrand Symposium Logic Colloquium '81," (ed. J. Stern), North-Holland, Amsterdam, New York, Oxford, pp. 299–324.

A. M. TURING [1939], *Systems of logic based on ordinals*, Proc. London Math. Soc. 45, 161–228.

C. E. M. YATES [1970], *Initial segments of the degrees of unsolvability, Part I: A survey*, in "Mathematical Logic and Foundations of Set Theory," (Y. Bar-Hillel, ed.), North-Holland, Amsterdam, pp. 63–83.

Logic, Methodology and Philosophy of Science IX
D. Prawitz, B. Skyrms and D. Westerståhl (Editors)

THE IMPACT OF MODEL THEORY ON THEORETICAL COMPUTER SCIENCE

J.A. MAKOWSKY

Department of Computer Science
Technion - Israel Institute of Technology, Haifa, Israel

1. Introduction

The purpose of this survey is to give an account of those aspects of model theory which we think are relevant to theoretical computer science. The exposition here follows rather closely the presentation given at the conference itself. A very expanded version of this paper appears as [Mak92b]. We give a general outline of the evolution of model theory, which will serve as an exposition of the major themes. For each of them we sketch its relevance to Theoretical Computer Science.

We assume the reader is familiar with the basics of First Order Logic, Computability Theory, Complexity Theory and Basic Algebra. Whenever possible we shall refer to textbooks and monographs rather than the original papers. Only material not treated in standard texts will be quoted in the original (or by referring to a subsequent paper which contains the result in the most readable form).

Logic and model theory are relatively old disciplines which enjoy renewed interest. They can serve as one explanatory paradigm for foundational problems in theoretical computer science. But the gap between the traditional logicians and mathematicians and the working computer scientists is first of all cultural in the sense of R. Wilder's [Wil81]. His studies deserve special attention especially when one has in mind the evolution and development of programming languages, operating systems, user interfaces and other paradigms of computing, but also in addressing foundational questions, cf. [Mak88].

Wilder's studies clearly show several phenomena: that the evolution of concepts to widely accepted norms of practice takes much longer and needs more than just the availability of such concepts; that the evolution of concepts is not due to individuals but is embedded in one (or several competing)

cultural systems which are themselves embedded in host cultural systems; that nevertheless the fame and prestige of the protagonists of science and scientific progress do play an important, possibly also counterproductive rôle; that cultural stress and cultural lag play a crucial rôle in the evolution of concepts; that periods of turmoil are followed by periods of consolidation after which concepts stabilize; that diffusion between different fields usually will lead to new concepts and accelerated growth of science; that environmental stresses created by the host culture and its subcultures will elicit observable response from the scientific culture in question; and, finally, revolutions may occur in the metaphysics, symbolism and methodology of computing science, but not in the core of computing itself. Wilder has developed in [Wil81] a general theory of 'Laws' governing the evolution of mathematics, from which I have adapted the above statements. It remains a vast research project to assimilate Wilder's theory into our context, but it is an indispensable project if we want to adjust our expectation of progress in computing science to realistic hopes. Wilder's work also sheds some light into the real problems underlying the so called 'software crisis': The cultural lag of programming practice behind computing science and the absence of various cultural stresses may account for the abundance of programming paradigms without the evolution of rigorous standards of conceptual specifications.

2. Theoretical computer science

As we discuss here applications of model theory to computer science we have to clarify what we intend both by model theory and theoretical computer science. Concerning Computer Science we take a pragmatic approach. Any mathematically modelled situation which captures any issue arising in the dealings with computers is a possible topic for computer science. This includes hardware, software, data modelling, interfaces and more. Some of the more classical fields of theoretical computer science have already matured into well established subdisciplines. Among them we find computability theory, algorithmics, complexity theory, database theory, data and program specification, program verification and testing etc. However, we feel that a certain confusion in the definitions of these fields is obfuscating the issues involved. It very much depends whether our point of view is *method–oriented* or *application–oriented.* Computability and complexity theory deal with the clarification of our notion of what is computable. This represents a clear case of a well defined method–oriented subdiscipline of computer science and the foundations of mathematics. Database theory on the other hand is a field which grew from an application–oriented approach. From a method–oriented point of view, database theory tends to fall apart into subfields, such

as *finite model theory*, *operating systems*, *file systems*, *user interfaces* and *algorithmics*, where each of these transcend the boundaries of the database applications. Scientifically speaking, the ad hoc collection of methods bound together by a vaguely defined common application is unsatisfactory. It is justified only for didactic purposes such as training application–oriented engineers. But such training is detrimental to a deeper understanding of the craft and the science and leads to chaotic duplicity of research and research subcultures each disguised in its own terminology and provincialisms.

In this paper we try to exhibit a method and a scientific framework, *model theory* and discuss typical problems whose discussion in this framework is beneficial to our understanding.

3. The set theoretic modelling of syntax and semantics

Model theory is the mathematical (set theoretical) study of the interplay between Syntax and Semantics. Historically it has its roots in the various attempts of reducing first mathematics to logic (Frege, Hilbert), then logic to number theory (Skolem, Gödel) and finally, of modelling logic within set theory (Tarski, Vaught). The first two reductions were motivated by the fundamental questions of the foundations of mathematics, whereas the latter accepts Bourbaki's view that set theory is the foundational framework of mathematics. It is this latter approach which forms the background of model theory proper. Let us elaborate this further: We take some Naive Set Theory for granted and attempt to model all objects of mathematical study within this Set Theory. Without having to bother too much about the choice of set theory we can model the natural numbers, finite strings, finite graphs within set theory. We accept the axiom of choice as a fact of life. With this we can model also most of the concepts of classical algebra (field theory, ring theory, group theory, but not necessarily cohomology theory) within set theory. The natural numbers, fields, graphs are mathematical structures which serve as the prime examples for models of logical theories. We usually think of models of a logical theory rather than of a single model, and the models form usually a proper class (the class of all groups, rings, etc.). If we restrict ourselves to finite mathematical structures we can additionally consider recursive sets of models or sets models of lower complexity classes (Logarithmic Space or Polynomial Time recognizable classes of models). Next we observe that logical theories are just sets of formulas and that formulas can be viewed again as either strings over some alphabet or as some kind of labeled trees. Most people think of formulas as inherently finite objects, but infinite formulas (then better viewed as trees) are easily conceivable. So formulas and sets of formulas can also be modelled in our

set theory. If we think of finite formulas as strings it makes sense to bring in also concepts of recursion theory and complexity theory.

The basic relationship between sets of formulas and models is the satisfaction relation. We view it here as a ternary relation $M(\Sigma, \mathcal{A}, z)$, where Σ is a set of formulas, $\mathcal{A}$ is a structure, i.e. a generalized algebra over some vocabulary (similarity type) and z is an assignment function mapping free variables of the formulas Σ into elements of the universe of the structure $\mathcal{A}$. If $M(\Sigma, \mathcal{A}, z)$ holds for every z we simply write $\mathcal{A} \models \Sigma$ and say that $\mathcal{A}$ is a model of Σ. The characteristic function of the satisfaction relation is often called meaning function. The meaning function can also be modelled in set theory.

3.1. First order structures

It is customary to model algebraic structures as sets equipped with functions and relations. This view has its origins in algebra as understood in the 19th century. A structure consists of a set, the *universe*, equipped with some relations, functions and constants, which model the *primitives*.

A *group* then is a set equipped with a binary function, called *multiplication*, a unary function, called the *inverse*, and a constant called the *unit element*. An ordered group is additionally equipped with a binary relation, called the *order relation*. In similar ways we can define *fields*, *rings* or the structure of *arithmetic on the natural numbers*. In computer science other *data structures* are defined similarly, such as *words, stacks, lists, trees, graphs, Turing machines etc.* A word of length n over the alphabet $\{0, 1\}$ can be viewed as a set of n elements with a binary relation which linearly orders that set and a unary relation, which indicates which places in the word are occupied by the letter 1. A graph is just a set with a binary relation. In each case it is required that the functions, relations and constants satisfy some interrelating properties which make it into a group (field, word, graph, Turing machine etc.).

Sometimes, it is more practical to model structures with several underlying sets, as in the case of vector spaces. These sets form several universes and are called *sorts*. We then speak of *many–sorted structures*. A Turing machine consists of two sets: a set of states and a set of letters; a binary relation between states and letters; a unary relation, the set of final states; and a constant, the initial state. Many–sorted structures allow us to model also concepts which involve sets of sets, such as topologies, families of subgroups or whatever comes to ones mind. This last statement is not just a sloppy way of saying something vague. It really expresses a belief, or rather experience, that everything which can be modelled in set theoretic terms with finitely

many basic concepts can be modelled by such structures.

In modern terms a structure is a tuple of sets of specified characteristics. The primitive concepts have *names* and these names form again a set, called the *vocabulary.* A structure then is an *interpretation* of a vocabulary. More precisely, a (first order) vocabulary τ is a set of sort symbols, function symbols, relation symbols and constant symbols. The function, relation and constant symbols have an *arity* which specifies the number and sorts of the arguments and values. The arity is mostly assumed to be *finite.* In this way we can naturally associate with a vocabulary τ the proper class of all τ–structures, which we denote by $STR(\tau)$.

3.2. The choice of the vocabulary

The notion of a τ–structure evolved naturally in mathematics, more precisely in algebra. Groups and fields are usually described as sets with operations, ordered fields are sets with operations and relations. The choice of the basic operations is in no way trivial. Should we add the inverse operation as basic or not? In the case of arithmetic we have the successor relation, addition and multiplication. The first order theory of arithmetic is undecidable, but if we leave out multiplication, it becomes decidable. This is a dramatic change. Subtraction is definable by a first order formula, so leaving it out or adding it, does not affect decidability. But it does affect the set of substructures.

In the case of graphs the modelling issue is more subtle. It is customary to describe a graph as a set with an incidence relation. Thus there is quantification over vertices but not over edges. If we choose to allow quantification over edges we change the notion of structure. To what extent this matters has been studied by Courcelle in a series of papers [Cou90a, Cou90b]. Finite graphs can also be described by their incidence matrix, which does not fit the notion of a τ–structure in a natural way. However, we can consider the incidence matrix itself as a τ–structure in many ways.

Logic and model theory take the notion of τ–structures for granted. How to choose the particular vocabulary depends on many extra–logical issues. Discussing some of these issues is a discipline in itself called Data Modelling. The issues discussed there come from data processing and data bases.

First order logic allows quantification only for elements of the underlying universe. This looks like a severe restriction, as in mathematics we quantify very often also over subsets and more complex objects. However, this restriction only affects the modelling issue. In set theory all objects are sets, and second order arithmetic can be formalized using first order τ–structures, where the universe consists of points and sets with a unary predicate distinguishing between them. It is in this sense that the notion of τ–structures is

as universal as the set theoretic modelling of mathematical situations.

More surprisingly, τ–structures can also capture situations of modal and temporal propositional logic. A propositional variable may be true in some moments of time and false in others. So let the universe of our discourse be time and propositions be unary predicates [Bur84]. This is almost obvious. In the case of modal logic it needed Kripke's ingenuity to make use of this idea [BS84]. The universe now is a set, the set of all possible worlds or situations, propositions are again unary predicates, but the relationship between possible worlds is described by an accessibility relation. From here, it is natural to continue and consider several accessibility relations (to model for example the distinction between the legally and the morally possible). In the theory of program verification this was used to model the behaviour of abstract programs (Dynamic Logic [Har84]). In AI this approach was extended further to model reasoning about knowledge [Eme90]. The interested reader will find more also in section 7. and [Ga92]. For reasons of space we shall not treat these issues much further in this paper.

The point we want to emphasize here is that the framework of τ–structures is flexible enough to model everything which can be modelled in mathematics, more precisely in set theory. The choice of vocabulary is sometimes difficult and guided by various issues, including user friendliness, explicitness and technicalities of the field of application.

3.3. Logics

The most prominent logic is First Order Logic. Although we argued that τ–structures are sufficiently general to model all situations which can be treated mathematically, First Order Logic has a limited expressive power. This means that a description in First Order Logic of a situation will allow what are called *non–standard models.* In other words, it will have models of that description which do not capture all the intended features.

Other logics we shall consider are Second Order Logic (allowing quantification over subsets and relations without making them into objects of the model), Monadic Second Order Logic (allowing quantification only over subsets), infinitary logics (allowing infinite conjunctions and disjunctions) and logics with generalized quantifiers. The latter is discussed in detail in [Mak92b].

A logic itself again can be modelled within set theory. It consists of a family of τ–formulas $Fm(\tau)$ with associated meaning functions M_τ subject to several conditions. The most fundamental among them is the *Isomorphism Condition* which asserts that isomorphic τ–structures cannot be distinguished by τ–formulas. The other conditions assert that the most ba-

sic operations such as conjunction, disjunction, negation, relativization and quantification over elements are well defined. Such logics are called regular logics. If negation is omitted we call the logics semi–regular. The model theory of such logics has been extensively studied, cf. [BF85].

For applications in computer science the relevant logics have two additional features: The set of τ–formulas Fm_τ is recursive for finite τ and the meaning functions M_τ are *absolute* for set theory, i.e., they do not depend on the particular model of Zermelo-Fränkel set theory we are working in. If we additionally require that the tautologies of such a logic are recursively enumerable, we call such a logic a *Leibniz Logic*. It now follows from work of Lindström and Barwise that every Leibniz Logic is in some precise sense equivalent to First Order Logic, cf. [BF85]. In other words, a proper extension of First Order Logic is either not regular or not absolute or its tautologies are not recursively enumerable. If we restrict ourselves to finite structures the latter is unavoidable even for First Order Logic (by Trakhtenbrot's theorem), but then the satisfiable formulas are recursively enumerable. Semi–regular logics on finite structures where the satisfiable formulas are recursively enumerable have many applications to computer science and are studied in [Mak92b].

4. The birth of model theory

Model theory deals with the mathematical study of the satisfaction relation or its characteristic function, the meaning function. For a specific syntactic system which we call logic, the meaning function singles out the pairs of first order structures and formulas which we interpret as asserting that the given formula holds in the given structure. Any mathematically proven statement about the meaning function is a model theoretic theorem.

4.1. The fundamental theorems

The first result of mathematical logic which could be called model theoretic was the famous Löwenheim-Skolem Theorem:

THEOREM 4.1 (LÖWENHEIM-SKOLEM THEOREM) *Let Σ be a set of formulas of first order logic such that there is an infinite $\mathcal{A}$ with $\mathcal{A} \models \Sigma$. Then there are models $\mathcal{B}$ of arbitrary infinite cardinalities such that $\mathcal{B} \models \Sigma$.*

The most basic model theoretic theorem is the compactness theorem for first order logic. We say that a set Σ of formulas is satisfiable if there is a structure $\mathcal{A}$ such that $\mathcal{A} \models \Sigma$. The compactness theorem now states that:

THEOREM 4.2 (COMPACTNESS THEOREM) *A set Σ of first order formulas is satisfiable iff every finite subset of Σ is satisfiable.*

It follows from Gödel's completeness theorem for countable Σ and was proven for arbitrary Σ by Mal'cev. A model theoretic proof of the completeness theorem was given independently by Hasenjäger, Henkin and Hintikka in 1949. This proof, most widely known as Henkin's method, was instrumental in shaping the further developments of logic and model theory.

The completeness theorem usually refers to some specific *deduction method* and states that a τ–formula ϕ is derivable from a set of τ–formulas Σ iff ϕ is a semantical consequence of Σ. The notion of semantical consequence is model theoretic. It says that for every τ–structure $\mathcal{A}$ and every assignment z such that $M(\Sigma, \mathcal{A}, z) = 1$ we also have $M(\phi, \mathcal{A}, z) = 1$. A purely model theoretic statement which captures the essence of the completeness theorem without reference to the particular deduction method is the following:

THEOREM 4.3 *For every recursive enumerable set Σ of τ–formulas the set of τ–formulas ϕ which are semantical consequences of Σ is recursive enumerable, and this uniformly in Σ.*

4.2. Definability questions

The next ten years of evolving model theory were marked by explorations of the compactness theorem and the Löwenheim-Skolem Theorem. The first of these explorations concerns definability questions, both negative and positive results.

On the negative side we have that many important mathematical concepts cannot be captured by first order formulas. Among them are the concept of well-orderings, connectivity of binary relations and Cauchy completeness of linear orders. This was first perceived as a blow to the foundation of mathematics, as it led to 'non–standard' models of the Natural Numbers, the Real Numbers and of Set Theory. However, A. Robinson realized that those non-standard models had their own usefulness for developing genuine first order mathematics. For theoretical computer science, non-standard models of number theory and set theory only recently started to play a rôle. We shall not discuss their use in this paper, but refer the reader to [ANS82, MS89, Pas90].

On the positive side we have Beth's theorem on implicit definitions and its various generalizations. Those theorems were mostly proven first by syntactic methods, but the model theoretic proofs found later make those theorems independent of the particular formalism of first order logic. Let Σ be a set of first order formulas over some vocabulary τ, and let P be an

n-ary relation symbol not in τ. We say that a formula $\phi(P)$ over $\tau \cup \{P\}$ defines P *implicitly* using Σ, if in each model $\mathcal{A}$ of Σ there is *at most one* interpretation of P. We say that the predicate implicitly defined by ϕ using Σ has an *explicit* definition if there is a formula $\theta(x_1, x_2, \ldots, x_n)$ over τ such that

$$\Sigma \cup \phi(P) \models \forall x_1, x_2, \ldots, x_n(\theta(x_1, x_2, \ldots, x_n) \leftrightarrow P(x_1, x_2, \ldots, x_n)).$$

Now Beth's theorem can be stated as follows:

THEOREM 4.4 (BETH) *Let Σ be a set of first order formulas and let $\phi(P)$ be an implicit definition of P using Σ. Then there is an explicit definition of P using Σ.*

Beth's theorem is trivially true for second order logic, and false for first order logic when restricted to finite structures. In the latter case, implicit definitions allow us to define classes of structures recognizable in $\mathbf{NP} \cap \mathbf{co\text{–}NP}$, whereas first order formulas define classes recognizable in $\mathbf{L}$ (Deterministic Logarithmic Space). Beth's theorem is mainly appealing as a closure property of a logic. There are surprisingly few genuine applications of Beth's theorem and its relatives. One of them, in the axiomatic treatment of specification theory, is relevant to theoretical computer science (cf. [MS92]). More recently Kolaitis has studied implicit definability on finite structures and related it to issues in complexity theory, [Kol90].

5. Maturing model theory

In the sixties model theory flourished around applications of the compactness theorem and around alternative proofs of it.

5.1. Preservation theorems

One line of explorations of the compactness theorem was initiated by Tarski. He observed that universal first order formulas are preserved under substructures. In other words, if Σ is a set of first order formulas in prenex normal form with universal quantifiers only and $\mathcal{A} \models \Sigma$ and $\mathcal{B} \subseteq \mathcal{A}$ is a substructure of $\mathcal{A}$ then $\mathcal{B} \models \Sigma$. The same is true for any Σ_1 equivalent to Σ. By an ingenious application of the compactness theorem he proved the converse of this observation:

THEOREM 5.1 (SUBSTRUCTURE THEOREM) *A set Σ of first order formulas is preserved under substructures iff Σ is equivalent to some set of universal formulas.*

The Substructure Theorem set a pattern for further investigations whose results are called preservation theorems. It led to similar syntactic characterizations for formulas preserved under unions of chains, homomorphisms, products, intersections and other algebraic operations. There are also some surprising interrelationships between a generalization of Beth's theorem and preservation theorems for a wide class of operations between structures, cf. [Mak85]. Some of these preservation theorems have variations and interpretations which are of importance in database theory [Mak84] and the foundations of logic programming, [Mak87]. Questions related to such preservation theorems also occur naturally in the compositional approach to model checking for various temporal logics, [Eme90]. The latter is a subdiscipline of program verification. It still remains an open avenue of research to find the preservation theorems which will be useful for model checking, in particular those preservation theorems which will reflect the compositionality of programs.

Horn formulas

Both in Relational Database Theory and Logic Programming, first order formulas form the syntactic background of the field. In both fields it was observed that certain syntactically defined classes formulas play a special rôle. For a detailed discussion of first order logic's rôle in database theory one may consult [Var88, Kan90] and the corresponding chapter in [Mak92a] The most prominent such class of formulas are called *Universal Horn formulas.* They also play a certain rôle in the Specification of Abstract Data Types.

DEFINITION 5.2 (HORN FORMULAS) *(i) A quantifierfree Horn formula is a formula of the form*

$$P_1 \wedge \ldots \wedge P_k \rightarrow P_0$$

where all the $P_i, i \leq k$ are atomic formulas.

(ii) A Universal Horn formula is a formula of the form $\forall x_1, \ldots, x_m \Phi$ with Φ a quantifier free Horn formula.

The classical theorem of model theory gives the following characterization of Universal Horn formulas.

THEOREM 5.3 (MAL'CEV) *Let K be a class of τ–structures which are exactly the models of a set of first order τ–formulas Σ. Then K is closed under substructures and products iff Σ is equivalent to a set of Universal Horn formulas.*

It is now tempting to try to use this characterization of Universal Horn formulas in order to explain their special properties in terms of Databases and Logic Programming. Fagin has done this in [Fag82] for the case of databases. Mahr and Makowsky have done this for the case of Specification of Abstract Datatypes [MM84] extracting the model theoretic content of [GTWW77]. The latter was based on ideas from Category Theory and Universal Algebra. A more general discussion may be found in [Mak84]. Clearly, neither the closure under substructures nor under products has any explanatory power per se in these contexts. It would be more satisfactory, if the formation of products and the closure under substructures could be replaced by some activity stemming from handling databases. This was achieved with moderately satisfactory results in [Mak81, MV86].

The predominant rôle Horn formulas play in Logic Programming can be explained syntactically by the similarity of Horn formulas to deterministic rules or instructions. Semantically, the situation is similar to Abstract Data Types in as much as one thinks of a unique minimal interpretation. An exact model theoretic analysis of Horn formulas in Logic Programming was proposed in [Mak87]. Its relevance for Negation by Failure was discussed in Shepherdson's [She84, She85, She88]. The exact formulation of this analysis is unfortunately not possible in this survey. An excellent exposition of special properties of Horn formulas is [Hod92]. Formulas preserved under relativization play a vital rôle in relational database theory, especially in connection with *safe* queries, cf. [Ull82, TS88, MV86]. Horn formulas preserved under intersections were analyzed in [Mak87]. Finally, formulas with monotone predicates can be characterized as formulas with positive occurrence of the predicate and play an important rôle in the theory of computable fixed points and related topics [Mos74]. The use of preservation theorems in Database Theory will be discussed in [Mak92a].

5.2. Ultraproducts and fast growing functions

With these early investigations centering around the compactness theorem and the preservation theorems an alternative proof of the compactness theorem was discovered using ultraproducts. The method of ultraproducts also lead to alternative proofs of preservation theorems and dominated research in model theory throughout the sixties (cf. [CK90]), but it had almost no impact on theoretical computer science. Although Kripke and Kochen [KK82] used bounded ultraproducts to give a model theoretic proof of the Paris-Harrington Theorem, Kanamori and McAloon [KM87] gave a model theoretic proof of this theorem without bounded ultraproducts. In the language of theoretical computer science this theorem can be stated as follows:

THEOREM 5.4 (PARIS, HARRINGTON) *There are programs (number theoretic functions) which*

(i) always terminate (are total) but

(ii) such that a termination proof does not exist within the formalization of Peano arithmetic.

A very picturesque version of this theorem is due to Kirby and Paris [KP82]. The function described there is a winning strategy for the fight of Hercules against the Hydra, where the Hydra grows n new heads after the nth blow it receives. The underlying theme of this theorem are *fast growing functions.* It is questionable whether the model theoretic proof of the Paris-Harrington 5.4 theorem really captures the essence of the matter completely. The original proof has a proof theoretic flavour and for various generalizations of this theorem no purely model theoretic proof is known. A prominent example is Friedman's theorem:

THEOREM 5.5 (H. FRIEDMAN) *There are programs (number theoretic functions) which*

(i) always terminate (are total) but

(ii) such that a termination proof does not exist within the formalization of various fragments of second order Arithmetic.

A technical and philosophical discussion of such theorems may be found in [HMŠS85]. A presentation of this theorem and related results accessible to computer scientists may be found in [Gal91].

5.3. Complete theories and elimination of quantifiers

Another line of early investigations was the study of complete theories. A set Σ of formulas (over a fixed vocabulary τ) is complete if for every formula ϕ either $\Sigma \models \phi$ or $\Sigma \models \neg\phi$. The original interest for complete theories stems from questions of decidability. A set of formulas Σ is decidable if its set of consequences is recursive.

THEOREM 5.6 *If Σ is recursive and complete then Σ is decidable.*

Proofs of completeness were often obtained using the method of elimination of quantifiers. Tarski used these ideas to show that there is a decision procedure for Elementary Geometry, which he identifies with the first order theory of real closed fields. This theorem led recently to interesting applications in

robotics. But the method of elimination of quantifiers has not yet received the attention it deserves among researchers in automated theorem proving. The state of art in automated theorem proving for elementary geometry is best discussed in [Cho88, SSH87].

Another way of proving completeness of first order theories is based on a simple but ingenious observation due to Vaught, which shows the power of model theoretic reasoning. Let Σ be a complete theory. If Σ has a model $\mathcal{A}$ which is finite, then it is unique up to isomorphism. If $\mathcal{A}$ is infinite, then by the Löwenheim-Skolem Theorem, Σ has models of arbitrary infinite cardinalities. Now, if all models of Σ of infinite cardinality κ are isomorphic, we say that Σ is κ-categorical. Note that if $\mathcal{A}$ and $\mathcal{B}$ are isomorphic then they satisfy the same first order sentences.

THEOREM 5.7 (VAUGHT) *If Σ is κ-categorical for some infinite κ and Σ has no finite models, then Σ is complete.*

Proof. Assume, for contradiction, that there is ϕ such that neither $\Sigma \models \phi$ nor $\Sigma \models \neg\phi$. As Σ has no finite models, using the Löwenheim-Skolem Theorem we can find models $\mathcal{A}$ and $\mathcal{B}$ such that $\mathcal{A} \models \Sigma \cup \{\phi\}$ and $\mathcal{B} \models \Sigma \cup \{\neg\phi\}$, both of cardinality κ. But then $\mathcal{A}$ is isomorphic to $\mathcal{B}$, which contradicts the fact that $\mathcal{A} \models \phi$ and $\mathcal{B} \models \neg\phi$. ∎

Classical mathematical results establish categoricity of a few natural first order theories. Hausdorff and Cantor showed that any two countable dense linear orderings are isomorphic, and a similar argument shows the same for countable atomless boolean algebras. Steinitz showed that any two uncountable algebraic closed fields of characteristic zero of the same cardinality are isomorphic. So Vaught's theorem quickly establishes that these theories are complete and therefore decidable.

6. Spectrum problems

The study of categoricity of first order theories was the driving force behind the deepest results of model theory. Ryll-Nardzewski, Svenonius and Engeler independently characterized ω-categorical theories, and Morley proved the following generalization of Steinitz' theorem:

THEOREM 6.1 (MORLEY) *If Σ is categorical for some uncountable κ then Σ is categorical for every uncountable κ.*

If Σ is not categorical, then it is natural to look at the following: Let Σ be a set of formulas and denote by $I(\Sigma, \kappa)$ the number of non-isomorphic models of cardinality κ. $I(\Sigma, \kappa)$ is called the *spectrum* of Σ. The study of $I(\Sigma, \kappa)$

for infinite κ was initiated by Morley and Vaught (cf. [CK90]). A complete analysis of the infinite case dominated the research efforts in model theory and culminated in Shelah's theorem [She90]:

THEOREM 6.2 (SHELAH'S SPECTRUM THEOREM)
For uncountable κ $I(\Sigma, \kappa)$ is non-decreasing in κ and, in fact either

(i) $I(\Sigma, \kappa) = 2^{\kappa}$ *or*

(ii) $I(\Sigma, \omega_\alpha) < BETH_{\omega_\alpha}(card(\alpha))$.

The infinite spectrum and its ramifications are the core of a highly sophisticated development in model theory called stability theory. Although it is of extreme mathematical depth and beauty I can so far see no fruitful interplay between stability theory and computer science.

Instead of $I(\Sigma, \kappa)$ for finite κ, we shall look at the finite cardinal spectrum $Spec(\Sigma)$ of finite sets of formulas Σ. $Spec(\Sigma)$ is the set of natural numbers n such that there is a finite model of Σ of cardinality n. (The study of $Spec(\Sigma)$ was initiated by Scholz. For the historic remarks cf. [Fag90]). In contrast to stability theory, the study of the finite cardinal spectrum $Spec(\Sigma)$ led to very interesting interactions between model theory and complexity theory, through the pioneering work of Büchi, Fagin and Immerman (cf. [Bü60, Fag74, Imm87]).

Büchi studied the interplay between Monadic Second Order Logic and automata theory. He looked at words over a finite alphabet as finite linearly ordered structures with unary predicates. Recall that a set of words is regular if it is recognizable by a finite automaton. His theorem states:

THEOREM 6.3 (BÜCHI) *A set of words is regular iff it is definable by an existential formula of monadic second order logic.*

Fagin studied the finite spectrum and was led to the following theorem:

THEOREM 6.4 (FAGIN) *A set of finite structures is in* **NP** *iff it is definable by an existential (full) second order sentence.*

Let ϕ be a first order formula over a vocabulary τ. We note that $Spec(\phi)$ can be viewed as the set of finite models of Φ over the empty vocabulary, where Φ is obtained from ϕ by existentially quantifying all the predicate symbols of τ. So Fagin's theorem generalizes both the spectrum problem as well as Büchi's theorem.

Immerman characterized similarly sets of ordered finite structures in **L**, **NL**, **P**. We shall discuss the interplay between model theory and complexity theory in the last section.

7. Beyond first order logic

In this survey we already have come across features which go beyond first order logic. We have tacitly introduced quantification over relations in Büchi's theorem, and we have mentioned the semantic restriction to finite structures. These mark the two independent directions generalizations might take: More sentences vs. more complex models.

The model theoretic study of richer logics over τ-structures in the usual sense was initiated in the late fifties independently by A. Tarski and his students, and E.Engeler for infinite first order formulas, and by A. Mostowski for generalized quantifiers. The book [BF85] contains an excellent bibliography and historic account. From a naive model theoretic point of view it is natural to ask whether for those generalized logics the compactness theorem and the Löwenheim-Skolem theorem are still true. For infinite formulas compactness fails trivially. It was also observed that in all the examples of generalized quantifiers studied one of the two usually failed. In 1966 P. Lindström published a paper which was hardly noticed till 1970. In it the following fundamental result was stated and proved:

THEOREM 7.1 (LINDSTRÖM) *Let $\mathcal{L}$ be a regular logic over τ-structures which both satisfies the compactness theorem and the Löwenheim-Skolem theorem. Then $\mathcal{L}$ is, up to semantic equivalence, first order logic.*

A logic is *regular* if it is closed under boolean operations, quantification, relativization and does not distinguish between isomorphic τ-structures. This theorem was followed by intense investigations of model theories of particular logics and the evolution of a framework for 'abstract model theory'. The fruits of these investigations were collected in the monumental volume [BF85].

In 1965 S. Kripke initiated the model theoretic study of logics different from classical first order or propositional logic, such as intuitionistic logics, modal logics and temporal logics. His main idea was to look at, say propositional modal or temporal logic, as a special case of first order logic. A Kripke-structure is a first order structure with a binary relation for accessibility to possible states (worlds in the case of modal logic, points in time in the case of temporal logic). Propositions then are unary predicates in Kripke-structures. The modal and temporal operators (necessarily/possibly, always/sometimes) now become first order definable. The axioms of modal or temporal logic shape the accessibility relation. In this way Kripke was able to state precisely the semantics of modal logic and prove, for the first time, completeness theorems. To illustrate this let us state here the case of the modal system T, which captures the unproblematic aspects of 'necessity'.

The formula $\Box\phi$ is read as 'necessarily ϕ'. The system T contains all substitutions of propositional tautologies, the axioms $\Box(\phi \to \psi) \to (\Box\phi \to \Box\psi)$ and $\Box\phi \to \phi$, and the two deduction rules Modus Ponens and from ϕ infer $\Box\phi$.

THEOREM 7.2 (KRIPKE) *A modal formula ϕ is provable in T iff ϕ is true in all Kripke-structures with a reflexive accessibility relation.*

We speak of temporal logic when the accessibility relation is a partial order, in the most natural case, a discrete linear order. The formula $\Box\phi$ is now read as 'always ϕ'. It is natural to ask whether the introduction of one temporal operator (or for that matter, modal operator) suffices, or whether there are many hitherto undiscovered temporal operators. Obviously we have operators corresponding to 'next'. 'previously' 'always in the future', 'always in the past', 'ϕ until ψ' and 'ϕ since ψ'. We note that all these operators are first order definable over linearly ordered Kripke-structures. H. Kamp now proved the following remarkable

THEOREM 7.3 (KAMP) *Let $TO(p_1, \ldots, p_n)$ be an n-ary temporal operator which is first order definable over discrete, complete linear orderings. Then $TO(p_1, \ldots, p_n)$ is definable from the operators 'next', 'previously', 'sometimes in the future', 'sometimes in the past', 'until' and 'since'. As a matter of fact, 'until' and 'since' suffice.*

The theorems of Kripke and Kamp are two prime examples of model theoretic theorems in non–standard logic. The underlying techniques, however, are applicable in a much wider context and have not yet been systematically developed. Good surveys are [Bur84, BS84, RS23].

Both types of generalizations of first order logic, more formulas and richer semantic structures, found rich applications in theoretical computer science. Engeler was the first to observe that infinitary logic can serve as a framework to formulate the input/output behaviour of programs. His approach was considered awkward. V. Pratt and D.Kozen used a Kripke-like semantics for an approach to axiomatize the input/output behaviour of programs, which was finally called 'Dynamic Logic'. This was received enthusiastically. However, it was soon observed that the two approaches were equivalent. Burstall suggested modal and Pnueli temporal logic for the axiomatic description of program behaviour. Kripke-structures are also abundant in foundational research in AI, especially in the theory of knowledge.

8. The hidden method

One model theoretic tool of central importance does usually not appear in the statement of theorems, but mostly in their proofs. This is the 'back-and-

forth' characterization of n-equivalent structures, i.e. structures satisfying the same sentences of quantifier rank n. This characterization originated in the early work of R. Fraïssé and was popularized in an influential paper by A. Ehrenfeucht. Ehrenfeucht also generalized the method to monadic second order logic, and further generalizations for infinitary logic and logics with generalized quantifiers and predicate transformers were developed subsequently, cf. [BF85]. In [Mak92b] one may find an extensive discussion of this method, which we call Ehrenfeucht-Fraïssé games. Here we only list some of its application.

Originally, Ehrenfeucht-Fraïssé games were used to prove that certain concepts are not definable by first order formulas even if restricted to finite structures. Among such concepts we find the connectivity and planarity of graphs. The deepest and most surprising application of Ehrenfeucht-Fraïssé games occurs in the proof of Lindström's theorem. A close analysis of this proof also shows that Beth's theorem can be proven using this method, as well as various preservation theorems. Ehrenfeucht's generalization of the method to monadic second order logic can be used to give a model theoretic proof of Büchi's theorem. It was used in [FR79] to establish lower and upper bounds for the complexity of decidable theories such as Presburger Arithmetic and the theory of two successors functions. And finally, it can be also used to prove the 0–1 law for first order logic over finite structures, due independently to R. Fagin and Glebskiĭ, Kogan, Ligon'kiĭ and Talanov.

9. 0–1 laws

To state the 0–1 Theorem, let τ be a vocabulary without function symbols and let ϕ be a first order τ-formula. We think of a structure of size n as having the universe $\{0, 1, \ldots, n-1\}$. Let $S_\tau(n)$ be the number of τ-structures of size n. Recall that $I(\phi, n)$ is the number of different structures of size n satisfying ϕ. Let $P(n, \phi)$ be the fraction of $I(\phi, n)$ and $S_\tau(n)$.

THEOREM 9.1 (0–1-LAW OF FIRST ORDER LOGIC) *For every τ without function symbols and every first order τ-formula ϕ the limit*

$$\lim_{n\to\infty} P(n, \phi)$$

is well defined and is either 0 *or* 1.

DEFINITION 9.2 (ALMOST TRUE FORMULAS) *A First Order Formula ϕ is almost true if* $\lim_{n\to\infty} P(n, \phi) = 1$.

In contrast to First Order Validity over finite structures, which is undecidable by Trakhtenbrot's classical theorem, the set of first order sentences true in almost all structures is decidable. In fact, Grandjean proved [Gra83]:

THEOREM 9.3 (GRANDJEAN) *Assume that τ has no function symbols. The problem of deciding, whether a first order τ–formula ϕ is almost true, is* **P**–*Space complete.*

0–1 Laws were investigated also for extensions of First Order Logic. For a further discussion of similar theorems the reader should consult [Com87, Fag90] and the literature quoted therein. Striking applications of 0–1 Laws in Computer Science are still missing. They may emerge in the context of Average Case Complexity Theory [Gur91], Graph Algorithms [GS87] and the like.

10. Model theoretic aspects of P $\neq$?NP

In this final section we discuss some model theoretic aspects of the question **P** $\neq$?**NP**.

10.1. Non–definability

The first observation ist the following consequence of Fagin's theorem 6.4.

THEOREM 10.1 **Co** – **NP** = **NP** *iff every second order formula of Second Order Logic is equivalent over finite structures to a Σ_1–formula, i.e. a second order formula of the form $\exists R\phi(R)$ with ϕ first order.*

In other words, to show that **Co** – **NP** $\neq$ **NP** it suffices to exhibit a second order formula which is not equivalent over finite structures to a Σ_1–formula.

This seems very difficult, although plausible. A more amenable problem might be to prove that **NL** $\neq$ **NP**. Here we have more hope for two reasons: **NL** (Nondeterministic LogSpace) has been identified by Immerman [Imm87] as the complexity class captured by the logic **TC** obtained from First Order Logic by adding transitive closure operators for $2n$–ary relations. Furthermore, for **TC** an Ehrenfeucht–Fraïssé type game has been defined in [Cal89, CM91]. The existence of similar games as introduced in this paper, already follows from successive papers cumulating in [MM85]. There, they are defined for logics with generalized quantifiers rather than predicate transformers.

The games naturally induce a sequence of equivalence relations $\equiv_n^{TC}$ between structures which we call n–isomorphic for **TC**. In [Cal89] Calò proves soundness and completeness of these equivalence relations in the sense that two structures are n–isomorphic for **TC** iff they satisfy the same formulas of quantifier depth n.

As an application of our work [CM91] we can state a necessary and sufficient conditions for separating the complexity classes **L**, **NL**, **P** and **NP** respectively which is of pure model theoretic character. In the case of $\mathbf{NL} \neq \mathbf{NP}$ this condition can be stated as follows:

Let $HALFCLIQUE$ be the set of ordered graphs which contain a clique of half its size. Let HAM be the set of ordered graphs which contain a hamiltonian path. Note that $HALFCLIQUE$ and HAM are **NP**–complete, cf. [GJ79].

THEOREM 10.2 $\mathbf{NL} \neq \mathbf{NP}$ *iff there is a sequence of pairs of ordered graphs* G_n, H_n *such that*

(i) $G_n \equiv_n^{TC} H_n$ *and*

(ii) $G_n \notin HALFCLIQUE$ *but* $H_n \in HALFCLIQUE$.

The same holds for $HALFCLIQUE$ *replaced by* HAM *or any other* **NP***–complete problem.*

The construction of such families of graphs may be very hard and possibly requires probabilistic methods similar to the ones used in [AF90]. The following result nevertheless sheds some light on the problem.

THEOREM 10.3 $HALFCLIQUE$ *and* HAM *are not definable in Monadic Second Order Logic (with arbitrary alternation of quantifiers) and hence not definable in* $\mathbf{TC}^1$, *the logic obtained from* **TC** *by restricting the transitive closure to binary relations.*

In [dR87] it is only proved that HAM is not definable in existential Monadic Second Order Logic.

This gives us a quick example for interesting families of pairs of ordered finite structures which are n-isomorphic in $\mathbf{TC}^1$.

COROLLARY 10.4 *There are functions* $f, g : \mathbf{N} \mapsto \mathbf{N}$ *such that for every* $n \in \mathbf{N}$ $f(n) \neq g(n)$ *and the words* $a^{f(n)}b^{f(n)}$ *and* $a^{f(n)}b^{g(n)}$ *are* n*-isomorphic (equivalent) in* $\mathbf{TC}^1$.

It would be interesting to estimate the growth rate of the functions f, g. Note that this corollary is a model theoretic analogue of the Pumping Lemma of Formal Language Theory.

10.2. Non–provability

We now address the question whether it could be possible that $\mathbf{P} \neq \mathbf{NP}$ is not provable in some formal system of arithmetic such as Peano Arithmetic or Predicative Analysis. The following variation of the Paris–Harrington Theorem is due to S. Ben–David [BDH91]:

THEOREM 10.5 (S. BEN–DAVID) *There is a language (sets of words recognizable by Turing machines) L_{PH} such that*

(i) L_{PH} is in **Co–NP**;

(ii) L_{PH} is not context free, but

(iii) it is not provable in Peano Arithmetic that L_{PH} is not regular.

The language L_{PH} is very simple. Its words consist of sequences of n a's followed by $r(a)$ b's where r is some fastgrowing function such as the Ramsey function needed in the Paris–Harrington Theorem. This particular language seems to be a good candidate to prove an analogue of corollary 10.4 for **TC**, which would establish that L_{PH} is not in **NL**.

Recently, S. Ben–David has analyzed these results further and related them to discuss the prospect of $\mathbf{P} \neq \mathbf{NP}$ not being provable in some formalized system such as Peano Arithmetic or fragments of Second Order Arithmetic [BDH91]. The key notion here are functions *extremely close to polynomials* where extremely close depends on the growth rate of functions not provably total in the formal system in question. His theorem states the following:

THEOREM 10.6 (S. BEN–DAVID) *If $\mathbf{P} \neq \mathbf{NP}$ is not provable in some fragment of second order Arithmetic* **S** *then every problem P in* **NP** *can be solved by an algorithm with run time upper bound* **S***–extremely close to a polynomial.*

Acknowledgements: I am indebted to many colleagues who encouraged me at several stages to pursue my research of model theoretic methods in computer science. Among them I would like to mention Erwin Engeler, Eli Shamir, Shimon Even, Vaughan Pratt, Catriel Beeri, Saharon Shelah, Jonathan Stavi, David Harel and Yuri Gurevich. I am also indebted to S. Ben–David for valuable discussions and his permission to quote his yet unpublished results. I am also indebted to our graduate students Arie Calò, Yaniv Bargury, Avy Sharell, Yachin Pnueli, Eli Dichtermann, who at some point or another helped me in the preparation of this text.

References

[AF90] M. AJTAI AND R. FAGIN. *Reachability is harder for directed than for undirected finite graphs.* Journal of Symbolic Logic, 55.1:113–150, 1990.

[ANS82] H. ANDRÉKA, I. NEMETI, AND I. SAIN. *A complete logic for reasoning about programs via non-standard model theory, parts I and II.* Theoretical Computer Science, 17:193–212 and 259–278, 1982.

[BDH91] S. BEN-DAVID AND S. HALEVI. *On the independence of P versus NP.* Technical report, Technion–Israel Institute of Technology, Haifa, Israel, 1991.

[BF85] J. BARWISE AND S. FEFERMAN, editors. *Model-Theoretic Logics.* Perspectives in Mathematical Logic. Springer Verlag, 1985.

[BS84] R.A. BULL AND K. SEGERBERG. *Basic modal logic.* In D. Gabbay and F. Günthner, editors, Handbook of Philosophical Logic, volume 2, chapter 1. D. Reidel Publishing Company, 1984.

[Bü60] J.R. BÜCHI. *Weak second-order arithmetic and finite automata.* Zeitschrift für mathematische Logik und Grundlagen der Mathematik, 6:66–92, 1960.

[Bur84] J. BURGESS. *Basic tense logic.* In D. Gabbay and F. Günthner, editors, Handbook of Philosophical Logic, volume 2, chapter 2. D. Reidel Publishing Company, 1984.

[Cal89] A. CALÒ. *The expressive power of transitive closure.* M.Sc. Thesis, Faculty of Computer Science, Technion–Israel Institute of Technology, Haifa, Israel, 1989.

[Cho88] SHANG-CHING CHOU. *Mechanical Geometry Theorem Proving.* Mathematics and its Applications. D. Reidel Publishing Company, 1988.

[CK90] C.C. CHANG AND H.J. KEISLER. *Model Theory.* Studies in Logic, vol 73. North–Holland, 3rd edition edition, 1990.

[CM91] A. CALÒ AND J.A. MAKOWSKY. *The Ehrenfeucht–Fraïssé games for transitive closure.* to appear, 1991.

[Com87] K.J. COMPTON. *A logical approach to asymptotic combinatorics I: First–order properties.* Advances in Mathematics, 65:65–96, 1987.

[Cou90a] B. COURCELLE. *Graph rewriting: An algebraic approach.* In J. van Leeuwen, editor, Handbook of Theoretical Computer Science, volume 2, chapter 5. Elsevier Science Publishers, 1990.

[Cou90b] B. COURCELLE. *The monadic second-order theory of graphs I: Recognizable sets of finite graphs.* Information and Computation, 85:12–75, 1990.

[dR87] M. DE ROUGEMONT. *Second-order and inductive definability on finite structures.* Zeitschrift für mathematische Logik und Grundlagen der Mathematik, 33:47–63, 1987.

[Eme90] E.A. EMERSON. *Temporal and modal logic.* In J. van Leeuwen, editor, Handbook of Theoretical Computer Science, volume 2, chapter 16. Elsevier Science Publishers, 1990.

[Fag74] R. FAGIN. *Generalized first-order spectra and polynomial time recognizable sets.* In R. Karp, editor, Complexity of Computation, pages 27–41. American Mathematical Society Proc, 7, Society for Industrial and Applied Mathematics, 1974.

[Fag82] R. FAGIN. *Horn clauses and database dependencies.* Journal of ACM, 29.4:952–985, 1982.

[Fag90] R. FAGIN. *Finite model theory - a personal perspective.* In S. Abiteboul and P.C. Kannelakis, editors, ICDT'90, volume 470 of Lecture Notes in Computer Science, pages 3–24. Springer Verlag, 1990.

[FR79] J. FERRANTE AND C.W. RACKOFF. *The Computational Complexity of Logical Theories*, volume 718 of Lecture Notes in Mathematics. Springer Verlag, 1979.

[Ga92] D. GABBAY AND AL., editors. *Handbook of Logic in Artificial Intelligence and Logic Programming*, volume 1-4. Oxford University Press, 1992.

[Gal91] J.H. GALLIER. *What is so special about Kruskal's theorem and the ordinal γ_0 ? A survey of some results in proof theory.* Annals of Pure and Applied Logic, 53:199–260, 1991.

[GJ79] M.G. GAREY AND D.S. JOHNSON, editors. *Computers and Intractability.* Mathematical Series. W.H. Freeman and Company, 1979.

[Gra83] E. GRANDJEAN. *Complexity of the first–order theory of almost all structures.* Information and Control, 52:180–204, 1983.

[GS87] Y. GUREVICH and S. SHELAH. *Expected computation time for hamiltonian path problem.* SIAM Journal for Computing, 16.3:486–502, 1987.

[GTWW77] J. GOGUEN, J.W. THATCHER, E.G. WAGNER, AND J.B. WRIGHT. *Initial algebra semantics and continuous algebras.* Journal of ACM, 24:68–95, 1977.

[Gur91] Y. GUREVICH. *Average case completeness.* Journal of Computer and System Sciences, 42:346–398, 1991.

[Har84] D. HAREL. *Dynamic logic.* In D. Gabbay and F. Günthner, editors, Handbook of Philosophical Logic, volume 2, chapter 10. D. Reidel Publishing Company, 1984.

[HMŠS85] L.A. HARRINGTON, M.D. MORLEY, A. ŠČEDROV, AND S.G. SIMPSON, editors. *Harvey Friedman's Research in the Foundations of Mathematics*, volume 117 of Studies in Logic and the Foundations of Mathematics. North Holland, 1985.

[Hod92] W. HODGES. *Logical features of horn clauses.* In D. Gabbay and al., editors, Handbook of Logic in Artificial Intelligence and Logic Programming. Oxford University Press, 1992.

[Imm87] N. IMMERMAN. *Languages that capture complexity classes.* SIAM Journal on Computing, 16(4):760–778, Aug 1987. Also appeared as a preliminary report in Proceedings of the 15th Annual ACM Symposium on the Theory of Computing.

[Kan90] P.C. KANNELAKIS. *Elements of relational database theory.* In J. van Leeuwen, editor, Handbook of Theoretical Computer Science, volume 2, chapter 17. Elsevier Science Publishers, 1990.

[KK82] S. KOCHEN AND S. KRIPKE. *Non–standard models of Peano arithmetic.* In V. Strassen E. Engeler, H. Läuchli, editor, Logic and Algorithmic, An international Symposium held in honour of E. Specker, pages 275–296. L'enseignement mathématique, 1982.

[KM87] A. KANAMORI AND K. MCALOON. *On Gödel incompleteness and finite combinatorics.* Annals of Pure and Applied Logic, 33:23–41, 1987.

[Kol90] P.G. KOLAITIS. *Implicit definability on finite structures and unambiguous computations.* In FOCS'90, pages 168–180. IEEE, 1990.

[KP82] L. KIRBY AND J. PARIS. *Accessible independence results for Peano arithmetic.* Bulletin of the London Mathematical Society, 14:285–293, 1982.

[Mak81] J.A. MAKOWSKY. *Characterizing database dependencies.* In ICALP'81, volume 115 of Lecture Notes in Computer Science, pages 86–97. Springer Verlag, 1981.

[Mak84] J.A. MAKOWSKY. *Model theoretic issues in theoretical computer science, part I: Relational databases and abstract data types.* In G. Lolli and al., editors, Logic Colloquium '82, Studies in Logic, pages 303–343. North Holland, 1984.

[Mak85] J.A. MAKOWSKY. *Compactness, embeddings and definability.* In Model-Theoretic Logics, Perspectives in Mathematical Logic, chapter 18. Springer Verlag, 1985.

[Mak87] J. A. MAKOWSKY. *Why Horn formulas matter for computer science: Initial structures and generic examples.* Journal of Computer and System Sciences, 34.2/3:266–292, 1987.

[Mak88] J.A. MAKOWSKY. *Mental images and the architecture of concepts.* In R. Herken, editor, The Universal Turing Machine. A Half-Century Survey. Oxford University Press, 1988.

[Mak92a] J.A. MAKOWSKY. *Database theory.* In S. Abramsky, D. Gabbay, and T. Maibaum, editors, Handbook of Logic in Computer Science, volume 5. Oxford University Press, 1992.

[Mak92b] J.A. MAKOWSKY. *Model theory and computer science: An appetizer.* In S. Abramsky, D. Gabbay, and T. Maibaum, editors, Handbook of Logic in Computer Science, volume 1, chapter I.6. Oxford University Press, 1992.

[MM84] B. MAHR AND J.A. MAKOWSKY. *Characterizing specification languages which admit initial semantics.* Theoretical Computer Science, 31:49–60, 1984.

[MM85] J.A. MAKOWSKY AND D. MUNDICI. *Abstract equivalence relations.* In Model-Theoretic Logics, Perspectives in Mathematical Logic, chapter 19. Springer Verlag, 1985.

[Mos74] Y. MOSCHOVAKIS. *Elementary Induction on Abstract Structures.* Studies in Logic, vol 77. North–Holland, 1974.

[MS89] J. A. MAKOWSKY AND I. SAIN. *Weak second order characterizations of various program verification systems.* Theoretical Computer Science, 66:299–321, 1989.

[MS92] T. MAIBAUM and M. SADLER. *Axiomatizing specification theory.* In S. Abramsky, D. Gabbay, and T. Maibaum, editors, Handbook of Logic in Computer Science, volume 2. Oxford University Press, 1992.

[MV86] J.A. MAKOWSKY AND M. VARDI. *On the expressive power of data dependencies.* Acta Informatica, 23.3:231–244, 1986.

[Pas90] A. PASZTOR. *Recursive programs and denotational semantics in absolut logics of programs.* Theoretical Computer Science, 70:127–150, 1990.

[RS23] M. RYAN AND M. SADLER. *Valuation systems and consequence relations.* In D. Gabbay and al., editors, Handbook of Logic in Computer Science. Oxford University Press, 1992/3.

[She84] J.C. SHEPHERDSON. *Negation as failure: A comparison of Clark's completed data base and Reiter's closed world assumption.* Journal of Logic Programming, 1:51–81, 1984.

[She85] J.C. SHEPHERDSON. *Negation as failure II.* Journal of Logic Programming, 3:185–202, 1985.

[She88] J.C. SHEPHERDSON. *Negation in logic programming.* In J. Minker, editor, Foundations of deductive data bases and logic programming, chapter 1. Morgan Kaufmann Publishers, 1988.

[She90] S. SHELAH. *Classification Theory and the number of non–isomorphic models.* Studies in Logic and the Foundations of Mathematics. North Holland, 2 edition, 1990.

[SSH87] J.T. SCHWARTZ, M. SHARIR and J. HOPCROFT, editors. *Planning, Geometry, and Complexity of Robot Motion.* Ablex Series in Artificial Intelligence. Ablex Publishing Corporation, 1987.

[TS88] R.W. TOPOR AND E.A. SONENBERG. *On domain independent databases.* In J. Minker, editor, Foundations of deductive data bases and logic programming, chapter 6. Morgan Kaufmann Publishers, 1988.

[Ull82] J.D. ULLMAN. *Principles of Database Systems.* Principles of Computer Science Series. Computer Science Press, 2 edition, 1982.

[Var88] M. VARDI. *Fundamentals of dependency theory.* In E. Börger, editor, Trends in Theoretical Computer Science, chapter 5. Computer Science Press, 1988.

[Wil81] R.L. WILDER. *Mathematics as a Cultural System.* Pergamon Press, 1981.

Logic, Methodology and Philosophy of Science IX
D. Prawitz, B. Skyrms and D. Westerståhl (Editors)

A DECIDABLE QUANTIFIED DEFEASIBLE LOGIC

DONALD NUTE

Artificial Intelligence Programs and Department of Philosophy
The University of Georgia, Athens, GA 30602, U.S.A.
dnute@uga.cc.uga.edu

1. Decidable and undecidable formal systems

For our purposes, a formal system consists of a formal language and a proof theory. Familiar examples are sentential logic and first-order logic (FOL). A well-formed formula (wff) φ in the formal language of a formal system Σ is derivable from a set Γ of such wffs (in symbols, $\Gamma \vdash_\Sigma \varphi$) iff there is a proof of φ from Γ in the proof theory of the system. A formal system Σ is *monotonic* iff for any wff φ and any sets of wffs Γ and Δ in Σ such that $\Gamma \subseteq \Delta$, if $\Gamma \vdash_\Sigma \varphi$, then $\Delta \vdash_\Sigma \varphi$. Of course, Σ is *nonmonotonic* iff there are φ and $\Gamma \subseteq \Delta$ such that $\Gamma \vdash_\Sigma \varphi$ and $\Delta \nvdash_\Sigma \varphi$. Familiar nonmonotonic systems include circumscription (McCarthy 1980, 1986; Lifschitz 1985, 1986), default logic (Reiter 1980), nonmonotonic logic (Doyle and McDermott 1980; Moore 1985), and autoepistemic logic (Moore 1984; Konolige 1988). In this paper we will say that a formal system is *decidable* iff for any finite set of wffs Γ, the set of wffs provable from Γ is recursive. A formal system is *semi-decidable* iff for any finite set of wffs Γ, the set of wffs provable from Γ is recursively enumerable. Sentential logic is decidable. FOL and Horn Clause Logic (HCL) are not decidable, but they are semi-decidable.

Nonmonotonic extensions of semi-decidable theories may also be semi-decidable. For example, consider an FOL language containing all the integers as constants and a monadic predicate F. To the proof theory of FOL, we add the rule that $F(n)$ is derivable from Γ if Γ has n many members. The resulting system is an extension of FOL, but it is nonmonotonic: $F(0)$ is derivable in this system from $\emptyset$, but it is not derivable from $\{F(1)\}$. Since it is decidable whether $F(n)$ is derivable from any finite set of wffs of this language, and since FOL is semi-decidable, this extension of FOL is also semi-decidable.

Nevertheless, most interesting nonmonotonic extensions of FOL are not

semi-decidable. This is because for any one of these systems Σ, there are wffs φ and ψ and some set of wffs Γ of Σ such that to prove φ from Γ in Σ, it is necessary to demonstrate that there is no proof of ψ from Γ in Σ. If any wff of an undecidable fragment of Σ can play the role of ψ in this requirement, then since the set of wffs which are not derivable from Γ is in general not recursively enumerable, Σ cannot be semi-decidable. A similar argument is not possible for a decidable system like sentential logic, and we should expect a nonmonotonic extension of a decidable system also to be decidable.

If a system is at least semi-decidable, it is possible to build a sound and complete theorem prover for it. So we can build sound and complete theorem provers for FOL and HCL. But we cannot in principle build a sound and complete theorem prover for a system that is not even semi-decidable. This suggests that we should look for nonmonotonic extensions of interesting decidable fragments of FOL. These should be not only semi-decidable but actually decidable. This is the approach taken here.

The system of nonmonotonic logic developed here is called *defeasible* because it involves conditional rules which can be blocked or defeated in some way. It is possible in this and similar systems to have a consistent theory which contains such a rule, its antecedent, and the denial of its consequent. This distinguishes defeasible formalisms from circumscription, default logic, autoepistemic logic, and other nonmonotonic formalisms. Defeasible formalisms have been developed by Pollock (1987), Loui (1987), Geffner (1989), Geffner and Pearl (1990) and others. The system presented here evolved from a system described in (Nute 1991). Besides showing that this system is adequate for many standard examples of nonmonotonic reasoning and that it is decidable, I will also show that it has a property similar to one which has been associated with the term *cumulativity* in the literature on nonmonotonic reasoning.

2. Strict logic

The monotonic basis for our defeasible logic will be called *strict* logic. A strict language L is generated by a recursive set B_L of predicates and constants called the *basis* of the language. Atomic formulae of L are formed from members of B_L and some countable set of variable expressions in the usual way. A negation symbol $\sim$ is introduced, and φ is a literal of L if φ is either an atomic formula of L or the negation of an atomic formula of L. A ground literal is a literal in which no variable expression occurs. Where φ is an atomic formula, we say φ and $\sim\varphi$ are the complements of each other. We represent the complement of any literal ψ as $\neg\psi$.

Next we introduce a metalinguistic symbol $\rightarrow$ and use it in the definition of a strict rule. Then we define a strict theory and present the simple proof theory for strict logic.

DEFINITION 1 *A* strict rule *is a triple* $(A, \rightarrow, \varphi)$ *where* A *is a set of literals and* φ *is a literal. We represent* $(A, \rightarrow, \varphi)$ *as* $A \rightarrow \varphi$ *and* $(\{\varphi\}, \rightarrow, \psi)$ *as* $\varphi \rightarrow \psi$.

DEFINITION 2 *A* strict theory *is a pair* (K, R) *where* K *is a set of literals and* R *is a set of strict rules.*

DEFINITION 3 *A* ground substitution *is a function* θ *which assigns to each variable* v *a constant* $\theta(v)$. *Let* θ *be a ground substitution,* $F(t_1, ..., t_n)$ *an atomic formula,* φ *a wff, and* Γ *a set of wffs. Then* $F(t_1, ..., t_n)\theta = F(t_1\theta, ..., t_n\theta)$, $(\sim\varphi)\theta = \sim\varphi\theta$, *and* $\Gamma\theta = \{\psi\theta : \psi \in \Gamma\}$.

DEFINITION 4 t *is a* strict proof *of* φ *from* (K, R) *iff* t *is a finite labeled tree,* φ *is a ground literal, and* (K, R) *is a strict theory such that*

1. *the top node of* t *is labeled* φ, *and*
2. *for every node* n *in* t, *either*
   (a) *there is* $\psi \in K$ *and an ground substitution* θ *such that* n *is labeled* $\psi\theta$ *or*
   (b) *there is* $A \rightarrow \psi \in R$ *and a ground substitution* θ *such that* n *is labeled* $\psi\theta$ *and* n *has a child labeled* $\alpha\theta$ *for each* $\alpha \in A$.

DEFINITION 5 φ *is* strictly derivable from (K, R) *(in symbols,* $(K, R) \vdash_S \varphi$*) iff there is a strict proof of* φ *from* (K, R).

Intuitively, a strict rule $A \rightarrow \varphi$ is a metarule that says φ is to be inferred from A; i.e., it is an inference rule. Soon we will introduce defeasible rules as well. The intended interpretation of both strict and defeasible rules in defeasible logic is as honest-to-goodness rules of inference or policies for belief fixation and revision (Israel 1980, Nute 1992). They play a different role than that of the literals in a strict language. Literals represent contents of beliefs; rules represent policies for accepting beliefs. Thus, we have a two-tiered language where rules are metalinguistic expressions in the higher tier denoting literals in the lower tier. Our proof theory for both strict and defeasible logic can be viewed as telling us what it means to comply with some set of strict or some set of strict and defeasible rules. Thus, our

proof theory provides *compliance conditions* for these kinds of inference rules (Nute 1988, 1992).

Despite the intended interpretation of strict rules just explained, strict logic is equivalent to a fragment of FOL. In this fragment, the only wffs are literals and expressions of the form $\varphi \supset \psi$ where φ is a conjunction of literals and ψ is a literal. The system has no axioms, and its only rule is a modified *modus ponens* which says to infer ψ from φ_1,...,φ_k, and $\varphi_1 \wedge ... \wedge \varphi_k \supset \psi$. Using this rule we would define linear proofs in the usual way. It is more convenient to define proofs for strict logic and its successor, defeasible logic, as trees because this simplifies the decidability argument and because it mirrors the search strategy of an automated theorem prover for the systems.

Strict logic is both monotonic and decidable.

3. Some examples of defeasible reasoning

Now we look at some examples which will motivate both the formal language and the proof theory for our logic of defeasible reasoning. Some of these examples have appeared often in the literature; others are new and demonstrate features of defeasible reasoning that have received little attention.

EXAMPLE 1 (NIXON DIAMOND) *Quakers are normally pacifists but republicans typically are not pacifists. Nixon is both a Quaker and a republican.*

Notice first that we have general principles here which may admit of exceptions. The uses of the terms 'normally' and 'typically' indicate this explicitly. We will call general principles like this, as well as their singular counterparts, *defeasible rules*. We might rephrase the first of these as 'If something is a Quaker, then normally it is also a pacifist' or as 'Take a thing's being a Quaker as evidence that it is a pacifist'. What conclusion can we draw about whether Nixon is a pacifist based on the information given in this example? None at all. We have here evidence supporting contradictory conclusions and no criteria for deciding between them. We might say that in the case of Nixon, our rules concerning both Quakers and republicans are defeated. Or perhaps more precisely, we can say that our application of each of these rules to Nixon is defeated. Moreover, we can say that each of the rules is defeated by the other.

EXAMPLE 2 (TWEETY TRIANGLE) *Birds normally fly; but penguins are birds and penguins normally don't fly. Tweety is a penguin.*

Here again, we have two competing principles. But in this case, we can resolve the conflict and conclude tentatively that Tweety apparently does

not fly. Notice that unlike the two defeasible rules in this example, the principle that penguins are birds is not defeasible. It is a strict rule. How do we resolve the conflict between 'Birds normally fly' and 'Penguins normally don't fly'? By using the information that penguins are birds. We know what normal or typical birds do, but we also know that penguins are special kinds of birds. Since they don't fly, we know in fact that one way in which they are special is in this respect. Put another way, the information that Tweety is a penguin is more specific than the information that Tweety is a bird, and any rule for penguins will be more specific than any rule for birds in general. This example suggests that when we have two competing defeasible rules and we can determine that one is more specific than the other, we prefer the more specific rule.

EXAMPLE 3 *College students normally don't have full-time jobs. But college students normally are adults and adults typically have full-time jobs. Joe is a college student.*

This example is similar to the Tweety Triangle, but here the connection between college students and adults is itself defeasible. Nevertheless, we are inclined to conclude that rules about college students are more specific than rules about adults and to resolve the conflict by concluding that Joe evidently does not have a full-time job.

EXAMPLE 4 *As we noted in the previous example, college students normally don't have full-time jobs even though college students are normally adults and adults typically have full-time jobs. Furthermore, people with jobs normally pay income tax and college students normally don't. Fred is a college student but he also has a full-time job.*

This example shows some of the complications that arise when we try to use the notion of specificity to resolve conflicts between rules. Here we have conflicting rules about who pays income tax. The conditions of both rules are satisfied by Fred. Although we have a chain of rules leading from being a college student to having a full-time job, we also know that Fred is an atypical college student with regard to his employment. So we do not say that the rule for college students is more specific than the rule for people with jobs. We can draw no tentative conclusion about whether Fred pays taxes based on this collection of rules and information. Our formalization of specificity should be correct for examples of this sort.

EXAMPLE 5 *Normally, something that looks red is red. But something that looks red under red light might not be red. Observing my car under red light, it looks red.*

The second principle in this example is unlike any we have seen before. Unlike either a strict or a defeasible rule, it does not tell us that some proposition is evidence for something else. Instead, it tells us that under certain conditions, something does not count as evidence which ordinarily would count as evidence. In our other examples, we can say that each of the conflicting rules *rebuts* its competitor. But our 'might' rule does not rebut its competitor since it does not give us evidence for its own conclusion. Rather it *undercuts* its competitor. We should not conclude, tentatively or otherwise, that my car either is or is not red on the basis of the facts and principles comprising this example. We will call a rule like this a *defeater.*

Notice that simply because the corresponding positive principles are defeasible, they allow that a republican might be a pacifist, a bird might not fly, a college student might have a full-time job, etc. But it would be peculiar to respond to an argument that presumably Tweety flies since Tweety is a bird and birds normally fly by objecting that a bird might not fly. The proper response to such an objection would be something like, "That's true. That's why I said that birds *normally* fly – because there are exceptions. But unless you give me some reason for thinking that Tweety is an exception, I am justified in tentatively concluding that Tweety flies." The point is that being a bird is not a reason to call into doubt a claim that something flies; but being sick or being very young is a reason to call into doubt a claim that a bird flies. We voice defeaters only to undercut defeasible rules, to point out circumstances under which the rules should not be applied.

EXAMPLE 6 *Native speakers of Pennsylvania Dutch are native speakers of (a dialect of) German. Native speakers of Pennsylvania Dutch are normally born in Pennsylvania. People born in Pennsylvania are born in the United States. Native speakers of German normally are not born in the United States. Hermann is a native speaker of Pennsylvania Dutch.*

One of our conflicting rules is strict while the other is defeasible. But the antecedent of the strict rule is itself a consequent of a defeasible rule, while the antecedent of the defeasible rule is a consequent of a strict rule. We should conclude that Hermann definitely is a native speaker of (a dialect of) German and that he apparently was born in Pennsylvania. Intuitively, we also conclude that apparently Hermann was born in the United States. This is because we prefer strict rules over defeasible rules, even when the commitment to the condition of the strict rule is only tentative. This also suggests that, as a general rule, we need only look at the conditions of the conflicting rules and the relations between them rather than at the arguments supporting these conditions.

EXAMPLE 7 *Football fans normally drink beer and Mormons normally do not drink beer. Joel is a Mormon football fan. Mormon football fans normally don't drink beer.*

Without the final principle in this example, it has the same form as the Nixon Diamond. But rules for Mormon football fans are clearly more specific than rules for football fans in general. So we conclude, at least tentatively, that Joel does not drink beer.

A problem with this example is that we need separate rules for Mormons and for Mormon football fans. We need the first rule to support conclusions about Mormons when we have no idea whether or not they are football fans. We need the second rule to rebut our general principle for football fans. There is a construction in English that allows us to represent the situation more elegantly.

EXAMPLE 8 *Football fans normally drink beer, but Mormons normally don't drink beer even if they are football fans. Alzina is a Mormon and Joel is a Mormon football fan.*

A single rule allows us to conclude that neither Alzina nor Joel drinks beer. The rule about Mormons does not require us to establish that Alzina is a football fan before we can conclude that she does not drink beer. However, the 'even–if' clause in the rule allows us to discount the football fan rule as a rebutting defeater. And this is the usual role of 'even–if' in English: to discount potential defeaters. We will incorporate this treatment of 'even–if' into our defeasible logic.

EXAMPLE 9 *Normally local residents are permitted to vote in local elections even if they do not have local jobs. But convicted felons are normally not permitted to vote in local elections even if they are local residents. Knuckles is a local resident and a convicted felon.*

Here are competing rules each of which has an even-if condition. Intuitively, whether Knuckles works locally is irrelevant and Knuckles apparently is not permitted to vote in local elections.

EXAMPLE 10 *Presumably, the water from the faucet is potable. But if the water from the faucet has a peculiar odor, it may not be potable. The water from the faucet has a peculiar odor.*

A weak initial premise is stated as a presumption. We also have an undercutting defeater for this presumption whose antecedent condition is satisfied. The intuitively correct conclusion is no conclusion. We cannot conclude even tentatively that the water from the faucet either is or is not potable.

EXAMPLE 11 (YALE SHOOTING PROBLEM) *At time t_0, a gunman holds a loaded gun and her victim is alive. At time t_1, she accurately aims the gun at her victim and fires. We presume that features of situations tend to remain unchanged from one time to another. We also accept the rule that if a loaded gun is accurately aimed at a person and fired at some time, then that person normally is not alive at any succeeding time.*

This example, a simplified version of an example in (Hanks and McDermott 1987), involves a simple case of temporal reasoning. The question is whether the intended victim is alive at time t_2. The intuitive answer is that he is not. The importance of the Yale Shooting Problem is that other nonmonotonic formalisms such as circumscription, default logic, nonmonotonic logic, and autoepistemic logic have problems getting it right. These approaches emphasize violation of a minimum number of defeasible principles. In the example, we can violate the temporal persistence principle with regard to the gun remaining loaded from time t_0 to time t_1, we can violate the persistence principle with regard to the victim remaining alive from time t_1 to time t_2, or we can violate the rule about the normal effects of accurately shooting a loaded gun at someone with regard to the intended victim. The victim's survival varies according to which principle we choose to violate.

With these examples as models to guide us, we will develop a defeasible extension of strict logic.

4. A language for defeasible logic

Starting with the formal language for strict logic, we add three new logical symbols @, $\Rightarrow$ and $\rightsquigarrow$. We use these to formulate a new kind of wff and two new kinds of rules.

DEFINITION 6 *A* tentative conclusion *is an expression of the form $@\varphi$ where φ is any literal.*

We read $@\varphi$ as 'Apparently, φ'. Other commonly used qualifiers are 'evidently', 'probably', and 'presumably', but we will reserve 'presumably' to mark tentative initial premises and take 'apparently' as the qualifier of choice for expressing conclusions that may rest on defeasible reasoning.

Literals and tentative conclusions comprise the wffs in our formal language. A ground wff is one in which no variable expression occurs. Besides the wffs, we have a second level of language made up of rules. These rules express policies about when we should believe the propositions expressed by the literals in our language. Besides strict rules, our language includes defeasible rules and defeaters.

DEFINITION 7 *A* defeasible rule *is an ordered quadruple* $(A, B, \Rightarrow, \varphi)$ *where* A *and* B *are sets of literals and* φ *is a literal. We represent* $(A, B, \Rightarrow, \varphi)$ *as* $A|B \Rightarrow \varphi$. *A defeasible rule* $A|B \Rightarrow \varphi$ *is a* presumption *iff* $A = \emptyset$. *We represent* $A|\emptyset \Rightarrow \varphi$ *as* $A \Rightarrow \varphi$, $\emptyset|B \Rightarrow \varphi$ *as* $|B \Rightarrow \varphi$, *and* $\emptyset|\emptyset \Rightarrow \varphi$ *as* $\Rightarrow\varphi$. *Where* A *or* B *has only one member, we omit the curly braces.*

We read $A|B \Rightarrow \varphi$ as 'Take everything in A being true as evidence for φ, even if some or all members of B are true', or colloquially as 'If all of A, then even if some or all of B, φ'. We read $\Rightarrow\varphi$ as 'Presumably, φ', and $|A \Rightarrow\varphi$ as 'Presumably, φ even if A'.

DEFINITION 8 *A* defeater *is a triple* $(A, \rightsquigarrow, \varphi)$ *where* A *is a set of literals and* φ *is a literal. We represent* $(A, \rightsquigarrow, \varphi)$ *as* $A \rightsquigarrow \varphi$ *and we represent* $(\{\varphi\}, \rightsquigarrow, \psi)$ *as* $\varphi \rightsquigarrow \psi$.

We read $A \rightsquigarrow \varphi$ as 'Take everything in A being true as a reason to doubt $\sim\varphi$', or colloquially as 'If all of A, then it might be that φ'. Defeaters do not give us evidence for anything. Their role is to inform us of circumstances under which we should doubt any evidence we have for some conclusion.

DEFINITION 9 T *is a* defeasible theory *iff there exist a finite set* K *of literals and a finite set* R *of rules such that* $T = (K, R)$.

Obviously, every strict theory is also a defeasible theory. Notice that tentative conclusions do not occur in defeasible theories. A tentative conclusion is used to indicate that some literal is supported by some theory. The inclusion of a literal in a defeasible theory indicates that the literal is accepted without possibility of defeat. The only way to withdraw commitment from a literal included in a defeasible theory is to remove the literal itself. However, we may on occasion wish to include a tentative initial premise in a theory, a premise that might be overridden or defeated when reviewed in light of other premises and rules in the theory. We do this, not by adding a tentative conclusion to the theory, but by adding a presumption.

The formal language we have described is clearly limited and does not have the power to express many kinds of statements or inference rules. Nevertheless, it is strong enough to represent all of the examples of the last section. Now we need a proof theory that will also allow us to derive the intuitively correct conclusions in these examples.

5. The monotonic core of defeasible logic

While strict logic provides the underlying monotonic foundation for our defeasible logics, some basic changes in the proof theory are required. The

proofs in strict logic are trees whose nodes are labeled by literals. The proofs in defeasible logics are also labeled trees, but the labels are more complex.

DEFINITION 10 *t is a* proof tree *iff t is a finite labeled tree and for every node n in t there is a defeasible theory (K, R) and a ground literal or ground tentative conclusion φ such that n is labeled (K, R, φ^+) or n is labeled (K, R, φ^-).*

In the system we are developing, it is sometimes necessary to show that something is not derivable in order to show that something else is derivable. Intuitively, a node labeled by (K, R) and φ^+ indicates that φ is derivable from (K, R), while a node labeled by (K, R) and φ^- indicates that φ is *demonstrably not derivable* from (K, R). Of course, we will never want the same wff to be both derivable and demonstrably not derivable from the same defeasible theory. Our definition of a defeasible logic will incorporate this requirement.

If we are to use specificity as a means of resolving conflicts between different rules, we must sometimes show that the antecedent conditions of one rule are derivable from the antecedent conditions of another. Thus, our subderivations may be based on different defeasible theories and it is necessary to label the nodes in our proof trees with the defeasible theories upon which those particular nodes depend.

DEFINITION 11 *Where n is a node in a proof tree and n is labeled $(K, R, \varphi^{+/-})$, a substitution θ is* restricted to n *iff the range of θ is restricted to the set of constants occurring in (K, R) and φ.*

Of course, not just any proof tree will constitute a defeasible proof. The nodes in trees must be supported in appropriate ways by their children just as the nodes in a strict proof are supported by their children. In fact, one way that a node may be supported by its children derives directly from the proof theory for strict logic. Thus we have the following condition which a node might satisfy, thereby justifying it as a proper "step" in a defeasible proof.

$\mathbf{M}^+$: Node n is labeled (K, R, φ^+) and either there is $\psi \in K$ and a substitution θ restricted to n such that $\psi\theta = \varphi$ or there is $A \rightarrow \psi \in R$ and a substitution θ restricted to n such that $\psi\theta = \varphi$ and for every $\alpha \in A$, n has a child labeled $(K, R, \alpha\theta^+)$.

Since our system is defeasible, we must also have a way to show that a literal is not derivable from a defeasible theory. This principle must be a strong negation of $\mathbf{M}^+$.

$\mathbf{M}^-$: Node n is labeled (K, R, φ^-), there is no $\psi \in K$ and substitution θ restricted to n such that $\psi\theta = \varphi$, and for every $A \rightarrow \psi \in K$ and substitution θ restricted to n such that $\psi\theta = \varphi$, n has a child labeled $(K, R, \alpha\theta^-)$ for some $\alpha \in A$.

If we can derive a literal in strict logic, then we certainly want the corresponding tentative conclusion to be derivable in our defeasible logic. This gives us another principle. As with $\mathbf{M}^+$, this new principle $\mathbf{E}^+$ has its negative counterpart: $@\varphi$ should not be derivable when $\neg\varphi$ is derivable. The only exception (because of $\mathbf{E}^+$) is when φ is also derivable.

$\mathbf{E}^+$: Node n is labeled $(K, R, @\varphi^+)$ and n has a child labeled (K, R, φ^+).

$\mathbf{E}^-$: Node n is labeled $(K, R, @\varphi^-)$, n has a child labeled (K, R, φ^-), and n has a child labeled $(K, R, \neg\varphi^+)$.

DEFINITION 12 $\mathbf{M} = \{\mathbf{M}^+, \mathbf{M}^-, \mathbf{E}^+, \mathbf{E}^-\}$.

We will define a defeasible logic as a set of conditions on proof trees. We will want to include $\mathbf{M}$ in the set of conditions that constitute any defeasible logic. We can think of $\mathbf{M}$ as comprising the *monotonic core* of defeasible logic. But we will want our logics to satisfy another important condition: it should not be possible for any wff φ and any theory (K, R) both to show that φ is derivable and that φ is demonstrably not derivable from (K, R). Before doing this, we will formally define what we mean when we say a wff is either derivable or demonstrably not derivable from a defeasible theory relative to some set of conditions on proof trees.

DEFINITION 13 *Where Σ is a set of conditions on proof trees, (K, R) is a defeasible theory, and φ is a wff, t is a* Σ-proof for $\varphi^{+/-}$ from (K, R) *iff t is a proof tree, the top node of t is labeled $(K, R, \varphi^{+/-})$ and for every node n in t, there is a condition in Σ which n satisfies.*

DEFINITION 14 *Where Σ is a set of conditions on proof trees, (K, R) is a defeasible theory, and φ is a wff,* φ is Σ-derivable from (K, R) *(in symbols, $(K, R) \vdash_\Sigma \varphi$) iff there is a Σ-proof for φ^+ from (K, R), and* φ is demonstrably not Σ-derivable from (K, R) *(in symbols, $(K, R) {}_\Sigma\dashv \varphi$) iff there is a Σ-proof for φ^- from (K, R).*

DEFINITION 15 Σ *is a* defeasible logic *iff Σ is a set of conditions on nodes in proof trees, $\Sigma \supseteq \mathbf{M}$, and there is no defeasible theory T and wff φ such that both $T \vdash_\Sigma \varphi$ and $T {}_\Sigma\dashv \varphi$.*

THEOREM 1 **M** *is a defeasible logic.*

A proof for the propositional counterpart to this theorem is found in (Nute 1991). The proof proceeds by induction on the depth of a proof tree and can easily be adapted to the quantified case.

Literals and tentative conclusions are not equivalent in defeasible logic. For example, from the empty theory $(\emptyset, \emptyset)$ we can construct a proof in **M** for φ^- but not for $@\varphi^-$.

6. The quantified defeasible logic $\mathbf{QD_e}$

$\mathbf{E}^+$ lets us infer that a literal is tentatively or defeasibly derivable from the fact that it is strictly derivable, and $\mathbf{E}^-$ lets us infer that a literal is demonstrably not tentatively derivable from the fact that it is not strictly derivable and its complement is strictly derivable. In either case, we draw a conclusion weaker than the premise(s) it is based on. Other conditions for proof trees permit us to infer that a literal is or is not defeasibly derivable without first deriving that it or its complement is strictly derivable.

There are two different cases to consider. The first is the case where the antecedent of a strict rule is only defeasibly derivable. The second is the case where the antecedent of a defeasible rule is strictly or defeasibly derivable.

Consider the case of the strict rule

$$\{\sim Married(x), Male(x), Adult(x)\} \rightarrow Bachelor(x)$$

and suppose we can derive the tentative conclusions

$$@\sim Married(john)$$

$$@Male(john)$$

$$@Adult(john)$$

Then we have reason to draw the tentative conclusion $@Bachelor(john)$. Under what circumstances should this inference be blocked? First and most obviously, we would not draw this conclusion if we *know* that John is not a bachelor, i.e., if $\sim Bachelor(john)$ is strictly derivable. (Under these circumstances, we would also know that John is either married, not a male, or not an adult, but neither strict logic nor defeasible logic allows this kind of *contrapositive* inference.) But there is another situation where we might want to block the inference. Let's add the strict rule

$$Infant(x) \rightarrow \sim Bachelor(x)$$

to our theory and suppose that we can also derive the tentative conclusion $@Infant(john)$. A complete model of the relevant portion of our conceptual scheme would include rules that tell us something can't be both an infant and an adult, but we will assume that these rules are missing from our theory. What we have, then, are two conflicting strict rules and we can defeasibly derive the antecedents of both. Something has clearly gone wrong either in our knowledge representation or in our information, but we may not be able to pinpoint the problem. What we can do, though, is a bit of damage control. In this situation, we can refrain from inferring either $@Bachelor(john)$ or $@{\sim}Bachelor(john)$. This amounts to saying that a strict rule can be defeated by another strict rule if its antecedent is only defeasibly derivable. We will call a defeasible logic *semi-strict* if strict rules may defeat each other in this way. Otherwise, we will say that a defeasible logic is *strict*. These terms are defined with greater precision in (Nute 1991).

$\mathbf{S}_\mathbf{s}^+$: Node n is labeled $(K, R, @\varphi^+)$, n has a child labeled $(K, R, \neg\varphi^-)$, there is $A \rightarrow \psi \in R$ and a substitution θ restricted to n such that $\psi\theta = \varphi$ and n has a child labeled $(K, R, @\alpha\theta^+)$ for each $\alpha \in A$, and for each $B \rightarrow \chi \in R$ and substitution π for n such that $\chi\pi = \neg\varphi$ there is $\beta \in B$ and a child of n labeled $(K, R, @\beta\pi^-)$.

There is no corresponding condition $\mathbf{S}_\mathbf{s}^-$. Instead, the strong negation of $\mathbf{S}_\mathbf{s}^+$ must be incorporated into whatever condition we have for applying defeasible rules.

In formulating conditions for applying defeasible rules, we must keep in mind our motivating examples. Defeasible rules are always defeated by competing strict rules. They are also defeated by competing defeasible rules or defeaters unless they are more specific than their competitors. There are several ways to characterize specificity and some of these are stated precisely in (Nute 1991). Here we will present only one form of specificity.

In showing specificity, we derive the conditions of one rule from the conditions of another. That is, we treat the conditions of the more specific rule as the set of literals in a defeasible theory for the purposes of the subderivation. However, we use all the rules available in the original theory with the exception of those defeasible rules with empty antecedent conditions.[1] Thus, we label nodes in proof trees with theories consisting of the antecedents of rules and the following subset of the set R of rules in our original theory:

[1]The reason we exclude defeasible rules with empty antecedents (presumptions) is that if $\Rightarrow\varphi$ is in our theory, then by using this presumption we would be able to show that $@\varphi$ is derivable from $(\emptyset, R)$. For example, $\{\varphi, \psi\} \Rightarrow \chi$ would be no more specific than $\psi \Rightarrow \chi$ if our rule set contained only these two rules and $\Rightarrow\varphi$. This is discussed more fully in (Nute 1991).

$$R* = R - \{\emptyset | Z \Rightarrow \nu : \emptyset | Z \Rightarrow \nu \in R\}$$

A complication not considered in (Nute 1991) is how we use even-if conditions in deciding specificity. I propose the following requirements for applying defeasible rules with even-if conditions.

$\mathbf{D_e^+}$ Node n is labeled $(K, R, @\varphi^+)$, n has a child labeled $(K, R, \neg\varphi^-)$, there is $A|B \Rightarrow \psi$ and a substitution θ restricted to n such that

1. $\psi\theta = \varphi$;
2. for each $\alpha \in A$, n has a child labeled $(K, R, @\alpha\theta^+)$;
3. for each $\Gamma \rightarrow \chi \in R$ and substitution π restricted to n such that $\chi\pi = \neg\varphi$, there is $\gamma \in \Gamma$ and a child of n labeled $(K, R, @\gamma\pi^-)$; and
4. for each $\Delta|E \Rightarrow \mu \in R$ or $\Delta \rightsquigarrow \mu \in R$, and substitution ρ restricted to n such that $\mu\rho = \neg\varphi$, either
   (a) there is $\delta \in \Delta$ and a child of n labeled $(K, R, @\delta\rho^-)$, or
   (b) there is $\alpha \in A \cup B$ and a child of n labeled $(\Delta\rho, R*, @\alpha\theta^-)$, and for each $\delta \in \Delta$, there is a child of n labeled $((A \cup B)\theta, R*, @\delta\rho a^+)$.

$\mathbf{D_e^-}$ Node n is labeled $(K, R, @\varphi^-)$ and either n has a child labeled $(K, R, \neg\varphi^+)$ or

1. there is no $\psi \in K$ and substitution θ restricted to n such that $\psi\theta = \varphi$;
2. for each $A \rightarrow \chi \in R$ and substitution θ restricted to n such that $\chi\theta = \varphi$, either
   (a) there is $\alpha \in A$ and a child of n labeled $(K, R, @\alpha\theta^-)$, or
   (b) there is $B \rightarrow \mu \in R$ and a substitution π restricted to n such that $\mu\pi = \neg\varphi$ and for each $\beta \in B$, n has a child labeled $(K, R, @\beta\pi^+)$;and
3. for each $\Gamma|\Delta \Rightarrow \nu \in R$ and substitution ρ restricted to n such that $\nu\rho = \varphi$, either
   (a) there is $\gamma \in \Gamma$ and a child of n labeled $(K, R, @\gamma\rho^-)$, or
   (b) there is $Z \rightarrow \eta \in R$ and substitution σ restricted to n such that $\eta\sigma = \neg\varphi$, and for every $\zeta \in Z$, n has a child labeled $(K, R, @\zeta\sigma^+)$, or

(c) there is $\Lambda|X \Rightarrow \upsilon \in R$ or $\Lambda \rightsquigarrow \upsilon \in R$, and substitution τ restricted to n such that $\upsilon\tau = \neg\varphi$, n has a child labeled $(K, R, @\lambda\tau^+)$ for each $\lambda \in \Lambda$, and either there is $\lambda \in \Lambda$ and a child of n labeled $((\Gamma \cup \Delta)\rho, R*, @\lambda\tau^-)$, or for each $\gamma \in \Gamma \cup \Delta$ there is a child of n labeled $(\Lambda\tau, R*, @\gamma\rho^+)$.

DEFINITION 16 $\mathbf{QD_e} = \mathbf{M} \cup \{\mathbf{S_s^+}, \mathbf{D_e^+}, \mathbf{D_e^-}\}$.

THEOREM 2 $\mathbf{QD_e}$ *is a defeasible logic.*

7. The examples revisited

In this section, we will see how the different examples listed earlier can be represented as defeasible theories and what conclusions are derivable in $\mathbf{QD_e}$ for these theories.

EXAMPLE 1 (NIXON DIAMOND) Let

$$K = \{Quaker(nixon), Republican(nixon)\}$$

and

$$R = \{Quaker(x) \Rightarrow Pacifist(x), Republican(x) \Rightarrow \sim Pacifist(x)\}.$$

Then (K, R) $\mathbf{QD_e}\dashv @Pacifist(nixon)$ and (K, R) $\mathbf{QD_e}\dashv @\sim Pacifist(nixon)$.

EXAMPLE 2 (TWEETY TRIANGLE) Let $K = \{Penguin(tweety)\}$ and

$$R = \{Bird(x) \Rightarrow Flies(x), Penguin(x) \Rightarrow \sim Flies(x), \\ Penguin(x) \rightarrow Bird(x)\}.$$

Then $(K, R) \vdash_{\mathbf{QD_e}} @\sim Flies(tweety)$.

EXAMPLE 3 Let $K = \{College(joe)\}$ and

$$R = \{College(x) \Rightarrow \sim HasJob(x), College(x) \Rightarrow Adult(x), \\ Adult(x) \Rightarrow HasJob(x)\}.$$

Then $(K, R) \vdash_{\mathbf{QD_e}} @\sim HasJob(joe)$.

EXAMPLE 4 Let $K = \{College(fred), Job(fred)\}$ and

$$R = \{College(x) \Rightarrow \sim HasJob(x), College(x) \Rightarrow Adult(x), \\ Adult(x) \Rightarrow HasJob(x), HasJob(x) \Rightarrow PaysTax(x), \\ College(x) \Rightarrow \sim PaysTax(x)\}.$$

Then (R, K) $\mathbf{QDe}\dashv$ $@PaysTax(fred)$ and (R, K) $\mathbf{QDe}\dashv$ $@\sim PaysTax(fred)$.

EXAMPLE 5 Let $K = \{LooksRed(car), UnderRedLight(car)\}$ and

$$R = \{LooksRed(x) \Rightarrow Red(x), \{LooksRed(x), UnderRedLight(x)\} \\ \rightsquigarrow \sim Red(x)\}.$$

Then (K, R) $\mathbf{QDe}\dashv$ $@Red(car)$ and (K, R) $\mathbf{QDe}\dashv$ $@\sim Red(car)$.

EXAMPLE 6 Let $K = \{DutchSpeaker(hermann)\}$ and

$$R = \{DutchSpeaker(x) \rightarrow GermanSpeaker(x), \\ DutchSpeaker(x) \Rightarrow BornInPA(x), BornInPA(x) \rightarrow BornInUSA(x), \\ GermanSpeaker(x) \Rightarrow \sim BornInUSA(x)\}.$$

Then $(K, R) \vdash_{\mathbf{QDe}} @BornInUSA(hermann)$.

EXAMPLE 7 Let $K = \{Mormon(joel), Fan(joel)\}$ and

$$R = \{Fan(x) \Rightarrow DrinksBeer(x), Mormon(x) \Rightarrow \sim DrinksBeer(x), \\ \{Mormon(x), Fan(x)\} \Rightarrow \sim DrinksBeer(x)\}.$$

Then $(K, R) \vdash_{\mathbf{QDe}} @\sim DrinksBeer(joel)$.

EXAMPLE 8 Let $K = \{Mormon(alzina), Mormon(joel), Fan(joel)\}$ and

$$R = \{Fan(x) \Rightarrow DrinksBeer(x), Mormon(x) | Fan(x) \Rightarrow \sim DrinksBeer(x)\}.$$

Then $(K, R) \vdash_{\mathbf{QDe}} @\sim DrinksBeer(alzina)$ and $(K, R) \vdash_{\mathbf{QDe}} @\sim DrinksBeer(joel)$.

EXAMPLE 9 Let $K = \{Resident(knuckles), Felon(knuckles)\}$ and

$$R = \{Resident(x)|\sim HasJob(x) \Rightarrow Voter(x),$$
$$Felon(x)|Resident(x) \Rightarrow \sim Voter(x)\}.$$

Then $(K, R) \vdash_{\mathbf{QD_e}} @\sim Voter(knuckles)$.

EXAMPLE 10 Let $K = \{Odor(water)\}$ and

$$R = \{\Rightarrow Potable(water), Odor(water) \rightsquigarrow \sim Potable(water)\}.$$

Then $(K, R) \ _{\mathbf{QD_e}}\dashv @Potable(water)$ and $(K, R) \ _{\mathbf{QD_e}}\dashv @\sim Potable(water)$.

EXAMPLE 11 (YALE SHOOTING PROBLEM) Let

$$K = \{Loaded(t_0), Alive(t_0), Fired(t_1), Next(t_0, t_1), Next(t_1, t_2)\}$$

and

$$R = \{\{Loaded(t), Next(t, t')\} \Rightarrow Loaded(t'),$$
$$\{Alive(t), Next(t, t')\} \Rightarrow Alive(t'),$$
$$\{Loaded(t), Fired(t), Next(t, t')\}|Alive(t) \Rightarrow \sim Alive(t')\}.$$

Then $(K, R) \vdash_{\mathbf{QD_e}} @\sim Alive(t_2)$.

We include one temporal persistence principle for each "stative" predicate. These have the general form

$$\{F(x_1, ..., x_k, t), Next(t, t')\} \Rightarrow F(x_1, ..., x_k, t').$$

Causal rules are change rules and normally compete with the persistence principle for some stative predicate. They say that if some event occurs at some moment, then some state holds at the next moment even if that state did not hold before. Causal rules have the general form

$$A|\sim\varphi \Rightarrow \varphi.$$

Including the complement of the consequent of the causal rule as an even-if condition makes the causal rule more specific than the persistence principle governing the stative predicate in the consequent. This solution to the Yale Shooting Problem is superior to the solution proposed in (Nute 1990) which did not use even-if conditions.

To our earlier list of examples, we will add one new, artificial example that illustrates a peculiarity of $\mathbf{QD_e}$.

EXAMPLE 12 *Let* $K = \{\varphi, \psi, \zeta, \eta\}$ *and let*

$$R = \{\varphi|\psi \Rightarrow \chi, \zeta|\eta \Rightarrow \sim\chi, \{\varphi, \psi\}| \Rightarrow \zeta, \{\zeta, \eta\}| \Rightarrow \varphi\}.$$

Then we get $(K, R) \vdash_{\mathbf{QD_e}} @\chi$ and $(K, R) \vdash_{\mathbf{QD_e}} @\sim\chi$. This is because the specificity condition built into our proof theory makes each of $\varphi|\psi \Rightarrow \chi$ and $\zeta|\eta \Rightarrow \sim\chi$ more specific than the other. Careful examination of the defeasible theory that gives rise to this "weak" contradiction shows that it is in fact intuitively contradictory. I propose that this is not a fault of the proof theory and that any defeasible theory that generates a non-asymetrical specificity relation on its own rule set is intuitively defective. Nevertheless, the present treatment of specificity for even-if rules is a first, rough approximation that may require refinement.

8. Decidability

Suppose t is a proof tree, the top node $\top$ in t is labeled $(K, R, \varphi^{+/-})$, and every node in t satisfies some condition in $\mathbf{QD_e}$.

We produce a new proof tree t_1 from t by the following procedure. We traverse t in breadth-first fashion. As we visit each node n, we find a condition $\mathbf{C}$ in $\mathbf{QD_e}$ which n satisfies and we prune all children of n not required by $\mathbf{C}$.

There is a limit on the number of children a node in t_1 can have. This limit is determined by the conditions in $\mathbf{QD_e}$ and by the defeasible theory that labels the top node $\top$ in t_1. By an inductive argument we see that if $\top$ is labeled (K, R, φ), then every node in t_1 is labeled by (K, R, ψ) for some wff ψ or by $(\Gamma, R*, \psi)$ for some wff ψ and some set of wffs Γ such that Γ is either an antecedent condition of a rule in R or the union of the antecedent and even-if conditions of a defeasible rule in R. Let i = the number of strict rules in R, j = the number of defeasible rules and defeaters in R, and $m = max\{k :$ $\Delta|E \Rightarrow \chi \in R$ *or* $\Delta \rightsquigarrow \chi \in R$, *and* k *is the number of members of* $\Delta(\cup E)\}$. Then a node that satisfies $\mathbf{D_e^+}$ can have no more than $1 + i + m(j + 2)$ children. Each of the other conditions in $\mathbf{QD_e}$ has its own limit. Let w be the maximum of these. Then each node in t_1 has at most w children.

Another inductive argument shows that every node in t_1 is labeled $(K, R, \psi^{+/-})$ where ψ is a wff formed from the predicates and constants occurring in (K, R) and φ, or by $(\Gamma, R*, \psi^{+/-})$ where ψ is as before and Γ is a finite set of literals formed through substituting constants in (K, R) and φ into the conditions for some rule in R. Then since K and R are finite, there are finitely many different labels that nodes in t_1 can have. Let d be the number of possible labels for nodes in t_1.

We prune t_1 again using the following procedure. We look to see if t_1 has a branch of length greater than d. If it does, then there must be some pair of nodes n and n' in this branch such that $n \neq n'$, n is beneath n', and n and n' have the same label. Find two nodes like this and replace the subtree with top node n' by the subtree with top node n. Clearly, all the nodes in the resulting tree still satisfy some condition in $\mathbf{QD_e}$. We apply the same procedure to the new tree and continue in this way until we have produced a tree which does not have any branch longer than d. Call this tree t_2.

We see immediately that t_2 can have at most

$$\sum_{i=0}^{d-1} w^i$$

nodes. Assuming that any condition we might impose on proof trees will limit the number of non-gratuitous children a node satisfying the condition might have, we get a general result.

THEOREM 3 (DECIDABILITY) *If Σ is a defeasible logic, then there is a function f_Σ which assigns to each defeasible theory (K, R) and wff φ an integer $f_\Sigma(K, R, \varphi)$ such that $(K, R) \vdash_\Sigma \varphi$ iff there is a proof tree with at most $f_\Sigma(K, R, \varphi)$ nodes whose top node is labeled (K, R, φ^+) and each of whose nodes satisfies some member of Σ, and $(K, R)\ {}_\Sigma\dashv \varphi$ iff there is a proof tree with at most $f_\Sigma(K, R, \varphi)$ nodes whose top node is labeled (K, R, φ^-) and each of whose nodes satisfies some member of Σ.*

From this theorem we get decidability since we can in principle construct all possible relevant proof trees for a given (K, R) and φ in finite time.

For any defeasible theory (K, R) and any wff φ, we can decide whether φ is either derivable or demonstrably not derivable from (K, R) in $\mathbf{QD_e}$. But this is not equivalent to saying that φ must be either derivable or demonstrably not derivable from (K, R) in $\mathbf{QD_e}$. For example, we have neither $(\emptyset, \{\varphi \Rightarrow \varphi\}) \vdash_{\mathbf{QD_e}} @\varphi$ nor $(\emptyset, \{\varphi \Rightarrow \varphi\})\ {}_{\mathbf{QD_e}}\dashv @\varphi$.

9. Cumulativity

Gabbay (1984) proposed a property for the consequence relation of a nonmonotonic system to replace the property of monotonicity. This property is usually referred to as *cumulativity* in the literature. A nonmonotonic system is cumulative if, whenever φ and ψ are both nonmonotonically derivable from a theory T, ψ is also nonmonotonically derivable from $T \cup \{\varphi\}$. One way to interpret cumulativity for defeasible logic is to expect $(K \cup \{\varphi\}, R) \vdash_{\mathbf{QD_e}} @\psi$ whenever $(K, R) \vdash_{\mathbf{QD_e}} @\varphi$ and $(K, R) \vdash_{\mathbf{QD_e}} @\psi$. $\mathbf{QD_e}$ is not cumulative

in this sense. Where (K, R) is defined as in Example 12, $(K, R) \vdash_{\mathbf{QD_e}} @\chi$ and $(K, R) \vdash_{\mathbf{QD_e}} @{\sim}\chi$, but $(K \cup \{\chi\}, R) \not\vdash_{\mathbf{QD_e}} @{\sim}\chi$. However, we do have the following suite of related results for $\mathbf{QD_e}$.

THEOREM 4 (MONOTONICITY) *For any sets K and K' of literals, any sets R and R' of rules, and any literal φ, if $(K, R) \vdash_{\mathbf{QD_e}} \varphi$, then $(K \cup K', R \cup R') \vdash_{\mathbf{QD_e}} \varphi$.*

THEOREM 5 (STRICT CUMULATIVITY) *For any defeasible theory (K, R), any literal φ, and any wff ψ, if $(K, R) \vdash_{\mathbf{QD_e}} \varphi$ and $(K, R) \vdash_{\mathbf{QD_e}} \psi$, then $(K \cup \{\varphi\}, R) \vdash_{\mathbf{QD_e}} \psi$.*

THEOREM 6 (DEFEASIBLE CUMULATIVITY) *For any defeasible theory (K, R), any literal φ, and any wff ψ, if $(K, R) \vdash_{\mathbf{QD_e}} @\varphi$ and $(K, R) \vdash_{\mathbf{QD_e}} \psi$, then $(K, R \cup \{\Rightarrow\varphi\}) \vdash_{\mathbf{QD_e}} \psi$.*

There is an important difference in $\mathbf{QD_e}$ between deriving a literal (strictly) and deriving a tentative conclusion, and there is an important difference between a fact and a presumption. These differences are reflected in our cumulativity results. These results are proved by a straightforward induction on the depth of a proof. Notice that since no presumptions occur in $R*$, $R* = (R \cup \{\Rightarrow\varphi\})*$. This means that when we add φ to the literals of a theory or add $\Rightarrow\varphi$ to the rules of a theory, the specificity relation does not change for the other rules in the theory. And of course a presumption will be less specific than any competing rule with a non-empty antecedent.[2]

10. Implications for implementation

Versions of defeasible logic that do not include the current treatment of even-if conditions have been implemented as extensions of the logic programming language Prolog (Nute and Lewis 1986). Each of these versions of d-Prolog (for defeasible Prolog) is a sound but not a complete theorem prover for the corresponding defeasible logic. These programs are incomplete because they do not use the fact that there is a maximal necessary number of nodes for proof trees to limit the exhaustive depth-first search for a proof of a query from a theory.

How can our decidability proof guide the construction of a sound and complete theorem prover for defeasible logic? First, notice that we do not need to be concerned about the pruning step in our argument that eliminates gratuitous nodes in a proof tree. Any reasonable theorem prover will

[2]Prompted by a discussion with Jürgen Dix at the conference, this section was added to the paper after the conference.

only generate nodes that are required by some condition in the logic. The important idea in the argument is that if a proof is possible, then there is a proof in which no branch has a length greater than some maximum that can be computed from the theory and the query. All that needs to be done, then, is to keep track of the current level as the theorem prover tries to build the proof tree in a recursive, depth-first fashion. The theorem prover must fail and backtrack whenever the maximum necessary depth is exceeded. In this way, the theorem prover will never engage in a fruitless attempt to construct an infinitely long branch because it has run into some circularity in the theory from which it is trying to construct the proof. This technique amounts to a cheap method for loop-checking.

This method has been implemented so far for strict logic, and efforts are currently underway to develop sound and complete theorem provers for a large family of defeasible logics similar to the one described here. Decision procedures for several versions of defeasible logic should be implemented by the time this paper appears. The number defined here as a limit to the necessary depth of a proof is both roughly calculated and large. A direction for further work is to find limits that will generally be smaller than the limit used here. Finding better limits will depend on analysis of the details of the defeasible theory upon which the computation will be based. Any limit will have to be recomputed whenever facts or rules are added to the theory. Another question is whether this analysis of the theory and the associated computation of the limit on proofs can be carried out in a modular fashion, i.e., in a way that does not require a complete reanalysis of the entire theory every time new items are added or deleted.

As a method for knowledge representation and automated reasoning involving incomplete or uncertain information, defeasible logics offer an alternative to numerical methods such as probabilistic, certainty factor, or fuzzy approaches. They also offer an alternative to other qualitative approaches such as circumscription or default logic. I expect that they will prove superior to these methods for some artificial intelligence applications and that they will complement these methods in other applications. The chief disadvantage of the defeasible system described here is its limited expressive power. Its chief advantages include its naturalness, its applicability to a wide range of problems of the kind that have been cited in the literature, and its decidability. The limit on expressive power can be relaxed by permitting functions, but the price is that the resulting system is no longer even semi-decidable.

References

GABBAY, D. 1985. *Theoretical foundations for non-monotonic reasoning in expert systems.* In K. Apt (ed.), Logics and Models of Concurrent Systems. Springer-Verlag, New York.

GEFFNER, H. 1989. *On the logic of defaults.* AAAI-89. Morgan Kaufmann, Los Altos, California.

GEFFNER, H. AND PEARL, J. 1990. *A framework for reasoning with defaults.* In Kyburg, H., Loui, R. and Carlson, G. (eds.), Knowledge Representation and Defeasible Reasoning. Studies in Cognitive Systems. Kluwer Academic Publishers, Boston.

HANKS, S. AND MCDERMOTT, D. 1987. *Nonmonotonic logic and temporal projection.* Artificial Intelligence 33:379-412.

ISRAEL, D. 1980. *What's wrong with non-monotonic logic?* AAAI-80. Morgan Kaufmann, Los Altos, California.

KONOLIGE, K. 1988. *On the relation between default theories and autoepistemic logic.* Artificial Intelligence 35:343-382.

LIFSCHITZ, V. 1985. *Computing circumscription.* IJCAI-85. Morgan Kaufmann, Los Altos, California.

LIFSCHITZ, V. 1986. *Pointwise circumscription: preliminary report.* AAAI-86. Morgan Kaufmann, Los Altos, California.

LOUI, R. 1987. *Response to Hanks and McDermott: temporal evolution of beliefs and beliefs about temporal evolution.* Cognitive Science 11:303-317.

MCCARTHY, J. 1980. *Circumscription - a form of non-monotonic reasoning.* Artificial Intelligence 13:27-39.

MCCARTHY, J. 1986. *Applications of circumscription to formalizing common sense knowledge.* Artificial Intelligence 28:89-116.

MCDERMOTT, D. AND DOYLE, J. 1980. *Non-monotonic logic I.* Artificial Intelligence 13:41-72.

MOORE, R. 1984. *Possible-worlds semantics for autoepistemic logic.* Proceedings of the 1984 Non-monotonic Reasoning Workshop. AAAI, Menlo Park, California.

MOORE, R. 1985. *Semantical considerations on non-monotonic logic.* Artificial Intelligence 25:75-94.

NUTE, D. 1988. *Defeasible reasoning and decision support systems.* Decision Support Systems 4:97-110.

NUTE, D. 1990. *Defeasible logic and the frame problem.* In H. Kyburg, R Loui, and G. Carlson (eds.), Knowledge Representation and Defeasible Reasoning. Studies in Cognitive Systems. Kluwer Academic Publishers, Boston.

NUTE, D. 1991. *Basic defeasible logic.* In L. Fariñas-del-Cerro and M. Pentonnen (eds.), Intentional Logics for Logic Programming, Oxford University, in press.

NUTE, D. 1992. *Inference, rules, and instrumentalism.* International Journal of Expert Systems Theory and Applications, in press.

NUTE, D. AND LEWIS, M. 1986. *A user's manual for d-PROLOG.* ACMC Research Report 01-0017. The University of Georgia, Athens, Georgia.

POLLOCK, J. 1987. *Defeasible reasoning.* Cognitive Science 11:481-518.

REITER, R. 1980. *A logic for default reasoning.* Artificial Intelligence 13:81-132.

Logic, Methodology and Philosophy of Science IX
D. Prawitz, B. Skyrms and D. Westerståhl (Editors)

NON-CLASSICAL LOGIC AND ONTOLOGICAL NON-COMMITMENT, AVOIDING ABSTRACT OBJECTS THROUGH MODAL OPERATORS

JOHN P. BURGESS

Department of Philosophy
Princeton University, Princeton, NJ 08544-1006 USA
jburgess@pucc.princeton.edu

1. Introduction

Mathematical objects and facts are supposedly not *concrete* but rather *abstract*, immutable and impassive, unchanging with times and independent of contingencies. Correspondingly, mathematical language makes no use of *temporal* and *modal* distinctions. Consequently, mathematical logic includes among its operators none expressing such distinctions.

To apply mathematical logic, with just its classical operators of negation, conjunction, disjunction, and existential and universal quantification ($\sim$, $\wedge$, $\vee$, $\forall$, $\exists$), to non-mathematical language, one must resort to *regimentation*. One must imitate the approach to time and motion and to contingency and chance taken in mathematical physics and mathematical statistics, where change with times and dependence on contingencies are represented by timeless and non-contingent relations to certain special *index* objects, "times" (or "instants" or "stages") and "contingencies" or ("cases" or "worlds").

For example, something like:

(1) A man was drowned and then hanged.

is regimented as something like:

(2) there [exists] a t and there [exists] a t' such that t [is] an index and t' [is] an index and t [is] earlier than t' and t' [is] earlier than the present and there [exists] an x such that x is a man and x [is drowned] at t and x [is hanged] at t'

and then as something like:

(3) $\exists t \exists t' (t < t' \wedge t' < a \wedge \exists x (Fx \wedge Gxt \wedge Hxt'))$

Here the brackets indicate that all verbs are to be understood as tenseless.

Analogously, something like:

(4) A hospital might have been built.

is regimented as something like:

(5) There {exists} a u and there {exists} an x such that u {is} an index and x {is} a hospital and x {is built} in u

and then as something like:

(6) $\exists u \exists x (Fx \wedge Gux)$

Here the braces indicate that all verbs are to be understood as moodless.

In natural language—English will always be taken as the example, though most of what is said should apply to other Germanic languages, and much of it to other Indo-European languages—temporal and modal distinctions are sometimes expressed through such phrases as "at that time" or "in that contingency", but usually they are expressed *noncommittally*, without overtly quantifying over such index objects as "times" or "contingencies". Such distinctions are often expressed through the verbal inflections or auxiliaries of past, present, and future *tense*, and indicative, subjunctive, and conditional *mood*. Or they may be expressed through temporal or modal adverbs or conjunctions.

An alternative to regimentation is provided by *autonomous* logics of tense and mood. Their development was partly motivated by *nominalistic* concerns, a desire to avoid commitment to such presumably abstract objects as indices, and partly motivated by *linguistic* concerns, a desire to treat modal distinctions in a formal language in a way less divergent from the way they are treated in natural language.

Thus temporal logic, developed as a logic of tense, has connectives $\mathcal{P}$ and $\mathcal{F}$ for past and future, "has (sometime) been" and "will (sometime) be", and dual connectives $\mathcal{H}$ and $\mathcal{G}$ for "has always been" and "will always be". Also modal logic, though originally developed as a *metalogical* logic, a logic of the (epistemic) notion of consistency and the dual notion of validity,

has been subsequently interpreted as a mood logic, its operator $\Diamond$ and the dual operator $\Box$ being understood as expressing instead the (metaphysical) notion of possibility in the sense of "might have been" and the dual notion of necessity in the sense of "couldn't have failed to be". Both temporal and modal logics have been intensively cultivated, along with a variety of related *intensional* logics.

In tense logic, (1) is turned into something like:

(7) past tense (there exists a man such that (past tense (he is drowned)) and he is hanged)

and then into something like:

(8) $\mathcal{P}(\exists x(Fx \wedge \mathcal{P}(Gx) \wedge Hx))$

In modal logic, (4) is turned into something like:

(9) possibly (there exists an x such that x is a hospital and x is built)

and then into something like:

(10) $\Diamond \exists x(Fx \wedge Gx)$

A series of completeness theorems for temporal and modal logic establish that the same arguments are validated by the regimented and the autonomous approaches whenever both are applicable. The regimented approach is, however, applicable to more arguments than the autonomous approach: Far more can be formally symbolized using overt quantifications over indices, than can be using only the connectives mentioned above, or even various supplementary temporal and modal operators that have been introduced by their successors. (Equivalence in expressive power with the regimented approach can be achieved for the autonomous approach only by using certain operators that lack obvious non-committal natural language counterparts.)

Nominalism has been connected with intensional logic in not one but two ways. For not only has a desire to avoid commitment to abstract index objects motivated the introduction of such logics, but also the desire to avoid commitment to abstract mathematical objects has motivated ambitious applications of modality, straining the limits of what can be easily expressed

in the formal languages of mainstream modal logics, thus motivating the introduction of novel modal logics.

For there is a substantial literature devoted to ambitious attempts to establish the *dispensability* of mathematical objects for scientific theorizing using modal logic: What is attempted in these applications is to establish a strategy for converting a standard scientific theory T involving mathematical objects but not modal operators into an empirically equivalent alternative theory T^* involving modal operators but not mathematical objects. Sometimes other non-classical logics are brought in instead of or in addition to modal logic.

The issues that arise when such ambitious applications of non-classical logic are examined from the standpoint of philosophy of mathematics are numerous and diverse. They include, along with many other and quite different ones, issues about overt versus covert, or *surface* versus *deep* commitments; and these are of interest not only for philosophy of mathematics, but also for linguistics. But aside from all such issues, modal nominalist strategies are of interest from the standpoint of philosophical logic, since when uses of non-classical, specifically of modal, logics in order to avoid ontological commitments, especially to abstract objects, are examined, divergences between the treatment of modal distinctions in formal and in natural languages are encountered.

In the present note, two modal nominalist strategies, inspired in broad outline by strategies in the literature, though not faithful in fine detail to them, will be sketched and then examined from the standpoint of philosophical logic, ignoring issues of philosophy of mathematics. In connection with each, divergences between natural language and the formal languages of mainstream systems of modal logic will be encountered. Most of these have analogues also in temporal logic, but these temporal analogues will not be considered explicitly.

2. First strategy

The two strategies to be sketched share several features. First, both start with a theory T, the formalization of some standard scientific theory involving mathematical objects; and both suppose this input theory may be assumed to take the following form: T will be a theory based on classical logic, say in a version with existential quantification a primitive logical operator and universal quantification as a defined logical operator; and T will be a two-sorted theory, with two styles of variables. One style, $x, y, z, \ldots$, will be reserved for concrete, physical objects of some sort(s) or other(s), which for definiteness will be called *ponderables*. The other style, $\xi, \upsilon, \zeta, \ldots$,

will be reserved for abstract, mathematical objects.

The language of T will have finitely many (non-logical) primitives, all predicates (and none constants). Some of these will be *of the first kind*, having places only for variables of the first sort. Some will be *of mixed kind*, having places for variables of both sorts. Some will be *of the second kind*, having places only for variables of the second sort, expressing abstract mathematical relations. Formulas of the language of T will be said to be *of the first kind* if they involve only predicates of the first kind; and the empirical consequences of T will be expressible by formulas of this kind. Formulas will be said to be *of mixed kind* if they involve predicates of mixed kind or if they involve predicates of more than one kind. Formulas will be said to be *of the second kind* if they involve only predicates of the second kind, expressing abstract mathematical facts. T will have only finitely many (non-logical) postulates (or schemes).

Since it is widely accepted that enough pure mathematics for scientific applications is provided by *analysis*, the orthodox theory of the real numbers (with the natural numbers as a distinguished subsystem), it will be assumed that the only primitives and postulates of the second kind are those of analysis. (Whether this should be so widely accepted is one of those issues pertaining more to philosophy of mathematics than to philosophical logic that are being ignored here.)

Thus typical examples of primitives of the three kinds might be:

(1) x weighs less than y does

(2) ξ is how much x weighs [in arbitrary but fixed, though here unspecified, units]

(3) ξ is less than υ is

The leading idea of both strategies will be to end with a theory T^*, the formalization of an alternative nominalist theory involving modal operators. The output theory T^* is to have the same primitives of the first kind as T, and is to have same formulas of the first kind as consequences as does T, so that T and T^* will be empirically equivalent.

The leading idea is to replace those postulates in the original theory that assert the actual existence of mathematical objects by some postulates in the final theory that will only assert the possible existence of some sort of objects. Ontological commitments to mathematical objects, commitments as to the actuality of their existence, commitments to the effect that they do exist, are to be avoided in favor of what may be called *dynatological* commitments, commitments as to the possibility of the existence of objects of some

sort, commitments to the effect that various objects of some sort might each have existed (and perhaps also by what may be called *syndynatological* commitments, commitments as to the compossibility of existence of objects of some sort, commitments to the effect that various objects of some sort might all have co-existed).

But objects of what sort? On the one hand, it seems part of the very concept of the abstractness of mathematical objects that for them no distinction between actual and possible existence can be made. Hence, the objects whose possible existence will be asserted in the replacement theory cannot just be the mathematical objects whose existence was asserted in the original theory, and arguably cannot be any sort of abstract objects at all.

On the other hand, interaction within a universal causal system seems part of the very concept of concreteness. Hence, if the objects whose possible existence will be asserted in the replacement theory are to be concrete objects, not numbers themselves but *surrogates* for them, it must be conceded that if those surrogate objects had existed, then the ponderable objects that actually exist might not have been just as they actually are: The old objects presumably would have been different owing to causal interactions with the new objects.

Hence almost inevitably any modal nominalist strategy will have to, and the two strategies to be sketched will, allow for *cross-comparisons* between surrogate objects ξ that there might have been (as they would have been) and ponderable objects x that there are (as they are). Such *hypothetical-actual* cross-comparisons are easily expressible and quite common in natural language, as when one compares the prison that was built on some site with the hospital that might have been built there, had America been more like Scandinavia. (Note that this last counterfactual clause itself involves a kind of cross-comparison, between how Sweden and its neighbors are and how the United States might have been.) So also are *hypothetical-hypothetical* cross-comparisons, as when one compares the hospital that might have been built there with the school that might have been built there: These will be needed in a strategy that seeks to avoid strong syndynatological commitments, that seeks to avoid the assumption that the various surrogates that each severally might have existed all jointly might have co-existed. They are not, however, easily expressible in the formal languages of mainstream modal systems, and will provide a major example of the divergence between common natural languages and mainstream formal languages.

There are several variants of the usual formulation of analysis. The first strategy begins by replacing analysis by one of these: The variant in question takes as its objects not real numbers, but rather infinite sequences of binary digits, which may be construed as *numerals* for numbers. Replacing the

usual formulation of analysis by this variant in T, one obtains a variant T' with primitives like:

(1) x weighs less than y does

(2) ξ marks how much x weighs

(3) ξ marks less than υ does

The first modal nominalist strategy to be outlined here assumes that there might have existed physically constituted objects construable as functioning linguistically as numerals in the sense of concrete *tokens* rather than of abstract *types*.

The first modal nominalist strategy produces an alternative theory T^* by transforming the theory T' as follows:

(4) each existential quantification of the first kind $\exists x$ is replaced by an *actuality* quantification $\exists' x$ read "there actually does exist an x ..."

(5) each existential quantification of the second kind $\exists \xi$ is replaced by a *possibility* quantification $\exists^\circ \xi$ read "there possibly might have existed a ξ ..."

and primitives F, G, H like (1), (2), (3) above are replaced by actual-actual, hypothetical-actual, and hypothetical-hypothetical primitives F'', $G^{\circ\prime}$, $H^{\circ\circ}$ like:

(1*) x (actually) weighs less than y (actually) weighs

(2*) ξ (necessarily) would have marked (if it had existed) how much x (actually) weighs

(3*) ξ (necessarily) would have marked (if it had existed) less than υ (necessarily) would have marked (if it had existed)

To establish the claim that this ultimate T^* is empirically equivalent to the intermediate T' and hence to the original T, it would suffice to establish a *metatheorem* to the effect that the above replacements transform arguments valid classical-logic arguments into valid modal-logic arguments. The logical issues involved in this claim will be examined first from an intuitive, natural language standpoint, and then from a technical, formal language standpoint.

As regards (4), and (1*) and (2*), note that what is at issue is solely what ponderables exist and what they are like: What ponderables would or might have existed if things had been other than as they are, or what ponderables that do exist would or might have been like if things had been other than as they are, is not at issue. (2*) is an instance of hypothetical-actual cross-comparison similar to the hospital-prison example above. The qualifying adverb "actually" is optional, actuality being already sufficiently indicated by use of the indicative mood in "exists" and "weighs", but it has been inserted for emphasis.

As regards (5), note that what is at issue is solely what tokens might have existed, if things had been other than as they are: What tokens do exist is not at issue. An assertion of the form, "there might have been something such that ...it ..." or " $\exists^{\circ}\xi \ldots \xi \ldots$ " is an assertion about the possible-if-not-actual existence of ordinary sorts of objects, not an assertion about the existence of extraordinary sorts of possible-if-not-actual objects: Such assertions do not assert the actual existence of anything, any more than the assertion that there existed, centuries ago, a man who was drowned and then hanged implies that he still exists (presumably as some sort of spook haunting the burial mounds of the old city). As regards (2*) and (3*), their lack of *existential import* is emphasized by the insertion of the optional qualifying clause "if it had existed". Moreover, while T implies the actual co-existence of many numbers, T^* implies nothing about the compossible co-existence of many tokens. As regards (3*), its lack of *co-existential import* is emphasized by the insertion of optional qualifying clauses "if it had existed" separately for each of ξ, υ rather than jointly for both. (3*) is an instance of hypothetical-hypothetical cross-comparison, as in the example above of the hospital and the school.

As regards (2*) and (3*) note also that it is assumed that how a token functions linguistically is *essential*, not *accidental*, to its identity: If ξ would have functioned linguistically differently from how υ would have done, then ξ would have been distinct from υ, even if ξ would have been constituted physically very similarly to how υ would have been. Moreover, it is only the linguistic function, not the physical constitution, of tokens that is at issue. Hence it is only essential, and not accidental, features of a token that are at issue: It is always a question of what a token (necessarily) would have been like, not of what it (possibly) might have been like. The contrasting qualifying adverbs "necessarily" and "possibly" are optional, the contrast being already sufficiently indicated by the contrast between the auxiliary verbs "would" and "might" with "marked", but it has been inserted for emphasis.

These glosses given, one might proceed to attempt the technical devel-

opment and intuitive justification of a symbolic logic of the actuality and possibility quantifiers $\exists'$ and $\exists^{\circ}$, and of actual-actual, hypothetical-actual, and hypothetical-hypothetical predicates like F'', $G^{\circ\prime}$, $H^{\circ\circ}$ above, and then establish the metatheorem enunciated above. But such a project, though feasible, will not be attempted here. Rather, what will be considered here will be whether, and if so how, one might attempt symbolize the natural language modal notions glossed above in one of the conventional formal languages of mainstream modal logics. The choice of an appropriate modal predicate logic involves several subsidiary choices.

First, if one is to obtain a modal predicate logic by combining a modal sentential logic with a suitable predicate logic, one must choose an appropriate predicate logic. Classical predicate logic seems not the appropriate choice, for there is a divergence between classical predicate logic and natural language, in that the former has built-in assumptions of existential import absent from the latter in modal contexts. Hence one must choose an alternative predicate logic without built-in assumptions of existential import, a *free* predicate logic (perhaps in a version or variant without built-in assumptions of actual existence but with built-in assumptions of possible existence). Indeed, one must choose among several versions of free predicate logic. Fortunately, most versions give the same results when applied in the restricted situation under consideration in the first strategy, where there are no constants, and where only with the actual properties of actual objects $x, y, z, \ldots$ of one sort and the essential properties of hypothetical objects $\xi, \upsilon, \zeta, \ldots$ of another sort are under consideration. Hence the choice of version need not be indicated in detail in the present sketch.

Second, one must choose an appropriate modal sentential logic to be combined with free predicate logic. For there are many systems, though the important ones differ only in their treatment of iterated modalities. Since verbs in natural language can be inflected for mood just once, allowing iteration of modal operators is itself another divergence from natural language (more appropriate when modal logic is interpreted as a metalogical logic than when it is interpreted as a mood logic). Fortunately this divergence can be partly remedied by adopting the strongest mainstream system, *S5*. For though *S5* still allows the introduction of iterated modalities, also allows for their elimination, by including *rigidity* axioms to the effect that a formula already modalized is unaffected by further modalization:

(6) $\Box\Diamond p \leftrightarrow \Diamond p$

Third, one must choose appropriate enrichments of the $\Box$, $\Diamond$–modal sentential logic chosen, for the following reason: In natural language, if the verb

in a main clause is inflected for some non-indicative mood, one often remains free to inflect or not the verbs in subordinate clauses, and the present or absence of inflection is usable to express distinctions. To illustrate, contrast:

(7a) If the estate had not been entailed, then the other children would have been as well off as the eldest son would (then) have been.

with:

(7b) If the estate had not been entailed, then the other children would have been as well off as the eldest son (now) is.

Since breaking the entail would have equalized the distribution, but not enlarged the size, of the inheritance, presumable (7a) is true, and (7b) is false. Such distinctions are important in the strategy under consideration. However, mainstream formal languages diverge from natural language in that an operator applied as the main connective to a sentence automatically governs all its subsentences. This divergence can be partly remedied by enriching the $\Box$, $\Diamond$–language with appropriately chosen further operators, notably, an operator for restoring the indicative mood in subordinate clauses or subsentences, the operator @ for "actually" or "*now*" in the non-temporal, modal sense illustrated in (7b). Appropriate axioms, including rigidity axioms like:

(6′) $\Diamond @p \leftrightarrow @p$

must also be chosen, though these will not be indicated in more detail in the present sketch.

Fourth, with the above choices appropriately made, only a start towards formalizing the theory T^* can now be made: The actuality and possibility quantifiers $\exists'$ and $\exists^\circ$ can be symbolized as $@\exists$ and $\Diamond\exists$. But then it remains to choose an appropriate formalization of actual-actual, hypothetical-actual, and hypothetical-hypothetical predicates F'',$G^{\circ\prime}$, $H^{\circ\circ}$ like (1*), (2*), (3*) above. Mainstream predicate logic treats every two-place predicate as if it involved a single transitive verb with its subject and object nouns. This is a divergence from natural language where, especially in the case of comparative predicates, there are often two verbs, to which modal inflections can be separately applied. An actual-actual comparative predicate F'' like (1*) above can perhaps be symbolized as $@F$, and a hypothetical-hypothetical predicate $H^{\circ\circ}$ like (3*) above can perhaps be symbolized as $\Box H$, but what is

to be done with a hypothetical-actual predicate $G^{o\prime}$ like (2*) above? Conventional systems in the mainstream literature do not allow attaching operators $\Box$ and @ separately to the two places of a predicate!

Since what are at issue are only predicates each place of which is modalized either by "necessarily would" or by "actually does", and since in natural language, once one such modalization has been added, no further ones may be, one could perhaps simply symbolize the predicates as F, G, H,... and add rigidity axioms akin to (6) above:

$$(6'') \quad \forall x \forall y(\Diamond Pxy \leftrightarrow Pxy)$$

(and similarly for ξ,υ-variables).

When *S5* with the @ operator is combined with free logic, and rigidity axioms are assumed, the required metatheorem can indeed then be proved, though details will be omitted in the present sketch. But this is a devious formalization of what arguably ought to be straightforwardly formalizable, its deviousness being a measure of the divergence of mainstream formalism from natural language.

3. Second strategy

There are many further variants of the usual formulation of analysis. The second strategy begins by considering a series of such variants. Variant (a) has two styles of variables, one X, Y, Z,... for natural numbers, and another ξ, υ, ζ,... for sets thereof. Its primitives are the usual *order* and *sum* and *product* primitives of *arithmetic*, and the *membership* primitive $\in$ of set theory. Its postulates are the usual postulates of arithmetic, and the usual postulates of what may be called *elemental* set theory: These are the *extensionality* postulate (sets having exactly the same elements are identical), the *comprehension* scheme (for each formula P, the axiom that there exists a set having as elements exactly those elements for which P holds), and some appropriate *choice* axiom or scheme. The key to proving that this new theory is (at least) as strong as the old is this: First, natural numbers can represent decimal fractions, with $2^u \cdot 3^v \cdot 5^w$ representing $(-1)^u \cdot v \cdot 10^{-w}$; and second, sets of decimal fractions can represent real numbers.

Variant (b) takes the variables ξ,υ,ζ,... to range over dyadic *relations* rather than monadic sets, appropriately changing the extensionality, comprehension, and choice postulates. As to the primitives and postulates of arithmetic, only those pertaining to order need be assumed, since those pertaining to sum and product become definable and derivable.

Variant (c) drops the assumption that the variables X, Y, Z, ... range

specifically over natural numbers. The unspecified objects over which X, Y, Z, ... range may be called *individuals*. Also the order primitive and the order postulates for that primitive are dropped in favor of a postulate to the effect that there exists a *progression*, a dyadic relation on the individuals for which the order postulates hold. This postulate is just one of many, all equivalent given an appropriate choice postulate, asserting the *infinity* of the set of individuals.

Variant (d) takes the variables ξ, υ, ... to range over sets again (but now sets of individuals), and adds variables Ξ, Υ, ... for *classes* in the sense of sets of sets of individuals, with a primitive for membership of a set in a class in addition to the primitive for membership of an individual in a set, and with the postulates of elemental set theory replaced by the exactly analogous postulates of elemental class theory. The key to proving that this new theory is (at least) as strong as the old is this: A symmetric dyadic relation on individuals can be represented by the class of those doubleton sets of individuals $\{X, Y\}$ whose two elements are related. A total order relation on individuals can be represented by the class of those sets of individuals that are initial segments of the total order. An arbitrary dyadic relation on individuals can be represented by a quadruple consisting of a total order relation on individuals; a set of individuals, namely, the set of individuals that are self-related; and two symmetric relations on individuals, namely, the relation X bears to Y if and only if the lower of the two in the total order in question bears the original relation to the higher, and the analogous relation with low and high reversed. Hence quantification over arbitrary dyadic relations can be replaced by fourfold quantification over sets and classes.

Variant (e) modifies the comprehension postulates of elemental set (and class) theories so as to disallow the empty set (and class), producing what may be called *positive* elemental set (or class) theory. The key to proving that this new theory is (at least) as strong as the old is this: Taking a pair of non-void sets ξ, υ, to represent the empty set $\{\}$ if ξ is a singleton set and to represent the set υ otherwise, arbitrary sets can be represented by pairs of non-void sets; moreover, this idea can be extended to classes.

Variant (f) drops the variables X, Y, Z, ... ranging over individuals: An individual X can be represented by its singleton set $\{X\}$. The primitive $\in$ of membership between individuals and sets is replaced by the primitive $\subseteq$ of inclusion between sets. The postulates of (positive) elemental set theory are replaced by the postulates of what may be called (positive) *inclusive* set theory (more usually called the theory of atomic Boolean algebras with a completeness scheme).

Any one of these variants can replace analysis as originally formulated in

the original in T, making corresponding changes in the mixed primitives. Thus a mixed primitive like:

(2) ξ is how much x weighs

would at stage (a) become:

(2a) X represents about how much x weighs [in the sense that X is of form $2^u \cdot 3^v \cdot 5^w$ and $(-1)^u \cdot v \cdot 10^{-w}$ is how much x weights to the nearest 10^{-w} unit]

At stage (b) this would remain the same, but at stage (c) it would become:

(2c) [ξ is a progression and] X represents, with respect to ξ, about how much x weighs

which introduces a new place for the relation parameter ξ. Accordingly, each postulate P of the original T will by stage (c) be replaced by a postulate of the form:

$$\forall\xi(\xi \text{ is a progression} \rightarrow P^c(\xi))$$

At stage (d) there would be further change:

(2d) [Ξ represents a progression and] X represents, with respect to Ξ, about how much x weighs

and accordingly each postulate P of the original T will by stage (d) be replaced by a postulate of the form:

$$\forall\Xi(\Xi \text{ represents a progression} \rightarrow P^d(\Xi))$$

At stage (e) there would be no change, but at stage (f) there would be a final change:

(2f) [Ξ represents a progression in which ξ represents an initial segment and] ξ represents, with respect to Ξ, about how much x weighs

Thus T may be replaced by a variant t' with primitives like:

(1′) x weighs less than y does

(2′) ξ represents, with respect to Υ, about how much x weighs

(3′) ξ is included in υ

(4′) ξ is an element of Υ

with the postulates pertaining to abstract objects being those of positive inclusive set theory and of positive elemental class theory, with a postulate of infinity.

Now two non-classical, but also non-modal, logics may usefully be introduced, *mereology* and *plural* logic.

Mereology treats the primitive:

(3″) ξ is part of υ

as a logical primitive, much as classical logic treats identity as a logical primitive. Accordingly, it treats certain postulates pertaining to this primitive as logical postulates, not counted among the (non-logical) postulates of specific theories formulated within the general framework of this logic. No list of logical postulates claims or can claim to be complete, but the most usual list consists precisely of the exact analogues of the postulates of positive inclusive set theory.

Adopting mereology, then, T' can be replaced by a variant T'' in which variables ξ, υ, ... range over concrete objects, for definiteness to be called *bodies*, (3″) replaces (3′), variables Ξ, Υ, ... range over sets of bodies, and the only postulates pertaining to abstract objects are those of positive elemental set theory.

Plural logic allows, in addition to singular quantifiers $\exists\xi$ or "there exists something ξ", also plural quantifiers $\exists\exists\xi\xi$ or "there exist some things the ξ's". Predicates may have plural as well as singular places; in particular, in addition to the logical primitive $\xi = \upsilon$ or "ξ is the same as υ " there is a logical primitive $\xi == \upsilon\upsilon$ or:

(4‴) ξ is one of the υ's

Plural logic treats certain postulates pertaining to this primitive as logical postulates, not counted among the (non-logical) postulates of specific theories formulated within the general framework of this logic. No list of logical

postulates claims or can claim to be complete, but the most usual list consists precisely of the exact analogues of the postulates of positive elemental set theory.

Adopting plural logic, T'' can be replaced by a variant T''' in which the variables Ξ, Υ, ... are dropped, quantification over sets of bodies being replaced by plural quantification over bodies, and $(4')$ replaced by $(4''')$, and there are no postulates pertaining to abstract objects, though there is still a postulate of infinity, to the effect that there exist infinitely many bodies.

The (non-logical) primitives of T''' then are just ones like $(1')$ above and ones like:

$(2''')$ υ represents, with respect to the ξ's, about how much x weighs

Accordingly, each postulate P of the original T will be replaced in T''' by a postulate of the form:

$$\forall\forall\xi\xi(N(\xi\xi) \rightarrow P'''(\xi\xi))$$

The second modal nominalist strategy proposes an alternative theory T^* by transforming the theory T''' as follows: The postulate that there do exist infinitely many bodies or equivalently, some bodies the ξ's are *nested* in the sense of forming a progression under the part relation, is replaced by the postulate that there might have (co-)existed some bodies the ξ's that were nested, $\Diamond\exists\exists\xi\xi N(\xi\xi)$, while each postulate P of the original of the original T will in T^* be replaced by a postulate of the form:

$$\Box\forall\forall\xi\xi(N(\xi\xi) \rightarrow P^*(\xi\xi))$$

where P^* is obtained from P by transforming it as follows:

(5) each existential quantification of the first kind $\exists x$ is replaced by an *actuality* quantification $\exists' x$ read "there now, actually does exist an x ..."

(6) each existential quantification of the second kind $\exists\xi$ is replaced by a *consequentiality* quantification $\exists^+\xi$ read "there then, consequently would have existed a ξ ..." (and similarly for plural quantifications $\exists\exists\xi\xi$)

and primitives F and G like $(1')$ and $(2''')$, above are replaced by actual-actual and hypothetical-actual primitives F'' and $G^{+\circ\prime}$:

(1*) x (now, actually) weighs less than y (now, actually) weighs

(2*) υ (then, consequently) would have marked with respect to the ξ's (if they had existed) how much x (now, actually) weighs

In attempting to formalize this second strategy, one issue that arose in the attempting to formalize the first strategy no longer arises: One is now assuming the compossible co-existence of nested bodies, and the use of hypothetical-hypothetical cross-comparisons to avoid syndynatological commitments is not required. Moreover, one is not assuming the possible existence of tokens specifically, but rather only the compossible co-existence of bodies generally.

Inversely, one new issue arises that did not arise in attempting to formalize the first strategy. This issue may be illustrated by an example related to the inheritance example above. In its temporal version it runs:

(7) When he was in power, those who (now) criticize him (then) praised him.

In its modal version it runs:

(8) If he had been in power, those who (now) criticize him (then) would have praised him.

The "now" and "then" in (7) are used in their temporal senses of "currently" and "contemporaneously", but are optional, since what they express is already sufficiently indicated by the sequence of tenses (past-present-past). The "now" and "then" in (8) are used in non-temporal, modal senses of "actually" and "consequently", but are optional, since what they express is already sufficiently indicated by the sequence of moods (subjunctive-indicative-conditional). Just as "then" in the temporal sense marks the event of the second clause of the consequent as contemporary with the event of the antecedent, so does "then" in the modal sense mark the hypothesis of the second clause as consequent upon the hypothesis of the antecedent.

What is new in the second strategy is that, in addition to the operator for restoring the indicative mood in subordinate clauses or subsentences, the operator @ for "actually" or "now" in the non-temporal, modal sense illustrated in (8), one needs also an operator for imposing the conditional mood in subordinate clauses or subsentences, an operator $\not\in$ for "consequently"

or "*then*" in the non-temporal, modal sense illustrated in (8). With such operators, (8) could be formalized as something like:

$$(9) \quad \Box(Px \rightarrow @\forall y(Qxy \rightarrow \not\in Rxy))$$

Here @ in effect cancels the modality $\Box$, and $\not\in$ restores it. Similarly, in formalizing T^*, actuality and consequentiality quantifiers can be symbolized as $@\exists$ and $\not\in\exists$. (In the particular examples (7) and (8), it may be possible to avoid the use of "then" operators by rearranging the clauses of the sentence, but in general, in other, more complex examples, it is not.) Further details will be omitted from the present sketch.

4. Acknowledgments and conclusion

In its logical, though not its other, features, the strategy of §2 above is directly inspired by the strategy expounded in the first half of CHIHARA [1990] (who in turn traces the logical features of his work back to work of Ernest Adams). Charles Chihara, however, though he regards "worlds" as fictitious, in the work cited discusses the logical features of his strategy in terms of "worlds", and does not indicate at any length how the discussion might or should be reworded in noncommittal natural language. The second half of the same work is a survey of the literature of nominalism of the past couple of decades, and to it the reader is directed for background on this topic in philosophy of mathematics, and references to original sources.

The strategy of §3 above is indirectly inspired by the "modal structuralism" of HELLMAN [1989] (a work whose modal aspects are traceable back to work of Hilary Putnam, and whose structuralist aspects are traceable back to work of Paul Benacerraf). (The structuralism comes in in the above sketch at the transition form a theory about natural numbers to a theory about objects of some indeterminate sort that form a progression.) Geoffrey Hellman, however, seems to assume, contrary to what has been suggested above, that many more concrete objects than do exist might have existed without the concrete objects that do exist having been affected by them or having been other than as they are. As to the non-classical but non-modal logics involved, mereology was introduced long ago by Stanisław Leśniewski, and has long played a role in work on nominalism. Plural logic was introduced much more recently by George Boolos. LEWIS [1991] is a good source for each. Moreover this work of David Lewis is the only source for how the two can usefully be combined. The appendix to that work contains a trick of Allen Hazen (a reduction of "polyadic second-order logic" to "monadic third-order logic") that was used above in the preliminary transformations

of analysis from one version to another (the other transformations that were used are folklore among specialists). The standard reference work on intensional logic is GABBAY & GUENTHNER [1984], consisting of survey articles on its various branches by several workers active in the field, with references to the original literature (including that of temporal logic since its origins in the work of Arthur Prior, and that of modal logic since its origins in the work of C. I. Lewis). The first-named editor, Dov Gabbay, has been an enthusiastic proponent of operators beyond the original $\mathcal{P}$, $\mathcal{F}$, $\mathcal{H}$, $\mathcal{G}$, and $\Box$, $\Diamond$. The subsequent work most important in the present context has been that of Harold Hodes, especially HODES [1984], a good source for information on operators like @, and the only source for operators like $\not\in$. (This work also contains a notable simile likening the relation of these two kinds of operators to the relation between the "return" and the "backspace" on a keyboard). These operators were not even mentioned in the chapter on basic modal logic in the standard reference work cited, not yet at the time it was written having been much discussed in the literature. The analogous temporal operators were barely mentioned in passing, "now" having been introduced earlier by Hans Kamp, and "then" by Frank Vlach, but they were not (*mea culpa*) given the attention they deserved in the chapter on basic tense logic. From these sources the state of mainstream literature can be judged.

If the present author had to judge the state of the mainstream literature—had to list the main observations of the present note in order from the currently most to the currently least widely acknowledged—the list would go something like this:

Modal predicate logic must...

... be based on free, not classical, predicate logic.

... be based on a system of modal sentential logic, permitting the elimination of iterated modalities.

... include an indicative-mood, actuality, modal "now", or "return" operator.

... include a conditional-mood, consequentiality, modal "then", or "backspace" operator.

... allow cross-comparative predicates, whose places are separately qualified by different modalities.

References

CHIHARA, CHARLES S., *Constructibility and Mathematical Existence*, Clarendon Press, Oxford, 1990.

GABBAY, D. and GUENTHNER, F., editors, *Handbook of Philosophical Logic*, volume II: Extensions of Classical Logic, D. Reidel Publishing Company, Dordrecht, 1984.

HELLMAN, GEOFFREY, *Mathematics Without Numbers: Towards a Modal-Structural Interpretation*, Clarendon Press, Oxford, 1989.

HODES, HAROLD, *On Modal Logics Which Enrich First Order S5*, Journal of Philosophical Logic, volume 13 (1984), pages 123-149.

LEWIS, DAVID, *Parts of Classes*, Basil Blackwell, Oxford, 1991.

Logic, Methodology and Philosophy of Science IX
D. Prawitz, B. Skyrms and D. Westerståhl (Editors)

RUSSELLIAN PROPOSITIONS

JUDY PELHAM
ALASDAIR URQUHART *

Department of Philosophy, University of Toronto, Toronto, Ontario, Canada

Bertrand Russell, in the first decade of this century, held an unconventional view of propositions. He took them to be complex abstract entities resembling logical formulas in their basic structure, but differing from formulas in that they may contain physical objects as constituents. The aim of this paper is to give an account of Russell's notion of a proposition during the period 1903-06, and to explore the extent to which the logic which coexisted with that account of propositions is feasible.

The period 1903-06 lies between Russell's completion of *The Principles of Mathematics* (Russell 1903) and the beginning of the writing of *Principia Mathematica* (Whitehead and Russell, 1910). During this time Russell worked on a form of type-free logic, which he called the "no-classes theory" or "substitutional theory" (Russell 1905b,1906b), as a resolution to his paradox (or "the contradiction", as he called it). Russell's view of propositions given in the *Principles* changed and developed with his work on the paradoxes. Our elaboration of Russell's notion of proposition is based on the formal evidence of the substitutional theory, as it was worked out in unpublished manuscripts of 1905-06. This theory is based on Russell's conception of propositions as structured non-linguistic entities, and their fundamental logical properties.

Russell's ideas about propositions have come back into favour of late in connection with theories of direct reference (Kaplan 1986,1989) and the situation theory developed by Barwise, Perry, Etchemendy and others (Barwise and Etchemendy 1987). Recent authors, although they have found their inspiration in Russell, have not usually claimed to provide a historically accurate account of Russell's early ideas about propositions. The reconstruction attempted in this paper is intended to be faithful (as far as this is possible) to Russell's original intentions.

*Research partially supported by the Social Sciences and Humanities Research Council of Canada.

1. Complexity and constituents

In the *Principles* Russell draws the following picture of the universe. It consists of *terms* (synonymously, units, individuals or entities), a very broad ontological category including everything that "may be an object of thought, or may occur in any true or false proposition, or can be counted as *one*" (1903, p. 43). Examples would include: the relation 'loves', unicorns, the centre of mass of the universe, the class of bald French kings, as well as ordinary tables, chairs, persons and so on. This broad definition of term accorded with Russell's view of logic as the general science of reasoning which applied to all things whatsoever. Of these terms, some may be logically simple, while others are logically complex, containing other terms as constituents; in particular, a proposition about a term contains that term as a constituent (1903, p. 45).

The terms which concern the logician are all of finite complexity; on the question of infinitely complex propositions, Russell makes the following remarks (1903, pp. 145-146):

> Now, for my part, I see no possible way of deciding whether propositions of infinite complexity are possible or not; but this at least is clear, that all the propositions known to us (and, it would seem, all propositions that we *can* know) are of finite complexity. It is only by obtaining such propositions about infinite classes that we are enabled to deal with infinity; and it is a remarkable and fortunate fact that this method is successful. Thus the question whether or not there are infinite unities must be left unresolved; the only thing we can say, on this subject, is that no such unities occur in any department of human knowledge, and therefore none such are relevant to the foundations of mathematics.

This passage shows that it would be incorrect to interpret a universally quantified Russellian proposition as an infinite conjunction; on the contrary, it is (on Russell's view) a crucial property of universal quantification that it allows us to express infinitely many facts by finite means.

During the time he was writing the *Principles* and while he worked on possible solutions to the contradiction, Russell changed his views on the constituents of propositions, and indeed, on whether or not a proposition itself was an entity. But he held throughout to the view that physical objects were constituents of the propositions about them. In this, his account of propositions contrasts sharply with that of Frege. According to Frege's well known account, a linguistic expression expresses a sense, which in turn determines the denotation (if any) of the expression. The denotation of a

proper name is determined by its sense; the sense of a complete sentence is a function of the senses of its parts (Frege 1892). The constituents of the sense of a sentence are themselves senses of the (syntactical) constituents of the sentences. This is expressed very clearly in a letter Frege wrote to Russell on November 13 1904 (Frege 1980, p. 163):

> Truth is not a component part of a thought, just as Mont Blanc with its snowfields is not itself a component part of the thought that Mont Blanc is more than 4000 metres high. But I see no connection between this and what you go on to say: 'For me there is nothing identical about two propositions that are both true or both false'. The sense of the word 'moon' is a component part of the thought that the moon is smaller than the earth. The moon itself (i.e. the meaning of the word 'moon') is not part of the sense of the word 'moon'; for then it would also be a component part of that thought.

Russell's reply of 12 December 1904 brings out his contrasting position in a dramatic way (Frege 1980, p. 169):

> I believe that in spite of all its snowfields Mont Blanc itself is a component part of what is actually asserted in the proposition 'Mont Blanc is more than 4000 metres high'. We do not assert the thought, for this is a private psychological matter: we assert the object of the thought, and this is, to my mind, a certain complex (an objective proposition, one might say) in which Mont Blanc is itself a component part. If we do not admit this, then we get the conclusion that we know nothing at all about Mont Blanc.

Russell makes it clear in the passage which follows this quotation that his view is founded on the idea that (in the case of proper names at least) there is no sense to be distinguished. In the case of a name like 'Socrates', there is only the idea (which is psychological) and the denotation; a proper name denotes an object without the need for a mediating sense. The same motivation led Kaplan to his revival of Russell's account (Kaplan 1989, p. 483), based on the *Principles*.

It is sometimes said (see, for instance, Kaplan 1989, p. 496) that Russell abandoned the account of propositions given in the *Principles* in his paper on the theory of descriptions (Russell 1905a). However, Russell did not abandon the view with his theory of descriptions, he simply modified it. It should be remembered that Russell was not primarily interested in an analysis of natural language, as are contemporary writers on theories of direct reference. If we suppose that Russell's view of the structure of propositions is designed

to explain the way in which language operates, it may seem reasonable that when Russell gives up denoting concepts as a counterpart in the structure of propositions to the definite descriptions which occur in sentences, he is giving up the idea of structured propositions. The theory of descriptions abandons the idea that there is a constituent of the proposition expressed by the sentence "The King of France is bald" which corresponds to the words "The King of France", but it does not follow from this that Russell has given up his earlier theory. What he has given up is an analysis which parallels language. As we show below, the substitutional theory of 1905-06 cannot be understood without adopting the idea that both physical and abstract objects can occur as constituents of propositions.

In the *Principles*, Russell provides an analysis of propositions considered as the objects of thought. The sentences of language may (if we are lucky) correspond with our thoughts in such a way as to express them clearly, but the structure of thought is the antecedent study. The structured nature of propositions which emerges from the account of the *Principles* does not emerge as the second tier of an analysis of language; it exists as an analysis of the elements of thought in their own right. Russell changed his views about the nature of propositions as a result of internal difficulties in his theory, especially those engendered by the contradiction. He initially took the grammar of ordinary language to be his guide to the structure of propositions (1903, p. 42), but as time went on he came to think that the logical structure of propositions was less and less like the grammatical structure of the sentences expressing them.

Throughout the period 1903-05, Russell held that objects are paradigms of constituents of propositions. Propositions as the objects of belief contain the objects of acquaintance about which the belief is held. His original position in the *Principles* included sets and relations among terms; but this led to the contradiction. Later modifications took the form of eliminating a given type of entity as a primitive element of the system, in favour of re-defining it in terms of other, simpler, entities. He eliminated classes in favour of propositional functions, denoting concepts (which in the *Principles* included such things as are expressed by the phrase 'any thing') in favour of quantifiers. Finally, in the substitutional theory, he attempted a radical reduction in which propositions and their constituents are the only primitive entities, while classes and relations are defined contextually using an apparatus of substitution. In making these modifications, Russell was not so much adopting a metaphysical position as *trying out* certain logical possibilities to see if they would allow the consistent construction of mathematics. While certain metaphysical views were tenaciously held by Russell (such as the universality of logic) others (such as the ontological status of certain classes

of entities) were readily adopted or abandoned according to the logical needs of the moment.

Some writers hold that the idea that physical objects can be constituents of propositions is bizarre, or intuitively unacceptable. For example, Peter Hylton, in an informative article on the substitutional theory, says of this doctrine: 'This view is so counter-intuitive that it may be hard to believe that Russell meant it literally, but the evidence that he did so is overwhelming' (Hylton 1980, p. 28). However, it is clear that an exactly parallel situation holds in standard set theory. In an applied set theory allowing physical objects as individuals, the set { Mikhail Gorbachev, George Bush, $\sqrt{2}$ } has as constituents both physical objects and an abstract object, all of which quite happily coexist as constituents of the set. Clearly, Russell's idea is no more (and no less) counter-intuitive than the corresponding idea in standard set theory.

Russell had definite epistemological reasons for holding his view of propositions, as the above quotation shows. Discussions of this view often mix epistemological and logical issues, even (in some cases) attempting to redefine the notion of propositional constituent in psychological or epistemological terms (Sainsbury 1986). The view taken here is that the theory of propositions and their constituents is purely logical, and that it can be investigated without reference to the epistemological arguments offered in its justification. We shall therefore not discuss the difficult and controversial issues connected with Russell's principle of acquaintance, instead taking the basic concepts of Russell's theory as logical primitives.

2. Russell's foundational project 1903-05

In 1903, Russell began work on emendation of his logical system to eliminate the contradiction. Possible solutions to the contradiction were constrained by two things: Russell's desire to avoid type distinctions (at least for logical entities like propositions), and by the goal of deriving mathematics on the basis of this logic. During the period from 1903 to 1906, Russell tried out a large number of different approaches, although all of these ended either with a contradiction, or the impossibility of deriving arithmetic. Many of these approaches survive in Russell's unpublished papers (most of the logical papers from this period are to appear in Volumes 4 and 5 of *The Collected Papers of Bertrand Russell*). In the attempts of 1903-04, Russell sought to avoid the contradiction by eliminating classes as general abstract terms, introducing them only when they could be seen to arise from a particular function; he adopted something like Frege's course of values operator, defining classes contextually by propositional functions.

Unfortunately, these attempts in turn foundered on contradictions, and for a time Russell's strategy was to distinguish those functions that led to a contradiction, and those that did not. He had various methods of trying to draw this distinction during the period 1903-04; at different times he distinguishes 'quadratic' (paradoxical) from 'simple' functions, or 'reducible' from 'irreducible' functions. The cluster of ideas which employ this general strategy may be called the zig-zag theory (or, rather, zig-zag theories) in keeping with Russell's classification in his paper Russell 1906a; there he says (Russell 1973, pp. 145-146) that the zig-zag theory is based on "the suggestion that propositional functions determine classes when they are fairly simple, and only fail to do so when they are complicated and recondite." Russell found all of his attempts to carry out these classifications of functions problematic, and in 1905 he began exploring the idea central to the substitutional theory, namely that neither functions nor relations nor classes should be taken *prima facie* as terms, but that the work they had performed in the construction of arithmetic should be taken over by simpler entities.

The basic idea of the substitutional theory is that classes, relations and propositional functions should not be quantified over (i.e. considered as terms), but should be eliminated in favour of the notion of a matrix consisting of a proposition and a constituent of the proposition. For example the pair consisting of the proposition "Mikhail Gorbachev is a communist" and the man Mikhail Gorbachev can stand for the set of communists. The proposition does not have to be true; the designated constituent (Gorbachev in this case) simply plays the role of a dummy or place holder. The membership relationship can be defined by substitution; an object b is a member of the class represented by the pair p, a if the result of substituting b for a in p is a true proposition. Russell's intention was to build up the notions of classes and functions he required through iterations of this method. If the substitution of b for a in p yields q, then Russell writes this as: $p(b/a)!q$, or $p/a; b!q$. Within this last expression for the proposition, the expression 'p/a' is an incomplete symbol for a class. The definition of the number 0 (for example) is obtained by considering the proposition $(x)(q).p(x/a)!q. \supset . \sim q$. This says that the result of substituting x for a in p always results in a false propostion. Thus (p/a) represents a concept with no instances. On the account in the *Principles*, the number zero is defined as the *set* of all such concepts. In the substitution theory, we represent the number zero by the matrix $\{(x). \sim (p/a; x)\}/(p, a)$, where '$(p/a; x)$' is a definite description standing for "the result of substituting x for a in p". In this way Russell hoped to construct a consistent type-free theory which would allow him to prove the truths of arithmetic. For further details on Russell's theory, and its historical background, the reader should consult Grattan-Guinness 1974,

1977, Hylton 1980, Landini 1987, 1989.

The development of the substitutional theory was never completed (it underwent several revisions) and it was eventually abandoned, for reasons we discuss below. But the basic intuitions underlying the theory seem to arise from Russell's understanding of the notion of a proposition. Let us take for example the notion of a constituent. In the substitution theory Russell uses this concept in the statement of his axioms, writing 'p *in* q' if p is a constituent of q. In fact, as we shall see, in his formal development Russell defined the concept of a constituent in terms of the basic four-place substitution relation. However, there is a sense in which the concept of constituent was a basic and primitive idea in Russell's thinking about the substitutional theory. Some of the properties that Russell wanted to prove for his notion are plausible properties of a general notion of constituent; for example, the relation (p *in* q) is transitive, and antisymmetric. These considerations seem to indicate that Russell had an antecedent conception of the structure of propositions which he moulded the substitution theory to fit. The present paper adopts these intuitions as primitive in the reconstruction of the theory.

Russell's operation of substitution, although formally resembling the corresponding syntactical operation, is an operation which substitutes *objects* for *objects*, not syntactical items for other syntactical items. An analogy with set theory may be useful here (as in considering other aspects of the substitutional theory). In a set theory with a ground type of individuals, a mapping defined on the individuals induces a mapping on the sets built from those individuals. For example, if a permutation maps a into b, b into a and c into itself, then the induced mapping maps the set $\{a, c\}$ into $\{b, c\}$. The substitution operation in Russell's theory is an operation of the same kind, a purely structural or logical operation which obtains between logical entities.

3. Formal definition of Russellian propositions

This section contains a formal definition of Russellian propositions and the basic relations between them. The definition employs the more general notion of *propositional form*; propositional forms differ from propositions in that they can contain free variables. Since the variables in a propositional form occupy the positions of objects, it is tempting to think of them as standing for arbitrary objects, in the sense of Fine 1985. However, this would be contrary to Russell's own philosophy of logic. Russell unambiguously declares his stout opposition to arbitrary objects in the *Principles* (1903, pp. 90-91), and retained this view throughout his philosophical career. It is bet-

ter to think of a propositional form as an "incomplete" or "unsaturated" object like Frege's functions (Frege 1893, pp. 5-7) in which the variables are simply place markers. We have avoided Russell's own terminology "propositional function" because he sometimes uses this phrase for what we call propositional forms, sometimes for what we would usually call the functions determined by these forms (compare Frege's contrast between functions and their courses of values).

In any case, the propositional forms can be considered as intermediate objects which are introduced mainly as a convenience in defining propositions; they do not form part of the domain of quantification, and so are not truly entities.

We assume as given a non-empty set I of logically simple individuals; we use the letters $a, b, c, \ldots, a_0, b_0, c_0, a_1, b_1, c_1, \ldots$ to stand for such individuals. In addition, we assume an infinite set of variables: the letters $p, q, r, s, t, \ldots, x, y, z, \ldots$ are used as variable symbols. In addition, we assume a set of basic predicates, including a four-place predicate S for substitution, and Id, the identity predicate. The logical notation adopted is largely that of Whitehead and Russell 1910, including the use of dots in place of parentheses. In the case of identity, we write $\alpha = \beta$ in place of $Id(\alpha, \beta)$.

The set of propositional forms built from I is defined recursively as follows:

1. A variable standing alone is a propositional form;

2. If Σ is a k-place predicate, and $\alpha_1, \ldots, \alpha_k$ propositional forms or members of I, then $\Sigma(\alpha_1, \ldots, \alpha_k)$ is a propositional form;

3. If α, β are propositional forms or members of I then $\sim \alpha$ and $\alpha \vee \beta$ are propositional forms;

4. If α is a propositional form, and v a variable, then $(v)\alpha$ is a propositional form;

5. All propositional forms are obtained by repeated application of the preceding four rules.

A *proposition* is a propositional form containing no free variables; we denote the set of propositions built from I by the expression $Prop(I)$. The class of *objects* is the class $Prop(I) \cup I$ (recall that we do not count propositional forms as objects).

It may seem odd that although we do not count individuals as propositions, negations and disjunctions of individuals are propositions. However, this represents Russell's own views during this period. His view at this time

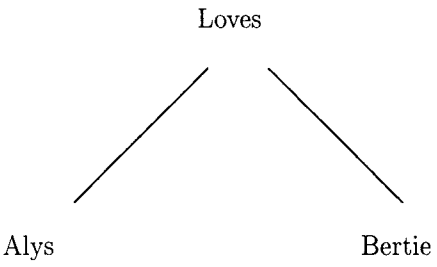


Figure 1: An atomic proposition

was that $\sim p$ was always a proposition, even when p was not; it was to be interpreted as "p is not true". Here, of course, Russell followed Frege in requiring that logical variables range over all conceivable objects, and that logical operations should be defined everywhere.

In view of the fifth condition in the truth definition, any propositional form has an associated formation tree, which shows how it is built up from the basic elements. We stipulate that a propositional form is uniquely determined by its formation tree, with the proviso that we identify two forms if one can be obtained from the other by change of bound variables. This definition of identity of Russellian propositions coincides with that given by Alonzo Church (Church 1984). As will be seen below, this "fine-grained" explication of Russell's notion of propositional identity receives support from some details of the substitutional theory.

It is useful to identify propositional forms with their formation trees, diagramming them as trees in which the leaves are labelled either with individuals or variables, and the interior nodes with predicates or logical operators. For example, the proposition expressed by "Alys loves Bertie" could be diagrammed as in Figure 1. In the figure we have shown the predicate "loves" as a logical operator rather than an object because in 1905, in contrast to the position in the *Principles*, Russell did not take properties or relations to be terms; the aim of the substitution theory was to avoid such an assumption by the apparatus of substitution. More complicated propositions can be diagrammed in a similar way. For example, the quantified proposition

$$(x) : x \vee \sim (S(x, x, x, x) = Mont\ Blanc)$$

can be represented as in Figure 2. The tree associated with an individual consists simply of that individual itself.

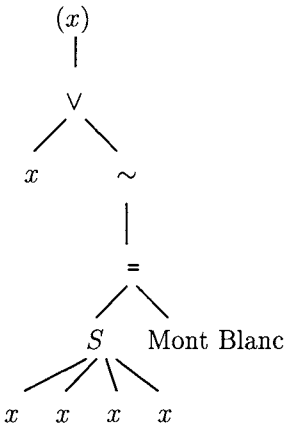


Figure 2: A universal proposition

An individual can have multiple occurrences in a proposition. This initially seems surprising because although we can create multiple copies of a symbol, it is harder to imagine multiple copies of an individual. However, a little thought reveals that the same phenomenon arises in set theory, where the number 1 occurs as both the first and second components of the ordered pair $\langle 1, 1 \rangle$. The solution to the puzzle in the case of propositions is the same as in the case of set theory: the concept "occurs in" is to be envisaged as a one-many relation both in the case of individuals and complex entities.

We define an individual or propositional form α to be a constituent of a propositional form β if α is identical with a subtree of β. If α, β are individuals or propositional forms, then the result of substituting β for γ in α is the individual or propositional form which results by replacing every occurrence of the constituent γ which satisfies two conditions by an occurrence of β; we write the resulting propositional form as $\alpha(\beta/\gamma)$. The conditions are: (1) no variable free in γ is bound in the occurrence of γ in α, and (2) no variable free in β is bound in the occurrence of β resulting from the substitution. The substitution operation can be visualized as a pruning and grafting operation; the occurrences of γ are removed from α, and replaced by occurrences of β. As a particularly simple case, we can replace one individual by another. For example, if Alys and Dora are Russell's first and second wives respectively, then Alys(Dora/Alys) = Dora.

It is an elementary exercise to define the class of Russellian propositions

as set-theoretical entities in a version of standard set theory containing individuals. In this case, a proposition could be defined as an equivalence class of sequences of symbols and individuals, and the constituent relation would be defined in terms of the membership relation. This is the course followed by Barwise and Etchemendy (1987, Chapter 4), in the more general context of Aczel's theory of hypersets. However, we avoid this course here, because we wish to emphasize the ontological primacy of propositions; in Russell's theory sets and relations are mere *façons de parler.*

4. A truth definition without quantification

For Russellian propositions not containing variables, the definition of truth is unproblematic, and is implicit in the description of the structure of propositions given in the preceding sections. Let us state the definition formally.

We consider only propositions in $Prop(I)$ not containing bound variables. For atomic propositions other than substitution and identity propositions we may assign truth-values arbitrarily. For the remaining atomic propositions, and truth-functionally compound propositions, we assign values according to the following rules, where $\alpha, \beta, \gamma, \delta$ stand for objects.

a. $\alpha = \beta$ is true if and only if α and β are identical; otherwise $\alpha = \beta$ is false.

b. $(\alpha \; in \; \beta)$ is true if and only if α is identical with a constituent of β; otherwise $(\alpha \; in \; \beta)$ is false.

c. $S(\alpha, \beta, \gamma, \delta)$ is true if and only if $\alpha(\beta/\gamma)$ and δ are identical.

d. $\alpha \vee \beta$ is true if and only if α is true or β is true; otherwise, $\alpha \vee \beta$ is false.

e. $\sim \alpha$ is true if and only if α is not true; otherwise, $\sim \alpha$ is false.

In the first three clauses of the truth definition, it is understood that identity of complex objects, their constituents and the substitution operation are to be construed as referring to the tree structure of the complexes in question. Thus, for example, in the first item, if α and β are propositions, then they are identical if the formation tree of β can be obtained from that of α by change of bound variables.

Although individuals are not assigned a truth-value, the complex objects constructed from them are; we are following Frege and the Russell of 1905 in their insistence that the truth-value assignment function be defined on the entire universe of objects (with the exception of simple individuals). So, for example, the complex object $\sim$(Bertrand Russell) has the value True, since

the author of *An Outline of Intellectual Rubbish* is not assigned the value True (compare Frege 1893, §6). It follows from this that any propositional form which is an instance of a tautology is true under any assignment.

If α is a propositional form which contains only free variables, then we say that α is *valid* if every proposition which results by substituting objects for the variables in α is true under every assignment. With this definition of validity in hand, we can investigate which propositions are valid, and compare the results with the evidence in Russell's own manuscripts. We adopt Russell's own notational conventions in writing $\alpha(\beta/\gamma)!\delta$ or $\alpha\frac{\beta}{\gamma}!\delta$ for $S(\alpha, \beta, \gamma, \delta)$.

It is easy to check that the following propositional forms are all valid:

1. $p\frac{x}{p}!x$
2. $p\frac{x}{x}!p$
3. $p\frac{x}{a}!q\,.\,p\frac{x}{a}!r.\supset.q = r$
4. $r\ in\ p\,.\,p\ in\ q.\supset.r\ in\ q$
5. $p\ in\ q\,.\,q\ in\ p.\supset.p = q$
6. $a \neq \sim p\,.\,p\frac{x}{a}!q.\supset.(\sim p)\frac{x}{a}!(\sim q)$

The propositions given above are among the primitive propositions adopted by Russell in an attempt at formalizing an early version of the substitution theory, dated December 22 1905 (Russell 1905b). They correspond to Russell's primitive propositions *12.21.211.201.24.241.212 (Russell in fact lists *12.211 as $p\frac{x}{x}!x$, but this appears to be a slip of the pen). The fifth proposition is particularly interesting, as it shows that Russell's concept of proposition is one which (in contrast to the currently fashionable idea of a proposition as a set of possible worlds) does not allow identification of logically equivalent propositions. For example, let us add to our language constants $\mathbf{T}$ and $\mathbf{F}$ for constant true and false propositions. Then $(\mathbf{T} \supset \mathbf{F})$ is equivalent to $\mathbf{F}$, while $(\mathbf{F} \supset \mathbf{F})$ is equivalent to $\mathbf{T}$. If we identify these equivalent pairs of propositions, then we obtain: $(\mathbf{T}\ in\ \mathbf{F})$ and $(\mathbf{F}\ in\ \mathbf{T})$, hence by the fifth proposition, $\mathbf{T} = \mathbf{F}$, an undesirable outcome. This observation underlines the fine-grained nature of Russell's conception of proposition; even apparently harmless identifications lead to a logical collapse.

Our results so far show that (if we confine the analysis to propositions not containing quantifiers) the substitutional theory of late 1905 is coherent, and that our semantical analysis is in accord with Russell's ideas, in so far as they are reflected in the primitive propositions which he set down in his

earliest attempt at an axiom system. This is encouraging, and furthermore demonstrates that objections to Russell's views directed simply against the idea that physical objects can be constituents of propositions are invalid. However, in presenting Russell's theory, we cheated a little; we took $(\alpha = \beta)$ and $(\alpha \; in \; \beta)$ to be atomic propositions. In fact, in the manuscript of December 1905 these are defined from the substitution predicate by means of quantifiers. Thus a fully accurate representation of Russell's theory, even for the propositions already listed, demands a definition of truth for quantified Russellian propositions. As will be seen in the next section, this leads to extremely difficult problems.

5. Truth for quantified propositions and a paradox

The extension of the definition of truth to Russellian propositions containing quantifiers appears initially to be quite straightforward. The following clause constitutes the natural extension of our earlier definition.

f. A proposition $(x).\alpha$ is true if and only if $\alpha(\beta/x)$ is true for any object β; $(x).\alpha$ is false if $\alpha(\beta/x)$ is false for some object β.

This amounts (in a sense) to a substitutional interpretation of the quantifiers, but it should not be confused with the interpretation which usually goes under that name. What is now called the "substitutional interpretation" involves substituting the *names* of individuals for variables, not the individuals themselves (Marcus 1961,1962, Quine 1961, Dunn and Belnap 1968, Parsons 1971). However, in Russell's scheme, the truth of a quantified proposition is evaluated by substituting the *objects themselves* for the free variable in the propositional form obtained by removing the binding quantifier. This interpretation is both substitutional (since it is defined by substitution in a propositional form), and referential (since it directly involves the objects in the universe of quantification). It is misleading to conflate Russell's account of quantification with that of the modern writers on substitutional theories. Russell is much closer to Quine in his insistence that in quantifying over objects we are implicitly referring to the objects themselves, not the names (or putative names) of objects.

In the Russellian interpretation of quantification, it is understood that the objects which may be substituted for variables include both the initial set of individuals and the complex objects constructed from them by the logical operations. Unfortunately, in the absence of type restrictions, this interpretation leads to paradox, if we assume that a truth-value can be assigned to all Russellian propositions. Russell discovered a basic paradox in the original version of his substitutional theory in 1906. Subsequently, he

modified the theory in various ways, but in the end was unable to escape the paradox.

Russell's substitutional paradox is most easily explained by starting from the assumption that there is an assignment of the values True and False to all Russellian propositions which satisfies the conditions (**a** - **f**) listed above. Let us assume in addition that the set of individuals I contains three distinct objects a_0, b, c, which we assume to be not logically complex (they could of course be physically complex). We define the proposition P_0 as:

$$(\exists p, a) : a_0. = . p\frac{b}{a}!c \ : \ (\exists z). p\frac{a_0}{a}!z \ . \ \sim z.$$

Now let R be the proposition $P_0(P_0\frac{b}{a_0}!c/a_0)$ which results from P_0 by substituting $P_0\frac{b}{a_0}!c$ for all occurrences of a_0 in P_0. Since we assumed that b, c are not logically complex, the only occurrences of a_0 in R are the two displayed above, so that R is the proposition:

$$(\exists p, a) : P_0\frac{b}{a_0}!c \ . = . p\frac{b}{a}!c \ : \ (\exists z). p\frac{P_0\frac{b}{a_0}!c}{a}!z \ . \ \sim z.$$

If the proposition R is false, then any proposition obtained by substitution in the propositional form which results by removing the initial quantifiers from R is also false. Hence, substituting P_0 for p and a_0 for a, we find that the following proposition is false:

$$P_0\frac{b}{a_0}!c = P_0\frac{b}{a_0}!c \ . \ (\exists z). P_0\frac{P_0\frac{b}{a_0}!c}{a_0}!z \ . \ \sim z.$$

Since the first conjunct is true, the second is false, so that the proposition

$$P_0\frac{P_0\frac{b}{a_0}!c}{a_0}!R \ . \ \sim R$$

is false. Since, by the definition of R, the first conjunct is true, $\sim R$ is false, that is, R is true, contrary to assumption. However, if R is true, then (by our definition of identity for propositions), the only objects p, a which satisfy the first conjunct in the matrix of R are P_0 and a_0, so that the following proposition is true:

$$(\exists z) : P_0\frac{P_0\frac{b}{a_0}!c}{a_0}!z \ . \ \sim z.$$

However, R is the unique object z satisfying the first conjunct, so that $\sim R$ is true. This is a contradiction.

The tricky substitution which produces the paradoxical proposition R bears a striking resemblance to the similar trick used by Gödel in constructing his undecidable sentence (Gödel 1931). In fact, the resemblance is not accidental, since Gödel discovered the undecidability result by attempting to give a substitutional interpretation of quantification over the real numbers, which failed because of the emergence of antinomies connected with truth and definability like the Liar and Richard's paradoxes (Wang 1981).

The paradox also bears a resemblance to the contradiction first pointed out by Tarski in theories of truth based on quotation-functions (Tarski 1956, pp. 159-162), and discussed by Binkley, Linsky, Marcus and Kripke (Binkley 1970, Linsky 1972, Marcus 1972, Kripke 1976). In her comments on the contradiction, Marcus observes that the derivation of the paradox implicitly violates the requirement for a recursive definition of truth that the definiens be less complex than the definiendum (Marcus 1972, pp. 246-247). Kripke observes than when this restriction is observed, extension of the truth definition to more inclusive languages leads to a theory closely resembling Russell's ramified theory of types for propositions (Kripke 1976, p. 368).

The contradiction revealed by the substitutional paradox is of a fundamental sort; it does not directly involve truth, quotation contexts or self-reference, as in other semantical paradoxes. When Russell first faced the paradoxes, he thought he could solve the problems posed by them by making *ad hoc* modifications to a set-theoretical apparatus built upon a logical foundation which could be assumed as given. The substitutional paradox, in contrast, uses only elementary relations between Russellian propositions in addition to basic logical concepts. It led in the end to the radical reconstruction of logic in the ramified theory of types.

6. Russell's way out

The substitutional paradox compelled Russell to modify his originally rather simple theory, making it more and more complicated. In unpublished manuscripts of 1906, Russell attempted to evade the contradiction by placing restrictions on the substitutions permitted in the theory. The reader should consult Landini's recent article for the details of some of these attempts (Landini 1989); we do not describe them here because they proved abortive.

In a letter (Russell 1907) dated January 22 1907, Russell wrote to Ralph Hawtrey: "I forgot to send you the paradox which killed the substitution-theory. Here it is." After giving the paradox more or less as it is given above, Russell concludes: "In trying to avoid this paradox, I modified the substitution-theory in various ways, but the paradox always reappeared in

more and more complicated forms." The final solution to the substitutional paradox and related antinomies was to be found only in 1907 with the creation of the ramified theory of types. Although in its first published version (Russell 1908) this theory retains a few traces of the substitutional theory from which it arose, it should be considered as a new departure, and not as the last version of the earlier attempts. In the final section we consider an alternative to ramification.

7. A partial truth predicate for Russellian propositions

Although the ramified theory of types provides a formally satisfactory solution to the paradoxes, its adoption went against the grain of Russell's deeply held logical convictions, in particular the idea that logic should be universal. In the ramified theory the universal quantifier no longer has the unrestricted scope which it had in the logic of the *Principles*, or in the 1905 version of the substitutional theory. From Russell's original viewpoint, this can hardly be accounted a satisfactory resolution, since it violates the universality which in the *Principles* constitutes the essence of logic. Indeed, it is hard not to feel a good measure of sympathy with Russell's opinion that the doctrine that propositions are of different types is "harsh and highly artificial" (1903, p. 528).

It is possible to resolve the substitutional contradiction while retaining a good deal of the 1905 substitutional theory by adopting some of the techniques recently used in the theory of truth by Kripke, Gupta, Herzberger and others (Kripke 1975, Gupta 1982, Herzberger 1982). Here semantically closed languages containing their own truth predicate are constructed by allowing the truth assignments to sentences to be partial. The assignments are constructed by starting from a collection of basic sentences, and adding new sentences by transfinite iteration until a fixed point is reached; the fixed point is used to define an assignment (which is necessarily partial, because the Liar paradox cannot have a truth-value).

We can follow essentially the same approach in attempting to reconstruct Russell's type-free substitutional theory. In doing this, it is necessary to rewrite the truth conditions for the propositional operators in order to accomodate the partiality of the assignment. Accordingly, we replace the conditions **d.** and **e.** above by the conditions:

d*. $\alpha \vee \beta$ is true if and only if α is true or β is true; $\alpha \vee \beta$ is false if and only if α is false and β is false;

e*. $\sim \alpha$ is true if and only if α is true; $\sim \alpha$ is false if and only if α is true.

These clauses correspond to the strong senses of the propositional connectives given by Kleene (Kleene 1952, p. 334). The remaining clauses of the truth definition remain as before.

The existence of a partial truth-value assignment to the set of Russellian propositions built from a set of individuals I can now be demonstrated by following the construction used in the theory of truth. We assume that Russellian propositions are modelled as set-theoretical entities built from the individuals in I, so that $Prop(I)$ forms a set. If X is a subset of the set $Prop(I) \times \{0,1\}$, we define $\Gamma(X)$ to be the union of X and the following sets:

1. $\{\langle \alpha = \beta, 1\rangle : \alpha, \beta$ identical objects$\} \cup$
   $\{\langle \alpha = \beta, 0\rangle : \alpha, \beta$ non-identical objects$\}$;

2. $\{\langle S(\alpha,\beta,\gamma,\alpha(\beta/\gamma)), 1\rangle : \alpha, \beta, \gamma$ objects $\} \cup$
   $\{\langle S(\alpha,\beta,\gamma,\delta), 0\rangle, \delta$ not identical with $\alpha(\beta/\gamma)\}$;

3. $\{\langle \alpha \vee \beta, 1\rangle : \langle \alpha, 1\rangle \in X$ or $\langle \beta, 1\rangle \in X\} \cup$
   $\{\langle \alpha \vee \beta, 0\rangle : \langle \alpha, 0\rangle \in X$ and $\langle \beta, 0\rangle \in X\}$;

4. $\{\langle \sim \alpha, 1\rangle : \langle \alpha, 0\rangle \in X\} \cup \{\langle \sim \alpha, 0\rangle : \langle \alpha, 1\rangle \in X\}$;

5. $\{\langle (x)\alpha, 1\rangle : \langle \alpha(\beta/x), 1\rangle \in X,$ for all objects $\beta\} \cup$
   $\{\langle (x)\alpha, 0\rangle : \langle \alpha(\beta/x), 0\rangle \in X,$ for some object $\beta\}$.

Since Γ is a monotone operator, it follows by a standard result in the theory of inductive definition (Moschovakis 1980, p. 404) that there is a subset of $Prop(I) \times \{0,1\}$ which is a least fixed point of Γ, denoted by Γ^∞. Thus $\Gamma(\Gamma^\infty) = \Gamma^\infty$, and Γ^∞ is the intersection of all subsets Y of $Prop(I) \times \{0,1\}$ satisfying $\Gamma(Y) \subseteq Y$. Let us call a subset Z of $Prop(I) \times \{0,1\}$ *consistent* if there is no α in $Prop(I)$ such that $\langle \alpha, 1\rangle \in Z$ and $\langle \alpha, 0\rangle \in Z$. The operator Γ preserves consistency, hence since $\emptyset$ is consistent, and consistency is preserved under union, the set Γ^∞ is consistent. It follows that we can define a partial assignment of truth values to $Prop(I)$ by the rule: a proposition α has the value True if $\langle \alpha, 1\rangle \in \Gamma^\infty$, while it has the value False if $\langle \alpha, 0\rangle \in \Gamma^\infty$. It is not hard to verify that this assignment in fact verifies the list of modified conditions for truth and falsehood given above.

We can now compare the results of this revised approach with the evidence in Russell's manuscripts. In the assignments constructed by the fixed point method, the proposition $(x).\beta\frac{x}{\alpha}!\beta$ always has a truth-value, which coincides with the truth-value of $\sim$ (α *in* β). Hence, (α *in* β) can be defined (as in Russell's manuscripts) by $\sim (x).\beta\frac{x}{\alpha}!\beta$. Similarly, $(\alpha = \beta)$ can be defined as $(x).x\frac{\alpha}{x}!\beta$; this is the definition adopted by Russell in an undated manuscript which appears to have been written in 1906 (Russell 1906c).

If we define classes contextually in the way sketched by Russell, then we find that the set theory given by this construction contains a universal class. Furthermore, it is not hard to see that the universal class is necessarily infinite. The argument for this conclusion in the present setting is a variant of the classic arguments of Bolzano and Dedekind (see Russell 1903, p. 357); for any object a, the objects $\sim a$, $\sim\sim a$, $\sim\sim\sim a$, ... are all distinct. (Anderson in his formulation of Russellian intensional logic (1989) deduces the Axiom of Infinity by the same argument.)

This result is of course one highly desired by Russell; in 1904 he argued at length against C.J. Keyser that the Axiom of Infinity is a truth of logic (Russell 1904), while in 1906 he gave a proof of it in the context of the substitutional theory (Russell 1906d), based on the sequence of propositions $a = u, (a = u) = u, ((a = u) = u) = u, \ldots$, where a, u are distinct objects. With the adoption of the ramified theory of types, the Axiom of Infinity was no longer derivable, and Whitehead and Russell had to assume it as an explicit hypothesis whenever it was needed in *Principia Mathematica* (although the argument of Anderson mentioned above shows that the Axiom of Infinity is derivable in Church's formulation of Russell's intensional logic, even though this logic is based on a ramified theory of types). In the present reconstruction we have reproduced some of the key features of Russell's theory of propositions of 1905, although at the cost of altering the logic. How much of the logicist programme can be reconstructed in this framework remains a subject for further research.

8. Acknowledgments

Our thanks to Peter Apostoli, John Corcoran, Nicholas Griffin and Ruth Barcan Marcus for helpful comments on earlier versions of this essay. Passages from unpublished manuscripts and letters of Russell are quoted with the permission of the Bertrand Russell Archives with whom the copyright resides.

References

For unpublished manuscript material by Russell, references such as 'RA 220.010940b' give the archival numbers of material held in the Bertrand Russell Archives, McMaster University, Hamilton, Ontario, Canada.

ALMOG, J., PERRY, J. and WETTSTEIN, H. (eds.), 1989, *Themes from Kaplan*, Oxford University Press.

ANDERSON, C.A., 1989, *Russellian Intensional Logic* in: Almog, Perry and Wettstein, pp. 67-103.

BARWISE, J. and ETCHEMENDY, J., 1987, *The Liar: An Essay on Truth and Circularity*, Oxford University Press, New York and Oxford.

BINKLEY, R., 1970, *Quantifying, Quotation, and a Paradox*, Noûs, 4, 271-277.

CHURCH, A., 1984, *Russell's Theory of Identity of Propositions*, Philosophia Naturalis, 21, 513-522.

DUNN, J.M. AND BELNAP, N.D., 1968, *The substitution interpretation of the quantifiers*, Noûs, 2, 177-185.

FINE, K., 1985, *Reasoning with Arbitrary Objects*, Basil Blackwell, Oxford.

FREGE, G., 1892, *Über Sinn und Bedeutung*, Zeitschrift für Philosophie und philosophische Kritik, n.s. 100, 25-50. English translation in Frege 1984.

FREGE, G., 1893, *Grundgesetze der Arithmetik, begriffsschriftlich abgeleitet*, Vol. 1, Verlag Hermann Pohle, Jena.

FREGE, G., 1980, *Philosophical and Mathematical Correspondence*, eds. Gabriel, Hermes, Kambartel, Thiel, Veraart; translated by Hans Kaal. University of Chicago Press, Chicago.

FREGE, G., 1984, *Collected Papers on Mathematics, Logic and Philosophy*, Basil Blackwell, Oxford.

GÖDEL, K. ,1931, *Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I*, Monatshefte für Mathematik und Physik, 38, 173-198.

GRATTAN-GUINNESS, I., 1974, *The Russell archives: Some new light on Russell's logicism*, Annals of Science, 31, 387-406.

GRATTAN-GUINNESS, I., 1977, *Dear Russell, Dear Jourdain*, Columbia University Press, New York.

GUPTA, A., 1982, *Truth and Paradox*, J. of Philosophical Logic, 11, 1-60.

HERZBERGER, H., 1982, *Notes on naive semantics*, J. of Philosophical Logic, 11, 61-102.

HYLTON, P., 1980, *Russell's substitutional theory*, Synthese, 45, 1-31.

KAPLAN,D., 1986, *Opacity*, in The Philosophy of W.V. Quine, Open Court Publishing Company, La Salle, Illinois, pp. 229-289.

KAPLAN,D., 1989, *Demonstratives. An Essay on the Semantics, Logic, Metaphysics, and Epistemology of Demonstratives and Other Indexicals*, in: Almog, Perry and Wettstein, pp. 481-563. (Circulated in mimeographed form since 1977.)

KLEENE, S.C., 1952, *Introduction to Metamathematics*, Van Nostrand, Princeton, N.J.

KRIPKE, S.A., 1975, *Outline of a Theory of Truth*, J. of Philosophy, 72, 690-716.

KRIPKE, S.A., 1976, *Is There a Problem about Substitutional Quantification?*, in: Evans and McDowell, eds., Truth and Meaning, Oxford University Press, 325-419.

LANDINI, G., 1987, *Russell's substitutional theory of classes and relations*, History and Philosophy of Logic, 8, 171-200.

LANDINI, G., 1989, *New evidence concerning Russell's substitutional theory of classes*, Russell, n.s. 9, 26-42.

LINSKY, L., 1972, *Two Concepts of Quantification*, Noûs, 6, 224-239.

MARCUS, R.B., 1961, *Modalities and intensional languages*, Synthese, 13, 303-322.

MARCUS, R.B., 1962, *Interpreting quantifiers*, Inquiry, 5, 252-259.

MARCUS, R.B., *Quantification and Ontology*, Noûs, 6, 240-250.

MOSCHOVAKIS, Y.N., 1980, *Descriptive Set Theory*, North Holland Publishing Company, Amsterdam.

PARSONS, C., 1971, *A plea for substitutional quantification*, J. of Philosophy, 68, 231-237. Reprinted in Parsons 1983, 63-70.

PARSONS, C., 1983, *Mathematics in Philosophy*, Cornell U.P., Ithaca, New York.

QUINE, W.V., 1961, *Reply to Professor Marcus* , Synthese, 13, 323-330. Reprinted in Quine 1966, 175-182.

QUINE, W.V., 1966, *The Ways of Paradox and other essays*, Random House, New York.

RUSSELL, B., 1903, *The Principles of Mathematics*, Cambridge University Press. Second edition, George Allen and Unwin, 1937.

RUSSELL, B., 1904, *The Axiom of Infinity*, The Hibbert Journal, 2, pp. 809-812. Reprinted in Russell 1973, 256-259.

RUSSELL, B., 1905a, *On Denoting*, Mind, n.s. 14, pp. 479-493. Reprinted in Russell 1973, 103-119.

RUSSELL, B., 1905b, *On Substitution*, unpublished manuscript (RA 220.010940b). To appear in Volume 5 of the Collected Papers of Bertrand Russell (ed. G.H. Moore).

RUSSELL, B., 1906a, *On some difficulties in the theory of transfinite numbers and order types*, Proceedings of the London Mathematical Society, series 2, 4, pp. 29-53. Reprinted in Russell 1973, 135-164.

RUSSELL, B., 1906b, *On the Substitutional Theory of Classes and Relations*, received by the London Mathematical Society 24 April 1906; read before the Society 10 May 1906. First printed in Russell 1973, 165-189.

RUSSELL, B., 1906c, *Substitution*, unpublished manuscript (RA 220.010940), To appear in Volume 5 of the Collected Papers of Bertrand Russell (ed. G.H. Moore).

RUSSELL, B., 1906d, *On 'Insolubilia' and their Solution by Symbolic Logic*, in Russell 1973, 190-214. First published in French translation as Les Paradoxes de la Logique, Revue de Métaphysique et de Morale, 14, 627-50.

RUSSELL, B., 1907, Letter to Ralph Hawtrey, dated January 22 1907. Copy in Russell Archives (REC. ACQ. 394).

RUSSELL, B., 1908, *Mathematical Logic as based on the Theory of Types*, American J. of Mathematics, 30, 222-262. Reprinted in Russell 1956, 59-102.

RUSSELL, B., 1956, *Logic and Knowledge*, ed. R.C. Marsh, George Allen and Unwin, London.

RUSSELL, B., 1973, *Essays in Analysis*, ed. Douglas Lackey, George Braziller, New York.

SAINSBURY, M., 1986, *Russell on acquaintance*, in G. Vesey ed., Philosophers Ancient and Modern, Cambridge University Press, pp. 219-244.

TARSKI, A., 1956, *Logic, Semantics, Metamathematics*, Oxford University Press.

WANG, H., 1981, *Some facts about Kurt Gödel*, J. Symbolic Logic, 46, 653-659.

WHITEHEAD, A.N. and RUSSELL, B., 1910, *Principia Mathematica*, Volume I, Cambridge University Press.

Logic, Methodology and Philosophy of Science IX
D. Prawitz, B. Skyrms and D. Westerståhl (Editors)

ACCEPTING FAILURE IN DYNAMIC LOGIC

KRISTER SEGERBERG

Dept. of Philosophy, Uppsala University, Villavägen 5, 75236 Uppsala, Sweden

1. Background

If one wants a good semantic representation of a program over a universe of possible total states, then one does well to consider the paths that are associated with α. Those paths may be divided into two groups: those that correspond to complete computations according to the program and those that do not. The latter group can be further divided into two subgroups: the finite and the infinite. Thus with each program α one can associate three sets of paths,

$$\begin{aligned} H(\alpha) &= \text{the set of } \textit{halt} \text{ paths,} \\ I(\alpha) &= \text{the set of } \textit{infinite} \text{ paths,} \\ F(\alpha) &= \text{the set of } \textit{fail} \text{ paths.} \end{aligned}$$

In this way each program α defines what we might call a *signature* $(H(\alpha), I(\alpha), F(\alpha))$. One might go further and identify programs with the same signature. To do so would not be completely justified. To take a trivial example, consider an impossible program ω—a program that can never be carried out. Evidently, $H(\omega) = \emptyset$. It is easy to see that the programs ω and $\omega;\omega$ (and, in general, $\omega;\omega;\ldots;\omega$) have the same signature, yet syntactically they are different programs. However, the proposed identification seems harmless for many purposes or even beneficial.

Working with paths is complicated. By comparison, what goes on in dynamic logic (in the narrow sense of the term) is much simpler. There a program α is represented by a binary "accessibility" relation $R(\alpha)$, two points x and y being related by $R(\alpha)$ if there is a computation according to α starting at x and terminating at y. This relation is readily defined in terms of the former paragraph. If p is any path, let us agree to write $p(0)$ for the first element of p and $p(\sharp)$ for the last element (if there is one). Then

we can define

$$R(\alpha) = \{(x, y) : \exists p \in H(\alpha)(p(0) = x \ \& \ p(\sharp) = y)\}.$$

This is simpler than the first modelling on two counts. One is that of the paths in $H(\alpha)$ only the start point and the halt point are retained. The other is that the sets $I(\alpha)$ and $F(\alpha)$ play no rôle at all.

Let us call the two kinds of semantics (both of which would seem to go back to Vaughan Pratt) *path semantics* and *relational semantics*, respectively. In relational semantics it might seem natural to identify two programs if they give rise to the same accessibility relation. However, such identification would not be as harmless as it was in path semantics. In the literature one sometimes sees the claim that the formalism of dynamic logic is "capable of expressing" operators such as IF - THEN - ELSE -. It is indeed true that a program IF A THEN α ELSE β has the same set of halt paths as the program $((?A); \alpha) + ((?\neg A); \beta)$ of dynamic logic ("either first to verify that A is true and then to run α, or first to verify that A is false and then to run β"). Therefore it is also true that the two programs determine the same accessibility relation in relational semantics, and that, for any formula C, the formula

$$[\text{IF } A \text{ THEN } \alpha \text{ ELSE } \beta]C \equiv [((?A); \alpha) + ((?\neg A); \beta)]C$$

is valid in that semantics. Yet the two programs need not be identical in terms of path semantics. For example, the trivial program IF $\top$ THEN $?\top$ ELSE $?\top$ has no fail paths, whereas $< x >$ is a fail path of $((?\top); (?\top)) + ((?\bot); (?\top))$, for every point x.

Suppose now that one wanted to strike a balance between the sensitivity of the path semantics and the convenience of the relational semantics. One such compromise, to be studied in this paper, would be to adhere to the perspective of dynamic logic in accepting $R(\alpha)$ as an abstraction of or substitute for $H(\alpha)$ but insist on some compensation for giving up $I(\alpha)$ and $F(\alpha)$. The minimum compensation would seem to be the possibility of expressing the possibility that a certain program α will always halt if started at a certain point x—in such a case let us say that x is *normal* with respect to α or that α is *safe* at x. This would motivate the introduction of a set $N(\alpha)$ as a new semantic primitive. In terms of the path semantics that set is readily definable:

$$N(\alpha) = \{x : \forall p(p \in I(\alpha) \cup F(\alpha) \Longrightarrow p(0) \neq x)\}.$$

But the converse does not hold—from a relation $R(\alpha)$ and a set $N(\alpha)$ one is not able, in any interesting case, to reconstruct the sets $H(\alpha)$, $I(\alpha)$ and $F(\alpha)$.

By accepting the importance of failure and providing some means for discussing it, the new semantics becomes able to handle certain operators which are sensitive to the existence or not of fail or infinite paths. One such operator is IF - THEN - ELSE -, and in a planned future paper we hope to give an analysis of that and some related operators. Another example is the delta-operator which the author studied in [5] and which is the topic of the present paper. The idea was to introduce an operator, δ, taking propositions (reporting states-of-affairs) to programs. Syntactically, δA is a well-formed expression (a "term") whenever A is a formula. Semantically, δA may be thought of as the action of seeing to it that A, or the bringing about that A.

How is one to model the new concept? The suggestion in [5] was to cast δA as the locally maximal reliable way of seeing to it that A. From the vantage point of any particular point, that a program is a reliable way of seeing to it that A means that every computation from that point according to the program halts at a point where A holds. Maximality is achieved by viewing δA as the "sum" or "union" of all those ways—as the most inclusive way under the circumstances of reliably seeing to it that A. Unfortunately, the system presented in [5] is unsatisfactory, something that has been pointed out by Timothy J. Surendonk and S. K. Thomason, independently of one another ([7], [8]): the formalism does not properly reflect the motivation behind it. The author's own diagnosis is that the defect in [5] is that fail paths are neglected—only halt paths are taken into account. In other words, the simplified perspective of ordinary propositional dynamic logic does not work when the delta-operator is added. An example due to Thomason makes this very clear: by taking only halt paths into account we invite the unwanted consequence that $\alpha; ?A$ ("first to run α, and then to verify that A is true") must be considered a way of doing A, for every α; for the formula $[\alpha; ?A]A$ is universally true. The point is that in a situation when α has been completed but A is not true, the program $\alpha; ?A$ fails—no halt path is ever forthcoming.

The motivational details have received more detailed investigation in [6]. Here we will restrict ourselves to providing a satisfactory formal semantics and giving a provably complete axiomatization of the resulting logic.

The expressions of our formal language are divided into two distinct categories, *terms* and *formulae*, which are interrelated in intricate ways. There is a denumerable supply of propositional letters, and they all count as formulae (our *basic formulae*). There are the usual Boolean operators (connectives), forming formulae from formulae. There are no program letters (in this paper); the *basic terms* are the expressions $?A$ and δA where A is any formula. Moreover, $\alpha + \beta$ and $\alpha; \beta$ and α^* are terms whenever α and β are terms. For every term α, $[\alpha]$ is a "box operator", so $[\alpha]A$ is a formula whenever α

is a term and A is a formula. Finally, we introduce a new formula making operator OK operating on terms: OK α is a formula whenever α is a term. From these remarks a formal definition is readily derivable. We will drop parentheses somewhat casually but in ways that should not invite confusion.

The special features of this paper are the delta-operator δ, explained above, and the new operator OK, explained below. Informally, for OK α one may read "α is safe" or (colloquially) "α is OK". Operators identical or closely related to OK have been studied by other authors under different names. Most notably, in [2] Robert Goldblatt introduced an operator t from programs to formulae, reading $\mathrm{t}(\alpha)$ as "the assertion that execution of α always terminates" ([2] p. 101). Goldblatt's interest in this operator stemmed from his insight that it is needed in order to formalize E. W. Dijkstra's concept of 'weakest precondition', described by Dijkstra as "the condition that characterizes the set of all initial states such that activation will certainly result in a properly terminating happening leaving the system in a final state satisfying a given postcondition" (quoted from [2] p. 90). Suppose that A is a given formula (the postcondition) and α a given term. Goldblatt suggests that the condition Dijkstra calls for—written $wp(\alpha, A)$—is to be defined as $wp(\alpha, A) = \mathrm{t}(\alpha) \wedge [\alpha]A$; or in the terminology of this paper:

$$wp(\alpha, A) = \text{OK } \alpha \wedge [\alpha]A.$$

This suggestion seems clearly right, at least within the framework considered here. If so, our work below takes on additional interest.

2. Formal semantics

Let U be a given set of points. By an *action* over U we mean an element $a \in \mathfrak{P}(U \times U) \times \mathfrak{P}U$. The two natural projection functions are denoted by R and N, respectively:

$$\begin{aligned} R &: \mathfrak{P}(U \times U) \times \mathfrak{P}U \longrightarrow \mathfrak{P}(U \times U), \\ N &: \mathfrak{P}(U \times U) \times \mathfrak{P}U \longrightarrow \mathfrak{P}U. \end{aligned}$$

Say that Ra is the *accessibility relation* corresponding to a and that Na is the *normal region* of a (if $x \in Na$ we say that a is *safe* at x and that x is *normal* with respect to a). Note that, for every action a,

$$a = < Ra, Na > .$$

We say that $(A, \oplus, \odot, ^*, ?)$ is an *algebra of actions* over U if A is a set of actions over U and $\oplus$ and $\odot$ and * are operations on A (binary, binary, unary, respectively) and $?$ is a function from $\mathfrak{P}U$ to A satisfying the following conditions:

(R1) $\quad R(a \oplus b) = Ra \cup Rb,$

(R2) $\quad R(a \odot b) = Ra \mid Rb,$

(R3) $\quad R(a^\star) = (Ra)^*,$

(R4) $\quad R(?S) = \Delta \uparrow S,$

(N1) $\quad N(a \oplus b) = Na \cap Nb,$

(N2) $\quad N(a \odot b) = Na \cap \{x : \forall y((x,y) \in Ra \Longrightarrow y \in Nb)\},$

(N3) $\quad N(a^\star) = \{x : \forall y((x,y) \in (Ra)^* \Longrightarrow y \in Na)\},$

(N4) $\quad N(?S) = S,$

(NR) $\quad Na \subseteq \{x : \exists y(x,y) \in Ra\}.$

Here $\mid$ stands for relative product and * for the ancestral. We have used Δ for the diagonal set $\{(x,x) : x \in U\}$ and $\Delta \uparrow S$ for the set $\Delta \cap (S \times S)$, in other words, $\{(x,x) : x \in S\}$. More notation: where r is a binary relation it is sometimes convenient to use the notation $r(x)$ or rx for the set $\{y : (x,y) \in r\}$. With this convention conditions (N2), (N3) and (NR) can be rewritten in the following more manageable form:

(N2) $\quad N(a \odot b) = Na \cap \{x : Ra(x) \subseteq Nb\},$

(N3) $\quad N(a^\star) = \{x : (Ra)^*(x) \subseteq Na\},$

(NR) $\quad Na \subseteq \{x : Ra(x) \neq \emptyset\}.$

Whenever $(A, \oplus, \odot, ^\star, ?)$ is an algebra of actions over U there is another auxiliary function of importance, namely, the function $D : \mathfrak{P}U \longrightarrow \mathfrak{P}(U \times U) \times \mathfrak{P}U$ defined as follows: for any $S \subseteq U$, $DS =< RDS, NDS >$, where

$$\begin{aligned} RDS &= \{(x,y) : \exists a \in A((x,y) \in Ra \ \&\ x \in Na \ \&\ Ra(x) \subseteq S)\}, \\ NDS &= \{x : \exists y(x,y) \in RDS\}. \end{aligned}$$

Let us say that $(A, \oplus, \odot, ^\star, ?)$ is a *D-algebra* if the range of D is included in A; that is, if $DS \in A$, for all $S \subseteq U$.

Turning now to logic, let us shift the perspective slightly. We define $(U, A, \oplus, \odot, ^\star, ?)$ as a *standard frame* if $(A, \oplus, \odot, ^\star, ?)$ is a D-algebra. As usual, a *valuation* in U is a function from the set of propositional letters to $\mathfrak{P}U$. A structure $\mathfrak{M} = (U, A, \oplus, \odot, ^\star, ?, V)$ is a *standard model* if $(U, A, \oplus, \odot, ^\star, ?)$ is a standard frame and V is a valuation in U. The task of semantics

is to provide a definition of the *meaning* or *intension* in $\mathfrak{M}$ of every term or formula in the object language; let us write $\|\alpha\|^{\mathfrak{M}}$ and $\|A\|^{\mathfrak{M}}$ for that concept, where α is any term and A is any formula (we omit the superscript when confusion is unlikely to result). Our definition goes as follows.

(int1) $\|P\| = V(P)$, if P is a propositional letter,

(int2) $\|A \wedge B\| = \|A\| \cap \|B\|$, $\|\neg A\| = U - \|A\|$, etc.,

(int3) $\|?A\| = (\Delta \uparrow \|A\|, \|A\|)$

(int4) $\|\delta A\| = D\|A\|$,

(int5) $\|\alpha + \beta\| = \|\alpha\| \oplus \|\beta\|$,

(int6) $\|\alpha; \beta\| = \|\alpha\| \odot \|\beta\|$,

(int7) $\|\alpha^*\| = \|\alpha\|^\star$,

(int8) $\|[\alpha]A\| = \{x : (R\|\alpha\|)(x) \subseteq \|A\|\}$,

(int9) $\|\text{OK } \alpha\| = N\|\alpha\|$.

Notice that the intension of a formula is a proposition, while the intension of a term is an action. If A is a formula, we say that A is *true* at x if $x \in \|A\|$. As usual, a formula is *valid* in a frame if true at all the points of the frame. The completeness problem is to provide an axiomatization of the set of formulae valid in all standard frames. This problem we will now address.

3. Proposed axiomatization

Let us say that a logic is *normal* if it is closed under the rules

(MP) if A and $A \supset B$ are theses, then so is B,

(RM) if $A \equiv B$ is a thesis, then so is $[\alpha]A \equiv [\alpha]B$,

and if it also contains as axioms

(AX0) all tautologies

as well as all instances of the following schemata:

(AX1) $[\alpha](A \wedge B) \equiv ([\alpha]A \wedge [\alpha]B)$,

(AX2) $[\alpha]\top,$

(AX3) $[\alpha + \beta]A \equiv ([\alpha]A \wedge [\beta]A),$

(AX4) $[\alpha;\beta]A \equiv [\alpha][\beta]A,$

(AX5) $[\alpha^*]A \supset A,$

(AX6) $[\alpha^*]A \supset [\alpha]A,$

(AX7) $[\alpha^*]A \equiv [\alpha^*][\alpha^*]A,$

(AX8) $A \supset ([\alpha^*](A \supset [\alpha]A) \supset [\alpha^*]A),$

(AX9) $[?A]B \equiv (A \supset B),$

(AX10) $[\delta A]A,$

(AX11) $[\delta A]B \supset ([\delta B]C \supset [\delta A]C),$

(AX12) $[\delta A]B \supset (A \supset B),$

(AX13) $\text{OK } \alpha + \beta \equiv (\text{OK } \alpha \wedge \text{OK } \beta),$

(AX14) $\text{OK } \alpha;\beta \equiv (\text{OK } \alpha \wedge [\alpha] \text{ OK } \beta),$

(AX15) $\text{OK } \alpha^* \equiv [\alpha^*] \text{ OK } \alpha,$

(AX16) $\text{OK } ?A \equiv A,$

(AX17) $\text{OK } \delta A \equiv < \delta A > \top,$

(AX18) $\text{OK } \delta A;\delta B \supset \text{OK } \delta B,$

(AX19) $\text{OK } \delta A;\delta B \supset ([\delta B]C \supset [\delta A;\delta B]C).$

LEMMA 3.1 *The following rules are derivable in any normal logic:*

(RD) *If* $A \equiv B$ *is a thesis, then so is* $[\delta A]C \equiv [\delta B]C.$

(RN) *If* $A \equiv B$ *is a thesis, then so is* $\text{OK } \delta A \equiv \text{OK } \delta B.$

We omit the proof but observe that (AX10) and (AX11) are used to prove that (RD) holds, while (AX17) ("both ways") is used for (RN). The lemma is useful if one wants to establish the derivability of yet another rule, that of Replacement of Provable Equivalents:

THEOREM 3.2 *The following rule is derivable in any normal logic:*

(RPE) *If $A \equiv B$ is a thesis and C and C' are formulae which differ only in one of them having an occurrence of A where the other has an occurrence of B, then $C \equiv C'$ is also a thesis.*

We omit the lengthy proof (cf. [6]).

LEMMA 3.3 *$A \supset \text{OK}\ \delta A$ is a thesis of any normal logic:*

Proof. Putting $\bot$ for B in (AX12) gives us $A \vdash < \delta A > \top$ after truth-functional simplification. Hence $A \vdash \text{OK}\ \delta A$, by (AX17). □

The class of normal logics is closed under intersection. The smallest normal logic is of course the intersection of all normal logics.

THEOREM 3.4 *All theses of the smallest normal logic are valid in all standard frames.*

Proof. A full proof is by induction on the length of formal proofs and consists in checking all axioms and the two rules. We are content to give just two examples. Let $\mathfrak{M} = (U, A, \oplus, \odot, ^\star, ?, V)$ be a standard model.

(AX12): Suppose that $x \in \|[\delta A]B\|$ and $x \in \|A\|$. Then, for all y, if $(x, y) \in R\|\delta A\|$ then $y \in \|B\|$. To show that $x \in \|B\|$ it will be enough to show that $(x, x) \in R\|\delta A\|$. Note that $?\|A\| \in A$ and that $?\|A\| = (\Delta \uparrow \|A\|, \|A\|)$. Hence $(x, x) \in R\|?A\|$ and $x \in N\|?A\|$; and, trivially, if $(x, z) \in R\|?A\|$, for any z, then $z \in \|A\|$. Consequently, $(x, x) \in R\|\delta A\|$, as wanted.

(AX18): Suppose that

(0) $\quad x \in \|\text{OK}\ \delta A; \delta B\|.$

Throughout this proof, for any formula C, we will move freely and without comment from $\|\delta C\|$ to $D\|C\|$ and back. From (0) it follows that

(1) $\quad x \in N\|\delta A; \delta B\|.$

Consequently,

(2) $\quad x \in ND\|A\|,$

(3) $\quad \forall u((x, u) \in RD\|A\| \Longrightarrow u \in ND\|B\|).$

From (1) and the definition of D we infer the existence of some element y such that

(4) $\quad (x, y) \in RD\|A\|.$

By (3) and (4),

$$(5) \qquad y \in ND\|B\|.$$

From (5) and the definition of D we infer the existence of some element z such that

$$(6) \qquad (y,z) \in RD\|B\|.$$

By (3) and (6), $(x,z) \in RD\|A\| \mid RD\|B\|$. Hence

$$(7) \qquad (x,z) \in R\|\delta A; \delta B\|.$$

Let w be any element such that $(x,w) \in R\|\delta A; \delta B\|$. Then there is some v such that $(x,v) \in R\|\delta A\|$ and $(v,w) \in R\|\delta B\|$. By the definition of D, $(v,w) \in RD\|B\|$ implies that $w \in \|B\|$. This argument shows that

$$(8) \qquad \forall w((x,w) \in R\|\delta A; \delta B\| \Longrightarrow x \in \|B\|).$$

Since $\mathfrak{M}$ is standard, $\|\delta A; \delta B\| \in A$. Therefore, in view of (1), (7) and (8), the definition of D yields $(x,z) \in RD\|B\|$. Consequently, $x \in ND\|B\|$. But then $x \in \|\text{OK } \delta B\|$, as we wanted to show. □

4. Fischer/Ladner closure

Like other proofs in this area, our proof starts with the set of all maximal, consistent sets of formulae and then "divides out" by a certain finite set Ψ of formulae (or "filters" the big collection of maximal consistent sets through the finite set Ψ). The latter set has to be closed under certain conditions, named after M.J.Fischer and R.E.Ladner (see [1]). In our case the Fischer/Ladner conditions will be these:

(FL0) If $o(A_0, \ldots, A_{n-1}) \in \Psi$, for some n-ary formula-making operator o on formulae, then $A_0, \ldots, A_{n-1} \in \Psi$.

(FL1) If $[?A]B \in \Psi$, then $A \in \Psi$.

(FL2) If $[\delta A]B \in \Psi$, then $A \in \Psi$.

(FL3) If $[\alpha + \beta]A \in \Psi$, then $[\alpha]A \in \Psi$ and $[\beta]A \in \Psi$.

(FL4) If $[\alpha; \beta]A \in \Psi$, then $[\alpha][\beta]A \in \Psi$.

(FL5) If $[\alpha^*]A \in \Psi$, then $[\alpha][\alpha^*]A \in \Psi$.

(FL6) If OK $?A \in \Psi$, then $A \in \Psi$.

(FL7) If OK $\delta A \in \Psi$, then $A \in \Psi$.

(FL8) If OK $\alpha + \beta \in \Psi$, then OK $\alpha \in \Psi$ and OK $\beta \in \Psi$.

(FL9) If OK $\alpha;\beta \in \Psi$, then OK $\alpha \in \Psi$ and $[\alpha]$OK $\beta \in \Psi$.

(FL10) If OK $\alpha^* \in \Psi$, then $[\alpha^*]$OK $\alpha \in \Psi$.

Let FLΨ, the *Fischer/Ladner closure* of Ψ, be the smallest extension of Ψ that satisfies conditions (FL1)-FL10). The main result of this section is that FLΨ is finite whenever Ψ is (Theorem 4.2). In order to conduct our discussion at a reasonable level of rigour we now define an ordering of the expressions of our object language that can be used as an induction order in a proof. The notion to be defined is that of *immediate predecessor*, for which we use the symbol $\lhd$.

(IP0) If P is a propositional letter, then there is no expression E such that $E \lhd P$;

(IP1) $A \lhd A \wedge B$ and $B \lhd A \wedge B$; $A \lhd \neg A$; etc.;

(IP2) $\alpha \lhd \alpha + \beta$ and $\beta \lhd \alpha + \beta$;

(IP3) $\alpha \lhd \alpha;\beta$ and $\beta \lhd \alpha;\beta$;

(IP4) $\alpha \lhd \alpha^*$;

(IP5) $\alpha \lhd [\alpha]A$ and $A \lhd [\alpha]A$;

(IP6) $A \lhd ?A$;

(IP7) $A \lhd \delta A$;

(IP8) $\alpha \lhd$ OK α.

(IP9) For any expressions E and F, $E \lhd F$ only by virtue of one of the conditions (IP1)-(IP8).

Precedence is the transitive closure of immediate precedence. It is clear that all and only the propositional letters occurring in an expression precede it.

LEMMA 4.1 *Let E be any expression and Q any propositional letter not in E. Then the following conditions are satisfied:*

(i) *If E is a formula A, then* FL$\{A\}$ *is finite.*

(ii) *If E is a term α, then* $\mathrm{FL}\{[\alpha]Q\}$ *and* $\mathrm{FL}\{\mathrm{OK}\ \alpha\}$ *are both finite.*

Proof. The proof, by induction on the complexity of E, is an elaboration of the corresponding proof in [1]. Notice that our argument is divided into a number of cases corresponding to the clauses in the definition of immediate precedence.

Basic step: E is a propositional letter P. It is clear that $\{P\}$ is closed under the Fischer/Ladner conditions—(FL0) holds trivially and the rest vacuously. Moreover, $\{P\}$ is certainly the smallest set with this property. Hence $\mathrm{FL}\{P\} = \{P\}$, indeed a finite set.

Induction step: E is complex. As induction hypothesis, assume that the lemma holds for the immediate predecessors of E. There are several cases; here we give four examples.

First suppose that E is a formula that is a Boolean combination of other formulae; for example, suppose that E is $A \wedge B$, for some A and B. Inspection shows that

$$\mathrm{FL}\{A \wedge B\} = \{A \wedge B\} \cup \mathrm{FL}\{A\} \cup \mathrm{FL}\{B\};$$

for the right hand side is closed under all the rules, and no smaller set is. Since $A \lhd A \wedge B$ and $B \lhd A \wedge B$, the induction hypothesis applies to A and to B. Hence $\mathrm{FL}\{A\}$ and $\mathrm{FL}\{B\}$ are finite, and therefore so is $\mathrm{FL}\{A \wedge B\}$. Other Boolean cases are similar.

For our second example, suppose that E is $\alpha;\beta$, for some terms α and β. Here we encounter a complication that requires new notation. If A and B are formulae and Q is a propositional letter, let us write $A(B/Q)$ for the formula resulting from A by uniform substitution of B for Q. If Σ is any set of formulae, then let us write $\Sigma(B/Q)$ for the set resulting from Σ by uniform substitution of B for Q; that is, the set $\{A(B/Q) : A \in \Sigma\}$. Now in the present case we have

$$\mathrm{FL}\{[\alpha;\beta]Q\} = \{[\alpha;\beta]Q\} \cup \mathrm{FL}\{[\alpha][\beta]Q\}.$$

We claim, omitting the proof, that

$$\mathrm{FL}\{[\alpha][\beta]Q\} = \mathrm{FL}\{[\alpha]Q\}([\beta]Q/Q) \cup \mathrm{FL}\{[\beta]Q\},$$

remarking only on the importance of the assumption that Q does not appear in α. Since $\alpha \lhd \alpha+\beta$ and $\beta \lhd \alpha+\beta$ both α and β are covered by the induction hypothesis. Consequently, $\mathrm{FL}\{[\alpha;\beta]Q\}$ is finite. Furthermore,

$$\mathrm{FL}\{\mathrm{OK}\ \alpha;\beta\} = \{\mathrm{OK}\ \alpha;\beta\} \cup \mathrm{FL}\{\mathrm{OK}\ \alpha, [\alpha]\mathrm{OK}\ \beta\}.$$

By the same argument as before,

$$\mathrm{FL}\{[\alpha]\mathrm{OK}\ \beta\} = \mathrm{FL}\{[\alpha]Q\}(\mathrm{OK}\ \beta/Q) \cup \mathrm{FL}\{\mathrm{OK}\ \beta\}.$$

Hence also $\mathrm{FL}\{\mathrm{OK}\ \alpha;\beta\}$ is finite.

For our third example, suppose that E is α^*, for some term α. Note that

$$\mathrm{FL}[\alpha^*]Q = \mathrm{FL}\{[\alpha][\alpha^*]Q\}.$$

Hence, since Q does not appear in α,

$$\mathrm{FL}[\alpha^*]Q = \mathrm{FL}\{[\alpha]Q\}([\alpha^*]Q/Q) \cup \{Q\}.$$

Since $\alpha \lhd \alpha^*$, $\mathrm{FL}\{[\alpha]Q\}$ is finite, by the induction hypothesis. It follows that $\mathrm{FL}[\alpha^*]Q$ is finite. Furthermore,

$$\mathrm{FL}\{\mathrm{OK}\ \alpha^*\} = \{\mathrm{OK}\ \alpha^*\} \cup \mathrm{FL}\{[\alpha^*]\mathrm{OK}\ \alpha\},$$

and

$$\mathrm{FL}\{[\alpha^*]\mathrm{OK}\ \alpha\} = \mathrm{FL}\{[\alpha^*]Q\}(\mathrm{OK}\ \alpha/Q) \cup \mathrm{FL}\{\mathrm{OK}\ \alpha\}.$$

We just saw that $\mathrm{FL}\{[\alpha^*]Q\}$ is finite. Moreover, since $\alpha \lhd \alpha^*$, $\mathrm{FL}\{\mathrm{OK}\ \alpha\}$ is finite, by the induction hypothesis. It follows that $\mathrm{FL}\{\mathrm{OK}\ \alpha^*\}$ is finite.

For our final example, suppose that E is $\mathrm{OK}\ \alpha$, for some term α. Since $\alpha \lhd \mathrm{OK}\ \alpha$, $\mathrm{FL}\{\mathrm{OK}\ \alpha\}$ is finite by the induction hypothesis. But this is exactly what we wanted to establish in this case. □

THEOREM 4.2 *If Ψ is any finite set of formulae, then $\mathrm{FL}\Psi$ is still finite.*

Proof. Suppose that $\Psi = \{A_0, \ldots, A_{n-1}\}$. Then $\mathrm{FL}\Psi = \mathrm{FL}\{A_0\} \cup \ldots \cup \mathrm{FL}\{A_{n-1}\}$. □

5. Preparing for filtration

Let L be a fixed normal logic. We write U_L for the set of all maximal L-consistent sets of formulae; we will use lower case letters x, y, z, etc. for elements of U_L. For each term α we define the *canonical accessibility relation* as

$$R_L(\alpha) = \{(x,y) : \forall C([\alpha]C \in x \Longrightarrow C \in y)\}.$$

We recall the following standard result, in effect due to E.J.Lemmon and Dana Scott:

$$[\alpha]C \in x \text{ iff } \forall y((x,y) \in R_L(\alpha) \Longrightarrow C \in y).$$

Let Ψ be a fixed finite nonempty set of formulae closed under the Fischer/Ladner conditions. In a natural way this set induces an equivalence relation on U_L, viz., by the definition $x \equiv y$ iff $x \cap \Psi = y \cap \Psi$. We write $\bar{x}$ for the set $\{x' : x \equiv x'\}$, the equivalence class of x. It is a fact of crucial importance that $U^\dagger$, the set $\{\bar{x} : x \in U_L\}$ of all equivalence classes, is finite. Let BCΨ be the set of Boolean combinations of formulae in Ψ. For every $A \in$ BCΨ, let $|A|^\dagger$ be the set $\{\bar{x} : A \in x\}$. Note that if $A, B \in$ BCΨ and $A \equiv B$ is a thesis of L, then $|A|^\dagger = |B|^\dagger$. Note also that for every subset $S \subseteq U^\dagger$ there is some $A \in$ BCΨ such that $S = |A|^\dagger$.

Our goal in this section is to construct a certain filtration-like structure over the space $U^\dagger$, a construction which will lead to completeness. In this section we take a first step towards that construction by defining, for each element x, a set PSAFE($\bar{x}$) of sets of paths starting at $\bar{x}$. As explained in Section 1, our relational semantics is different from the path semantics, but even in our cruder models it is possible to define certain paths.

Informally, let us focus on paths along which action proceeds in such a way that there is no need to risk failure. How can one be certain of avoiding failure in $U^\dagger$? Perhaps by imitating safe action in U. Doing C at u in U—running δC starting at u—is safe if OK $\delta C \in u$; or equivalently, according to (AX17), if $< \delta C > \top \in u$. In other words, doing C at u in U is safe if there is some v such that $(u, v) \in R_L(\delta C)$. One might hope, then, that the same condition will guarantee the safety of doing C at $\bar{u}$ in $U^\dagger$, except that for obvious reasons we must require that $C \in$ BCΨ. This would then be an example of what might be called an "atomic" action that is safe. Presumably chains or series of safe atomic actions would also be safe.

After this informal preamble let us now move on to formal construction. First some concepts concerning paths in $U^\dagger$. If $p =< \bar{z}_0, \ldots, \bar{z}_m >$ is a path and $\bar{u}$ is any element in $U^\dagger$, then $p^\frown\bar{u}$ is the path $< \bar{z}_0, \ldots, \bar{z}_m, \bar{u} >$. If $p = < \bar{z}_0, \ldots, \bar{z}_m >$ and $q =< \bar{u}_0, \ldots, \bar{u}_n >$ are paths, then pq is defined if and only if $\bar{z}_m = \bar{u}_0$; in that case, $pq =< \bar{z}_0, \ldots, \bar{z}_m, \bar{u}_1, \ldots, \bar{u}_n >$. Thus pq is defined only if $p(\sharp) = q(0)$. Again, if $p =< z_0, \ldots, z_m >$ and $q = < z_0, \ldots, z_n >$, we say that p is an *initial* of q if $m \leq n$; and if $m < n$ we say that p is a *proper initial* of q or that q *continues* p.

For every formula $C \in$ BCΨ we define a new binary relation over $U^\dagger$:

$$R^\dagger(\delta C) = \{(\bar{x}, \bar{y}) : \exists x' \equiv x \exists y' \equiv y (x', y') \in R_L(\delta C)\}.$$

Notice that if $C \equiv C'$ is a thesis of L, then $R^\dagger(\delta C) = R^\dagger(\delta C')$.

We define PSAFE($\bar{x}$) as the smallest set Σ of sets of paths in $U^\dagger$ that satisfies the following conditions. First, the one-element set whose only member is the one-element path consisting of just $\bar{x}$ is an element of Σ,

called the *trivial* element:

$$\{< \bar{x} >\} \in \Sigma.$$

In general, suppose that $\sigma \in \Sigma$. Let p be a certain path such that $p \in \sigma$ and suppose that $p(\sharp) = \bar{u}$, for some u. Let τ be a set of paths such that, for some formula $C \in \mathrm{BC}\Psi$,

$$\tau = (\sigma - \{p\}) \cup \{p^\frown \bar{v} : (\bar{u}, \bar{v}) \in R^\dagger(\delta C)\}.$$

Then $\tau \in \Sigma$. This completes the definition.

We make a number of assorted comments. The sequence $< \bar{x} >$ is a path starting at $\bar{x}$ representing zero atomic actions. Thus the trivial element $\{< \bar{x} >\}$ defends its membership in $\mathrm{PSAFE}(\bar{x})$ by playing the rôle of null element. By the same token, the inductive step of the definition of $\mathrm{PSAFE}(\bar{x})$ in effect defines a relation of succession: in the situation described we say that τ is an *immediate successor* of σ or that σ *immediately precedes* τ. Notice that $\sigma \in \mathrm{PSAFE}(\bar{x})$ only if there are $\sigma_0, \ldots, \sigma_m$ such that $\sigma_0 = \{< \bar{x} >\}$ and $\sigma_m = \sigma$ and, for all $i < m$, σ_i immediately precedes σ_{i+1}. In the latter case we say that the sequence $< \sigma_0, \ldots, \sigma_m >$ *generates* σ. We say that σ *produces* $A \in \mathrm{BC}\Psi$ if $A \in p(\sharp)$, for all paths $p \in \sigma$. Later we will find it convenient to use the binary relation *rel* σ, defined as follows:

$$\mathit{rel}\ \sigma = \{(\bar{x}, \bar{y}) : \exists p \in \sigma \bar{y} = p(\sharp)\}.$$

Thus, for each $\sigma \in \mathrm{PSAFE}(\bar{x})$ and $A \in \mathrm{BC}\Psi$, σ produces A if and only if $\forall z((\bar{x}, \bar{z}) \in \mathit{rel}\ \sigma \Longrightarrow A \in z)$.

We now prove some technical results that are fairly obvious but which will be important later. Without proof we observe that every set in $\mathrm{PSAFE}(\bar{x})$ is finite (even though $\mathrm{PSAFE}(\bar{x})$ itself may be infinite) and that for every set $\sigma \in \mathrm{PSAFE}(\bar{x})$ and every path $p \in \sigma$, $p(0) = \bar{x}$.

LEMMA 5.1 *Suppose that* $< \sigma_0, \ldots, \sigma_m >$ *generates* σ, *for some* $\sigma \in \mathrm{PSAFE}(\bar{x})$. *If* $p \in \sigma_i$, *for some* $i \leq m$, *then there is some* $q \in \sigma$ *such that* p *is an initial of* q.

Proof. Make the assumptions of the lemma. The proof is by backward induction on i. If $i = m$ the situation is trivial. Suppose that the lemma holds for $i+1$ and that $p =< \bar{z}_0, \ldots, \bar{z}_n >$ is a path such that $p \in \sigma_i$ (whence of course $\bar{z}_0 = \bar{x}$). If $p \in \sigma_{i+1}$, then the desired conclusion follows from the induction hypothesis. So suppose that $p \notin \sigma_{i+1}$. Then, by the definition of $\mathrm{PSAFE}(\bar{x})$, there is some element v and some formula C such that

$$p^\frown \bar{v} \in \sigma_{i+1}.$$

Then by the induction hypothesis there is some $q \in \sigma$ such that $p^\frown \bar{v}$ is an initial of q. But p is an initial of $p^\frown \bar{v}$, hence *a fortiori* an initial of q. □

LEMMA 5.2 *Suppose that* $\sigma \in \mathrm{PSAFE}(\bar{x})$. *Suppose also that for some* $p \in \sigma$ *there is a nontrivial set* $\tau \in \mathrm{PSAFE}(p(\sharp))$. *Let* $\theta = \{pq : q \in \tau\}$. *Then* $\theta \in \mathrm{PSAFE}(\bar{x})$.

Proof. Make the assumptions of the lemma. Suppose that $< \sigma_0, \ldots, \sigma_m >$ generates σ and that $< \tau_0, \ldots, \tau_n >$ generates τ. Since τ is nontrivial, $n > 0$. For each i such that $0 \leq i \leq n$, define $\nu_i = (\sigma - \{p\}) \cup \tau_i$. Then $\sigma_m = \nu_0$, and the sequence $< \sigma_0, \ldots, \sigma_m, \nu_1, \ldots, \nu_n >$ generates θ. □

COROLLARY 5.3 *Suppose that* $\sigma \in \mathrm{PSAFE}(\bar{x})$. *Suppose also that for each* $p \in \sigma$ *there is a nontrivial set* $\tau(p) \in \mathrm{PSAFE}(p(\sharp))$. *Let* $\theta = \{pq : p \in \sigma \ \& \ q \in \tau(p)\}$. *Then* $\theta \in \mathrm{PSAFE}(\bar{x})$.

Proof. The number of paths in σ is finite. The result follows by repeated applications of Lemma 5.2. □

6. Filtrations

In the model we are building there will be two kinds of basic actions (not to be confused with the atomic actions we have been dealing with above): those of type $|?A|^\dagger$ and those of type $|\delta A|^\dagger$. The former are defined in agreement with the general format adopted in Section 2. If $A \in \mathrm{BC}\Psi$, then

$$\begin{aligned} R|?A|^\dagger &= \{(\bar{x}, \bar{x}) : A \in x \cap \Psi\} = \Delta \uparrow |A|^\dagger, \\ N|?A|^\dagger &= \{\bar{x} : A \in x \cap \Psi\} = |A|^\dagger. \end{aligned}$$

On the other hand, if $A \notin \mathrm{BC}\Psi$, then $R|?A|^\dagger = N|?A|^\dagger = \emptyset$.

The relation $R|\delta A|^\dagger$—not to be confused with the relation $R^\dagger(\delta A)$—is defined as follows. If $A \in \mathrm{BC}\Psi$, then $R|\delta A|^\dagger$ is defined as the set of all pairs $(\bar{x}, \bar{y})$ such that, for some $\sigma \in \mathrm{PSAFE}(\bar{x})$,

(i) $(\bar{x}, \bar{y}) \in \mathit{rel}\ \sigma$,

(ii) σ produces A.

If $A \notin \mathrm{BC}\Psi$, then by definition $R|\delta A|^\dagger = \emptyset$. In either case we define $N|\delta A|^\dagger = \{\bar{x} : \exists y (\bar{x}, \bar{y}) \in R|\delta A|^\dagger\}$. Note that $N|\delta A|^\dagger = \emptyset$ if $A \notin \mathrm{BC}\Psi$.

We are now in a position to define the structure $(U^\dagger, A^\dagger, \oplus, \odot, ^\star, ?)$ promised above. Define the operations $\oplus$ and $\odot$ and * and $?$ as follows:

$$\begin{aligned} |\alpha|^\dagger \oplus |\beta|^\dagger &= |\alpha + \beta|^\dagger, \\ |\alpha|^\dagger \odot |\beta|^\dagger &= |\alpha; \beta|^\dagger, \\ (|\alpha|^\dagger)^\star &= |\alpha^*|^\dagger, \\ ?(|A|^\dagger) &= |?A|^\dagger. \end{aligned}$$

Let $A^\dagger$ be the set of all $|\alpha|^\dagger$ such that α is a term. It is clear that $A^\dagger$ is closed under the operations $\oplus$, $\odot$,* and ?; therefore, $(U^\dagger, A^\dagger, \oplus, \odot, ^\star, ?)$ is a frame. The main result of this section is that it is even standard (Theorem 6.3).

LEMMA 6.1 *Suppose that* $(\bar{x}, \bar{y}) \in R|\alpha|^\dagger$ *and* $\bar{x} \in N|\alpha|^\dagger$. *Then there is some* $\sigma \in \mathrm{PSAFE}(\bar{x})$ *such that*

(i) $\quad (\bar{x}, \bar{y}) \in \mathit{rel}\ \sigma,$

(ii) $\quad \forall z((\bar{x}, \bar{z}) \in \mathit{rel}\ \sigma \Longrightarrow (\bar{x}, \bar{z}) \in R|\alpha|^\dagger).$

Proof. Assume that $(\bar{x}, \bar{y}) \in R|\alpha|^\dagger$ and $\bar{x} \in N|\alpha|^\dagger$. The proof is by induction on α.

The basic step consists of two cases. In one α is of the form δA; this case is unproblematic, so we omit the proof. In the other α is of the form $?A$. The underlying assumption is that $(\bar{x}, \bar{y}) \in R|?A|^\dagger$ and $\bar{x} \in N|?A|^\dagger$. It follows that $\bar{x} = \bar{y}$ and $A \in x \cap \mathrm{BC}\Psi$. Since $U^\dagger$ is finite we can find a formula $C \in \mathrm{BC}\Psi$ that characterizes $\bar{x}$ is the following sense: for all $u \in U_L$,

$$\bar{u} = \bar{x} \text{ iff } C \in u.$$

Define $\sigma = \{(\bar{x}, \bar{z}) : (\bar{x}, \bar{z}) \in R^\dagger(\delta C)\}$. As $C \in \mathrm{BC}\Psi$, $\sigma \in \mathrm{PSAFE}(\bar{x})$. Suppose that $[\delta C]B \in x$. Since $C \in x$, (AX12) yields $B \in x$. This shows that $(x, x) \in R_L(\delta C)$. Hence $(\bar{x}, \bar{x}) \in \mathit{rel}\ \sigma$, so condition (i) is satisfied. Take any z such that $(\bar{x}, \bar{z}) \in \mathit{rel}\ \sigma$; that is, $(\bar{x}, \bar{z}) \in R^\dagger(\delta C)$. Then there is some $x' \equiv x$ and some $z' \equiv z$ such that $(x', z') \in R_L(\delta C)$. By (AX10), $C \in z'$, hence $\bar{x} = \bar{z}$. But $(\bar{x}, \bar{x}) \in R|?A|^\dagger$, so condition (ii) is also satisfied.

In the second part of the basic step α is of the form δA; we omit the unproblematic proof.

For the inductive step, assume as the induction hypothesis that the lemma holds for some terms α and β; we wish to prove that it holds for $\alpha + \beta$ and $\alpha;\beta$ and $\alpha^\star$ as well. In the first of these cases the proof is straightforward, and we omit it. The second case is more problematic. Here the underlying assumption is that $(\bar{x}, \bar{y}) \in R|\alpha;\beta|^\dagger$ and $\bar{x} \in N|\alpha;\beta|^\dagger$. It follows that there is some u such that on the one hand $(\bar{x}, \bar{u}) \in R|\alpha|^\dagger$ and $\bar{x} \in N|\alpha|^\dagger$ and on the other $(\bar{u}, \bar{y}) \in R|\beta|^\dagger$ and $\bar{u} \in N|\beta|^\dagger$. With the help of the induction hypothesis (applied twice) we deduce that there is some $\sigma \in \mathrm{PSAFE}(\bar{x})$ and some $\tau \in \mathrm{PSAFE}(\bar{u})$ such that

(1) $\quad (\bar{x}, \bar{u}) \in \mathit{rel}\ \sigma,$

(2) $\quad \forall z((\bar{x}, \bar{z}) \in \mathit{rel}\ \sigma \Longrightarrow (\bar{x}, \bar{z}) \in R|\alpha|^\dagger);$

(3) $\quad (\bar{u}, \bar{y}) \in \mathit{rel}\ \tau,$

(4) $\quad \forall z((\bar{u}, \bar{z}) \in \mathit{rel}\ \tau \Longrightarrow (\bar{u}, \bar{z}) \in R|\beta|^{\dagger}).$

Let us make some observations. Take any arbitrary path $s \in \sigma$ and suppose that $\bar{v} = s(\sharp)$. Then $(\bar{x}, \bar{v}) \in \mathit{rel}\ \sigma$, so $(\bar{x}, \bar{v}) \in R|\alpha|^{\dagger}$, by (2) . The fact that $\bar{x} \in N|\alpha; \beta|^{\dagger}$ implies that $\bar{v} \in N|\beta|^{\dagger}$. Hence there exists some element w such that $(\bar{v}, \bar{w}) \in R|\beta|^{\dagger}$. By the induction hypothesis, therefore, there is some set $\tau(\bar{v}) \in \mathrm{PSAFE}(\bar{v})$ such that

(5) $\quad (\bar{v}, \bar{w}) \in \mathit{rel}\ \tau(\bar{v}),$

(6) $\quad \forall z((\bar{v}, \bar{z}) \in \mathit{rel}\ \tau(\bar{v}) \Longrightarrow (\bar{v}, \bar{z}) \in R|\beta|^{\dagger}).$

We make a stipulation in the special case that $\bar{v} = \bar{u}$: that $\tau(\bar{u}) = \tau$. These observations show that we are facing a situation of the kind described in Lemma 5.3. Hence $\theta \in \mathrm{PSAFE}(\bar{x})$, where $\theta = \{pq : p \in \sigma\ \&\ q \in \tau(p(\sharp))\}$. We must show that

(©1) $\quad (\bar{x}, \bar{y}) \in \mathit{rel}\ \theta,$

(©2) $\quad \forall z((\bar{x}, \bar{z}) \in \mathit{rel}\ \theta \Longrightarrow (\bar{x}, \bar{z}) \in R|\alpha; \beta|^{\dagger}).$

That (©1) holds follows at once from (1) and (3). As for (©2), suppose that $(\bar{x}, \bar{z}) \in \mathit{rel}\ \theta$. Then there is some v such that $(\bar{x}, \bar{v}) \in \mathit{rel}\ \sigma$ and $(\bar{v}, \bar{z}) \in \mathit{rel}\ \tau(\bar{v})$. From (2) and (6), respectively, we conclude that $(\bar{x}, \bar{v}) \in R|\alpha|^{\dagger}$ and $(\bar{v}, \bar{z}) \in R|\beta|^{\dagger}$. Consequently, $(\bar{x}, \bar{z}) \in R|\alpha; \beta|^{\dagger}$, as we wanted.

In the third case, the underlying assumption is that $(\bar{x}, \bar{y}) \in R|\alpha^{*}|^{\dagger}$ and $\bar{x} \in N|\alpha^{*}|^{\dagger}$. Then there is some n and some elements $\bar{z}_0, \ldots, \bar{z}_n$ such that $\bar{z}_0 = \bar{x}$ and $\bar{z}_n = \bar{y}$ and, for all $i < n$, $(z_i, z_{i+1}) \in R|\alpha|^{\dagger}$ and $z_i \in N|\alpha|^{\dagger}$. An argument along the lines of the preceding case is now possible; we omit the details. □

COROLLARY 6.2 *For all* $A \in \mathrm{BC}\Psi$, $D|A|^{\dagger} = |\delta A|^{\dagger}$.

Proof. First suppose that $(\bar{x}, \bar{y}) \in RD|A|^{\dagger}$, for some $A \in \mathrm{BC}\Psi$. Then there is some term α such that

(1) $\quad (\bar{x}, \bar{y}) \in R|\alpha|^{\dagger},$

(2) $\quad \bar{x} \in N|\alpha|^{\dagger},$

(3) $\quad \forall z((\bar{x}, \bar{z}) \in R|\alpha|^{\dagger} \Longrightarrow \bar{z} \in |A|^{\dagger}).$

Thanks to (1) and (2), we may infer from Lemma 6.1 the existence of some set $\sigma \in \mathrm{PSAFE}(\bar{x})$ such that $(\bar{x}, \bar{y}) \in \mathit{rel}\ \sigma$ and $\forall z((\bar{x}, \bar{z}) \in \mathit{rel}\ \sigma \Longrightarrow (\bar{x}, \bar{z}) \in R|\alpha|^\dagger)$. By (3), then, $\forall z((\bar{x}, \bar{z}) \in \mathit{rel}\ \sigma \Longrightarrow \bar{z} \in |A|^\dagger)$. In other words, σ produces A. Consequently, $(\bar{x}, \bar{y}) \in R|\delta A|^\dagger$. This argument shows that $RD|A|^\dagger \subseteq R|\delta A|^\dagger$. The converse is easily seen to hold, so in fact $RD|A|^\dagger = R|\delta A|^\dagger$.

We use the result just obtained to complete the proof: $ND|A|^\dagger = \{\bar{x} : \exists y(\bar{x}, \bar{y}) \in RD|A|^\dagger\} = \{\bar{x} : \exists y(\bar{x}, \bar{y}) \in R|\delta A|^\dagger\} = N|\delta A|^\dagger$. □

THEOREM 6.3 *$(U^\dagger, A^\dagger, \oplus, \odot, ^\star, ?)$ is a standard frame.*

Proof. By Corollary 6.2, $A^\dagger$ is closed under the ?-operator. But we must also show that condition (NR) is satisfied. In other words, we must establish the following claim: for all terms α,

$$\bar{x} \in N|\alpha|^\dagger \Longrightarrow \exists y(\bar{x}, \bar{y}) \in R|\alpha|^\dagger.$$

But this claim is readily proved by induction on α. □

7. Completeness

At this point let us take stock of the situation. We have seen previously (Theorem 3.4) that the axiom system in Section 3 is consistent with respect the semantics in Section 2. In this section we will prove the converse: that the axiomatization is actually complete as well as consistent. This result will following from the following theorem.

Let $\mathfrak{M}^\dagger$ be defined as the standard model $(U^\dagger, A^\dagger, \oplus, \odot, ^\star, ?, V^\dagger)$, where, for every propostional letter P,

$$V^\dagger(P) = \begin{cases} \{\bar{x} : P \in x\} & \text{if } P \in \Psi, \\ \emptyset & \text{if } P \notin \Psi. \end{cases}$$

We mark intensions in $\mathfrak{M}^\dagger$ by a dagger. Let us write $\alpha\eta\Psi$ if the term α occurs in some formula in Ψ; that is, if $[\alpha]B \in \Psi$, for some B, or OK $\alpha \in \Psi$.

THEOREM 7.1 *In the model $\mathfrak{M}^\dagger$, for all formulae $A \in \mathrm{BC}\Psi$ and for all terms $\alpha\eta\Psi$, $\|A\|^\dagger = |A|^\dagger$ and $\|\alpha\|^\dagger = |\alpha|^\dagger$.*

The full proof would be by induction on the complexity of A and α; the induction order laid down in Section 4 can again be used. The basic step is when A is a propositional letter; the valuation $V^\dagger$ was designed with that

step in mind. The inductive step is broken up into a number of cases, most of which are straightforward. Two typical examples:

$$\begin{aligned} \|\delta A\|^\dagger & \quad \text{(by the definition of intension)} \\ &= D\|A\|^\dagger \text{ (by the induction hypothesis)} \\ &= D|A|^\dagger \text{ (by Corollary 6.2)} \\ &= |\delta A|^\dagger. \end{aligned}$$

$$\begin{aligned} \|\alpha + \beta\|^\dagger & \quad \text{(by the definition of intension)} \\ &= \|\alpha\|^\dagger \oplus \|\beta\|^\dagger \text{ (by the induction hypothesis)} \\ &= |\alpha|^\dagger \oplus |\beta|^\dagger \text{ (by definition of } \oplus) \\ &= |\alpha + \beta|^\dagger. \end{aligned}$$

Rather than giving the proof in full we note that only two cases are problematic: formulae of type $[\alpha]A$ and formulae of type OK α. For those, three partial results are crucial:

(A) For all $\alpha\eta\Psi$, if $(x, y) \in R_L(\alpha)$ then $(\bar{x}, \bar{y}) \in R|\alpha|^\dagger$.

(B) For all α, if $(\bar{x}, \bar{y}) \in R|\alpha|^\dagger$ and $[\alpha]B \in x \cap \Psi$ then $B \in y$.

(C) If OK $\alpha \in \Psi$, then $\bar{x} \in N|\alpha|^\dagger$ if and only if OK $\alpha \in x$.

Given the proof in [4] it is enough to prove (A) and (B) in the special cases that α is of type $?A$ or δA. Thus we have our work cut out for us: proving those special cases of (A) and (B) (Lemma 7.3 and Lemma 7.4 below) and proving all of (C) (Lemma 7.5).

Much of the difficulty in proving completeness by the filtration method has attached to proving (A) for the case that α is of the form δA. In fact, it was in order to deal with this case that the path construction described in Section 5 was introduced. Before proving the desired result, there is still one technical lemma to prove. Let us say that a path p is an *initial in* σ if p is an initial of some path $q \in \sigma$.

LEMMA 7.2 *Suppose that* $\sigma \in \mathrm{PSAFE}(\bar{x})$ *and that* σ *produces* A, *for some* $A \in \mathrm{BC}\Psi$. *Let* p *be any initial in* σ *and* u *an element such that* $\bar{u} = p(\sharp)$. *(i)* OK $\delta A \in u$. *(ii) If* C *is a formula such that* OK $\delta C \in u \cap \Psi$ *and* $p^\frown\bar{v}$ *is an initial in* σ *whenever* $(\bar{u}, \bar{v}) \in R^\dagger(\delta C)$, *then* OK $\delta C; \delta A \in u$.

Proof. We prove this result by strong backward induction. Make the general assumption of the lemma. As the induction hypothesis suppose that the result holds for all paths in σ of which p is an initial. Two cases.

In the first case, p has no continuations in σ. Then $p \in \sigma$, and so $A \in u$, since σ produces A. By Lemma 3.3, $A \supset \text{OK}\ \delta A$ is a thesis of our logic L. Hence $\text{OK}\ \delta A \in u$. This confirms condition (i). For (ii), assume that, for some formula C,

(1) $\quad \text{OK}\ \delta C \in u \cap \Psi$,

(2) $\quad \forall v((\bar{u}, \bar{v}) \in R^{\dagger}(\delta C) \Longrightarrow p^{\frown}\bar{v}$ is an initial in $\sigma)$.

As p is without continuations in σ, (2) reduces in this case to the condition that $\forall v(\bar{u}, \bar{v}) \notin R^{\dagger}(\delta C)$, which in turn implies $\forall v(u, v) \in R_L(\delta C)$. However, together with (1) and (AX17) this yields a contradiction. The conclusion is that, in this particular case, condition (ii) holds trivially.

In the second case, p has at least one continuation in σ. Tackling (ii) first, again assume that there is some C such that

(1) $\quad \text{OK}\ \delta C \in u \cap \Psi$,

(2) $\quad \forall v((\bar{u}, \bar{v}) \in R^{\dagger}(\delta C) \Longrightarrow p^{\frown}\bar{v}$ is an initial in $\sigma)$.

Take any w such that $(u, w) \in R_L(\delta C)$. Then $(\bar{u}, \bar{w}) \in R^{\dagger}(\delta C)$. Hence by (2), $p^{\frown}\bar{w}$ is an initial in σ. By the induction hypothesis, then, $\text{OK}\ \delta A \in w$. This argument shows that

(3) $\quad [\delta C]\ \text{OK}\ \delta A \in u$.

By (1), (3) and (AX14), $\text{OK}\ \delta C; \delta A \in u$. This establishes part of the lemma. The remainder is established by (AX18) by which it follows that $\text{OK}\ \delta A \in u$. □

LEMMA 7.3 *(i) If* $?A\eta\Psi$, *then* $(x, y) \in R_L(?A)$ *only if* $(\bar{x}, \bar{y}) \in R|?A|^{\dagger}$. *(ii) If* $\delta A\eta\Psi$, *then* $(x, y) \in R_L(\delta A)$ *only if* $(\bar{x}, \bar{y}) \in R|\delta A|^{\dagger}$.

Proof. (i) Suppose that $?A\eta\Psi$. Conditions (FL1) and (FL6) guarantee that $A \in \Psi$ and *a fortiori* $A \in \text{BC}\Psi$. Assume that $(x, y) \in R_L(?A)$. By one half of (AX9), $x = y$ and $A \in x$. Hence $\bar{x} = \bar{y}$ and $A \in x \cap \Psi$. Consequently, $(\bar{x}, \bar{y}) \in R|?A|^{\dagger}$.

(ii) Suppose that $\delta A\eta\Psi$. By (FL2) and (FL7), $A \in \Psi$ and *a fortiori* $A \in \text{BC}\Psi$. Assume that $(x, y) \in R_L(\delta A)$. Note that $\sigma = \{(\bar{x}, \bar{z}) : (\bar{x}, \bar{z}) \in R^{\dagger}(\delta A)\}$ is a nontrivial element of $\text{PSAFE}(\bar{x})$ and that $(\bar{x}, \bar{y}) \in \mathit{rel}\ \sigma$. Let z be any element such that $(\bar{x}, \bar{z}) \in \mathit{rel}\ \sigma$. Then $(\bar{x}, \bar{z}) \in R^{\dagger}(\delta A)$. Hence there are $x' \equiv x$ and $z' \equiv z$ such that $(x', z') \in R_L(\delta A)$. By (AX10), $A \in z'$. Since $A \in \Psi$, $\bar{z} \in |A|^{\dagger}$. This means that σ produces A. Consequently, $(\bar{x}, \bar{y}) \in R|\delta A|^{\dagger}$. □

LEMMA 7.4 *(i) Suppose that* $(\bar{x}, \bar{y}) \in R|?A|^\dagger$, *for some* A. *Then* $[?A]B \in x \cap \Psi$ *only if* $B \in y$. *(ii) Suppose that* $(\bar{x}, \bar{y}) \in R|\delta A|^\dagger$, *for some* A. *Then* $[\delta A]B \in x \cap \Psi$ *only if* $B \in y$.

Proof. (i) Suppose that $(\bar{x}, \bar{y}) \in R|?A|^\dagger$. Then $\bar{x} = \bar{y}$ and $A \in x$. Furthermore, suppose that $[?A]B \in x \cap \Psi$. By the other half of (AX9), $B \in x$. By (FL0), $B \in \Psi$. Moreover, $x \equiv y$. Hence $B \in y$.

(ii) Suppose that $(\bar{x}, \bar{y}) \in R|\delta A|^\dagger$. Hence $A \in \mathrm{BC}\Psi$. By definition there is some set $\sigma \in \mathrm{PSAFE}(\bar{x})$ such that $(\bar{x}, \bar{y}) \in \mathit{rel}\ \sigma$ and σ produces A. Take $p \in \sigma$ such that $p =< \bar{z}_0, \ldots, \bar{z}_n >$ and $\bar{z}_0 = \bar{x}$ and $\bar{z}_n = \bar{y}$. Assume that $[\delta A]B \in x \cap \Psi$. We claim that

(§) $\quad [\delta A]B \in z_i$, for all $i \leq n$.

The proof is by induction on i. For $i = 0$ it is enough to refer to the assumption. For the inductive step, assume for a certain $i < n$ that

(1) $\quad [\delta A]B \in z_i$.

There is some C such that $(\bar{z}_i, \bar{z}_{i+1}) \in R^\dagger(\delta C)$. Hence there are u and v such that

(2) $\quad u \equiv z_i$,

(3) $\quad v \equiv z_{i+1}$,

(4) $\quad (u, v) \in R_L(\delta C)$.

From (1), (2) and the assumption that $[\delta A]B \in \Psi$ it follows that

(5) $\quad [\delta A]B \in u$.

Since $< \bar{z}_0, \ldots, \bar{z}_{i+1} >$ is also an initial in σ we may apply Lemma 7.2:

(6) $\quad \mathrm{OK}\ \delta C; \delta A \in u$.

Hence, by (5), (6) and (AX19), $[\delta C; \delta A]B \in u$. Trivially by (AX4), $[\delta C][\delta A]B \in u$, whence $[\delta A]B \in v$ by (4). Thanks to (3) we may now infer that $[\delta A]B \in z_{i+1}$, which is the desired conclusion and which ends the proof of (§).

It follows from (§) that $[\delta A]B \in y$. But we already know that $A \in y$. By (AX12), $B \in y$. □

For the remaining piece, the claim (C), we give the full proof.

LEMMA 7.5 *If* $\mathrm{OK}\ \alpha \in \Psi$, *then* $\bar{x} \in N|\alpha|^\dagger$ *if and only if* $\mathrm{OK}\ \alpha \in x$.

Proof. By induction on α. First suppose that OK $?A \in \Psi$. By (FL6), $A \in \Psi$. Hence $A \in \mathrm{BC}\Psi$, so $N|?A|^\dagger = |A|^\dagger$. Consequently, $\bar{x} \in N|?A|^\dagger$ iff $\bar{x} \in |A|^\dagger$ iff $A \in x$ iff, by (AX16), OK $?A \in x$.

Next suppose that OK $\delta A \in \Psi$. By (FL7), $A \in \Psi$. First suppose that $\bar{x} \in N|\delta A|^\dagger$. Then there is some y such that $(\bar{x}, \bar{y}) \in R|\delta A|^\dagger$. Consequently, there is some $\sigma \in \mathrm{PSAFE}(\bar{x})$ such that $(\bar{x}, \bar{y}) \in \mathit{rel}\ \sigma$ and σ produces A. By Lemma 7.2, OK $\delta A \in x$. Conversely, if OK $\delta A \in x$, then $< \delta A > \top \in x$, by (AX17). Hence $(x, y) \in R_L(\delta A)$, for some y. By Lemma 7.3, $(\bar{x}, \bar{y}) \in R|\delta A|^\dagger$. Therefore, $\bar{x} \in N|\delta A|^\dagger$. This completes the basic step.

Inductive step: assume that the lemma holds for α and for β. Three cases. First suppose that OK $\alpha + \beta \in \Psi$. By (FL8), OK $\alpha \in \Psi$ and OK $\beta \in \Psi$. Now, $\bar{x} \in N|\alpha + \beta|^\dagger$ iff (by definition) $\bar{x} \in N|\alpha|^\dagger \cap N|\beta|^\dagger$ iff (by the induction hypothesis) OK $\alpha \in x$ and OK $\beta \in x$ iff (by (AX13) OK $\alpha + \beta \in x$.

Next suppose that OK $\alpha; \beta \in \Psi$. By (FL9), OK $\alpha \in \Psi$ and $[\alpha]$OK $\beta \in \Psi$. Then $\bar{x} \in N|\alpha; \beta|^\dagger$ iff (by definition) $\bar{x} \in N|\alpha|^\dagger$ and $\forall y((\bar{x}, \bar{y}) \in R|\alpha|^\dagger \Longrightarrow \bar{y} \in N|\beta|^\dagger)$ iff (by the induction hypothesis) OK $\alpha \in x$ and $\forall y((\bar{x}, \bar{y}) \in R|\alpha|^\dagger \Longrightarrow$ OK $\beta \in y)$ iff OK $\alpha \wedge [\alpha]$OK $\beta \in x$ iff (by (AX14)) OK $\alpha; \beta \in x$.

Before we embark on the final leg of the proof, we remind the reader of a result in [4]:

($\P$) whenever $[\alpha^*]B \in \Psi$ and $\forall y((\bar{x}, \bar{y}) \in (R|\alpha|^\dagger)^* \Longrightarrow \bar{y} \in |B|^\dagger)$, then $[\alpha^*]B \in x$.

With this in mind, suppose that OK $\alpha^* \in \psi$. By (FL10), $[\alpha^*]$OK $\alpha \in \Psi$. We have $\bar{x} \in N|\alpha^*|^\dagger$ iff (by definition) $\forall y((\bar{x}, \bar{y}) \in (R|\alpha|^\dagger)^* \Longrightarrow \bar{y} \in N|\alpha|^\dagger)$ iff (by the induction hypothesis) $\forall y((\bar{x}, \bar{y}) \in (R|\alpha|^\dagger)^* \Longrightarrow$OK $\alpha \in y)$ iff (by ($\P$)) $[\alpha^*]$OK $\alpha \in x$ iff (by (AX15)) OK $\alpha^* \in x$. □

References

[1] FISCHER, M. J. and LADNER, R. E., *Propositional modal logic of programs*. Proceedings of the 9th Annual A. C. M. Symposium on Theory of Computing , pp. 286-294. Boulder, Colorado, 1977.

[2] GOLDBLATT, ROBERT. *Axiomatising the logic of computer programming*. Lecture Notes in Computer Science, vol. 130. Berlin, Heidelberg, New York: Springer-Verlag, 1982.

[3] GOLDBLATT, ROBERT. *Logics of time and computation*. CSLI Lecture Notes, vol. 7. Stanford University: 1987

[4] SEGERBERG, KRISTER. *A completeness theorem in the modal logic of programs*. In T. Traczyk (ed.), Universal algebra and applications, pp. 31-46. Banach Center Publications, vol. 9. Warsaw: P.W.N., 1982

[5] SEGERBERG, KRISTER. *Bringing it about*. Journal of philosophical logic, vol. 18 (1989), pp. 327-347.

[6] SEGERBERG, KRISTER. *Action incompleteness*. Studia logica, vol. 51 (1992), pp. 533-550.

[7] SURENDONK, TIMOTHY J. *Making maximal reliable action maximal*. Theoria, to appear.

[8] THOMASON, S. K. *Dynamic logic and the logic of ability*. In XIII incontro: Logiche modali e temporali, Siena, 28-29-30-31 Maggio 1989, pp. 15-35. Atti degli incontri di logica matematica, vol. 6. Siena and Padova: Associazione Italiana di Logica e sue Applicazioni, s.a.

Logic, Methodology and Philosophy of Science IX
D. Prawitz, B. Skyrms and D. Westerståhl (Editors)

RELIABLE METHODS

KEVIN T. KELLY

Department of Philosophy, Carnegie Mellon University

1. Method and reliability

At the beginning of his recent book, *The Pursuit of Truth*, W. V. O. Quine asks how it can be that science arrives at accurately predictive theories of the world, given the meager inputs it has to work with.

> From impacts on our sensory surfaces, we in our collective and cumulative creativity down the generations have projected our systematic theory of the external world. Our system is proving successful in predicting subsequent sensory input. How have we done it?[1]

What kind of answer would be appropriate? In the generous spirit peculiar to philosophy, Quine gratefully consigns the matter to others, including neurologists, psychologists, psycholinguists, evolutionary geneticists, and historians of science. He might have included sociologists, anthropologists, and archaeologists as well. These disciplines are indeed charged with discovering how the brain or the society is actually disposed to do what it does.

But would such a story answer Quine's original question? What if (contrary to fact) a careful sociological or psychological analysis were to reveal that social relations determine what science produces, pretty much without any regard to the data? That would in a sense explain *how* we produced *what* we produced, but it surely would not explain our *success*: i.e. *why* we arrived at a *successfully predictive theory*.[2] Similarly, if we stick a small physics book into a toaster and depress the handle, the toaster will "discover" that branch of science (in a slightly scorched condition) in just a minute. But this toaster would not slake Quine's legitimate sense of wonder about scientific success. For the toaster (like the scientific society just

[1](Quine 90), p. 1.
[2]Assuming we have succeeded.

described) is just "lucky" to have arrived at a correctly predictive theory (supposing the inserted physics text is such), for if the laws in the book had been very wrong, the toaster method would be very wrong, and would remain wrong, no matter how much more evidence we collect and stuff into its other bread-slot. It would never recover and stabilize to a correctly predictive theory, or even to a nearly correct one. The toaster *happens* to be right, but it wouldn't *ever* have been right had things been different from what they are, even if we were to give it all the time in the world to correct its conclusions.

The toaster response to Quine's question is unsatisfying because it appeals to luck, the poorest of explanations.[3] And if the scientific society turns out to be more like the toaster than not, even a complete sociological theory of science would not address Quine's question. What is required is a method that is in some sense *guaranteed* by its very nature to arrive at a truly predictive theory. For if we actually use such a method, then it is guaranteed that we arrive, eventually, at a correctly predictive theory. This explanation is aimed at *success* rather than at the production of some particular conjecture or other.

But if we avoid reliance on luck, we must not succumb to the other extreme of naive foundationalism, in which it is demanded that the method be guaranteed to succeed over all logical possibilities. Every guarantee of reliability will require *some* material assumptions. The very conception of inquiry as a process headed for the truth presupposes that there is time, that inputs are received by the scientist, and that methods can determine dispositions to produce outputs in response to arbitrary inputs. Even more assumptions, restricting the possible character and order of the data in the limit, may be necessary. But some methods are guaranteed to succeed under weaker material assumptions while other methods require stronger ones. The fewer the assumptions behind the guarantee, the better the response to Quine's question. To put the matter in slightly different language, a method guaranteed to arrive at a correct theory under weaker material conditions is **more reliable** than a method whose guarantee demands stronger material conditions.

While methodology can be applied to the explanation of past scientific successes, as Quine suggests, there is a more important task: namely, to recommend new and improved methods to those who want to succeed. So while Quine's question is aimed at the reliability of actual human inductive behavior, it cannot help but lead, upon reflection, to more general, more philosophical, and more *logical* questions such as: do there exist more reliable methods? Are those methods feasible? In what sense of feasibility?

[3]Despite the fact that it may be the only true explanation in some cases.

Could equally reliable methods converge to the truth as fast or faster? What are the minimal assumptions required for reliability concerning a given inductive problem? Are there complete architectures for induction so that every reliably solvable inductive problem is solvable by a method with such an architecture? Such questions are exactly the sort with which this paper is concerned.

Between naive foundationalism and free-wheeling naturalism lies a conception of scientific method that is naturalistic in its frank admission of material preconditions on reliability, but logical and normative rather than empirical in its analysis of methods and in its demand that the material preconditions be minimized. It is the consequences of this conception of methodology that I wish to develop in this paper.

2. Confirmation

The point of view just enunciated is by no means universal in the present day. Many contemporary methodologists, both in philosophy and in statistics, do not conceive of scientific method as a reliable process at all. They view it as a set of principles of **confirmation**, or **evidential support**. Confirmation theorists are champions of the *here-and-now*, not of the *would have been later*. The relevant scientific question, according to them, is not whether science would have stabilized to a correctly predictive theory, but whether our current views are actually confirmed by our actual data, period. Confirmation theorists have tried to prove that questions about the process by which a conjecture is generated are necessarily psychological or at any event irrelevant to legitimate scientific interests.[4] Regardless where a hypothesis has come from, we are told, all we need to do is to check whether it is confirmed now. Tomorrow is another day, which we may not live to see. Other possible worlds are worlds we do not live in. The scientist and the statistician must face his problems here and now, without delay. All of this has a nice, practical, down-to-Earth ring to it— until we ask what *problem* it is (other than the problem of stamping hypotheses "confirmed") that confirming hypotheses in the here-and-now *solves*, and *how* confirmation solves it.

There is no better indication of the prevalence of this sort of thinking than the very next paragraph of Quine's book, *The Pursuit of Truth*. Here, Quine attempts to change the subject from the *reliability* of man's process for producing beliefs to the *confirmation* of these beliefs on the basis of available evidence.

[4](Popper 68), (Hempel 65).

> Within this baffling tangle of relations between our sensory stimulation and our scientific theory of the world there is a segment that we can gratefully separate out and clarify without pursuing neurology, psychology, psycho-linguistics, genetics, or history. It is the part where theory is tested by prediction. It is the relation of evidential support, and its essentials can be schematized by means of little more than logical analysis.

One small slip for Quine. One giant leap for methodology. The leap is a giant one because confirmation theorists tend either to ignore, to resent, or to evade the question of how confirmation contributes to our pursuit of truth (or at least of correctly predictive theories). For example, the question is just ignored by proponents of hypothetico-deductivism, inference to the best explanation, instance confirmation, consilience, simplicity, and common causes. Some confirmation theorists have at least called for such a connection without providing one.[5] Instead, they either argue (sociologically) that their theories are good sociological generalizations of historical cases, or (inductively) that such methods succeeded when used in the past, and hence ought to continue to succeed. Those who resent the question claim that everybody has first principles, and canons of confirmation are *our* first principles, so that questions connecting confirmation to reliability are ill-posed demands for external justifications of ultimate principles of justification.[6] Perhaps the most interesting strategy is to exchange the question for another. Many Bayesians, for example, replace questions of reliability in the limit with questions about rational preference orderings over acts in the short run. At least they have a clear, alternative view about what confirmation is *for*. But what it is for is frankly conceded to have nothing to do with whether or not a highly confirmed hypothesis is true.

None of these strategies addresses Quine's intriguing question concerning the prospects for reliable inquiry. This is not, however, to say that they are not answers to other, well-motivated questions. The role of institutional

[5] "There is nothing in this book that corresponds to an attempt to show that the methods I have described are justified or uniquely rational. There are arguments for the methods, arguments that purport to show that the strategy achieves our intuitive demands on confirmation relations better than do competing strategies, but these arguments do not show that the bootstrap strategy will lead us to the truth in the short run or the long run, or will lead us to the truth if anything can, or is required by some more primitive canon of rationality. There are such arguments for other confirmation theories, although none of them are wholly good arguments; perhaps it would be better to have a bad argument of one of these kinds for the bootstrap strategy than to have none at all." (Glymour 81), p. 377. Glymour has since taken his own advice: c.f. (Kelly and Glymour 89), (Kelly and Glymour 1990).

[6] (Horwich 91).

relations in the actual functioning of the actual scientific society is an interesting sociological topic in its own right. Inductive arguments of future success may be in some sense reassuring, if the circle is not too tight. Everyone is indeed entitled to his own first principles, so long as they engender a fruitful theoretical progeny. The rationality and representation of preference is a rich and interesting field. My point is only that reliability is one interesting issue among many, an issue of an especially logical and normative character.

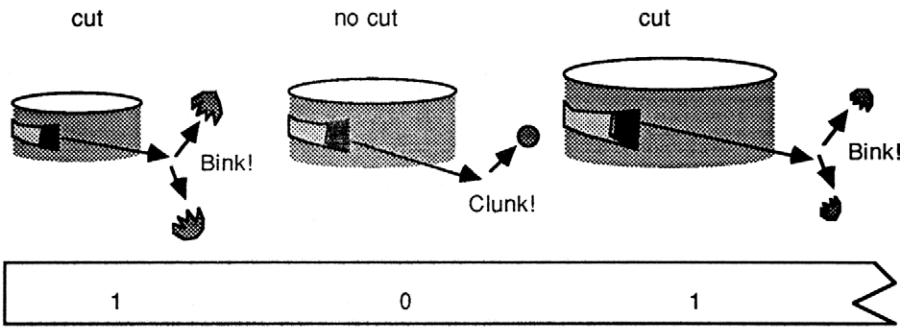
My approach to the logic of reliability draws from **computational learning theory**, a body of work more familiar to computer scientists than to philosophers. The name may have something to do with that fact. It sounds like an empirical study of how people learn; the sort of thing Quine has in mind under the rubric of **naturalized epistemology**. But in fact, it is a logical, *a prioristic* framework for addressing just the sorts of questions about the prospects of reliability that I have portrayed as arising out of Quine's question. Reliability and computability are primary concerns rather than mere afterthoughts, and thus the methodological pie is cut a bit differently by learning theorists than by philosophers or statisticians, as will be apparent.

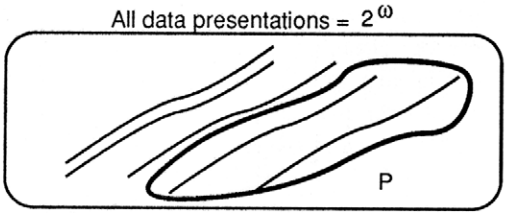
Learning theory may be traced to philosophical work by Kemeny[7] and Reichenbach.[8] It was developed by Putnam in a response to Carnap's confirmation theory[9] and in independent work in logical theory.[10] The ideas gained popularity among cognitive scientists in the work of the computer scientist E. M. Gold,[11] who applied it to the analysis of learnability in Chomsky's linguistic program. From there it was adopted by recursion theorists interested in artificial intelligence.[12] It has since developed into a recognized sub-discipline of computer science.[13]

3. Reliable hypothesis assessment

Imagine a debate between Descartes, Newton and Kant concerning the infinite divisibility of matter. Descartes insists that material substance is identical with geometrical extension, and hence is infinitely divisible. Newton replies that matter is composed of ultimately indivisible atoms floating

[7](Kemeny 53).
[8](Kelly 91b).
[9](Putnam 63).
[10](Putnam 65).
[11](Gold 67), (Gold 65).
[12](Angluin and Smith 82).
[13]For a sample of recent work, c.f. (Rivest *et. al.*, 89).





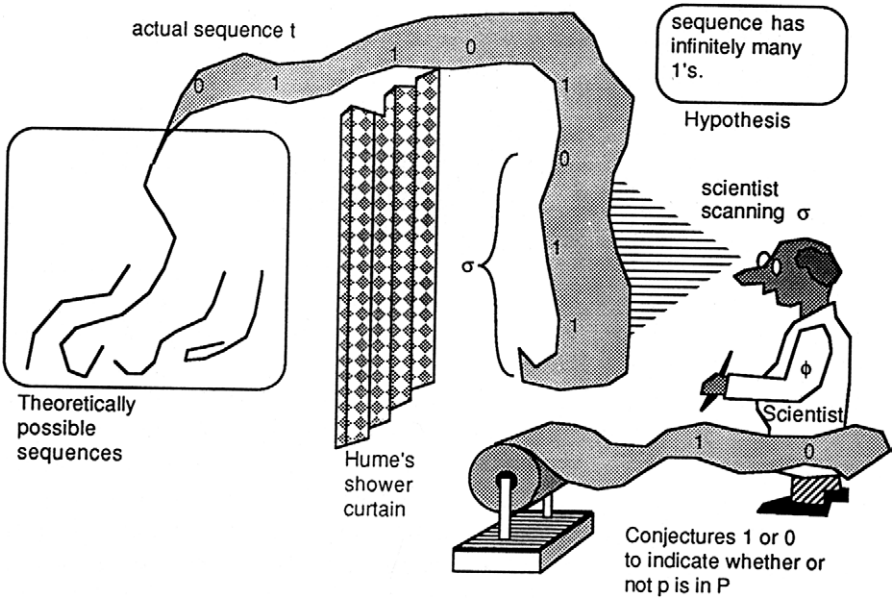
in a void. Kant dismisses the proceeding as nonsense, since the question lies beyond the scope of all possible experience.

Let's consider an intrepid scientist who overhears the discussion by these luminaries, and who decides to undertake a forthright, empirical approach to the question. Our scientist is determined to pursue the following procedure for eternity (a foundation is set up to survive him). He has a potentially infinite sequence of designs for ever more powerful cutting devices, ranging from knives to razors to radiation beams to particle accelerators of ever greater extent and energy. He starts out with a small bit of matter. At each stage, he takes the smallest bit of matter resulting from his previous attempt (whether successful or not) and attempts to cut it with the next more powerful device. When the cut succeeds he writes down 1 and when it fails he writes down 0. The result is an infinite tape of 1's and 0's, which we will take as his data.

Assuming that if a particle is divisible then some cutting device of sufficient energy will eventually cut it, we may think of the hypothesis that matter is infinitely divisible as the set of all infinite data sequences in which infinitely

many 1's (i.e. cut indicators) occur.

The scientist ϕ's situation is this. He sees ever larger initial segments of the infinite data presentation t that arises in the limit when his experimental protocol is followed in the manner described. t is only one of the (infinitely many) possible data presentations that might arise from the experimental set-up, for all the scientist knows. "Hume's shower curtain" prevents him from seeing the full, infinite, extent of t, or the hidden workings of the world that produce it. For each finite, initial segment σ of t, the scientist ϕ produces a conjecture as to the truth value of the hypothesis in question, where 1 indicates truth and 0 falsity.



This simple picture frames the ontology of the scientist's position. Now we move to the normative concept of reliable success. In logic and in computation theory, one standard of success is **decidability**. A machine M decides some set S of objects over some domain just in case the machine returns 1 on each domain element in S and 0 for each domain element not in S. Decidability has long been taken to have epistemic significance for a logical system, for it implies that eventually one will have the right answer.

Decidability does not imply any fixed time by which the correct answer will come in. But on the negative side, if no procedure can decide a problem, no procedure can decide it quickly enough to suit us, either.

We can also think of a scientist ϕ as deciding an empirical hypothesis P on the basis of some data stream t, where P is some subset of a collection K of infinite data presentations representing the prior background knowledge of the scientist. When the scientist is guaranteed to stop with the right answer after some time, we may think of him as eventually "knowing that he knows", and we say that he decides the hypothesis *with certainty*.

> Scientist ϕ **decides with certainty** hypothesis P assuming knowledge $K \Leftrightarrow$ for each possible $t \in K$, ϕ eventually stops reading t and produces 1 if $t \in P$, and eventually stops reading t and produces 0 otherwise.

> P is **decidable with certainty** $\Leftrightarrow$ there is a scientist ϕ that decides P with certainty assuming knowledge K.

Just as in the case of a decidable proof system, we don't know ahead of time *when* ϕ will find the truth, but ϕ is at least guaranteed to find the truth and to know when he has. This is arguably the notion of inductive success operative in Plato's discussion of the Meno paradox.[14]

We cannot always expect logical systems to be decidable by a given type of machine. In particular, the validity of first-order logic is not decidable by any Turing machine, a fact of importance in the epistemology of mathematics. A formal set S is **verifiable** by machine M just when there is a machine M so that M outputs 1 on input x if and only if $x \in D$. It is permitted for M to run on forever when $x \notin D$. Sets verifiable by Turing machines are said to be **recursively enumerable**, and the import of the familiar completeness theorem of first-order logic is that first-order logical validity is a recursively enumerable (or Turing-machine verifiable) relation. Dually, a formal set may be **refutable**, in the sense that some machine returns 0 on input D if and only if $x \in D$.

The analogous notions of hypotheses **verifiable** or **refutable with certainty** are familiar from the philosophy of science. Notice, however, that we define these notions, as computer scientists do, in terms of the existence of reliable methods rather than in terms of the logical form of the hypothesis. Thus, unlike the positivists, who identified verifiable hypotheses with existentially quantified hypotheses, we are committed to no such thing; for if the data is not true and complete, then there may be no scientist who can verify an existential hypothesis with certainty.

[14](Kelly and Glymour 91).

Scientist ϕ **verifies** hypothesis P **with certainty** given knowledge $K \Leftrightarrow$ for each possible $t \in K$, ϕ eventually stops reading t and produces 1 if and only if $t \in P$.

P is **verifiable with certainty** given knowledge $K \Leftrightarrow$ there is a scientist ϕ that verifies P with certainty given knowledge K.

Scientist ϕ **refutes** hypothesis P with **certainty** given knowledge $K \Leftrightarrow$ for each possible $t \in K$, ϕ eventually stops reading t and produces 1 if and only if $t \in P$.

P is **refutable with certainty** given knowledge $K \Leftrightarrow$ there is a scientist ϕ that refutes P with certainty given knowledge K.

The similarity between these criteria of scientific success and the corresponding criteria of computational success is evident. One such similarity is that a hypothesis is decidable with certainty if and only if it is both verifiable and refutable with certainty, as is readily seen. One important disanalogy between inquiry and computation is that formal decidability becomes trivial when computation is read so liberally that arbitrary functions are computable. This is not the case for empirical decidability, however, so the study of scientists of unbounded computational power is not trivial.

Empirical decidability with certainty, like decidability of a formal system, is a desirable goal, but it is rarely obtainable. More leniently, we may demand that the scientist stabilize to the truth without knowing for sure when he has done so.

Scientist ϕ **decides** hypothesis P **in the limit** given knowledge $K \Leftrightarrow$ for each possible $t \in K$, there is a time after which ϕ always outputs the correct truth value of P.

P is **decidable in the limit** given knowledge $K \Leftrightarrow$ there is a scientist ϕ that decides P given knowledge K.

As it turns out, the following, "one-sided" senses of limiting success are easier to analyze.

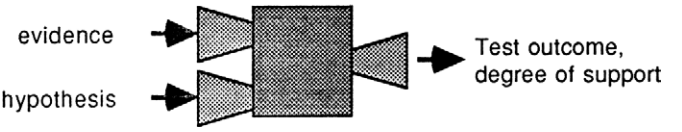
Scientist ϕ **verifies [refutes]** hypothesis P **in the limit** given knowledge $K \Leftrightarrow$ for each possible $t \in K$, $t \in P$ $[t \notin P]$ if and only if after some time, ϕ always outputs 1 [0].

P is **empirically verifiable [refutable] in the limit** given knowledge $K \Leftrightarrow$ there is a scientist ϕ that verifies [refutes] P in the limit given knowledge K.

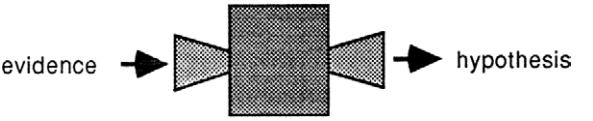
Decidability in the limit can then be studied as the conjunction of verifiability and refutability in the limit.

4. The logic of discovery

So far, we have mentioned only problems of hypothesis **assessment**, in which the scientist is assigned some hypothesis P to investigate. Such problems will be referred to as *assessment problems*.



From a broader perspective, the scientist is not interested in finding the truth value of a particular hypothesis, for if the correct value is 0, the scientist may not be able to predict anything. Rather, the scientist receives evidence, and his job is to stabilize to some "adequate" hypothesis, where adequacy is some desirable relation R holding between infinite data presentations and hypotheses. R may be taken to imply truth, verisimilitude, sufficient logical strength, informativeness, explanatory power, simplicity, or any of the other usual features often listed as theoretical virtues. Such problems will be referred to as **discovery problems**.



Confirmation theorists like to draw a sharp distinction between the problems of discovery and assessment. Assessment is a matter of *justifying* hypotheses according to *logical* standards of confirmation. On this view, the justificatory relation of confirmation holds or fails to hold as a matter of logic, quite independently of the psychological and social currents and eddies that brought the hypothesis in question to attention.[15] Thus discovery

[15](Laudan 80).

is alleged to be a subject for psychologists, rather than for logicians. Perhaps Quine's penchant to toss the question of reliability to psychologists is a tacit throwback to this well-worn positivistic doctrine.

The tendency to disparage the logic of discovery is not surprising among confirmation theorists, since confirmation theories address the problem of assessment, rather than that of discovery; assuming that they address any coherent *problem* at all. But from the point of view of methodological reliability, there is a seamless analogy between discovery and assessment. In either case, the question is reliability. Once again, we can define success so that the scientist knows when he has it right:

> ϕ **identifies** an R-adequate hypothesis **with certainty** given K $\Leftrightarrow$ for each $t \in K$ there is a hypothesis h correct for t such that after some time, ϕ outputs h and stops.

> An R-adequate hypothesis is **identifiable with certainty** given K $\Leftrightarrow$ some ϕ identifies an R-adequate hypothesis with certainty given K.

Or we can define success in such a way that the scientist eventually stabilizes to some particular adequate hypothesis without knowing when.

> ϕ **identifies** an R-adequate hypothesis **in the limit** given K $\Leftrightarrow$ for each $t \in K$ there is a hypothesis h correct for t such that after some time, ϕ outputs only h.

> An R-adequate hypothesis is **identifiable in the limit** given K $\Leftrightarrow$ some ϕ identifies an R-adequate hypothesis in the limit given K.

Now we have various notions of success, together with a rudimentary model of how data arrives to the scientist. Any specification of these matters will be referred to as an inductive **paradigm**[16] or **setting**. A given setting admits many different inductive **problems**. A **discovery problem** is given by a specification of the adequacy relation R and the background knowledge K. An **assessment problem** is given by a fixed hypothesis P to investigate, together with some choice of K.

[16]This term is due to (Osherson *et. al.* 86).

5. Four levels of reliabilist methodology

From this general perspective on assessment and discovery, there are three fundamental levels at which methodological questions may be posed. At level (0) we have questions about success or failure of *particular* methods in a few *particular* possible worlds.

(0): Does method ϕ succeed on data presentation t?

This level corresponds to the usual sort of "historical case analysis" that has been popular in recent decades in the philosophy of science. Indeed, P. K. Feyerabend's methodological case for "anything goes" rests on the argument that a few proposed methodological principles have been disobeyed with apparently successful results.[17]. In artificial intelligence it was recently common to see proposals for "learning machines" recommended on the basis of success on one or two standard "test cases". From a reliabilist perspective, however, such complaints and recommendations are equally faint. An optimally reliable method can make a mistake; an unreliable method like the toaster described above can succeed on a few, judiciously chosen examples. Neither result tells us much about the overall reliability of the method.

At level (1), we consider questions specific to a given *method*, so that the only relevant quantifier ranges over the possible data presentations in K. Such questions include

(1): How reliable is method ϕ? Is method ϕ more reliable than method ψ? Does knowledge K entail that ϕ is reliable?

The second level of generality focuses on inductive *problems* rather than on inductive methods, and quantifies over methods. Such questions include:

(2): Is there a reliable method given knowledge K? What kind of knowledge K is minimally necessary for reliability concerning P? What sense of convergence can reliably be achieved given only knowledge K? Is some ϕ optimally reliable?

Optimal reliability is defined relative to weak dominance, so that no method succeeds wherever an optimal method succeeds and somewhere else as well.

Third, we come to the very general sorts of questions that involve quantification over problems in a *paradigm*.

[17](Feyerabend 79).

(3): Is there a **complete architecture** for the paradigm? Is there a structural characterization for problem solvability in the paradigm? Are two given paradigms **equivalent**, or is one more **stringent** than the other?

An **inductive architecture**[18] is a special way of constructing or presenting inductive methods. An architecture is **complete** for a given paradigm just in case every problem in the paradigm has a solution of the required form. One frequently sees proposed architectures for discovery in which discovery procedures are to be built out of some sort of test connected to some sort of search. For example, Popper proposes the architecture of conjectures and refutations as the best possible (and only) architecture for discovery. Is it complete? If not, is there a non-trivial, alternative architecture that is? We will return to these questions later.

A **characterization** of solvability is a structural relation between K and P (in the case of discovery problems, between K and R) that holds whenever a reliable solution to the problem exists. Of course we want to avoid triviality, so the structural relation between K and P should not be defined in terms of scientists or reliability, but in terms of the respective mathematical structures of K and of P. Thus a characterization of this sort may be thought of as a reliabilist version of a **complete transcendental deduction**. Recall that for Kant, a transcendental deduction shows some condition to be necessary for knowledge. A complete transcendental deduction would provide necessary and sufficient conditions for knowledge. If knowledge is associated with stability and reliability, then a structural characterization of solvability fits this bill; but unlike Kant's transcendental deductions, these depend upon the operative definition of scientific success rather than upon the synthetic structure of human cognition. Thus, unlike Kant's transcendental program, this one makes mundanely clear both where transcendental deductions come from and how they can be genuine deductions.

Paradigms are **equivalent** when the same problems are solvable in each, whereas one is more **stringent** than another when the problems solvable in it are a proper subset of those solvable in the other. When one paradigm seems odd and another seems natural, it is interesting to discover that they are in fact equivalent. Equivalence implies that in some rarefied methodological coin, the two distinct standards of success are equally valuable. Stringency is interesting because when our pet problems are unsolvable in one paradigm, it is open to us to consider less stringent ones in which it is solvable.

Nothing prevents methodological questions of higher types, that quantify over paradigms, classes of paradigms, and so forth. But the questions I have

[18]Osherson, Stob and Weinstein (1986) use the term **strategy**.

found most engaging occur in levels (1) through (3) and particularly at level (3).

6. Methods and problems

Let us consider some examples of questions at levels (1) - (3), starting with the first. Recall our example problem, concerning the assessment of the hypothesis P that matter is infinitely divisible. Consider the trivial scientist ϕ who simply repeats the last entry on the tape. It is easy to see that ϕ refutes P in the limit when K is the set of all infinite 0-1 sequences. Moreover, ϕ is computable by a two-state finite state automaton, so refutation in the limit is especially easy in this case.

Next, consider the question whether P is decidable in the limit given K. This depends entirely on whether P is verifiable in the limit over K, since refutability in the limit has already been settled. The answer is negative; the proof being a simple *diagonal argument*. Let ϕ be an arbitrary scientist who hopes to succeed. Now consider a demon, who fools the scientist by feeding a 1, 0 and then all 0's while the scientist says 1, and by feeding 1 while the scientist says 0. Either ϕ changes his mind infinitely often or not. If so, then since the devil tosses in the pair 1, 0 infinitely often, P is true so ϕ fails, contradiction. If not, then ϕ stabilizes to the wrong answer, contradiction. Thus no ϕ succeeds, by reductio argument. So we have determined that P is refutable but not verifiable in the limit and hence not decidable in the limit. This amounts to a fairly tight classification of the intrinsic difficulty of the problem, since a trivial finite state machine succeeds at refutation in the limit, but cognitive gods cannot succeed at verification in the limit.

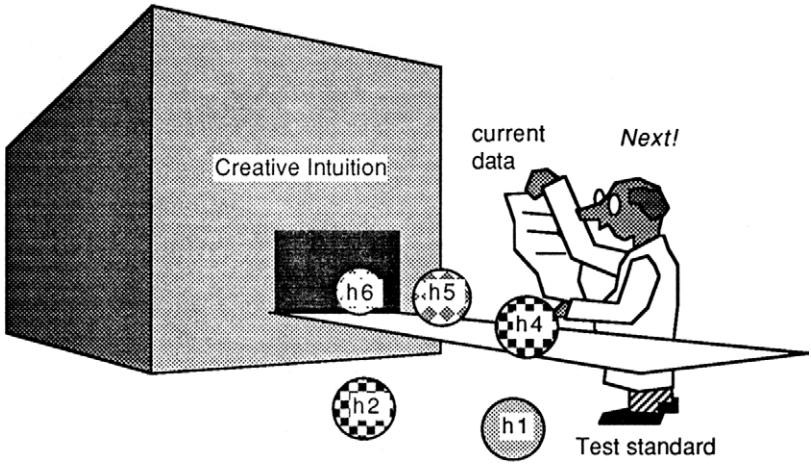
7. A complete architecture for discovery

Let's leap in abstraction to level (3) and consider the issue of complete architectures for discovery. Philosophers tend to analyze scientific inquiry into test or confirmation on the one hand and an unregulated module of invention or intuition on the other.

Popper has alleged that the method of bold conjectures and refutations is our best and only tool for grasping the truth[19]. Observe that the "method" of conjectures and refutations is not really a single method; it is a restriction on how to construct a method for a given problem, or, alternatively, the set of specific methods that are constructed in this way; that is, an *inductive architecture*.

[19](Popper 68).

So what does the *architecture* of conjectures and refutations amount to? If the "generate" part of such a method takes in data, then it is an arbitrary discovery method in its own right. But then generate and test architecture is trivially complete, for every discovery method is duplicated in reliability by itself with a test that passes anything tacked on the end. Presumably, a less trivial architecture is intended. Popper's idea seems to be, indeed, that the generator ranks hypotheses by audacity and simplicity *without looking at the data.* Then a consistency (refutation) test that does use the data is applied, and the current output of the generator is replaced by the next one when (and only when) the current one fails the consistency test. The current, non-discarded output of the generator is then conjectured until it is refuted in turn. The order of production by creative intuition may be assumed to reflect simplicity, power, or other arbitrary preferences[20]. Such procedures have also been recommended by Kemeny, Putnam, and Gold, among many others.



So we take Popper to claim that this non-trivially restrictive architecture of conjectures and refutations is complete, or at least that no other architecture is *more* complete than it is[21]. But both claims are false. For suppose

[20]This proposal was made by Kemeny (1953) and was investigated further by Putnam (1963).

[21]This is shorthand for saying that no other architecture has methods that solve all problems solved by conjectures and refutations methods, plus some more that are not.

K says that either a finite set of even numbers occurs in the data, or all natural numbers will come in the data. Suppose that each set has some natural number as an index, and an index is adequate for t just in case it indexes the set of numbers occurring in t. Now there are infinitely many finite sets of even numbers. But no index of such a set is adequate when the truth is that all numbers will be seen. So the generator must generate the hypothesis h that all numbers will be seen at some finite time, else the method will be wrong when this hypothesis is true. But wherever h is put in the enumeration of hypothesis indices, it must precede all but finitely many of the indices for finite, even sets. And since h cannot be refuted by observing even numbers, the conjectures and refutations method that uses the enumeration in question converges incorrectly to h when the truth is that only some finite set of even numbers occur in the data. But since the enumeration produced by "creative intuition" is arbitrary, no conjectures and refutations method identifies an adequate hypothesis over all of K.

Could any *method* succeed where Popper's *architecture* fails? Just conjecture the index for the current data until an odd number is seen, and then conjecture h. Generate-and-test architecture prevents us from using this method, or any method that works. But this fact does not deflate Popper's proposal unless we come up with an alternative, non-trivial *architecture* that shows us how to assemble tests and enumerations in a more general, and more complete manner.

There is such an architecture, which, moreover, is *demonstrably complete*. We may refer to it as **priority architecture**. A priority discovery method is factorable into a fixed enumeration and test procedure, but now instead of a refutation test, we think of the test as verifying each hypothesis in the limit. Now it will not do to throw away a hypothesis forever when the test says 0, as Popper recommends, for the test may change its mind back to 1 later. Instead, we employ an infinitely repetitive enumeration of hypotheses and initialize a pointer at the beginning of the enumeration. On evidence σ, we test the hypothesis currently pointed to on successive, initial segments of σ, moving the pointer one step to the right each time the hypothesis pointed to fails the test. The hypothesis pointed to when all of σ is read is then conjectured.[22]. This architecture is demonstrably complete,[23] and thus handles the example which has just been seen to defeat the architecture of conjectures and refutations. We conclude that conjectures and refutations is neither our best nor our only nor even a very good discovery architecture so far as reliability and completeness are concerned. Discovery architec-

[22]For a similar method applied to the assessment of first-order hypotheses, c.f. (Osherson, *et. al.* 91)

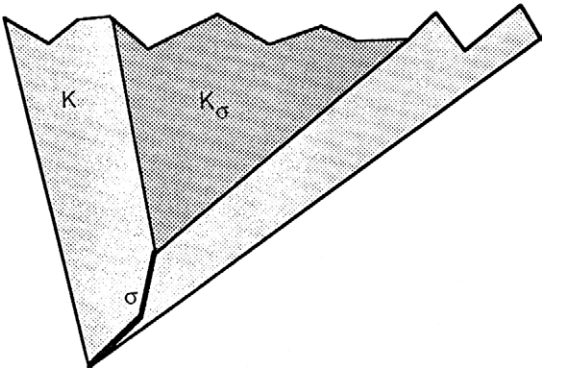
[23](Kelly 91).

tures must allow for **non-monotone** tests[24], which can be more reliable than **monotone** tests, and such tests require special handling of rejected conjectures in light of their ability to vacillate forever if the hypothesis is false.

8. Topological characterizations of hypothesis assessment

Let us move now to the subject of characterization theorems. We will come up with a way of building hypotheses out of data presentations in background knowledge K so that the hypothesis is decidable or verifiable in the limit or what not just in case it is built up in the appropriate way. Thus, the theorems will equate some notion of reliable success on the left-hand side with some bound on the complexity of the operation of building up the hypothesis from simple building blocks on the right-hand side. The appropriate building blocks and the operations of formation will be different, but highly analogous in the alternative cases of ideal (unrestricted) and computable scientific methods, thus yielding a smooth transition between ideal and computationally bounded epistemology. The general approach discussed here follows that of Gold (1965) and Putnam (1965).

The basic hypothesis building block will be called a K-**fan**. The K-fan with **handle** s is the set of all data presentations in K that extend a finite data segment σ, and will be denoted by K_σ.

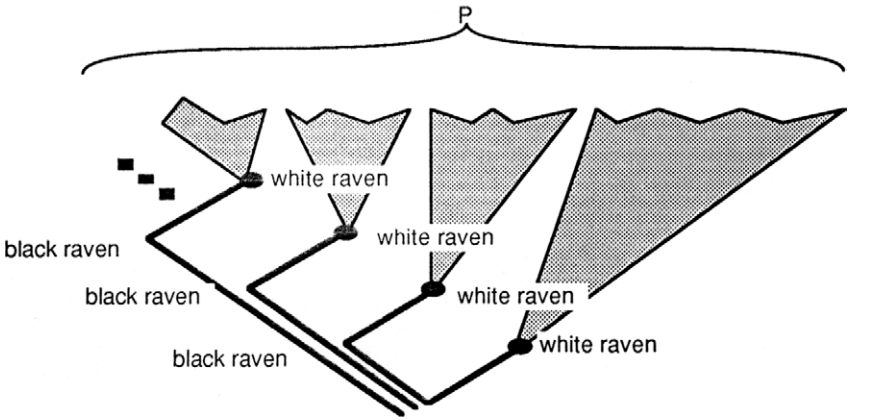


Since K_σ is a set of infinite data presentations, we may think of K_σ as a hypothesis, in just the manner that the hypothesis of infinite divisibility was

[24]A **non-monotone** test is one that does not verify or refute with certainty.

represented by a set of infinite data presentations in the infinite divisibility example above. K_σ is clearly empirically decidable, for once we see s in the data we know that K_σ is correct, and once some deviation from s is seen, we know that K_σ is incorrect. But this is not the only way to build an empirically decidable hypothesis. It is just the most elementary way.

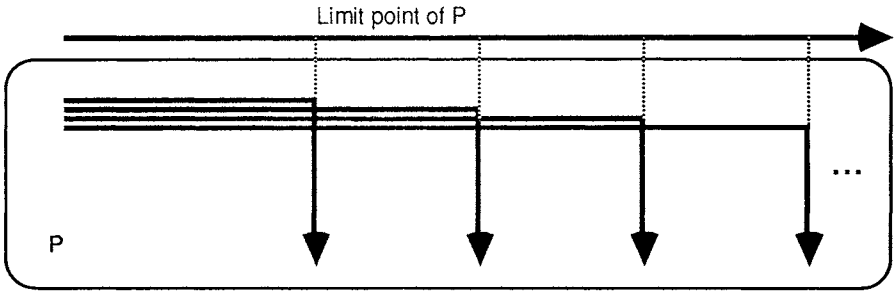
To build an empirical hypothesis that is verifiable with certainty, we simply form some arbitrary (and thus countable) union of K-fans.



In the diagram we may think of the endpoint of the handle of each fan as something the hypothesis says we will find. Thus, the hypothesis "some raven is white" is verified when and only when a white raven is observed. On the other hand, we can never be sure that no white raven will be seen. Such hypothesis will be said to be *K-open*. *K-closed* hypotheses are complements of K-open hypotheses, and may be thought of as the empirically refutable or "universal" hypotheses. In the diagram, the singleton consisting of the sequence that runs off to the left, in which only black ravens are seen, corresponds to the refutable hypothesis "all ravens are black". K-clopen hypotheses are both K-open and K-closed, and are exactly the hypotheses that are empirically decidable over K.

The topological terminology is not accidental, for K-fans together constitute a countable basis for the topology on K whose open sets are the K-open sets. When $K = \omega^\omega$, the resulting topological space is called the **Baire space**. When $K \subset \omega^\omega$, we have the Baire space **restricted to** K. From this topological perspective, the data presentation in which only black ravens occur is a topological **limit point** of the hypothesis "some raven is

white". This may sound strange, unless we recall that points in Baire space are actually infinite sequences, and the limit point of a set in Baire space is a sequence, each initial sequence of which is in the set.



A standard version of the problem of induction arises when no finite data sequence secures the truth of the hypothesis in question. This version of the problem of induction can now be viewed as a *topological property* of hypotheses relative to background knowledge. It arises when some data presentation in the hypothesis is a limit point of the complement of the hypothesis in K, or when some data presentation in the complement of the hypothesis is a limit point of $K-P$. An important role of measure theory in mathematics is to smooth out problematic sets by discarding troublesome points (like those that give rise to the problem of induction) in a set of measure 0. From this point of view, topology, rather than probability theory, is the appropriate setting for the analysis of the traditional problem of induction.

A familiar practice in the branch of mathematics known as **descriptive set theory**[25] is to build hierarchies of objects in which to classify their complexity. The *finite Borel hierarchy relative to* K may be defined inductively as follows:

$$\Sigma_1^{B,K} = \text{the } K\text{-open sets.}$$

$$\Pi_1^{B,K} = \text{the } K\text{-closed sets.}$$

Now, for each ordinal $n > 1$, define

$$\Sigma_n^{B,K} = \text{the set of all countable unions of elements of } \Pi_{n-1}^{B,K}.$$

$$\Pi_n^{B,K} = \text{the set of all countable intersections of elements of } \Sigma_{n-1}^{B,K}.$$

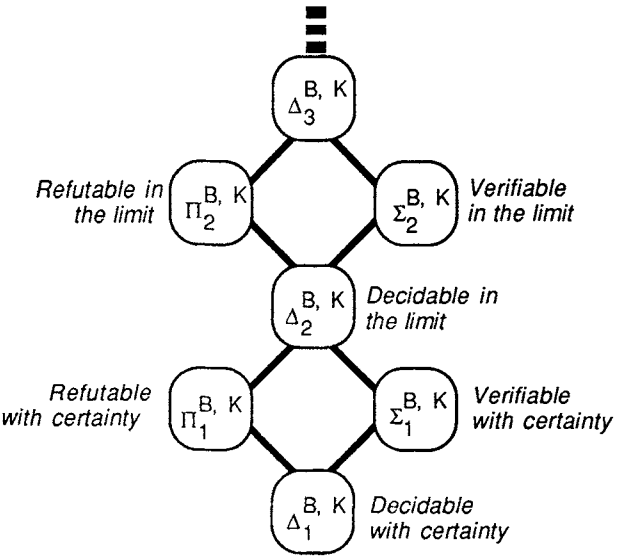
[25] (Moschovakis 80).

And for each $n \geq 1$, define

$$\Delta_n^{B,K} = \Sigma_n^{B,K} \bigcap \Pi_n^{B,K}.$$

Of particular interest to us are the complexity classes $\Pi_2^{B,K}$, $\Sigma_2^{B,K}$ and $\Delta_2^{B,K}$. These complexity classes characterize verifiability in the limit given K and refutability in the limit given K, respectively when the scientist is allowed to be an ideal agent.[26] $\Delta_2^{B,K}$ characterizes decidability in the limit. I summarize in the following diagram the characterizations of the various senses of success defined above:

THEOREM 1.



Thus we see in a simple paradigm how reliabilistic motives can yield a systematic, "ship-shape" epistemology in which there is a place for everything and everything can with mathematical certainty be put into its place. Analogous results can be given for the case of decidability with n mind-changes, which requires that the method ϕ change its mind at most n times before stabilizing to the truth in the limit.[27]

[26](Kelly 91). The results are relativized, topological, functional versions of results presented independently in (Putnam 65) and (Gold 65).
[27](Kelly 91)

The proofs of the characterization theorems all involve the construction of complete architectures for hypothesis assessment of the relevant sort and may be thought of as inductive completeness theorems for these architectures.

9. Turing computable inquiry

The learning-theoretic perspective on induction comes into its own when we turn to the inductive powers of computational agents. Bayesian confirmation theorists have theoretical problems with the very idea of using a computer, since the output of a program on a given input is a fact of arithmetic, and an ideally coherent Bayesian should have probability 1 on all such truths to begin with. One response is to demand coherence only over some simple, problem-specific meta-language over possible outputs of computers.[28] The simpler the meta-language that the Bayesian is driven to in order to preserve coherence in the face of computational limitations, the more trivial the constraints imposed by coherence become, unless auxiliary principles of direct inference conditional on logical relations are imposed.[29] But even if the meta-language is extremely simple, computational complexity can blow up in the number of atomic statements in the meta-language.[30] In sharp contrast, the reliabilist approach to induction outlined in this paper starts out with notions of scientific success explicitly analogous to the standards of success familiar in proof theory and in the theory of computation. This results in a seamless analogy between ideal and the computationally bounded norms rather than in a radical breach to be bridged by appeal to arbitrary and artificial meta-languages.

The learning-theoretic approach also develops a deep analogy between induction and computation. We may think of the Turing read-write head as a "stupid little scientist", and of the Turing tape as a "data presentation" produced by "formal experiments" carried out on the tape by the head. In a sense, the localized vision of the read-write head leads to the computational bounds posited by Church's thesis just as the localized sensory apparatus of the scientist leads to the bounds on reliability that raise Hume's problem.[31] So from the logical, reliabilist perspective sketched here, Church's thesis and Hume's problem are intimately related. That two such fundamental epistemological principles, one from the philosophy of mathematics and the other

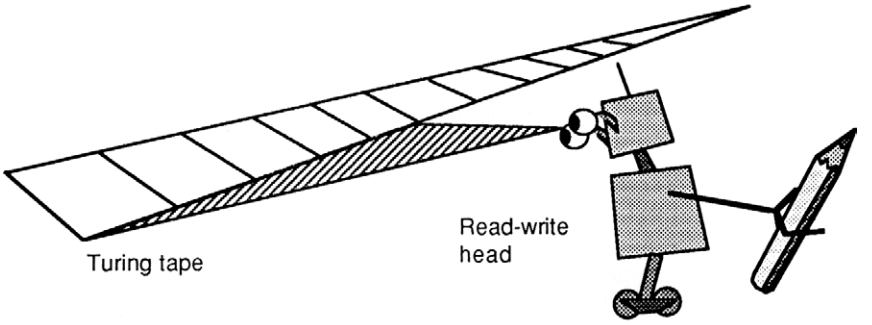
[28](Garber 83).

[29]e.g. Garber imposes the requirement that $P(A \mid B \,\&B \models A) = 1$.

[30]Deciding consistency in propositional logic is an NP-complete problem.

[31]For a rigorous characterization of Turing computability in terms of causal locality, c.f. (Gandy 80).

from the philosophy of science, should share a common root and structure from the reliabilist point of view is itself a fascinating philosophical result.



The analogy between Church's thesis and Hume's problem is all the more interesting when the two problems are rolled into one by the consideration of computable scientists. The characterization theorems look *exactly the same*, except that the relevant notion of complexity is given by the **arithmetical** or **Kleene hierarchy** rather than by the finite Borel hierarchy. The difference between the two hierarchies is that the former starts out with recursively enumerable sets rather than with open sets.[32] If one looks inside the characterization theorem, one sees that computable scientists turn computationally hard formal problems (such as the consistency of hypotheses with the data) into new, "internal" inductive inference problems studied in parallel with the external data presentation. Exactly the same sorts of methodological considerations arise for the internal, formal inductive problem and for the external, empirical inductive problem.[33]

The logical analysis of computational science raises an interesting range of questions that have been examined in detail by Osherson, Stob and Weinstein.[34] It turns out that since computers cannot handle complex logical relations, imposition of logical coherence and consistency norms on the performance of a mechanical scientist in the short run can have the consequence of making him less reliable than some computable scientist violating these norms could have been. An example of such a case is **consistency**: the requirement that only hypotheses consistent with the data be produced. Thus,

[32]Relativization to background knowledge K proceeds just as in the Borel case.
[33](Putnam 65), (Gold 65), (Kelly 91).
[34](Osherson, *et. al.*, 86).

from the reliabilist perspective, we have an inversion of the standard, philosophical perspective on methodology. The usual sorts of "methodological rules" are not thought of as tools for finding the truth, but as restrictions on the scientist's choice of a computational method for inquiry. At best such a restriction doesn't hurt the quest for truth, but the imposition of such restrictions is often damaging to reliability for computable scientists.[35] Indeed, this is the case concerning consistency. For a simple example, just let hypothesis P state that the first datum to be encountered is a member of the halting problem, and let $K = \omega^{\omega}$. P is verifiable with certainty by a computable method that tests the index i occurring first in the data for halting under increasing time bounds, saying 1 with certainty when the machine with index i halts on input i. On the other hand, no computable method can be consistent for this problem, on pain of being a recursive solution to the halting problem.

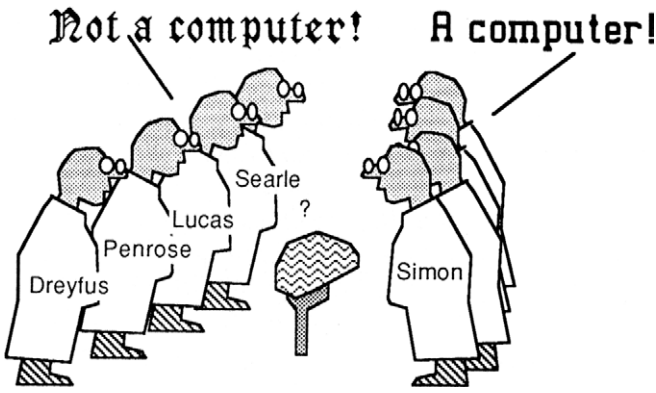
Consider a "Bayesian" scientist who conjectures only hypotheses with maximally probable posterior probability. Since posterior probability must go to 0 as soon as the hypothesis is refuted by the data, such an agent must be consistent in the sense just described. Thus the requirement that the conjecture at each stage be maximally probable with respect to some probability measure is also restrictive for computational agents.

10. The empirical paradox of cognitive science

The preceding discussion has been somewhat abstract. Let us consider the systematic application of learning theoretic ideas to a question familiar in cognitive science, namely, whether human behavior could be produced by a computational mechanism. Philosophers have attempted to give metaphysical, linguistic, and logical arguments against the computational thesis. H. Simon and other exponents of artificial intelligence have claimed that it is an empirical question whether or not cognition is computation, and that the evidence is good. For each new and supposedly "inspired" fragment of human "intuition", the hackers can come up with a big LISP program that seems to duplicate it. Then the mystery disappears.

H. Dreyfus (1972) has objected that piece-meal handling of "micro-worlds", one after the other, will never add up to a complete, synthetic computational account of human behavior as a whole. Or does the question

[35]This is one, principled reading of Feyerabend's familiar dictum "anything goes". But notice that while it is better to be able to choose among all possible methods than among those that satisfy some fixed architecture or methodological constraint, it is false that any method is as reliable or as efficient as any other. So in another sense, "not just anything goes".



transcend empirical method, as in the case of infinite divisibility? As it turns out, the question has an especially ironic status.

A hypothesis is **self-defeating** just in case if it is true then it is false: e.g. "every sentence with more than two words is false". By analogy, a hypothesis P is **empirically self-defeating** for scientists in class S just in case one of the following two situations obtains:

1. If P is true then P is not verifiable in the limit by scientists in S.

2. if P is false then P is not refutable in the limit by scientists in S.

For example, "this hypothesis is not verifiable in the limit by scientists in S" provides us with a trivial example of an empirically self-defeating hypothesis.

What about the hypothesis that human behavior is Turing computable? Our experimental setup is this. We feed different strings of inputs to a human subject and observe the response. By coding input sequences we can treat each such sequence as a natural number and then treat the subject's overall behavioral disposition as a recursive function.[36] These input-output pairs may be listed on a data tape for our psychologist. Thus, the hypothesis that human behavior is Turing-computable can be represented as the set of all infinite sequences that present the graphs of total recursive functions. Background knowledge is vacuous, so there is no known restriction on how

[36]It might fairly be objected that the scientist could not reliably distinguish between a new input sequence and a continuation of an old input sequence, but this only makes the problem harder, and we are principally interested in negative results here.

such a sequence may come in. With these assumptions we have the following results:[37]

PROPOSITION 2 *The hypothesis that human behavior is Turing computable is not refutable in the limit by us (or by anybody else).*

But we also have an ironic negative result:

PROPOSITION 3 *The hypothesis that human behavior is Turing computable is empirically self-defeating for us.*

The situation is this. The hypothesis that human behavior is Turing computable is verifiable in the limit by some ideal agent, but the fact that we are computable if the hypothesis is true implies that we cannot verify it in the limit, since no Turing computable scientist can.

But we can go somewhat further. A paradox is a statement that is true if and only if it is false. The famous example is *the liar*: "this statement is false". A hypothesis P is an **empirical paradox** for scientists in class S just in case one of the following two situations holds:

1. P is true $\Leftrightarrow$ P is not verifiable in the limit by scientists in S.

2. P is false $\Leftrightarrow$ P is not refutable in the limit by scientists in S.

Our previous example, "this hypothesis is not verifiable in the limit by scientists in S" is also an example of an empirical paradox for scientists in S.

The thesis that human behavior is Turing computable is not an empirical paradox for us, because a non-computable scientist might not be quite powerful enough to verify it in the limit. On the other hand, if we assume that any non-computable human is at least non-computable enough to solve the halting problem, then we have

PROPOSITION 4 *Assuming that uncomputable humans can solve the halting problem, the hypothesis that human behavior is Turing computable is empirically paradoxical for us.*

So the very first question about cognitive science is hopeless so far as getting to the truth about it is concerned. On the other hand, this does not mean that the investigation of *particular* cognitive hypotheses is hopeless. A **complete cognitive theory**, viewed as a computer simulation of a person's behavior, may be represented as a singleton $\{t\}$, where t is an

[37]For proofs of results in this section see (Kelly 91).

infinite, recursive data presentation. For each complete cognitive hypothesis $\{t\}$, a computable scientist can simply compute each prediction made by t and check it against the subject's output, rejecting $\{t\}$ forever when a discrepancy is found. Thus:

PROPOSITION 5 *Each recursive cognitive hypothesis $\{t\}$ is refutable with certainty by a Turing-computable scientist.*

So, perhaps the situation is not so bad after all. Perhaps cognitive scientists can arrive at a true cognitive theory by checking individual theories and tallying their results in some way. But as a matter of fact, they cannot, *if* human behavior is Turing computable! Again, there is a self-defeating character resulting from the self-referential character of the cognitive hypothesis:

PROPOSITION 6 *The complete cognitive truth is identifiable given the knowledge that human behavior is computable, but it is* not *identifiable by any computable scientist given the same knowledge.*[38]

But if human behavior is Turing computable then our behavior is as well, so we cannot identify the complete cognitive truth even if we were to know what we cannot reliably discover, namely, that human behavior is Turing computable. Notice the interplay between background knowledge, assessment, discovery, and self-reference to the scientist in this series of results.

Our discussion so far suggests that the relevant cognitivistic attitude is that human behavior is Turing computable. The psychologist J. R. Anderson[39] has suggested that primitive recursion ought to provide enough resources for cognitive theory. In this case, the picture looks quite different:

PROPOSITION 7 *The hypothesis that human behavior is primitive recursive is verifiable but not refutable in the limit by primitive recursive scientists.*

Thus there is no empirically self-refuting or empirically paradoxical character to this more restrictive cognitive hypothesis. This fact is also reflected in the prospects for cognitive discovery under this hypothesis:

PROPOSITION 8 *Given the knowledge that human behavior is primitive recursive, the complete cognitive truth is identifiable by a primitive recursive scientist.*

This raises further questions about whether other proposed bounds on cognitive power, such as pushdown or finite state automata, are empirically self-defeating for us.

[38](Gold 65) and (Putnam 63).
[39](Anderson 79).

11. Computable inquiry concerning uncomputable hypotheses

One apology for the working hypothesis that cognition is computable is that if we are in fact computable, we would have no way to test uncomputable hypotheses against the evidence. How does this suggestion hold up? Is it necessary for reliable, computable hypothesis assessment that the hypothesis in question be **effective** (i.e. that its predictions be derivable by computer)?

This question leads us to recognize deeper connections between the logic of inquiry and the theory of computation. The **basis theorems** of recursion theory[40] may be thought of as highly constructive analogues of the axiom of choice; a matter apparently far removed from mundane issues in empirical methodology. But as a matter of fact, these theorems provide a detailed picture of how the computational difficulty of deriving predictions from a theory relates to the computational difficulty of determining the truth value of the theory on the basis of evidence. That is to say, the basis theorems of recursion theory relate a theory's **deductive complexity** to its **inductive complexity**.[41] For example, a surprising result is that

PROPOSITION 9 *there exist complete cognitive theories that are in a sense*[42] *infinitely difficult to derive predictions from, that are nonetheless refutable with certainty by a computable scientist given no background knowledge.*[43]

Another surprising application is that this fact depends upon the number of different kinds of predictions the theory makes. If the hypothesis makes only finitely many different kinds of behavioral predictions, it is necessary for verification in the limit that predictions be computable.[44] Thus the example mentioned in Proposition 9 must involve infinitely many distinct predictions. These are precise answers to questions about the subtle balance between effectiveness and scientific method that confirmation theorists have not even asked.

12. Conclusion

Naturalists study how humans actually produce conclusions, a perfectly interesting project in its own right. Confirmation theorists probe and regiment intuitions about evidential relevance; a project of interest to many. But neither of these studies tells us something crucial, namely, how the cognitive

[40](Hinman 78).
[41]For details, c.f. (Kelly 91).
[42]That is, the function from time to prediction is not definable in arithmetic.
[43](Kelly 91).
[44](Kelly 91).

processes and regimented intuitions contribute to finding the truth. The purpose of this paper has been to exhibit an approach to methodology in which the scope of and prospects for reliable and computationally feasible inquiry are on center stage; a point of view in which the key decision is choosing a method, and in which both the intuitions of confirmation theory and the cognitive constraints of naturalism are conceived as side-constraints that merely impede our choice of the best methods for finding the truth. From this perspective, the relevant questions are as follows. Can our current methods be improved in their reliability or efficiency? How reliable can one possibly be? How should reliable tests be built into reliable discovery procedures? How badly do various methodological norms prevent us from being as reliable as we could have been without them? What are the necessary and sufficient structural conditions for reliable success for different classes of agents with different cognitive powers? What paradigms are equivalent in difficulty? How does computational complexity in hypotheses translate into computational complexity in scientific inquiry? Sociology and regimented introspection can neither answer these questions nor make them disappear. But these questions should be answered, and I have attempted to show how a truly logical inductive logic can answer many of them.

Acknowledgments

I am indebted to Clark Glymour, Cory Juhl, Peter Spirtes, and Teddy Seidenfeld for useful discussions concerning this material.

Bibliography

ANGLUIN, D. and C. SMITH (1982), *A Survey of Inductive Inference Methods*, Technical Report 250, Yale University, October.

ANDERSON, J. R. (1976), *Language, Memory, and Thought*, Hillsdale, NJ: Erlbaum.

DREYFUS, H. (1972), *What Computers Can't Do*, New York: Harper and Rowe.

FEYERABEND, P. K. (1979), *Against Method*, London: Verso.

GANDY, R. (1980), *Church's Thesis and Principles for Mechanisms*, In *The Kleene Symposium*, eds. J. Barwise, J. J. Keisler, and K. Kunen, Amsterdam: North Holland.

GARBER, D. (1983), *Old Evidence and Logical Omniscience in Bayesian Confirmation Theory*, in *Testing Scientific Theories*, J. Earman, ed.

GLYMOUR, C. (1980), *Theory and Evidence*, Princeton: Princeton University Press.

GOLD, E. M. (1967), *Language Identification in the Limit*, Information and Control 10 447-474.

GOLD, E. M. (1965), *Limiting Recursion*, Journal of Symbolic Logic, 30: 1. pp. 27-48.

HEMPEL, C. G. (1965), *Aspects of Scientific Explanation*, New York: Macmillan.

HINMAN, P. G. (1978), *Recursion-Theoretic Hierarchies*, New York: Springer.

HORWICH, P. (1991), *On the Nature and Norms of Theoretical Commitment*, Philosophy of Science, 58: 1.

KELLY, K. and C. GLYMOUR (1989), *Convergence to the Truth and Nothing but the Truth*, Philosophy of Science 56: 2.

KELLY, K. and C. GLYMOUR (1990), *Theory Discovery from Data with Mixed Quantifiers*, Journal of Philosophical Logic, Forthcoming.

KELLY, K. and C. GLYMOUR (1991), *Thoroughly Modern Meno*, Pittsburgh Studies in the Philosophy of Science, John Earman, ed., University of California Press, forthcoming.

KELLY, K. (1991), *Learning Theory and Descriptive Set Theory*, forthcoming, Journal of Logic and Computation.

KELLY, K. (1991B), *Reichenbach, Induction, and Discovery*, Erkenntnis, 35: 1-3.

KEMENY, J. (1953), *The use of simplicity in induction*, Philosophical Review, LXII, 391-408.

LAUDAN, L. (1980), *Why Abandon the Logic of Discovery?*, in Scientific Discovery, Logic, and Rationality, T. Nickles, ed., Boston: D. Reidel.

MOSCHOVAKIS, Y. (1980), Descriptive Set Theory, Amsterdam: North Holland.

OSHERSON, D., M. STOB and S. WEINSTEIN (1986), *Systems that Learn,* Cambridge: MIT Press.

OSHERSON, D. and S. WEINSTEIN (1991), *A Universal Inductive Inference Machine*, Journal of Symbolic Logic, 56: 2.

POPPER, K. (1968), *The Logic of Scientific Discovery*, New York: Harper.

PUTNAM, H. (1963), *'Degree of Confirmation' and Inductive Logic*, in The Philosophy of Rudolph Carnap, A. Schilpp (ed), Lasalle, Illinois: Open Court.

PUTNAM, H. (1965), *Trial and Error Predicates and a Solution to a Problem of Mostowski.* Journal of Symbolic Logic, 30: 1. pp. 49-57.

QUINE, W. (1990), *Pursuit of Truth*, Cambridge: Harvard.

RIVEST, R., D. HAUSSLER and M. WARMUTH, eds. (1989) *Proceedings of the Second Annual Workshop on Computational Learning Theory*, San Mateo: Morgan Kaufmann.

Logic, Methodology and Philosophy of Science IX
D. Prawitz, B. Skyrms and D. Westerståhl (Editors)

TAKING NATURALISM SERIOUSLY

PENELOPE MADDY

Department of Philosophy, University of California, Irvine

Once upon a time, students of scientific method were motivated by a desire to found science on some pre-scientific cornerstone. Philosophical observers from Descartes to Carnap hoped to uncover a source of pre-scientific certainty on which all scientific knowledge could be based. But whatever their inherent appeal, such projects have suffered a well-known history of unrelenting failure. Quine's response has been to note that this very desire to secure scientific foundations is driven by a skepticism that is itself a product of science. Without rudimentary physical science in the form of common sense about medium-sized physical objects, the notion of sensory illusion would collapse. '[S]ceptical doubts are scientific doubts,' he writes, 'Epistemology is best looked upon, then, as an enterprise within natural science. Cartesian doubt is not the way to begin.'[1]

The proper way to begin, from this point of view, is with the external reality of common sense as elaborated by science; in Quine's words, 'The naturalistic philosopher begins his reasoning within the inherited world theory as a going concern.'[2] Thus, naturalism: the 'abandonment of the goal of first philosophy ... the recognition that it is within science itself ... that reality is to be identified and described.'[3] The central epistemological question becomes – how does this particular species of primate manage to develop such a workable world view?[4]– and the methodologist's proper part of this question is the study of how that primate's scientific hypotheses are related to the evidence she cites in their support.

[1]W. V. Quine, 'The nature of natural knowledge,' in S. Guttenplan, ed., *Mind and Language*, (Oxford: Oxford University Press, 1975), pp. 67-81. The quotation comes from page 68.

[2]W. V. Quine, 'Five milestones of empiricism', in *Theories and Things*, (Cambridge, MA: Harvard University Press, 1981), pp. 67-72. The quotation comes from page 72.

[3]Quine, ibid, p. 72, and 'Things and their place in theories', in *Theories and Things*, pp. 1-23, especially p. 21. See also 'Epistemology naturalized', in *Ontological Relativity*, (New York: Columbia University Press, 1969), pp. 69-90.

[4]See 'The nature of natural knowledge', p. 68.

Quine himself has concentrated much of his naturalized epistemological attention on 'the learning of language and ... the neurology of perception,'[5] but he recognizes that more than this goes into the development of good scientific hypotheses and suggests that '[f]urther counsel is available anecdotally in the history of hard science'.[6] My aim here is to take the counsel of some such anecdotes. Along the way, I hope to shed some light on the practice of a truly naturalized methodology, but I suspect no one will be surprised to hear that my ulterior motive is to explore some implications for the philosophy of mathematics. More precisely, I am interested in the status of some fairly concrete statements about point sets that turn out to be independent of the current set theoretic axioms. But more on this in due time.

So how does this particular species of primate go about confirming a controversial scientific hypothesis? Among the more startling features of our contemporary world view is the belief that ordinary physical objects consist of largely empty space dotted with small particles too small for us to see. Why do we believe this? According to Quine, the molecular doctrine is supported by 'a convergence of indirect evidence'[7] which he divides into five benefits: simplicity, familiarity of principle, scope, fecundity, and agreement with experiment. Accepted scientific theories may enjoy these qualities to varying degrees – a surplus of scope and fecundity, for example, might compensate a deficiency in familiarity – but it is on the basis of these theoretical virtues that scientific hypotheses are to be judged.

The connection with the philosophy of mathematics is first drawn at this very general level. A scientific theory with generous portions of the theoretical virtues will often also include a good measure of mathematics: the temperature of a gas as a function of time, acceleration as a second derivative, the fundamental equations of Maxwell's theory. It seems impossible to remove the mathematical component of the theory while preserving a sufficiently virtuous version of the physical component. Indeed, as Putnam has emphasized, it is often difficult to see how a purely physical version of a sophisticated scientific theory can even be stated.[8] Thus, the required math-

[5] 'Five milestones', p. 72.

[6] *The Pursuit of Truth*, (Cambridge, MA: Harvard University Press, 1990), p. 20.

[7] W. V. Quine, 'Posits and reality', in *The Ways of Paradox*, revised edition, (Cambridge, MA: Harvard University Press, 1976), pp. 246-254. The quotation is from p. 246, and the five benefits are described on p. 247.

[8] See H. Putnam, 'Philosophy of logic', in *Philosophical Papers*, volume 1, second edition, (Cambridge: Cambridge University Press, 1979), pp. 323-357. Hartry Field, in *Science Without Numbers* (Princeton, NJ: Princeton University Press, 1980) and *Realism, Mathematics and Modality* (Oxford: Basil Blackwell, 1989), argues that it is possible, despite appearances, to remove the mathematics from our physical theories. For all their inge-

ematics should be accepted along with the physics; the theoretical virtues of the theory give us as much reason to believe its mathematical claims as its physical ones. In Putnam's words:[9]

> ...mathematical entities [are] indispensable for science ...this commits us to accepting the existence of the mathematical entities in question. This type of argument stems, of course, from Quine, who has for years stressed both the indispensability of ...mathematical entities and the intellectual dishonesty of denying the existence of what one daily presupposes.

All this is so familiar as to have passed into the realm of philosophical folklore: a hypothetico-deductive justification for physical theories; an indispensability defense for mathematical theories. These arguments function at an extremely high level of generality, in the rarefied world of scientific theory T and mathematical theory M. What I propose at this point is to lower the level of discussion considerably by attending to a particular historical case. In other words, I mean to consult historical anecdote. The case I have in mind is the same one Quine touched on above, that of atomic/molecular theory.[10]

The notion that matter is composed of tiny invisible bits goes back to the Greeks, but the beginning for the modern atomic hypothesis was Dalton's work in the first decade of the 19th century. During this period, Proust experimentally verified the Law of Definite Proportions – the proportions in which two substances combine do not vary continuously – and Dalton added the Law of Multiple Proportions – the definite proportions in which substances combine come in simple integral multiples. (So, for example, three grams of carbon combine with four grams of oxygen or with eight grams of oxygen, but with no amount in between. And eight is twice four.) Dalton

---

nuity, I think his efforts are not successful. (See my *Realism in Mathematics*, (Oxford: Oxford University Press, 1990), chapter 5, and 'Mathematics and Oliver Twist', *Pacific Philosophical Quarterly* 71 (1990), pp. 189-205.)

[9]Putnam, op. cit., p. 347. Putnam (and Quine) actually speak of the indispensability of *quantification over* mathematical entities, rather than the indispensability of the entities themselves, but this degree of precision is irrelevant here.

[10]Some of the fascinating details of this case were first brought to my attention by R. Miller's discussion in his recent book *Fact and Method*, (Princeton, NJ: Princeton, 1987), pp. 470-482. Grateful as I am for the stimulus, I can't fully endorse Miller's analysis of the case in terms of 'topic specific truisms'. For more detailed historical sources, see C. Glymour, *Theory and Evidence*, (Princeton, NJ : Princeton University Press, 1980), pp. 226-263, A. Idhe, *The Development of Modern Chemistry* (New York: Harper and Row, 1964), chapters 4 - 8, M. Nye, *Molecular Reality*, (New York: American Elsevier, 1972), and the classic J. Perrin, *Atoms*, first published in 1913, English translation by D. Hammick, (New York: van Nostrand, 1923). My account is based on these sources.

hypothesized that a sample of an elementary substance is actually made up of many tiny identical particles, that these remain unchanged through chemical reactions, and that a sample of a compound is made of many identical molecules, each composed of an identical combination of atoms from the constituent substances. This simple atomic hypothesis explains both laws of proportion.

In the same decade, Gay-Lussac discovered the Law of Combining Volumes: at a given temperature and pressure, the volumes of gases A and B that combine to form a given compound are in simple integral proportions. (E.g. Two volumes of hydrogen combine with one volume of oxygen to form two volumes of water.) In 1811, Avogadro theorized that equal volumes of gas (under similar conditions) contain equal numbers of Dalton's atoms, and that many elementary gases consist of diatomic molecules. This embellishment of atomic theory explains not only the Law of Combining Volumes, but also Boyle's Law of 1662 (pressure varies inversely with volume) and Charles's Law (gases expand equally when heated equally).

During the 1820s, various scientists realized that compounds with different chemical properties sometimes analyze into the same elements in the same proportions. An atomic explanation for this 'isomerism', as it is now called, soon followed: the same atoms can combine in different spatial relationships, and those spatial relationships influence the molecule's chemical behavior. In 1830s, Dumas noticed that a compound losing hydrogen while gaining chlorine did so in equal volumes, which led to his Law of Substitution. The notion that the substitution took place atom for atom could scarcely be avoided. Several decades of clues finally came together in the early 1850s, when Frankland added the concept of valence to the developing picture.

Despite this impressive string of successes, atomic theory during the first half of the 19th century was plagued by one very serious difficulty: the problem of determining relative atomic weights. Dalton had chosen hydrogen as his basis and calculated the relative weight of oxygen by measuring the amount of oxygen that combines with a given amount of hydrogen to form water. Obviously, no conclusion can be drawn from these measurements unless the chemical formula for water is already known. Dalton overcame this obstacle by assuming that the most common compound of two elements has a binary molecule, and thus, that water is HO. This simple error points up the problem: atomic weights can be calculated from combining weights and molecular formulas, and molecular formulas can be calculated from combining weights and atomic weights, but the early 19th century chemists knew only the combining weights.

Soon after Dalton, Berzelius devised his own table of atomic weights, based on different hypotheses and differing also in the assigned values. By 1820,

two new methods were added to this early guesswork: Mitshcerlich's Law of Isomorphism (similar crystalline structures result from the same number of atoms in the same arrangement) and Petit and Dulong's Law (the product of the specific heat and the atomic weight is a constant). Petit and Dulong produced another table of atomic weights that differed from those of Berzelius and Dalton. And in 1826, Dumas announced yet another method, based on his measurements of vapor densities.

Conflicts between the results achieved by these various methods led Dumas to conclude that atomic theory should be banished from chemistry. Though he apparently believed in atoms, Dumas came to reject the many hypotheses of atomic theory and to abandon the hope that they might produce a table of atomic weights confirmable by independent empirical tests. His dramatic statement reads:[11]

> If I were master, I would erase the word 'atom' from science, persuaded that it goes beyond experience; and never in chemistry ought we to go beyond experience.

Despite Dumas's stature, this admonition went unheeded. Compounds continued to be analyzed, molecular formulas proposed, and atomic weights conjectured.

Finally, in 1858, Cannizzaro did what Dumas had neglected to do: he distinguished carefully between molecule and atom.[12] With this simple clarification, a steadfast reinstatement of Avogadro's hypotheses, and the assumption that the smallest quantity of an element occurring in a molecule of a compound is its atomic weight, Cannizzaro was able to calculate a consistent table of atomic weights using vapor densities. He then compared these results with those achieved via specific heats, with admirable success, thus bringing order to atomic theory after decades of confusion.

Two years later, in 1860, around 140 of the world's most respected chemists convened in Karlsruhe to assess the status of the atomic theory. Cannizzaro presented his results, and reprints of his 1858 paper were distributed. Meyer describes his reaction as follows:[13]

> The scales seemed to fall from my eyes. Doubts disappeared and a feeling of quiet certainty took their place. If some years later I was myself able to contribute something toward clearing the situation and calming heated spirits no small part of the

[11] Quoted in Glymour, op. cit., p. 257.

[12] Gaudin had suggested this move back in 1826, but Glymour speculates that Dumas overlooked the idea out of a distaste for theory. See Glymour, op. cit., p. 254.

[13] Quoted by Idhe, op. cit., p. 229.

> credit is due to this pamphlet of Cannizzaro. Like me it must have affected many others who attended the convention. The big waves of controversy began to subside, and more and more the old atomic weights of Berzelius came to their own. As soon as the apparent discrepancies between Avogadro's rule and that of Dulong and Petit had been removed by Cannizzaro both were found capable of practically universal application, and so the foundation was laid for determining the valence of the elements, without which the theory of atomic linking could certainly never have been developed.

Meyer's own contribution, alluded to in this passage, began with his influential *Die modernen Theorien der Chemie* of 1864, which contains one of the first hints of the periodic table. In the words of one historian, the solution of the problem of atomic weights brought 'the atom into general acceptance as the fundamental unit of chemistry'.[14]

Around the same time, with the advent of the kinetic theory of heat, the influence of atomic thought spread into physics. In the hands of such thinkers as Maxwell and Boltzmann, the kinetic theory flowered, providing, among other things, the first calculations of absolute molecular magnitudes. Perrin's summary gives the flavor of these results:[15]

> ... each molecule of air we breathe is moving with the velocity of a rifle bullet; travels in a straight line between two impacts for a distance of nearly one ten-thousandth of a millimeter; is deflected from its course [five billion] times per second ... There are thirty [billion billion] molecules in a cubic centimeter of air, under normal conditions. Three thousand million of them placed side by side in a straight line would be required to make up one millimetre. Twenty thousand million must be gathered together to make up one thousand-millionth of a milligram.

By 1900, the atomic theory enjoyed all five theoretical virtues in abundance; its power and usefulness became more obvious with each experimental and conceptual advance.

At this point, historical anecdote deals a blow: despite the virtues of atomic theory, scientists did not agree on the reality of atoms. Though some antagonism toward the theory undoubtedly arose from general cultural factors – e.g. the rise of social and political 'idealism' at the end of the 19th

[14]Idhe, op. cit., p. 257.
[15]J. Perrin, op. cit., p. 82.

century[16]– a good portion arose inside the scientific community itself. In 1877, at a meeting of the French Académie des Sciences, Berthelot posed the rhetorical question, '... who has ever seen a gas molecule or an atom?'[17] Behind the rhetoric, and apart from worries about the conception of atoms as without parts, Berthelot simply opposed the appeal to entities inaccessible to direct experimental verification. Speaking of the early skeptics among chemists, Perrin makes a similar point,

> It appeared to them more dangerous than useful to employ a hypothesis deemed incapable of verification in the exposition of well-ascertained laws.[18]

By 1900, the hypothesis probably seemed less dangerous, but the desire for a direct test remained.

As an aside, we should note that a certain distrust of mechanical models in general had arisen toward the end of the century, inspired partly by Maxwell's success in divorcing his electromagnetic equations from the mechanical models he used to generate them. We've all heard Hertz's famous remark, 'Maxwell's theory is Maxwell's system of equations'.[19] The laws of thermodynamics, viewed as pure inductive generalizations from experience, were viewed by many as a model of scientific method. At one extreme, Ostwald proposed 'energetics', the doctrine that atoms are fictions and energy is fundamental. While admitting the historical fecundity of the atomic hypothesis, he set out to base its consequences on purely thermodynamic, experimental considerations. Others, particularly in England, were happy to continue using mechanical models, including the atomic theory, to generate testable equations, while they dismissed the question of how those models might relate to physical reality.

In any case, as we examine the skeptics's reactions to atomic theory from a naturalistic perspective, it becomes clear that the evaluation of this scientific hypothesis involved more than attention to the five theoretical virtues. In particular, the virtue closest to the favored idea of experimental verification must be the fifth – agreement with experiment – but the atomic theory had plenty of that. What it didn't have was some stronger sort of experimental success, something more 'direct', something that more conclusively 'verifies'. Without this, a sizable minority of the scientific community felt justified in

[16]See Nye, op. cit., p. 30.

[17]Quoted in Nye, op. cit., p. 7. J. Bernstein attributes a similar remark to Mach in his introduction to E. Mach, *Popular Scientific Lectures*, (La Salle, IL: Open Court, 1986), p. xiv.

[18]J. Perrin, op. cit., p. 15.

[19]Quoted in Nye, op. cit., p. 15.

withholding assent. To quote Ostwald's influential textbook of 1904:[20]

> ... the atomic hypothesis has proved to be an exceedingly useful aid to instruction and investigation, since it greatly facilitates the interpretation and use of the general laws. One must not, however, be led astray by this agreement between picture and reality and combine the two.

Even those who disagreed admitted that such skepticism was legitimate and even useful at the time.[21] For a philosopher to insist that the skeptics were mistaken because they failed to appreciate the evidential force of the five virtues would be an offense against naturalism.

The resolution of this impasse came soon after the comment of Ostwald just quoted. Describing the work that led to one of his remarkable series of papers published in 1905, Einstein writes:[22]

> Not acquainted with the earlier investigations of Boltzmann and Gibbs, which had appeared earlier and actually exhausted the subject, I developed the statistical mechanics and the molecular-kinetic theory of thermodynamics which was based on the former. *My major aim in this was to find facts which would guarantee as much as possible the existence of atoms of definite finite size.*

On the basis of his theoretical calculations, Einstein concluded that:[23]

> ... according to the molecular-kinetic theory of heat, bodies of microscopically-visible size suspended in a liquid will perform movements of such magnitude that they can be easily observed in a microscope. ... If the movement discussed here can actually be observed ... an exact determination of actual atomic dimension

[20] Quoted by Miller, op. cit., p. 473.

[21] See J. Perrin, op. cit., p. 216: 'the sceptical position ... was for a long time legitimate and no doubt useful.'

[22] A. Einstein, 'Autobiographical notes', in P. Schilpp, ed., *Albert Einstein: Philosopher-Scientist*, (La Salle, IL: Open Court, 1949), volume I, p. 47. Emphasis added.

[23] A. Einstein, 'On the movement of small particles suspended in a stationary liquid demanded by the molecular-kinetic theory of heat', first published in 1905, reprinted in his *Investigations on the Theory of Brownian Motion*, R. Furth, ed., (London: Methuen and Company, 1926). The quotation is from pp. 1-2. In the final sentence, I have substituted 'should' for 'had' and 'prove' for 'proved', so that the forward-looking final clause will match the rest of the sentence. (For the German, see Nye, op. cit., p. 139.) When this passage was written, the prediction had not yet been tested, and Einstein's later correspondence with Perrin suggests that Perrin's experiments displayed a level of precision Einstein had not thought possible (see Nye, op. cit., p. 147).

> is then possible. On the other hand, [should] the prediction of this movement [prove] to be incorrect, a weighty argument would be provided against the molecular-kinetic conception of heat.

The movement involved in this crucial test might be the so-called Brownian motion, but Einstein confessed 'the information available to me regarding the latter is so lacking in precision, that I can form no judgement in the matter.'

Meanwhile, thinkers more familiar with Brownian motion were convinced of its relevance. In a series of papers appearing between 1888 and 1895, Gouy argued that Brownian motion was caused by molecular movements and that it offered a potential confirmation of the kinetic theory of heat. In a letter written some years later, he remarks[24]

> From the historical point of view, one wonders today how the great founders of kinetic theory .... have not been able to see that Brownian movement places under the eyes the realisation of all their hypotheses!

The phrase 'under the eyes' is especially conspicuous when compared with Ostwald's complaint, published in the same year as Gouy's last paper, against the atomists's practice of[25]

> ...disturbing us with forces, the existence of which we cannot demonstrate, acting between atoms which we cannot see.

At an international Congress in 1904, Poincaré, another opponent of atomic theory, commented on Gouy's work along similar lines:[26]

> If this be so, we no longer have need of the infinitely subtle eye of Maxwell's demon; our microscope suffices us.

Indeed, Born uses the same terms to describe Einstein's work:[27]

> The fundamental step taken by Einstein was the idea of raising the kinetic theory of matter from a possible, plausible, useful hypothesis to a matter of observation, by pointing out cases where the molecular motion and its statistical character can be made visible.

[24] Quoted in Nye, op. cit., p. 21.
[25] Quoted by Nye, op. cit., p. 28.
[26] Quoted by Nye, op. cit., p. 38.
[27] M. Born, 'Einstein's statistical theories', in *Albert Einstein: Philosopher-Scientist*, pp. 163-177. The quotation comes from page 165.

Here was the promise of the longed for direct verification, but without experimental confirmation, it carried little conviction.

This task was undertaken by Perrin: [28]

> However seductive the hypothesis may be that finds the origin of the Brownian movement in the agitation of the molecules, it is nevertheless a hypothesis only. ...I have attempted to subject the question to a definite experimental test that will enable us to verify the molecular hypothesis as a whole.

Perrin based his first experiment on fairly transparent reasoning. Gas contained in a vertical column is more compressed lower down and more rarefied higher up simply due to gravity; the density of oxygen, for example, at 0° centigrade, will be reduced by half at a height of five kilometers. Using experimental techniques of unprecedented accuracy, Perrin measured the rate of rarification of tiny manufactured particles subject to Brownian movement in a dilute emulsion. In his own words:[29]

> Thus, once equilibrium has been reached between the opposing effects of gravity, which pulls the particles downwards, and of the Brownian movement, which tends to scatter them, equal elevations in the liquid will be accompanied by equal rarefactions. But if we find that we have only to rise 1/20 of a millimetre, that is, 100,000,000 times less than in oxygen, before the concentration of the particles becomes halved, we must conclude that the effective weight of each particle is 100,000,000 times greater than that of an oxygen molecule. *We shall thus be able to use the weight of the particle, which is measurable, as an intermediary or connecting link between masses on our usual scale of magnitude and the masses of the molecules.*

Perrin and his co-workers carried out experiments of this sort on particles of various sizes and compositions, suspended in various liquids, in various concentrations, and at various temperatures, and the numbers obtained for the absolute atomic weights and for Avogadro's number varied only slightly (e.g. between 65×10^{22} and 72×10^{22} for Avogadro's number).

Describing these results under the heading 'A Decisive Proof', Perrin relates that[30]

[28] J. Perrin, op. cit., p. 88-89.
[29] J. Perrin, op. cit., pp. 93-94.
[30] J. Perrin, op. cit., p. 104-105.

> ... [i]t was with the liveliest emotion that I found, at the first attempt, the very numbers that had been obtained from the widely different point of view of the kinetic theory. ... *Such decisive agreement can leave no doubt as to the origin of the Brownian movement.* ... The objective reality of the molecules therefore becomes hard to deny. At the same time, molecular movement has not been made visible. The Brownian movement is a faithful reflection of it, or, better, it is a molecular motion in itself, in the same sense that infra-red is still light.

Perrin went on to verify the rest of Einstein's predictions in a series of equally well-made experiments.

Perrin's results were published between 1908 and 1911, followed by his masterful popular exposition in *Atoms*, first published in 1913, and his conclusions were quickly accepted. To cite only the more dramatic reversals, in the 1908 preface to the fourth edition of his *Outline of Physical Chemistry*, Ostwald writes:[31]

> I have satisfied myself that we arrived a short time ago at the possession of experimental proof for the discrete or particulate nature of matter – proof which the atomic hypothesis has vainly sought for a hundred years, even a thousand years. The isolation and measurement of gases on the one hand, which the lengthy and excellent works of J.J. Thomson have crowned with complete success, and the agreement of Brownian movement with the demands of the kinetic hypothesis on the other hand, which have been proved through a series of researches and at last most completely by J. Perrin, entitle even the cautious scientist to speak of an experimental proof for the atomistic constitution of space-filled matter.

And commenting at the conclusion of a 1912 conference, Poincaré declared:[32]

> ... the long-standing mechanistic and atomistic hypotheses have recently taken on enough consistency to cease almost appearing to us as hypotheses; atoms are no longer a useful fiction; things seem to us in favour of saying that we see them since we know how to count them. ... The brilliant determination of the number of atoms made by M. Perrin have completed this triumph of atomism. ... The atom of the chemist is now a reality.

[31] Quoted in Nye, op. cit., p. 151.
[32] Quoted by Nye, op. cit., p. 157.

A few leading thinkers, most notably, Mach and Duhem, who both died in 1916, remained opposed to atomism despite the general trend of opinion, a fact for which their admirers are still apologizing.[33]

The case of atomic theory strikes me as fascinating and instructive for a number of reasons. First, as we've already noted, for many of the scientists in this story, the five theoretical virtues by themselves were not enough to establish the truth of a scientific hypothesis. Indeed, the counsel of this historical anecdote presents similar difficulties for many otherwise attractive general accounts of scientific method. Thus, for example, if science is guided by inference to the best explanation, atomic theory should have been accepted after 1860; it was, after all, the best explanation of chemical and physical phenomena, and to add the single phenomenon of Brownian motion to the list of those explained would not seem to justify such a dramatic change in status. The puzzle for such general theories is to distinguish between the situation in 1860, when the atom became 'the fundamental unit of chemistry', and that in 1913, when it was accepted as real. For example, some writers try to explain the evidential force of Perrin's experiments in terms of the convergence of many very different techniques in their estimates of Avogadro's number; Perrin lists thirteen in the *Atoms*. But don't the same considerations apply to Cannizzaro's success in 1860, when the many different techniques for computing atomic weights were all shown to agree?[34]

Another intriguing aspect of this case is the underlying theme of observation: first, the objection that molecules and atoms cannot be seen; then, the talk of seeing the random walk of molecules in Brownian motion. Here historical anecdote suggests that van Fraassen-style insistence on observability as the measure of reality was once a responsible scientific attitude. From the naturalistic perspective, the van Fraassenite errs in holding to this requirement on a priori grounds when it has long since been rejected on a posteriori, scientific grounds. Indeed, studies of the scientific, as opposed to the philosophical, use of the term 'observable' describe a remarkable expansion of application, so that nowadays, one reads comments like, 'Of these fermions, only the t quark is yet unseen'.[35] Suggestive as this line of thought may be, it should be noted that in their more careful moments, the scientists in our story would probably use a term like 'experimentally testable'

[33]See, for example, L. de Broglie's introduction to P. Duhem, *The Aim and Structure of Physical Theory*, (Princeton, NJ: Princeton University Press, 1954), or J. Bernstein's introduction to E. Mach, op. cit.

[34]I owe this point to R. Miller, op. cit., along with much of this paragraph.

[35]Quoted by I. Hacking in his *Representing and Intervening*, (Cambridge: Cambridge University Press, 1983), p. 182. See also D. Shapere, 'The concept of observation in science and philosophy', *Philosophy of Science* 49 (1982), pp. 485-525.

or 'experimentally verifiable' in place of 'observable'. I make no attempt to analyze this notion.

For my purposes here, the most important moral of the story is that the scientist's attitude toward contemporary scientific practice is rarely so simple as a uniform belief in some overall theory T. For example, Dumas believed that atoms exist but ought nonetheless to be excluded from chemistry, while Ostwald held atomic theory to be false but nevertheless useful for organizing chemical theory. In both cases, we find what is useful distinguished from what is true, with some parts of current theory falling under one heading and other parts falling under the other. Sometimes, as in the case of atoms, an especially potent experiment can move a theory from useful to true; in other cases, such as Maxwell's equations, a useful part is jettisoned when it has served its purpose; and in still others, such as the practices of the English school, parts of unknown truth value can be tolerated indefinitely, so long as they yield testable equations.

So, I take this historical anecdote as counseling us not to view a scientific practice in terms of a homogeneous theory T, but rather to carefully examine the various parts of that practice, to assess the levels of scientific commitment in more subtle gradations, to evaluate the bearing of evidence on various aspects of theory with more sensitivity to distinctions drawn between them by practitioners. The force of this counsel for standard thinking in the philosophy of mathematics is obvious: if science is viewed as a patchwork of hypotheses, models, approximations and experiments, if epistemic status varies from one patch to another, if we resist the temptation to lump everything together in one commodious theory T, then the truth of mathematics cannot simply be inferred from its appearance somewhere in the scientific fabric. After all, atoms pervaded chemistry well before Perrin. To understand how the success of science bears on the truth of mathematics, we must look more conscientiously at precisely where it appears in science and exactly how it functions there.

To bring this point home, it is enough to open an elementary physics text. In the analysis of water waves, for example, Feynman's introductory lectures remark that 'the ocean is considered infinitely deep'.[36] This is a perfectly reasonable way to proceed; otherwise, the mathematics would be unworkable. But notice that an indispensability argument based on this occurrence of mathematics in science would be laughable: we should believe in infinity because it plays a role in our best theory of water waves?! Similar examples abound. We use real-valued functions to represent such quantities as energy, charge and angular momentum, though we know them to be

[36]R. Feynman, R. Leighton, and M. Sands, *Lectures on Physics*, volume I, (Reading, MA: Addison-Wesley, 1963), chapter 51, p. 7.

quantized; subjects like fluid mechanics are firmly premised on the false assumption that matter is continuous. However indispensable they may be, none of these applications of continuum mathematics should convince us of the reality of the continuum.

Before developing this point further, let me pause a moment to describe a philosophical attitude toward mathematics that is common, though by no means universal, among physicists. I call this attitude 'philosophical' in the sense of 'non-scientific', that is, to indicate that it isn't backed by the sort of detailed empirical evidence these same physicists would cite in support of scientific claims. This distinction is somewhat rough-and-ready, but nevertheless important for the naturalist: it is the scientific, not the philosophical, practice of the scientist that counts as data for the naturalized methodologist.[37] Still, this philosophical attitude is worth noting.

So, for example, Feynman begins by describing mathematics as the language of physics, but he continues:[38]

> [M]athematics is *not* just another language. Mathematics is a language plus reasoning; it is like a language plus logic. Mathematics is a tool for reasoning ... a way of going from one set of statements to another. ... You state the axioms, such-and-such is so, and such-and-such is so. What then? The logic can be carried out without knowing what the such-and-such words mean.
>
> [Mathematics] is evidently useful in physics, because we have these different ways in which we can speak of things, and mathematics permits us to develop consequences ... For instance, I can say that [gravitational] force is directed toward the sun. I can also tell you ... that the planet moves so that if I draw a line from the sun to the planet, and draw another line at some definite period, like three weeks, later, then the area that is swung out by the planet is exactly the same as it will be in the next three weeks, and the next three weeks, and so on as it goes around the sun. I can explain both of those statements carefully, but I cannot explain why they are both the same. ... logic permits you to go from one to the other.

[37]To put the contrast crudely, we should take the scientist seriously when she says that this counts as evidence for that, but not (necessarily) when she claims to be writing down thoughts in the mind of God.

[38]R. Feynman, *The Character of Physical Law*, (Cambridge, MA: MIT Press, 1967), pp. 40-1, 45, 55. For another example, see A. Leggett, *The Problems of Physics*, (Oxford: Oxford University Press, 1987), pp. 28-30.

Mathematics, we're being told, is a matter of what follows logically from what. The physicist plugs into this logical grid by expressing his physical claims in mathematical language, after which logic generates consequences and interconnections. And again,

> Mathematicians are only dealing with the structure of reasoning, and they do not really care what they are talking about. ... But in physics you have to have an understanding of the connection of words with the real world.

We might just as well say that mathematics isn't about anything until the physicist interprets it. This attitude is encouraged when, as often happens, the same bit of mathematics applies to totally different phenomena – as, for example, the mathematics of a mass on a spring also works for electromagnetic oscillations – or when a pre-existing piece of pure mathematics is pulled off the shelf, as it were, and applied to a concrete situation for the first time.

In philosophical circles, a position like this is often called 'if-thenism'. Variations on this theme have been entertained by various well known philosophers of mathematics, among them Russell, who gave it up for Logicism, and Putnam, who gave it up for Platonism.[39] According to contemporary orthodoxy, the strongest objection to if-thenism is an indispensability argument: one cannot view science as literally true and mathematics as contentless because the very scientific statements one regards as literally true cannot be stated without the use of mathematical vocabulary.[40] But, if we are naturalists, and if scientists treat parts of their theories as non-literal even when they cannot dispense with them, then the inseparability of mathematics from science alone no longer carries the epistemic force the orthodox argument requires.[41] As far as this objection is concerned, if-thenism re-emerges as a live option.

Both this reexamination of if-thenism and our previous examples of bad indispensability arguments suggest that what needs to be determined is

[39] See H. Putnam, 'The thesis that mathematics is logic', in his *Philosophical Papers*, pp. 12-42. See also p. xiii.

[40] See H. Putnam, 'What is mathematical truth?' and 'Philosophy of logic', in his *Philosophical Papers*, pp. 60-78, and pp. 323-357.

[41] As a possible reply to the indispensability argument, Putnam considers 'fictionalism': the view that entities of a certain sort are merely 'useful fictions'. This is not unlike the pre-Perrinian skeptical attitude toward atoms. Putnam assumes that the only grounds for fictionalism are philosophical or theological and thus rejects fictionalism because 'it could not exhibit a better method for fixing our belief than the scientific method' (op. cit., p. 356). But the counsel of historical anecdote suggests that fictionalism is sometimes a proper part of the scientific method itself.

whether or not mathematics plays a role in scientific claims that are taken to be literally true. Observers of scientific method suggest that non-literal applications or 'models' most often occur within the framework of a more general background theory,[42] so perhaps such a general background theory is the best place to look for a literal application.

In contemporary physics, surely General Relativity qualifies as a fundamental theory, to be taken as literally true if anything is. This theory uses the mathematics of differential manifolds; indeed it asserts that space-time is such a manifold. If this is correct, then there exists, in physical reality, a continuous structure, namely, space-time. The real existence of such a structure would have profound implications for the foundations of set theory, in particular, for the status of some independent statements. Thus I have finally reached my ulterior goal. Let me say a few words about the type of independent statement I have in mind and the bearing of a physical continuum on their status.

Most people are familiar with Cantor's famous Continuum Hypothesis (CH), the claim that no infinite set of real numbers has cardinality intermediate between that of the natural numbers and that of the entire set of reals. To this day, no one knows whether or not Cantor was correct in this conjecture, but it has been shown that the question is independent of the standard set theoretic axioms (ZFC). Those of an if-thenist persuasion tend to take this result as a solution to the problem: the only way to make sense of the question 'Is CH true or false?' is to ask 'Does CH follow logically from ZFC or does its negation so follow?', and those two questions have been answered. To pursue the problem beyond this point would be to generate a pseudo-problem, much as Ostwald once accused the atomists of doing in their pursuit of facts about atoms.[43] On the other hand, those of a Platonistic bent, like Gödel, reject this reasoning, insist that ZFC is only part of the truth about the real world of sets, and hope that the truth-value of CH might one day be settled by the adoption of new set theoretic axioms.

Now the existence of a physical continuum by itself is not enough to provide a determinate truth value to the Continuum Hypothesis, for the obvious reason that the CH refers not simply to all reals or points but to all sets of reals or point sets. Rather that involving myself in the question of which point sets (or space-time regions) General Relativity will require over and above the continuum of points itself, let me shift attention to an independent question of a more concrete sort. To heighten the contrast with

[42]See P. Achinstein, *Concepts of Science* (Baltimore, MD: Johns Hopkins Press, 1968), chapter 7, or M. Redhead, 'Models in physics', *British Journal for the Philosophy of Science* 31 (1980), pp. 145-163.

[43]See Nye, op. cit., p. 17.

CH, it is worth noting how this other question arose.

Its roots are in the all-important concept of 'function'. Mathematical histories tell the story of how the simple notion of a curve or a continuous motion developed over the centuries, under pressure from physical problems like the vibrating string and heat flow and from foundational problems like those of the calculus, to the general idea of a completely arbitrary correspondence between real numbers. During the 19th century, mathematicians were confronted with an ever-weirder series of pathologies, for example, Dirichlet's shotgun function – zero on the rationals, one on the irrationals – a function nowhere continuous, without derivative or integral.

Any development this radical is bound to prompt skepticism, and by the turn of the century, there was considerable controversy over the propriety of this notion of an arbitrary function. The French analysts Baire, Borel and Lebesgue hoped to isolate a mathematically responsible fragment from the vast sea of non-continuous mappings by defining a hierarchy of functions of increasing but manageable complexity. As it turned out, the complexity of functions can be defined in terms of the complexity of sets of reals, as, for example, a function is continuous if and only if the pre-image of an open set is always open. Thus arose a hierarchy of sets of reals, beginning from open sets and proceeding via complements and countable unions, namely, the Borel sets.

The Borel sets turned out to be quite well-behaved, but not quite as well-behaved as Lebesgue thought. Lebesgue's analyses included a 'trivial' lemma to the effect that the projection of a Borel subset of the plane is a Borel subset of the line.[44] In fact, the projection of a Borel set can be more complex than Borel, but Lebesgue's uncharacteristic error wasn't discovered until ten years later, when it was exploited by the Russian school of Suslin and Luzin. Projection and complementation lead to a new hierarchy of projective sets of reals, more complex than Borel sets, but still fairly well-behaved at the lower levels.

For all its naturalness and quasi-constructive simplicity, this lively research program on the properties of manageable sets of real numbers was stalled by the 1930s. Consider, for example, the analytically useful property of Lebesgue measurability. Measurability for the Borel sets follows from the most elementary properties of Lebesgue measure, and the Russians quickly established it also for projections of Borel sets and their complements. But what about the sets that result from one more application of projection, the sets we now call $\sum_2^1$? The question of their measurability, along with other equally straightforward questions, remained frustratingly unanswered. The

[44]The projection of a two-dimensional set can be thought of as the shadow it casts on one coordinate axis.

reason for this dead end became clear only in the decades that followed: around 1940, Gödel showed that the measurability of $\sum_2^1$ sets could not be proved from ZFC, and around 1970, Solovay used Cohen's forcing methods to show that their non-measurability could not be proved either.[45]

This question – are $\sum_2^1$ sets Lebesgue measurable? – differs from the Continuum Problem in a number of relevant respects. First of all, rather than making a claim about all sets of reals, it concerns only a small fraction of those, the $\sum_2^1$ sets. These $\sum_2^1$ sets can be defined by simple formulas involving quantification only over reals, and they can also be characterized by the geometrically concrete operations of projection and complementation. Thus, if a physical continuum exists, it is hard to see why the space-time regions corresponding to $\sum_2^1$ point sets would not. As for the question being asked about these regions, Lebesgue measurability is a straight-forward mathematical concept that arose naturally in development of analysis. By contrast, the Continuum Question is about infinite cardinalities, a bold new notion introduced into mathematics only toward the end of the 19th century, one whose connection with the physical roots of analysis is much more attenuated. Again, if a physical continuum exists, the question of the Lebesgue measurability of its $\sum_2^1$ regions should be a legitimate one, with a determinate answer.

So, the bearing of General Relativity, along with its mathematical conception of space-time, on the status of this particular independent question is quite straightforward: if this physical theory is literally true, there ought to be a fact of the matter about whether or not all $\sum_2^1$ regions of space-time are Lebesgue measurable. From this perspective, the if-thenist would be wrong to dismiss the question as a pseudo-problem; the realist would be justified in pursuing a solution. This probably isn't the sort of justification many Platonists have in mind – it doesn't involve any non-physical reality – but it does give content to the problem in a naturalistically untroublesome way, that is, without raising questions about the nature of some non-physical reality, about our access to it, etc.

Having established its relevance to the foundations of set theory, let's return to General Relativity. We need to know if it should be considered literally true, or more particularly, if the mathematics of its treatment of space-time should be treated as literally true. To put it most simply: we need to know whether or not space-time is really continuous. Notice that this is not one of the usual questions raised in philosophical discussions of space-time;[46] it is without much recent philosophical history. There are

[45]See my *Realism in Mathematics*, chapter 4, for references and a more complete discussion.

[46]If space-time is relational rather than substantial, the question of whether or not it

those, however, who would classify it as unempirical, on the grounds that any measurement can be represented by a rational number, so no experiment could verify the continuous structure of anything.

Faced with such a question, as naturalists, our first instinct should be to examine the scientific literature for clues, to take once again the counsel of history, including, if I may put it this way, the counsel of contemporary history. Now I am no physics expert, so my inquiry here has so far been quite limited, but I can report a few observations. First, beginning with Einstein, there is some ambivalence toward the use of the continuum in fundamental physical theory:[47]

> Adhering to the continuum originates with me not in a prejudice, but arises out of the fact that I have been unable to think up anything organic to take its place.

This vague unease gives way to real discomfort when mathematical difficulties turn up in the theory of the electromagnetic field. In his *Lectures*, Feynman describes the problem this way:[48]

> Now we want to discuss a serious trouble – the failure of the classical electromagnetic theory. ... the concepts of electromagnetic momentum and energy, when applied to the electron or any charged particle ... are in some way inconsistent. ... There is an infinite amount of energy in the field surrounding a point charge ...

He describes various efforts to get around this difficulty, then continues:[49]

> We have already mentioned that it might be a waste of time to work so hard to straighten out the classical theory, because it could turn out that in quantum electrodynamics the difficulties will disappear or may be resolved in some other fashion. But the difficulties do not disappear in quantum electrodynamics. ... It

---

is continuous vanishes without being answered. Still, the usual ways of instituting relationalism will replace the question of whether or not space-time is continuous with that of whether or not fields are continuous, which serves our purposes just as well. See H. Field, 'Can we dispense with space-time?', in his *Realism, Mathematics, and Modality*, pp. 171-226, especially section 3.

[47]A. Einstein, 'Reply to Critics', in *Albert Einstein: Philosopher-Scientist*, volume two, p. 686. See also A. Fine, *The Shaky Game*, (Chicago, IL: University of Chicago Press, 1986), pp. 97-99.

[48]R. Feynman, R. Leighton, and M. Sands, *Lectures on Physics*, volume II, (Reading, MA: Addison-Wesley, 1964), chapter 28, pp. 1-2.

[49]Ibid, chapter 28, p. 10.

> turns out ... that nobody has ever succeeded in making a *self-consistent* quantum theory out of *any* of the modified theories. ... We do not know how to make a consistent theory – including the quantum mechanics – which does not produce an infinity for the self-energy of an electron, or any point charge. And at the same time, there is no satisfactory theory that describes a non-point charge. It's an unsolved problem.

The problem does not, however, stop the physicist:[50]

> ... it turns out that it is possible to sweep the infinities under the rug, by a certain crude skill, and temporarily we are able to keep on calculating.

Though this method was invented by Feynman himself, he has little affection for it:[51]

> Schwinger, Tomonaga, and I independently invented ways to make definite calculations ... (we got prizes for that). People could finally calculate with the theory of quantum electrodynamics! ... The shell game that we play ... is technically called 'renormalization'. But no matter how clever the word, it is what I would call a dippy process!

Finally, he elaborates his suspicion about what is going wrong:[52]

> I believe that the theory that space is continuous is wrong, because we get these infinities and other difficulties ... I rather suspect that the simple ideas of geometry, extended down into infinitely small space, are wrong.

Here we find the first hint that continuum mathematics may not be physically realized at the quantum level.

The difficulties become more acute and the hints more explicit and emphatic when physicists try to account for gravity on the quantum scale. Davies describes the situation this way:[53]

---

[50]R. Feynman, *The Character of Physical Law*, p. 156.

[51]R. Feynman, *QED*, (Princeton, NJ: Princeton University Press, 1985), p. 128.

[52]R. Feynman, *The Character of Physical Law*, p. 166-167.

[53]P. Davies, 'The new physics: a synthesis', in P. Davies, ed., *The New Physics*, (Cambridge: Cambridge University Press, 1989), pp. 1-6. The quotation comes from page 1.

> When I was a student in the 1960s ... [t]he four fundamental forces of nature ... were .. ill-understood at the quantum ... level. Only one of these forces, electromagnetism, had a consistent [renormalized] theoretical description. The weak force could not be properly understood, and many calculations of its effects gave manifest nonsense. ... The strong force appeared to be not a single force at all, but a complex tangle of perplexing interactions that seemed to have no simple underlying form. Gravitation was dismissed as irrelevant to particle physics, and the most strenuous attempts at providing it with a quantum description gave mathematical rubbish for almost all predictions.

He goes on to explain how, during the 1970s, the weak force was combined with quantum electrodynamics (QED), to produce a quantum theory of the electroweak force, the success of which inspired development of a theory of the strong force, called quantum chromodynamics (QCD). Then[54]

> With promising theories of three out of the four of nature's forces 'in the bag' the conspicuous odd man out is gravitation. Gravitation was the first of nature's forces to receive a systematic mathematical description ... but it continues to resist attempts to provide it with a quantum field description ... Direct attempts to quantise gravity in analogy with QED soon run into insuperable mathematical problems associated with the appearance of infinite terms in the equations. These 'divergences' have plagued all quantum field theories over the years, but the gauge nature of the other forces enables the divergences in their theories to be circumvented.

In other words, the quantum field theories of the other three forces can be renormalized, but quantum gravity resists this 'dippy process'. Davies concludes:

> So long as gravity remains an unquantised force there exists a devastating inconsistency at the heart of physics.

So the problem of quantizing General Relativity is an extremely important one, but the tricks that have worked in the past no longer manage to sweep the difficulties under the rug.

Why not? Isham suggests that the roots of the problem lie in conflicting notions of space-time:[55]

---

[54] P. Davies, op. cit., p. 3.

[55] C. Isham, 'Quantum gravity', in *The New Physics* , pp. 70-93. The quotation comes from page 70.

> ... gravitational effects are regarded as arising from a *curvature* in spacetime, and it is the reconciliation of this dynamical view of spacetime with the passive role it plays in quantum theory that constitutes the primary obstruction to the creation of a satisfactory quantum theory of gravity.

He continues:[56]

> It must be admitted that, at both the epistemological and ontological levels, our current understanding of space and time leaves much to be desired. In a gross extrapolation from daily experience, both special and general relativity use a model for spacetime that is based on the idea of a continuum, i.e. the position of a spacetime point is uniquely specified by the values of four real numbers (the three space, and one time, coordinates in some convenient coordinate system). But the construction of a 'real' number from integers and fractions is a very abstract mathematical procedure, and there is no *a priori* reason why it should be reflected in the empirical world. Indeed, from the viewpoint of quantum theory, the idea of a spacetime point seems singularly inappropriate: by virtue of the Heisenberg uncertainty principle, an *infinite* amount of energy would be required to localise a particle at a true point; and it is therefore more than a little odd that modern quantum field theory still employs fields that are functions of such points. It has often been conjectured that the almost unavoidable mathematical problems arising in such theories (the prediction of infinite values for the probabilities of physical processes occurring, and the associated need to 'renormalise' the theory ...) are a direct result of ignoring this internal inconsistency. Be this as it may, it is clear that quantum gravity, with its natural Planck length, raises the possibility that the continuum nature of spacetime may not hold below this length, and that a quite different model is needed.

Wheeler goes even further out on this limb:[57]

> The spacetime continuum? Even continuum existence itself? Except as an idealization neither the one entity nor the other can make any claim to be a primordial category in the description of nature.

[56] Ibid., p. 72.

[57] J. Wheeler, 'Information, physics, quantum: the search for links', in W. Zurek, ed., *Complexity, Entropy and the Physics of Information*, (Redwood City, CA: Addison-Wesley, 1990), pp. 3-28. The quotation comes from page 16.

Here Feynman's suspicion becomes an outright claim.

Of course much of this talk must be regarded as speculative, because the future shape of a theory of quantum gravity is still unknown, but we see here the real possibility that our assumptions about the continuity of space-time might have empirical ramifications, and indeed, that they might even turn out to be empirically false! Granted, any appeal to indispensability considerations to support mathematical claims brings with it an unavoidable trace of a posteriority, but it is one thing to allow that a claim is true only a posteriori and quite another to face the possibility that it is an a posteriori falsehood. Under the circumstances, it seems best to re-examine our consciences by asking what would follow if continuum mathematics were actually falsified.

First, I think we can safely predict that the calculus and higher analysis would not disappear from natural science. Just as Euclidean Geometry is still indispensable at non-relativistic dimensions, continuum mathematics would remain essential in large parts of science, from simple mechanics to fluid dynamics. What isn't obvious is how these falsified but indispensable theories should be understood. We are assuming they are not true theories of physical reality, and if we set aside the notion of a non-physical reality, we must conclude that they are not theories of any independent subject matter, and we are left to chose between various versions of anti-realism.

One such is the if-thenism mentioned earlier: mathematics is the study of what follows logically from what. So, for example, we might say Euclidean Geometry is the study of what follows from these axioms. Using Tarski's complete axiomatization, [58] we can assure that every elementary statement ostensibly about Euclidean space comes out with a determinate truth value on this reading. But, as we've seen, the situation is different for the continuum. If we take our theory of the continuum to be what follows logically from ZFC, there are questions about the continuum – e.g. are all $\sum_2^1$ sets Lebesgue measurable? – that it makes no sense to ask.

But set theorists do ask this question, and they look for new axioms to answer it. Even if we reject (for the moment) the realistic notion that they are looking for new axioms that are true of some independent subject matter, if-thenism seems unable to account for the many acknowledged constraints on this practice: not any old axiom will do.[59] Another anti-realist position, which I'll call fictionalism, may fair better in this regard. The fictionalist

[58]See A. Tarski, 'What is elementary geometry?', in J. Hintikka, ed., *The Philosophy of Mathematics*, (Oxford: Oxford University Press, 1969), pp. 164-175.

[59]This isn't the only problem with if-thenism, of course. See M. Resnik, *Frege and the Philosophy of Mathematics*, (Ithaca, NY: Cornell University Press, 1980), chapter 3, for discussion. (Resnik calls the position 'deductivism'.)

replaces the realist's analogy between mathematics and science with a new analogy between mathematics and imaginative story telling. Where the realist sees the justification of mathematical claims as analogous to theory confirmation in physics, the fictionalist sees the constraints on mathematical theorizing as analogous to broadly aesthetic criteria for good story telling. Both analogies involve differences as well as similarities: the realist admits that mathematics uses rigorous deductive reasoning in far greater measure than physics does, and the fictionalist allows that the constraints on mathematical story-telling are tighter, and depend more on the uses to which the story will be put, than the familiar aesthetic guidelines for good novel writing.

What interests us here is that this version of fictionalism differs from if-thenism on the key point: when a statement about the continuum is shown to be independent of our current story of the continuum, this does not mean there is no more to be said on the subject.[60] On the contrary, for the fictionalist, an independent question should inspire the mathematician to extend the story, which is in fact what happens. And the story cannot be continued just any old way; there are good and bad ways to do mathematics. The constraints are not purely aesthetic in the usual sense of the term, any more than the realist's justification is purely experimental in the usual sense of that term. They depend in part on what a particular mathematical theory is intended to do, but I won't try to elaborate here.[61]

I'm suggesting, then, that if continuum mathematics were to be falsified, the best course might be to adopt a fictionalist approach to our current theory of the continuum and to search for an appropriate extension of it to settle the independent questions.[62] If, on the other hand, continuum mathematics is literally true, if there is a physical continuum, then we should proceed as realists, doing our best to extend our theory by adding new true axioms. Now it isn't obvious that these two approaches would lead to the same theory, that the realist and the fictionalist will apply the same

[60]This fictionalism differs from C. Chihara's mythological platonism (see his *Ontology and the Vicious Circle Principle* (Ithaca, NY: Cornell University Press, 1973), pp. 61-75), which is designed to avoid a decision on CH. It also differs from H. Field's fictionalism, which takes the indispensability of continuum mathematics at face value and concludes that space-time is continuous.

[61]See M. Wilson's remarks on the 'hidden essentialism' of mathematicians in his 'Frege: the royal road from geometry', *Nous* 26 (1992), pp. 149-180, and the related work of K. Manders referenced there.

[62]I don't presuppose that there must turn out to be a unique appropriate extension. But 'appropriateness' must carry enough bite to rule out such easy answers as: both ZFC + CH and ZFC + not-CH are appropriate ways to extend ZFC. If this were enough, set theorists would not invest so much energy in devising and evaluating new axioms.

methods when they attempt to add new axioms. If not, then the determination of which of these two mathematical methodologies is correct – realist or fictionalist – would hinge on the answer to a physical question: is space-time continuous?

I'll leave you to further contemplate this rather odd conclusion at your leisure. Should the set theorist wait till the physicist completes a viable theory of quantum gravity before deciding which new axioms to adopt? If not, why not? Is there a general methodological theorem to the effect that realism and fictionalism will lead to the same conclusions? (I doubt it.) Is there something wrong with the line of reasoning that brought us to this pass? (Probably.) I suspect that the observations compiled here, if rearranged and looked at from another angle, might constitute a case against the indispensability arguments, or better, a case against the general view of mathematics engendered by the indispensability arguments, but I'll leave that thought for another day.[63] For now, I only hope to have shown that from a naturalist's perspective, the role of mathematics in science and the implications of that role for the foundations of set theory are more complex and subtle than has heretofore been appreciated.[64]

[63]I take this up in 'Indispensability and practice', Journal of Philosophy 89 (1992), pp. 275-289

[64]My thanks go to Gregory Chaitin, Yoshi Oono and Mark Wilson for references and advice on the issues surrounding renormalization and quantum gravity, and to the NSF (DIR-9004168) and UCI for their support.

Logic, Methodology and Philosophy of Science IX
D. Prawitz, B. Skyrms and D. Westerståhl (Editors)

RECENT PERSPECTIVES ON SIMPLICITY AND GENERALIZATION

PETER M. WILLIAMS

School of Cognitive and Computing Sciences, University of Sussex

The problem of generalising induction and its relation to simplicity are long standing issues in the methodology of science. These problems, in various guises, are familiar to philosophers of science, probabilists and statisticians as well as to empirical scientists. Recent developments in *computing* and *artificial intelligence* have brought them to the fore in a way that impinges directly on issues in the foundations of probability and induction.

1. Neural computation

Neural computation, sometimes called connectionism or parallel distributed processing, refers to a type of computing, or to a machine, that is unlike those hitherto in common use.[1] The model was suggested by biological nervous systems, as its name suggests, though contributions are being made by physicists, neurophysiologists and statisticians as well as computer scientists.

Aleksander [2] points out that conventional computing is based on *algorithms*. These are usually implemented on a von Neumann style of machine (an arithmetic-logic unit operating sequentially on data held in memory) and can be considered as representations of human knowledge. This means that conventional computing is restricted to tasks for which humans can find an algorithm. Living creatures, however, are not "programmed" by spelling out every step in a process but by *experience*. A child, for example, learns to recognise the character A not by an explicit geometric description but by being shown a sequence of positive and negative instances. Attempts to program a computer to recognise hand written characters by an exhaustive system of rules lead to combinatorial explosion as more and more exceptions and special cases have to be listed. A better approach is to allow the program or machine to develop its own internal representations of the necessary

[1] See [13] for a survey of current theory.

concepts and how to apply them.

One of the motivations of this approach is to gain an improved understanding of human competence in areas where humans currently perform better than computers. These relate especially to perceptual and high level cognitive abilities. From a practical engineering point of view, however, we can take the neural model as a basis on which to design systems aimed at performing some very specific task without being concerned whether this is exactly how humans perform it. These systems of *artificial neural networks* can combine the advantages of learning by experience with the precision and reliability of current computing machinery.[2]

2. Artificial neural networks

Artificial neural networks consist of a collection of connected processing elements. The connections form a directed graph and the network is called *feed-forward* if its graph contains no cycles. We restrict attention to feed-forward networks in this paper. A typical network is pictured in Figure 1.

The network determines a function from input to output. Input values are first latched into each of the input units and these are then passed through the network to emerge, after internal processing, as outputs at the output units. The behaviour of a typical unit is shown in Figure 2. Its output y is given by

$$y = \sigma(\theta + w_1 y_1 + \cdots + w_n y_n)$$

where $y_1, \ldots, y_n$ are outputs from similar units or else external inputs, θ is a *bias* attached to the unit and $w_1, \ldots, w_n$ are *weights* giving the connection strengths between units. σ is a thresholding or *transfer* function having a sigmoidal shape. A common choice, which will be used here, is the hyperbolic tangent function $\sigma(x) = \tanh(x)$, in which case the output range is between -1 and $+1$. This can be thought of as expressing the idea that if the total weighted input exceeds a certain threshold then the cell fires, otherwise not. The non-linearity introduced by the transfer function σ is the essential

[2]Good examples are the hand-written digit recognition system devised by Le Cun et al. [15], currently in use by the US Postal Service, and the system devised by Widrow [18] for simulating control of a reversing articulated vehicle. See [12] for further examples. Biological nervous systems are relatively slow in their individual processing elements but compensate by the use of massive parallelism. One of the goals of neural computing is to discover ways in which machines might also store and use experience in a way that exploits parallel architectures of a complexity comparable to that found in brains. Although this is still some way off, many insights of the neural analogy can be exploited using conventional processing architectures.

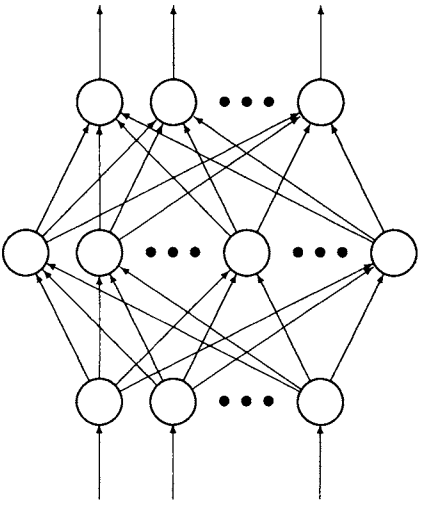


Figure 1: A 3 layer network. The lower layer comprises the input units and the upper layer the output units. The middle layer consists of internal or "hidden" units. There can be several internal layers. In general the connections between successive layers can be arbitrary.

novelty. If σ were the identity function, the overall network function would be the composition of linear (affine) mappings would itself be just a linear (affine) mapping.

3. Training sets

Feed-forward networks determine a functional mapping from input to output. Once the network architecture and transfer function are fixed, the mapping depends only on the weights and biases as parameters. Usually we have a specified set of input-output pairs that the network is required to associate. This constitutes a *training set.* The aim is to choose suitable weights and biases so that the network produces the desired target output for inputs in this set.

If there are m input units and n output units, a *training pair* consists of a pair $(\mathbf{x}, \mathbf{t})$ where $\mathbf{x} = (x_1, \ldots, x_m)$ is the input vector and $\mathbf{t} = (t_1, \ldots, t_n)$ is the target output vector. For an arbitrary choice of parameters, however, the actual output vector, $\mathbf{y} = (y_1, \ldots, y_n)$ say, will differ from the target output. This difference, thought of as an error, can be measured, for example, by the Euclidean distance between the actual and target output vectors. Suppose

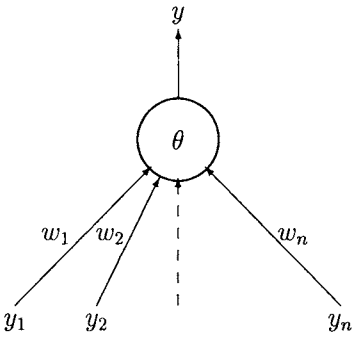


Figure 2: A typical unit.

there are P such training pairs $\{(\mathbf{x}_p, \mathbf{t}_p) : p = 1, \ldots, P\}$. Each individual pair gives rise to an error E_p with the total mean pattern error given by

$$E = \frac{1}{P} \sum_{p=1}^{P} E_p.$$

In the case of the Euclidean error measure

$$E_p = \frac{1}{2} \sum_{j=1}^{n} (y_{pj} - t_{pj})^2$$

where $\mathbf{y}_p = (y_{p1}, \ldots, y_{pn})$ is the actual output for the input $\mathbf{x}_p$. For a given training set, the error $E = E(\ldots, \theta, \ldots, w, \ldots)$ is just a function of the weights and biases. The aim is to choose these so as to minimise E.

The process of modifying the weights and biases in the network so as to minimise E can be thought of as one in which the network "learns" the desired association. Rumelhart et al. [20] propose a version of simple gradient descent. Considered as an unconstrained optimisation problem, however, faster methods are available using second-order information, see [5, 3, 25] for example. All methods have in common, however, the need to compute the *gradient* of the error function with respect to the variable parameters. The algorithm for doing this is referred to as "back-propagation".

4. Existence and uniqueness

Given a network architecture and a corresponding training set, one can ask (1) whether there exists an assignment of weights to the connections so that the network reproduces the association given by the training patterns and, if so, (2) whether there exist many such assignments. The first question

depends on the resources provided by the network architecture. The answer is that, provided the network has enough hidden unit, it can learn any reasonable mapping. Two layers are sufficient for an arbitrarily close approximation, provided enough hidden units are supplied, and the same is true for a single layer if the function is continuous, see for example [8, 14]. The problem, rather, is that there may be many assignments of weights that fit the data. Some may perform well on out-of-sample data and others very poorly. The problem is similar to that of polynomial interpolation. A high enough order polynomial will fit any set of data points, but it may behave erratically in between.

Consider the following example studied by Denker et al. [9]. A network is to be trained to classify sequences of 10 binary inputs according to whether or not they contain two or more clumps, where a clump is defined to be a consecutive sequence of +1's. Some sample inputs and the intended classifications are given in Table 1. There are 1024 possible inputs. If the network

Input	Output	Meaning
---+++----	-	1 clump
---++-+---	+	2 clumps
-+++++++++	-	1 clump
+++--++--+	+	3 clumps
----------	-	0 clumps

Table 1: Examples of the two-or-more clumps predicate.

is trained to classify a certain proportion correctly, will it generalise well to the "correct" rule on unseen samples? The problem is similar to the familiar one of completing a series of which only the first few terms are given.

This is a case of classification in which both the sample inputs and the intended outputs are precise or "noise-free". In applications using physical data, relating to mineral exploration or analysis of meteorological data for example, both inputs and outputs may be known only with a certain degree of error. In such cases, there is the usual risk of overfitting, i.e. of fitting not only the signal but also the noise. An overfitted solution is unlikely to generalise well beyond the data. This is typical of "ill-posed problems" in the sense of [21]. We now review two approaches to *stabilising* or *regularising* the solution.

5. Structural stabilisation

By analogy with curve fitting, an attempt can be made to obtain a smooth fit by limiting the complexity of the interpolant. A natural measure of complex-

ity for polynomials is *degree*. For feed-forward networks, a straightforward measure is the number of free parameters. It turns out that a more suitable measure is the *Vapnik-Chervonenkis (VC) dimension* [23, 22, 1]. In simple cases, however, such as a three layer network with a single output unit, these roughly coincide.

5.1. Theory

Using results of Vapnik and Chervonenkis, Baum and Haussler[4] have shown that, for $0 < \epsilon \leq \frac{1}{8}$,

> ... if $m \geq O(\frac{W}{\epsilon} \log \frac{N}{\epsilon})$ random examples can be loaded on a feed-forward network of linear threshold functions with N nodes and W weights, so that at least a fraction $1 - \frac{\epsilon}{2}$ of the examples are correctly classified, then one has confidence approaching certainty that the network will classify a fraction $1 - \epsilon$ of future test examples drawn from the same distribution. [4, p.151]

Ignoring the constant and logarithmic factors, this suggests that about 10 times as many examples as weights in the network would be needed to achieve an accuracy level of 90%, corresponding to $\epsilon = 0.9$. Conversely

> ... for fully-connected feedforward nets with one hidden layer, any learning algorithm using fewer than $\Omega(\frac{W}{\epsilon})$ random training examples will, for some distributions of examples consistent with an appropriate weight choice, fail at least some fixed fraction of the time to find a weight choice that will correctly classify more than a $1 - \epsilon$ fraction of the future test examples. [4, p.151]

These results assume that the node functions used are linear threshold functions, or at least Boolean valued. It has been conjectured that similar results fold for continuous real valued functions such as sigmoids.

The question arises whether these results give a practical guide to regularisation of network training. The following points are relevant.

(1) The Vapnik-Chervonenkis uniform convergence theorem assumes that the training examples are generated by some probability distribution P on a population X. The examples are to be picked at random according to some definite probability distribution. Although the results hold for any probability distribution, the same distribution must be used for generating the training examples as for generating the instances to which the network is subsequently applied. In most cases, including ones of current interest,

this is certainly not exactly or even approximately true and it is not easy to calculate the extent to which this affects the validity of the results.[3]

(2) The bounds provided are broad and inexact. The sufficient condition demonstrated for valid generalisation is in fact that $m \geq \frac{32W}{\epsilon} \log \frac{32N}{\epsilon}$. Baum and Haussler comment that 32 is likely to be an overestimate and that no serious attempt has been made to minimise it. Nor do they know if the log term is unavoidable. From a practical point of view, however, even a factor of 2 can be crucial when training examples are in short supply.

(3) The necessary condition demonstrated for valid generalisation is specific to a three-layer network for which a bound on the VC dimension can be calculated. The condition then is that $m \geq \frac{2\lfloor \frac{k}{2} \rfloor n-1}{32\epsilon}$, where k is the number of hidden units and n is the number of inputs. For large k and n, the quantity $2\lfloor \frac{k}{2} \rfloor n$ is approximately equal to the total number W of weights in the network, so that the necessary condition becomes $m \geq \frac{W}{32\epsilon}$. Furthermore, even this very weak condition has only been proved necessary for the *worst-case* distribution that is consistent with some function realisable on the network.

Although these interesting results do not yet provide practical numerical guides for regularisation, they emphasize that the ratio of weights to training examples should be as small as possible consistent with achieving a meaningful fit to the data.

5.2. Practice

In an attempt to implement the pruning of connections in a network in a practical way, Le Cun et al.[16] have proposed an algorithm for "optimal brain damage". It is based on the ideas (i) that connections with small weights have the least effect on the training error and can profitably be deleted and (ii) that

> a "simple" network ... is more likely to generalize correctly than a more complex network, because it presumably has extracted the essence of the data and removed the redundancy from it. [16, p.604]

Specifically the *saliency* of a weight w_i is defined to be $\frac{1}{2} w_i^2 \, \partial^2 E / \partial w_i^2$, assuming that all parameters are independent. Saliency is a better measure than simple magnitude since it is concerned with the exact effect of the deletion

[3] Giedymin [10] proposes random sampling as a condition for a good test of a scientific hypothesis in replying to an argument of Goodman [11] on simplicity. Unfortunately most of the population of interest is necessarily unavailable to random sampling when attempting to predict the future.

of a weight on the magnitude of the training error. It is proposed that, having chosen an initial network graph, the network should be trained until a reasonable solution is obtained. Some connections with low-saliency weights should be deleted and the network retrained. This process continues until an acceptable balance is obtained between network simplicity and training error. This is claimed to give better generalisation with fewer training examples needed.

6. Formal stabilisation

An alternative approach is based on the idea, advocated most notably by Tikhonov and Arsenin [21], that an extra term should be included in the objective function. This is designed to improve generalisation by smoothing the fit. The objective function will then be of the form

$$E = \alpha E_D + \beta E_R \qquad (\alpha > 0,\ \beta \geq 0) \tag{1}$$

where E_D is the data misfit and E_R is the regularising term. Both are functions of the free parameters. The ratio of α to β determines the trade-off between degree of fit and model complexity. The following discussion of choice of regularising term and parameters draws on ideas of Mackay [17] and Buntine & Weigend [7].

6.1. Maximum likelihood estimation

To discuss the choice of functions E_D and E_R and of the regularising parameters α and β, it is useful to recall the interpretation of fitting by least squares error as maximum likelihood estimation.[4] Suppose we are trying to predict the value of a real-valued quantity t on the basis of data x. The data is modelled as deviating from a predicted value $y = f(x, w)$ by some additive noise process

$$t = y + \nu$$

where f is the modelling function depending on parameters w. (All quantities may be vectors.) If ν is assumed to be zero-mean Gaussian noise with standard deviation σ, the likelihood density for a single observation t is

$$P(t \mid x, w) = \frac{1}{\sqrt{2\pi}\sigma} \exp -\frac{(t-y)^2}{2\sigma^2}. \tag{2}$$

Suppose now we have a sequence $\mathbf{x} = \{x_1, \ldots, x_N\}$ of data inputs and a corresponding sequence of predicted values $\mathbf{y} = \{y_1, \ldots, y_N\}$, where each y_n

[4]For references see, for example, [19, Ch.14].

is a function $f(x_n, w)$ of the data inputs x_n and the model parameters w. If the noise is modelled as independent between data items but with the same standard deviation σ, the likelihood density of a sequence of observations $\mathbf{t} = \{t_1, \ldots, t_N\}$ is

$$P(\mathbf{t} \mid \mathbf{x}, w) \propto \prod_{n=1}^{N} \exp - \frac{(t_n - y_n)^2}{2\sigma^2} = \exp - \sum_{n=1}^{N} \frac{(t_n - y_n)^2}{2\sigma^2}. \tag{3}$$

Recalling that each y_n is dependent on the parameters w, maximum likelihood estimation chooses w to maximise this quantity. This is the same as choosing parameters w to minimise

$$(1/\sigma^2) \sum_{n=1}^{N} \tfrac{1}{2}(t_n - f(x_n, w))^2$$

which can be identified with the first term of (1) for $\alpha = 1/\sigma^2$.

6.2. The Bayesian viewpoint

From the Bayesian point of view, however, it is not $P(\mathbf{t} \mid \mathbf{x}, w)$ that is our direct concern. Our real concern is with the posterior probability of the parameters given the training data, namely $P(w \mid \mathbf{x}, \mathbf{t})$. Using Bayes theorem

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

we have

$$P(w \mid \mathbf{x}, \mathbf{t}) \propto P(\mathbf{t} \mid \mathbf{x}, w) P(w \mid \mathbf{x}). \tag{4}$$

The first term on the right is already given by (3) so it remains to determine $P(w \mid \mathbf{x})$. This expresses a prior distribution over the possible weights in the network. A simple assumption is that these are distributed independently with a zero-mean Gaussian of standard deviation τ, independently of the data inputs $\mathbf{x}$. If there are W weights in the network this gives

$$P(w \mid \mathbf{x}) = P(w) \propto \prod_{i=1}^{W} \exp - \frac{w_i^2}{2\tau^2} = \exp - \sum_{i=1}^{W} \frac{w_i^2}{2\tau^2}. \tag{5}$$

Let us assume that the aim is to choose network weights w that are maximum *a posteriori* probable. In that case we want to maximise $P(w \mid \mathbf{x}, \mathbf{t})$. This is the same as maximising $\log P(w \mid \mathbf{x}, \mathbf{t})$ or, equivalently, minimising $-\log P(w \mid \mathbf{x}, \mathbf{t})$. Putting together equations (4), (3) and (5), the aim is therefore to minimise

$$\alpha \sum_{n=1}^{N} \tfrac{1}{2}(t_n - y_n)^2 + \beta \sum_{i=1}^{W} \tfrac{1}{2} w_i^2 \qquad (\alpha = 1/\sigma^2,\ \beta = 1/\tau^2) \tag{6}$$

which is exactly of the form of (1) for this choice of noise model and regulariser.

6.3. Choice of α and β

The first coefficient $\alpha = 1/\sigma^2$ represents the amount of noise we expect to find in the system. In some cases this might be known in advance, though in general we shall have little idea of its value. The amount of noise can depend on the model. If, for example, the aim is to predict likely gold occurrence on the basis of magnetic data alone, every factor that influences the actual extent of mineralisation and is not directly correlated with the magnetic field is a source of noise. The amount of noise in the system, in this sense, cannot be estimated *a priori*. Only when it is known how well the model can be made to fit the data, can we estimate the extent to which other factors might be influencing mineralisation. On the other hand, if a sufficiently complicated model is adopted, with a large number of free parameters, the data can be fitted exactly, suggesting misleadingly that the noise is zero. This is where regularisation comes in. However, in order to make use of the regularising term we need an estimate of the extent to which the data needs smoothing, which is expressed by the value of β. This cannot be estimated *a priori* until we have an idea of the noise level in the system, and we have come full circle.

Here is a way out. The noise level σ is unknown *a priori* but we may, nonetheless, have an opinion about its likely value expressed by a prior probability distribution $P(\sigma)$. In that case the quantity $P(\mathbf{t} \mid \mathbf{x}, w)$ in (4) can be calculated by

$$P(\mathbf{t} \mid \mathbf{x}, w) = \int_\sigma P(\mathbf{t} \mid \mathbf{x}, w, \sigma) P(\sigma \mid \mathbf{x}, w)\, d\sigma.$$

If it is assumed that knowledge of the parameters of the model and the data inputs alone does not affect opinions about the noise level, then $P(\sigma \mid \mathbf{x}, w) = P(\sigma)$ and the preceding equation simplifies to

$$P(\mathbf{t} \mid \mathbf{x}, w) = \int_\sigma P(\mathbf{t} \mid \mathbf{x}, w, \sigma) P(\sigma)\, d\sigma. \tag{7}$$

What can be said about $P(\sigma)$? All that is known *a priori* is that σ is a positive *scale* parameter. What is needed is a suitable *noninformative* prior.[5] A commonly preferred prior in such cases is $P(\sigma) \propto 1/\sigma$. In order to use equation (7) to calculate $P(\mathbf{t} \mid \mathbf{x}, w)$ it is necessary to know the full σ-dependency of $P(\mathbf{t} \mid \mathbf{x}, w, \sigma)$. For the Gaussian model we have

$$P(\mathbf{t} \mid \mathbf{x}, w, \sigma) \propto \frac{1}{\sigma^N} \exp -\frac{E_D}{\sigma^2}$$

[5]See [6, Sec.3.3] for discussion.

where

$$E_D = \sum_{n=1}^{N} \tfrac{1}{2}(t_n - y_n)^2.$$

Substituting in (7), and using a change of variable $x = E_D/\sigma^2$ to evaluate the integral, gives

$$P(\mathbf{t} \mid \mathbf{x}, w) \propto \int_\sigma \frac{1}{\sigma^{N+1}} \exp -\frac{E_D}{\sigma^2}\, d\sigma \propto E_D^{-N/2}. \tag{8}$$

The same argument can be applied to τ, or equivalently β, to obtain

$$P(w \mid \mathbf{x}) = P(w) = \int_\tau P(w \mid \tau) P(\tau)\, d\tau \propto E_R^{-W/2} \tag{9}$$

where

$$E_R = \sum_{i=1}^{W} \tfrac{1}{2} w_i^2.$$

6.4. Revised objective function

Returning to the Bayes equation (4) and substituting (8) and (9) we have

$$P(w \mid \mathbf{x}, \mathbf{t}) \propto E_D^{-N/2} E_R^{-W/2}.$$

The aim is to maximise this *a posteriori* probability. Taking negative logarithms, this means that we wish to minimise

$$Q = \frac{N}{2} \log E_D + \frac{W}{2} \log E_R. \tag{10}$$

Notice that the relative weight of the two terms depends on the ratio of the number N of data points to be fitted to the number W of free parameters of the model. As the number of training examples N increases for a fixed number W of weights, the need for stabilisation diminishes. Conversely, for a fixed number of training examples N, the penalty term increases, expressing the need for greater stabilisation.

It is instructive to compare the gradient of $E(w)$ defined in (1) with the gradient of $Q(w)$ defined in (10). The first quantity assumes fixed values for α and β given in advance. The second has marginalised these away by using non-informative priors. We have

$$\nabla E = \alpha \nabla E_D + \beta \nabla E_R \tag{11}$$

compared with

$$\nabla Q = \frac{N}{2E_D}\nabla E_D + \frac{W}{2E_R}\nabla E_R. \tag{12}$$

Recalling that $\alpha = 1/\sigma^2$ and $\beta = 1/\tau^2$, this is equivalent to adapting σ and τ during the optimisation process by

$$\sigma^2 = \frac{2E_D}{N} = \frac{1}{N}\sum_{n=1}^{N}(t_n - y_n)^2, \quad \tau^2 = \frac{2E_R}{W} = \frac{1}{W}\sum_{i=1}^{W} w_i^2$$

which are current estimates for the corresponding variances. It is also worth noting the relative weights of the two terms in (12). Because of the logarithms in (10) these now depend inversely on the current values of E_D and E_R. As the misfit E_D decreases, less relative importance is given to the regularising term E_R.

6.5. Alternative error models

The preceding section shows how the error function corresponds to a statistical model of the noise and the regularising term to a prior probability distribution over the free parameters. Alternatives to the model discussed above could be considered. The Gaussian error model is appropriate for repeated physical measurements under identical experimental conditions. When experimental conditions vary to a greater extent, however, but the standard deviation is still independent of the data inputs, a distribution with wider tails such as the Laplace distribution may be more appropriate.[6] This means replacing (2) by

$$P(t \mid x, w) = \frac{1}{2\Delta}\exp -\frac{|t - y|}{\Delta}. \tag{13}$$

Using the same technique as before with an uninformative prior over Δ, the function now to be minimised in place of (10) is

$$Q = N \log E_D + \frac{W}{2}\log E_R, \tag{14}$$

where

$$E_D = \sum_{n=1}^{N} |t_n - y_n|,$$

assuming that the same quadratic regulariser is used.

[6]See, for example, [22, pp.84–5] and [19, Sec 14.6].

6.6. Alternative regularisers

We have so far assumed the simple sum of squares of weights as the regularising term. Mackay [17] points out that this implies a prior that expects all weights to have the same typical size. But input and output weights would need to be rescaled in response to rescaling of input and output variables, so this assumption is inconsistent, i.e. not invariant under appropriate transformations. Suppose then that the weights are divided into m exclusive but not necessarily exhaustive classes $\mathcal{W}_1, \ldots, \mathcal{W}_m$, of respective sizes $W_1, \ldots, W_m$, and that the prior distribution over weights in $\mathcal{W}_i$ is zero-mean Gaussian with unknown standard deviation τ_i $(i = 1, \ldots, m)$. Assuming noninformative priors for τ_i, the same argument as before shows that the regularising term

$$\frac{W}{2} \log E_R$$

in (10) or (14) should be replaced by

$$\frac{W_1}{2} \log E_{R_1} + \cdots + \frac{W_m}{2} \log E_{R_m}$$

where

$$E_{R_i} = \sum_{w \in \mathcal{W}_i} \tfrac{1}{2} w^2 \qquad (i = 1, \ldots, m).$$

A simple instance is to take $m = 2$ with $\mathcal{W}_1 =$ output unit weights (all weights on a connection inputting to an output unit) and $\mathcal{W}_2 =$ hidden unit weights (all weights on a connection inputting to a hidden unit). This includes all the connection weights but no biases. From the Bayesian point of view, it amounts to imposing a uniform prior on all biases. Another reasonable model, in a layered network, would be to partition all connection weights by layer, making $n - 1$ classes for an n-layered network.

Another approach that has been proposed is to replace the

$$\sum \tfrac{1}{2} w^2$$

term in (6) by

$$\sum \tfrac{1}{2} \frac{w^2/w_0^2}{1 + w^2/w_0^2} \tag{15}$$

where w_0 is a typical weights size, e.g. $w_0 = 1$. The intention here is to force small weights to zero so as to reduce model complexity in line with the ideas with Section 5.1.. This has been claimed in [24] to enhance generalisation in applications to sunspot time series and currency exchange rates. Although it

appears less susceptible to a Bayesian analysis than the quadratic regulariser, it corresponds, broadly speaking, to a prior distribution with wider tails than the Gaussian. This suggests that a similar result could be obtained by using

$$\sum \tfrac{1}{2} \log(1 + w^2/w_0^2),$$

corresponding to the Cauchy distribution, in place of (15). Another possibility, based on the distribution with density

$$f(x) = \frac{1}{2\sigma} \operatorname{sech} \frac{\pi}{2} \left(\frac{x - \mu}{\sigma} \right),$$

would be to use

$$\sum \log \cosh(w/w_0).$$

In both cases the regulariser is obtained from the negative logarithm of the density.

7. Application

These ideas were applied recently to a problem in mineral exploration. The aim was to correlate gold occurrence with magnetic field anomalies. The training set was obtained from assay values of cores extracted from 341 drill holes. Elements of the training set consisted of pairs of which the first member was a representation of the local magnetic field at the drill hole location and the second was a typical assay down the hole.

Generalising ability was assessed by simple cross-validation. 71 of the training samples were extracted to form a test set by removing two geographically distinct regions, one of which was known as a result of drilling to be rich in ore and the other to be predominantly waste. The drill hole assays in these regions were not known to the network at the training stage. Questions of interest were (i) whether the trained network would have discovered the deposits in the favourable region and (ii) whether it would have discouraged costly drilling in the unfavourable region.

Training was carried out both with and without regularisation and was continued until a local minimum was reached. The objective function for the unregularised network was given by equation (1) with E_D as the quadratic error measure and $\beta = 0$. For the regularised network the objective function was that of equation (10), again with quadratic measures for E_D and E_R, but with the regularising term divided between output and hidden unit weights as described in Section 6.6.. Results are shown in Table 2. Both networks score well in the unfavourable region. In the favourable region, however, the

	Favourable		Unfavourable		
	Correct	Incorrect	Correct	Incorrect	
Unregularised	22	24	23	2	71
Regularised	40	6	23	2	71

Table 2: Distribution of correct and incorrect classifications of drill holes in favourable and unfavourable regions by regularised and unregularised networks.

unregularised network gives classifications that are no better than chance. The regularised network scores equally well in both regions.[7]

8. Conclusion

The perspective employed in this paper shows promise for understanding some familiar problems in the philosophy of science. We have examined issues involved in improving generalisation in certain types of neural network models. Regularisation techniques illuminate the relation between simplicity and generalisation for the models discussed which, although restricted, have wide practical application. It can be hoped that empirical concept formation and its relation to generalisability will be further illuminated by the ideas of neural computation in future.

References

[1] YASER S. ABU-MOSTAFA. *The Vapnik-Chervonenkis dimension: Information versus complexity in learning.* Neural Computation, 1:312–317, 1989.

[2] IGOR ALEKSANDER. *Myths and realities about neural computing architectures.* In M. Reeve and S. E. Zenith, editors, Parallel Processing and Artificial Intelligence, chapter 1. John Wiley & Sons, 1989.

[3] ROBERTO BATTITI. *Accelerated backpropagation learning: Two optimization methods.* Complex Systems, 3:331–342, 1989.

[4] ERIC B. BAUM and DAVID HAUSSLER. *What size net gives valid generalization?* Neural Computation, 1:151–160, 1989.

[5] SUE BECKER and YANN LE CUN. *Improving the convergence of back-propagation learning with second order methods.* In David Touretzky, Geoffrey Hinton, and Terrence Sejnowski, editors, Proceedings of the 1988 Connectionist Models Summer School, pages 29–37, 1988.

[6] J. O. BERGER. *Statistical Decision Theory and Bayesian Analysis.* Springer-Verlag, 1985.

[7] It may be an unusual feature of this prospect that the magnetic field correlates so well with gold occurrence. In general gravitational, geological and geochemical etc. data sets would need to be incorporated.

[7] WRAY L. BUNTINE and ANDREAS S. WEIGEND. *Bayesian back-propagation.* Submitted for Publication, July 1991.

[8] G. CYBENKO. *Approximation by superpositions of a sigmoidal function.* Mathematics of Control, Signals, and Systems, 2:303–314, 1989.

[9] JOHN DENKER, DANIEL SCHWARTZ, BEN WITTNER, SARA SOLLA, RICHARD HOWARD, LAWRENCE JACKEL, and JOHN HOPFIELD. *Large automatic learning, rule extraction, and generalization.* Complex Systems, 1:877–922, 1987.

[10] JERZY GIEDYMIN. *Strength, confirmation, compatibility.* In Mario Bunge, editor, The Critical Approach to Science and Philosophy, chapter 4. Collier Macmillan, 1964.

[11] NELSON GOODMAN. *Safety, strength, simplicity.* Philosophy of Science, 28:150–151, 1961.

[12] ROBERT HECHT-NIELSEN. *Neurocomputing.* Addison-Wesley, 1989.

[13] JOHN HERTZ, ANDERS KROGH, and RICHARD G. PALMER. *Introduction to the Theory of Neural Computation.* Addison-Wesley, 1991.

[14] K. HORNIK, M. STINCHCOMBE, and H. WHITE. *Multilayer feedforward networks are universal approximators.* Neural Networks, 2:359–366, 1989.

[15] Y. LE CUN, B. BOSER, J. S. DENKER, D. HENDERSON, R. E. HOWARD, W. HUBBARD, and L. D. JACKEL. *Backpropagation applied to handwritten zip code recognition.* Neural Computation, 1:541–551, 1989.

[16] YANN LE CUN, JOHN S. DENKER, and SARA A. SOLLA. *Optimal brain damage.* In David S. Touretzky, editor, Advances in Neural Information Processing Systems 2, pages 598–605. Morgan Kaufmann, 1990.

[17] DAVID J. C. MACKAY. *Bayesian Modelling and Neural Networks.* PhD thesis, Computation and Neural Systems, California Institute of Technology, Pasadena, CA, 1991.

[18] DERRICK NGUYEN and BERNARD WIDROW. *The truck backer-upper.* In INNC 90 Paris, pages 399–407. The International Neural Network Society, Kluwer Academic Publishers, 1990.

[19] WILLIAM H. PRESS, BRIAN P. FLANNERY, SAUL A. TEUKOLSKY, and WILLIAM T. VETTERLING. *Numerical Recipes in C.* Cambridge University Press, 1988.

[20] D. E. RUMELHART, G. E. HINTON, and R. J. WILLIAMS. *Learning internal representations by error propagation.* In D. E. Rumelhart and J. L. McClelland, editors, Parallel Distributed Processing, volume 1, chapter 8, pages 318–362. MIT Press, 1986.

[21] A. N. TIKHONOV and V. Y. ARSENIN. *Solutions of Ill-Posed Problems.* John Wiley & Sons, 1977.

[22] V. N. VAPNIK. *Estimation of Dependencies Based on Empirical data.* Springer-Verlag, 1982.

[23] V. N. VAPNIK and A. YA. CHERVONENKIS. *On the uniform convergence of relative frequencies of events to their probabilities.* Theor. Prob. Appl., 16:264–280, 1971.

[24] ANDREAS S. WEIGEND, DAVID E. RUMELHART, and BERNADO A. HUBERMAN. *Generalization by weight-elimination with application to forecasting.* In Richard P. Lippmann, John E. Moody, and David S. Touretzky, editors, Advances in Neural Information Processing Systems, pages 875–882. Morgan Kaufmann, 1991.

[25] P. M. WILLIAMS. *A Marquardt algorithm for choosing the step-size in backpropagation learning with conjugate gradients.* Technical report, University of Sussex, 1990.

Logic, Methodology and Philosophy of Science IX
D. Prawitz, B. Skyrms and D. Westerståhl (Editors)

THREE LEVELS OF INDUCTIVE INFERENCE

PETER GÄRDENFORS

Cognitive Science, Department of Philosophy, University of Lund, Sweden

1. Three perspectives on observations

One of the most impressive features of human cognitive processing is our ability to perform inductive inferences. Without any perceived effort, we are prepared, sometimes with great confidence, to generalize from a very limited number of observations.

One of the goals of cognitive science in general, and artificial intelligence in particular, is to provide computational models of different aspects of human cognition. So how can we mechanize induction? How can we even *hope* to capture the ease and assurance of the human inductive competence in a model confined by the thoroughness and strictness of computation?

It is commonplace that induction is going from single observations to generalizations. But this statement loses its air of triviality if one takes seriously, as I propose to do, the question of *what* an observation is. It is surprising that this question has received very little attention within the philosophy of science.[1] The key argument of this article is that there is no unique way of characterizing an observation. Indeed, I shall distinguish three levels of accounting for observations (or, since all levels may be adopted at the same time, they may as well be called perspectives):

1. The linguistic level: This way of viewing observations consists of describing them in some specified language. The language is assumed to be equipped with a fixed set of primitive predicates and the denotations of these predicates is taken to be known. As will be argued in Section 2, the linguistic approach is a central part of logical positivism.

2. The conceptual level: On this level observations are not defined in relation to some language but characterized in terms of some underlying 'conceptual space'. The conceptual space, which is more or less connected to perceptual mechanisms, consists of a number of 'quality dimensions'. Induction is here

[1] One notable exception is Shapere (1982). See Section 4.1.

seen as closely related to *concept formation*. According to the conceptual perspective, inductive inferences show prototype effects, in contrast to the linguistic perspective which operates on Aristotelian concepts (cf. Smith & Medin 1981).

3. The subconceptual level: Observations are here characterized in terms of inputs from sensory receptors. The observations are thus described as occurring before conceptualization. The inductive process is seen as establishing connections between various types of inputs. One currently popular way of modelling this kind of process is by using neural networks.

My main objective in this article is to argue that depending on which approach to observations is adopted, thoroughly different considerations about inductive inferences will come into focus.[2] In my opinion there is a multitude of aspects of inductive reasoning and not something that can be identified as *the* problem of induction. The upshot is that there is no canonical way of studying induction. What is judged to be the salient features of the inductive process depends to a large extent on *what* an observation is considered to be.

2. The linguistic level

2.1. Observation statements and the riddles of induction

The most ambitious project of analyzing inductive inferences during this century has been that of the logical positivists. According to their program, the basic objects of scientific inquiry are *sentences* or *statements* in some formal or natural language. An *observation* is a particular type of statement. The observational statements are supposed to be furnished to the reasoner by uncorrigible perceptual mechanisms.

Ideally, the scientific language is a version of first order logic where a designated subset of the atomic predicates represent *observational* properties and relations. These observational predicates are taken to be *primitive* notions. This means that when it comes to inductive reasoning, all observational predicates are treated in the same way. For example, Carnap (1950, Section 18B) requires that the primitive predicates of a language be logically independent of each other. The advantage of this, from the point of view of the positivists, is that induction then becomes amenable to *logical analysis* which, in the purist form, is the only tool admitted.

However, it became apparent that the methodology of the positivists led to serious problems for their analysis of induction. The most famous ones are

[2] I cannot talk about three ways of *describing* observations, because the very notion of 'describing' presumes the linguistic level.

Goodman's (1955) "riddle of induction" and Hempel's (1965) "paradox of confirmation". In Gärdenfors (1990), I have analyzed these problems using the conceptual approach to induction. One conclusion to be drawn from the analysis is that these problems show that the linguistic level is not sufficient for a complete understanding of inductive reasoning.

2.2. An example from machine learning

The most common type of knowledge representation within the AI tradition is 'propositional' in the sense that it is based on a set of rules or axioms together with a data base. In this representation, the 'facts' in the data base correspond to observations. The rules and the data base are combined with the aid of a theorem prover or some other inference mechanism to produce new rules or facts. The basic and the derived 'knowledge' is then the material on which a planning or problem solving program can operate.

The propositional form of knowledge representation used in mainstream AI is thus well suited to the positivist tradition. And when implementing inductive inference mechanisms on a computer, this has been the dominating methodology. A rather typical example of the linguistic perspective within AI is the chapter on induction in Genesereth and Nilsson (1987). They assume (pp. 161-162) that there is a set Γ of sentences which constitutes the *background theory* and a set Δ of data (which is to be generalized). It is required that Γ does not logically imply Δ. They then define a sentence ϕ to be an *inductive conclusion* if and only if (1) ϕ is consistent with $\Gamma \bigcup \Delta$ and (2) the hypothesis ϕ *explains* the data in the sense that $\Gamma \bigcup \{\phi\}$ logically entails Δ.[3]

In general, Genesereth and Nilsson view inductive inferences as problems of *concept formation*:[4]

> The data assert a common property of some objects and deny that property to others, and the inductive hypothesis is a universally quantified sentence that summarizes the conditions under which an object has that property. In such cases, the problem of induction reduces to that of forming the concept of all objects that have that property. (1987, p. 165).

They define a *concept-formation* problem as a quadruple $\langle P, N, C, \Lambda \rangle$, where P is a set of positive instances of a concept, N is a set of negative instances,

[3]Note that this criterion can only be seen as supplying necessary but not sufficient conditions. For example, for any sentence α such that Γ logically entails α, it holds that $\neg\alpha \vee \wedge\Delta$ (where $\wedge\Delta$ is the conjunction of all elements in Δ) is consistent with $\Gamma \bigcup \Delta$ and $\Gamma \bigcup \{\neg\alpha \vee \wedge\Delta\}$ logically entails Δ.

[4]For a similar approach, see Michalski and Stepp (1983).

C is a set of concepts to be used in defining the concept, and Λ is a language to use in phrasing the definition.

For example, consider the problem of identifying a class of cards from a regular card deck. The language for problems of this kind of problem is taken to be a standard first order language with a set of basic predicates like 'numbered', 'face', 'odd', 'jack', 'four', 'red', and 'spade'. The set P consists of those cards we know belong to the class and N consists of the cards we know are not members of the class. The 'conceptual bias' C determines which among the basic predicates are allowed to be used in forming the inductive rule determining the class. For example, only 'numbered', 'face', 'black' and 'red' may be allowed when describing the rule, so that 'bent' and 'played with the left hand', among others, are excluded. Λ, finally, is the 'logical bias' which restricts the logical form of the rule that determines the class. For instance, only definitions consisting of conjunctions of basic predicates may be allowed.

Using the notion of a concept-formation problem $\langle P, N, C, \Lambda \rangle$, Genesereth and Nilsson develop an algorithm for performing inductive inferences satisfying the constraints given by C and Λ. A central notion in their construction is that of the 'version space' for the concept-formation problem which consists of all rules that are satisfied by all the positive instances in P, but by no instance in N. The algorithm works by pruning the version space as new positive and negative instances are added.

Even though AI researchers have had some success in their attempts to mechanize induction, it is clear that their methodology suffers from the same general problems as the linguistic level in general. The enigmas of induction that have been unearthed by Goodman, Hempel and others are applicable also to the induction programs in recent mainstream AI.

Trying to capture inductive inferences by an algorithm also highlights some of the general limitations of the linguistic perspective. The programs work by considering the applicability of various logical combinations of the atomic predicates. But the epistemological origin of these predicates are never discussed. Even though AI researchers are not actively defending the positivist methodology, they are following it implicitly by treating certain predicates as observationally, or at least externally, given. However, the fact that the atomic predicates are assumed as granted from the beginning means that much inductive processing has already been performed.

I agree with Genesereth and Nilsson (1987) that induction is *one form* of concept formation, but their sense of concept formation is *much too narrow.* We not only want to know how observational predicates should be combined in the light of inductive evidence, but, much more importantly, *how the basic predicates are inductively established* in the first place. This problem has,

more or less, been swept under the rug by the logical positivists and their programming followers in the current AI tradition. Using logical analysis, the prime tool of positivism and AI, is of no avail for these forms of concept formation. In brief, the linguistic approach to induction sustains no creative inductions, no genuinely new knowledge, and no conceptual discoveries. To do this, we have to go below language.

3. The conceptual level

What I see as the source of the troublesome cases for the linguistic approach, like Hempel's and Goodman's riddles, is that if we use *logical relations* alone to determine which inductions are valid, the fact that all predicates are treated on a par induces *symmetries* which are not preserved by our understanding of the inductions: "Raven" is treated on an equal basis with "non-raven", "green" with "grue" etc. What we need is a non-logical way of distinguishing those predicates that may be used in inductive inferences from those that may not.

There are several suggestions for such a distinction in the literature. One idea is that some predicates denote "natural kinds" or "natural properties" while others don't, and it is only the former that may be used in inductions (cf. Quine 1969 and Gärdenfors 1990). Natural kinds are normally interpreted realistically, following the Aristotelian tradition, and thus assumed to represent something that exists in reality independently of human cognition. However, when it comes to inductive inferences it is not sufficient that the properties exist out there somewhere, but we need to be able to *grasp* the natural kinds with our minds. In other words, what is required to understand induction, as performed by humans, is a conceptualistic or cognitive analysis of *observations* of natural properties. Thus we are back at the problem of saying what an observation is, but now on the conceptual level.

3.1. Conceptual spaces

One of the primary functions of concepts is to structure the perceptual sensory inflow into categories that are useful for planning, reasoning and other cognitive activities. The concepts we use are not independent of each other but can be structured into *domains*: Spatial concepts belong to one domain, kinship relations to another, concepts for sounds to a third, and so on.[5]

[5]Cf. Langacker's (1986) use of 'domains' in cognitive semantics.

The epistemological framework for a domain of concepts I propose to call a *conceptual space*. A conceptual space consists of a number of *quality dimensions*. I have no exhaustive definition of a what a quality dimension is, but must confine myself to giving examples. Some of the dimensions are closely related to what is produced by our sensory receptors like space, pitch, temperature and color, but there are also quality dimensions that are of an abstract non-sensory character like time and dimensions of social relations. The dimensions of a conceptual space are taken to be cognitive and infra-linguistic in the sense that we (and other animals) can represent the properties of objects, for example when planning an action, without presuming an internal language in which these properties are expressed.

The notion of 'space' should be taken in the mathematical sense. It is assumed that each of the quality dimensions is endowed with certain *topological* or *metric* structures. For example, 'time' is a one-dimensional structure which we conceive of as being isomorphic to the line of real numbers.[6] Similarly, 'weight' is one-dimensional with a zero point, isomorphic to the half-line of non-negative numbers. The topological structure of the color space is described in Gärdenfors (1990). Some quality dimensions have a discrete structure, i.e., they merely divide objects into classes, e.g., the sex of an individual.[7]

Let us now turn to the problem of identifying observations on the conceptual level. Using the notion of conceptual spaces, an observation can be defined as *an assignment to an object of a location in a conceptual space*. For example, the observation that is described on the linguistic level as "x is red" is expressed on the conceptual level by assigning x a point in color space. Since natural languages only divide the color domain into a finite number of categories the information contained in the statements that x is red is much less precise than the information furnished by assigning x a location in color space. In this sense, the conceptual level allows much richer devices for reporting observations.

3.2. Concept formation

On the conceptual level one can distinguish between two types of inductive processes. One is closely related to *concept formation*: In Gärdenfors (1990),

[6]To some extent the representation of time is culturally dependent, so that other cultures have a different time dimension as a part of their cognitive structure. Cf. Gärdenfors (1992) for a discussion of how this influences the structure of language.

[7]Discrete dimensions may also have additional structure as, for example, in kinship or biological classifications. The topology of discrete dimensions is further discussed in Gärdenfors (1990).

I analysed 'natural properties' in terms of conceptual spaces. The key idea is that a natural property is identified with a *convex region* of a given conceptual space. Via the notion of 'convexity' the topological properties of the quality dimensions are utilized. A convex region is characterized by the criterion that for every pair o_1 and o_2 of points in the region, all points *between* o_1 and o_2 are also in the region. The definition presumes that the notion of 'between' is meaningful for the relevant dimensions. This is, however, a rather weak assumption which demands very little of the underlying topological structure.

On the basis of this criterion of natural properties, it is now possible to formulate a constraint on induction, which is helpful in solving the conundrums of the linguistic approach:

(C) Only properties corresponding to a convex region of the underlying conceptual space may be used in inductive inferences.

It is only proposed that convexity is a necessary condition, but perhaps not sufficient, for a property to count as natural and thus allowed in inductive inferences. I argue in Gärdenfors (1990) that criterion (C) solves many of the problems of induction that appear on the linguistic level. Furthermore, the criterion can also be used to explain the prototype effects that are exhibited by natural concepts (Rosch 1975, 1978, Gärdenfors 1991).

An assumption that is within reach now is that most basic words in natural languages denote convex regions in some conceptual space. (This assumption can be made even if we have no idea of what the dimensions are or how their topology looks like). From the assumption it follows that the assignment of meanings to the expressions on the linguistic level is far from arbitrary. On the contrary, the semantics (and to some extent even the grammar) of the linguistic constituents is severely constrained by the structure of the underlying conceptual space. This thesis is anathema for the Chomskian tradition within linguistics, but, as a matter of fact, it is one of the central tenets of the recently developed 'cognitive' linguistics.[8]

As another sign of the importance of the conceptual level, I submit that most of scientific theorizing takes place at this level. Determining the relevant dimensions involved in the explanation of a phenomenon is a prime scientific activity. And once the conceptual space for a theory has been established, theories, in the form of *equations*, that connect the dimensions can be proposed and tested.[9]

[8]Cf. Lakoff (1987) and Langacker (1986).

[9]For a discussion of the role of conceptual spaces in science, see Gärdenfors (1990) and (1991).

3.3. The origin of quality dimensions

The second kind of inductive process on the conceptual level concerns how the quality dimensions of the conceptual spaces are determined. There does not seem to be a unique origin of our quality dimensions. Some of the dimensions are presumably *innate* and to some extent hardwired in our nervous system, as for example color, pitch, and probably also ordinary space. These subspaces are obviously extremely important for basic activities like finding food and getting around in the environment.

But from the point of view of induction, the dimensions that are *learned* are of greater interest. Learning new concepts often involves expanding one's conceptual space with new quality dimensions. 'Volume' is an example here. According to Piaget's 'conservation' experiments with five year olds, small children do not make a distinction between the height of a liquid and its volume. The conservation of volume, which is part of its conceptual structure, is something that must be learned. In general, introducing *new* quality dimensions is a much more advanced form of induction than concept formation *within* a given conceptual space.

A similar process occurs within *science*. By introducing theoretically precise, non-psychological quality dimensions, a scientific theory may help us find new inductive inferences that would not be possible on the basis of our subjective conceptual spaces alone. As an example, consider Newton's distinction between weight and mass, which is of crucial importance for the development of his celestial mechanics, but which has no correspondence in human psychology. It seems to me that the cognitive construction involved in Newton's discovery of the distinction between mass and weight is of the same nature as when a child discovers the distinction between height and volume. Another example of a scientifically introduced dimension is the distinction between temperature and heat, which is central for thermodynamics. In contrast, human perception of heat is basically determined by the amount of heat transferred from an object to the skin rather than by the temperature of the object.

In order to give another illustration of how the scientific process is helpful in constructing the underlying conceptual space, thereby providing an understanding of how concepts are formed, I shall briefly present the phonetic identification of *vowels* in various languages. According to phonetic theory, what determines a vowel are the relations between the basic frequency F_0 of the sound and its formants (higher frequencies that are present at the same time). In general, the first two formants F_1 and F_2 are sufficient to identify a vowel. This means that the coordinates of two-dimensional space spanned by F_1 and F_2 (in relation to a fixed basic pitch F_0) can be used as a fairly ac-

curate description of a vowel. Fairbanks and Grubb (1961) investigated how people produce and recognize vowels in 'General American' speech. Figure 1 summarizes some of their findings.

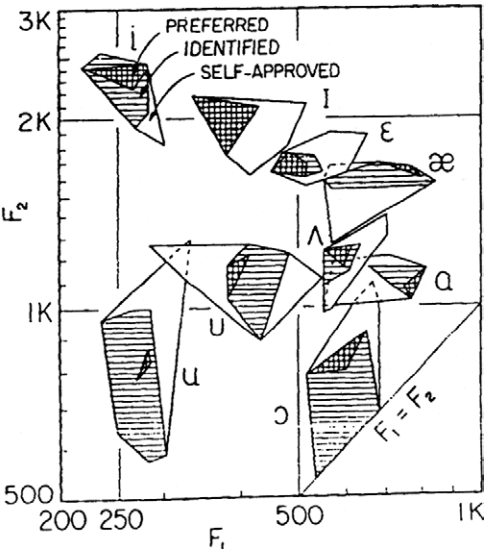


Figure 1: Frequency areas of different vowels in the two-dimensional space generated by the first two formants. Values in cps. (From Fairbanks and Grubb (1961))

The scales of the abscissa and ordinate are the logarithms of the frequencies of F_1 and F_2 (the basic frequency of the vowels was 130 cps). A *self-approved* vowel is one that was produced by the speaker and later approved of as an example of the intended kind. An *identified* sample of a vowel is one that was correctly identified by 75% of the observers. The *preferred* samples of a vowel are those which are "the most representative samples from among the most readily identified samples" (Fairbanks and Grubb 1961, p. 210).

As can be seen from the diagram, the preferred, identified and self-approved examples of different vowels form convex subregions of the space determined by F_1 and F_2 with the given scales. As in the case of color terms, different languages carve up the phonetic space in different ways (the number of vowels identified in different languages varies considerably), but I conjecture again that each vowel in a language will correspond to a convex region of the formant space.

The important thing to note in this example is that identifying F_1 and F_2 as the relevant dimensions for vowel formation is a phonetic *discovery*. We had the concepts of vowels already before this discovery, but the spatial analysis makes it possible for us to understand several features of the classifications of vowels in different languages.

The last examples of how science introduces new quality dimensions for concept formation highlight one fundamental problem for this second type of inductive process on the conceptual level: *Where do the dimensions and their topology come from?* According to Popper's terminology this kind of process belongs to the 'context of discovery'. Within traditional philosophy of science, it has in general been thought to be futile to construct a mechanistic procedure for generating scientific discoveries of this kind. However, when it comes to human learning and concept formation, the prospects may not be so hopeless after all. This will be the topic of next section where inductive processes below the conceptual level will be considered.

4. The subconceptual level

4.1. Observations by receptors

In the most basic sense an observation is what is received by our sensory organs. In this sense, an observation can be identified with what is received by a set of *receptors*. For human beings, these inputs are provided by the sensory receptors, but one can also talk of a machine having observations of this kind via some measuring instruments serving as receptors. The receptors provide 'raw' data in the sense that the information is not assumed to be processed in any way, neither in a conceptual space, nor in the form of some linguistic expression.

Within the philosophy of science, it is important to make a distinction between *perception* and *observation*. As Shapere (1982) points out, the term 'observation' plays a double role for the traditional philosopher of science. He writes:

> On the one hand, there is the *perceptual* aspect: "observation", as a multitude of philosophical analyses insist, is simply a special kind of perception, usually interpreted as consisting in the addition to the latter of an extra ingredient of focussed attention. ... On the other hand, there is the *epistemic* aspect of the philosopher's use of 'observation': the *evidential* role that observation is suppose to play in leading to knowledge or well-grounded belief or in supporting beliefs already attained. (Shapere 1982: 507-508)

Within the empiricist tradition of philosophy of science, the two uses of 'observation' have been confounded. However, in modern science it is obvious that it is the epistemic aspect of observation that is of importance. As Shapere (1982: 508) formulates it:

> Science is, after all, concerned with the role of observation as evidence, whereas sense-perception is notoriously untrustworthy .... Hence, with the recognition that information can be received which is not directly accessible to the senses, *science has come more and more to exclude sense-perception as much as possible from playing a role in the acquisition of observational evidence*; that is, it relies more and more on other appropriate, but dependable, receptors.

Given that we are focussing on the epistemic aspect of observations, let us then consider induction on the subconceptual level. How do we distill sensible information from what is received by a set of receptors? Or, in other words, how do we make the transition from the subconceptual to the conceptual and the linguistic levels? These questions indicate the kinds of inductive problems that occur on the subconceptual level.

The basic problem is that the information received by the receptors is too rich and unstructured. What is needed is some way of transforming and organizing the input into a form that can be handled on the conceptual or linguistic level. There are several methods for treating this kind of problem. Within psychology, various methods of *multidimensional scaling* have been developed.

For example, in Shepard's (1962a,b) algorithm, the input data is assumed to contain information about the relative distances between n points in some unknown space. The distances between the points are not expressed in metrical terms, but only given as a rank order of the $n(n-1)/2$ distances between the n points. Any such rank order can be represented in a space of $n-1$ dimensions. Shepard's algorithm starts out from a representation in such a space and then successively reduces the dimensionality until no further dimensions can be eliminated without a substantial disagreement between the rank order generated by the metric assignment and the original rank order. For many empirical areas the initial data can be reduced to a space with two or three dimensions.[10] These dimensions can then function as a basis for concept formation according to the outline in Section 3.2.

[10]Cf. Shepard (1962b) for several examples of the results of the procedure.

4.2. Induction with the aid of neural networks

In this subsection a different method for going from the subconceptual to the conceptual level will be outlined. The mechanisms for the method are based on *neural networks*. In a neural network, the receptors and the information they receive can be identified with a set of *input neurons* and their *activity values*. This set of values will be called the *input vector*. In interesting cases there is a large number of input neurons which means that the dimensionality of the input vector is very high. The purpose of an inductive method at this subconceptual level is to reduce the complexity of the input information in an efficient and systematic way.

The neural network model I will be outlining here is based on Kohonen's (1988) *self-organizing feature maps*. The distinguishing property of these maps is that they are able to describe the topological relations of the signals in the input vector using something like a conceptual space with a small number of dimensions. Basically, the mapping can be seen as reducing the dimensionality of the input vector.

A self-organizing feature map is a neural network which consists of an input vector that is connected to an output array of neurons. In most applications, this array is one- or two-dimensional, but in principle it could be of any number of dimensions. The essential property of the network is that the connections between the neurons in the array and the learning function are organized in such a way that *similarities* that occur among different input vectors are *preserved* in the mapping, in the sense that input vectors that have common features are mapped onto *neighbouring* neurons in the map. The degree of similarity between two input vectors is determined by some *distance* measure (which normally is the standard Euclidean metric, but many metrics are possible to use).

In other words, the mapping from the input vector to the array preserves the topological relations while reducing the dimensionality of the representation space. The low-dimensional 'feature map' that results as an output of the process can be viewed as a conceptual space in the sense of the preceeding section. The mapping is *generated* by the network itself via the learning mechanism of the network. In practice, it normally takes quite a large number of learning instances before the network stabilizes enough so that further changes can be ignored.[11]

The mechanism is best illustrated by a couple of artificial examples taken from Kohonen (1988). In figures 2 and 3 the input vectors were assumed to

[11]New learning by instances that do not follow the previous frequency pattern can always change the mapping function. This means that it is impossible to talk about a 'final' mapping function.

be uniformly distributed over a triangular area. In the network represented in figure 2, the output array was one dimensional, i.e., the output neurons were arranged along a line. The number of neurons on this line is fixed. Any input from the triangular space results in some activities in the neurons in this line. Figure 2 shows the inverse mapping of the input vectors which resulted in the highest activities of single neurons in the line, where each dot corresponds to an output neuron. As can be seen, the mapping preserves relations of similarity, and, furthermore, there is a tendency of the line trying to 'cover' as much as possible of the surface, in the sense that the distance between any point in the surface and the line being as small as possible.

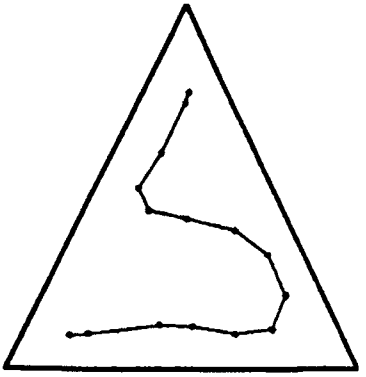


Figure 2: Distribution of weight vectors on a linear array of neurons. (From Kohonen (1988), p. 135.)

In figure 3, the corresponding network contains an output array that is two-dimensional, with the neurons arranged in a square. Figure 3 again shows the inverse mapping, indicating which neurons in the input space produce the greatest responses in the output square. As can be seen, the inverse mapping represents a deformation of the output array that preserves topological relations as much as possible.

Figure 4 shows an example of how the network self-organizes in learning a mapping. The initial values of the mapping were selected so that there was a random mapping from a circular region of the input triangle to a linear array of output neurons. The network was then fed with a number of input vectors, randomly selected from the full triangle. The sequence of figures

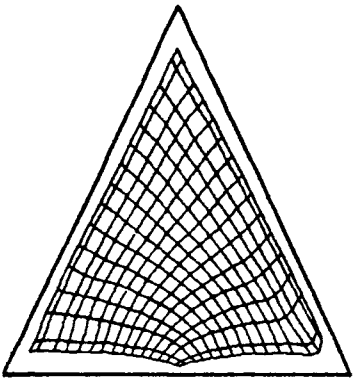


Figure 3: Distribution of weight vectors on a rectangular array of neurons. (From Kohonen (1988), p. 135.)

indicate how the mapping is improved over time, where the numbers below the figures represent the number of learning trials.

These examples are artificial in that we know the initial distribution of input vectors, which furthermore is of low dimensionality. In real applications, the dimensionality of the input space is high and its topology is unknown. However, it can be shown, at least when the output array is one-dimensional, that the mapping in the limit (i.e., after infinitely many learning instances) will preserve as much as possible of the topological structure of the input space.[12]

Kohonen's goal in using the maps is not limited to inductive inference only but representation of information in general. He writes:

> Economic representation of data with all their interrelationsships is one of the most central problems in information sciences, and such an ability is obviously characteristic of the operation of the brain, too. In thinking, and in the subconscious information processing, there is a general tendency to compress information by forming *reduced representations* of the most relevant facts, without loss of knowledge about their interrelationsships (Kohonen 1988, p. 119).

[12]For a more precise statement of this result and a proof see Kohonen (1988), pp. 145-148.

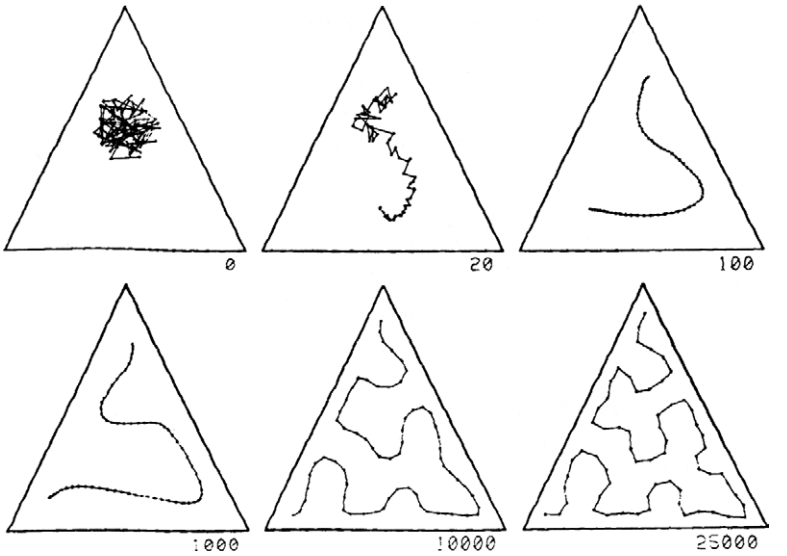


Figure 4: Distribution of weight vectors on a rectangular array of neurons. (From Kohonen (1988), p. 135.)

In collaboration with Christian Balkenius, I have started working on a general architecture for a neural network system that utilizes selforganizing feature maps to perform inductive inferences. The overall architecture of the inductive network is depicted in Figure 5. The input receptors are divided into a small number of subsets (in the figure there are two such subsets). The purpose of this division is to group together receptors that contain information about 'the same' feature, so for example, visual receptors belong to one group, auditory receptors to another etc. When the network is applied, the decision about how the set of receptors should be grouped must be made by the user. But this is about the only thing she has to decide except for some parameter settings, the network then performs the rest of the inductive inferences.

The input vectors are then mapped onto one Kohonen surface each. In figure 5 these are depicted as one-dimensional lines, but they may as well be two- or three-dimensional surfaces. In the figure, there are only two Kohonen surfaces, but they may, of course, be more than two depending on how the input receptors are grouped into subspaces. One of the surfaces may be a purely classificatory space, representing 'names' of the categories that are

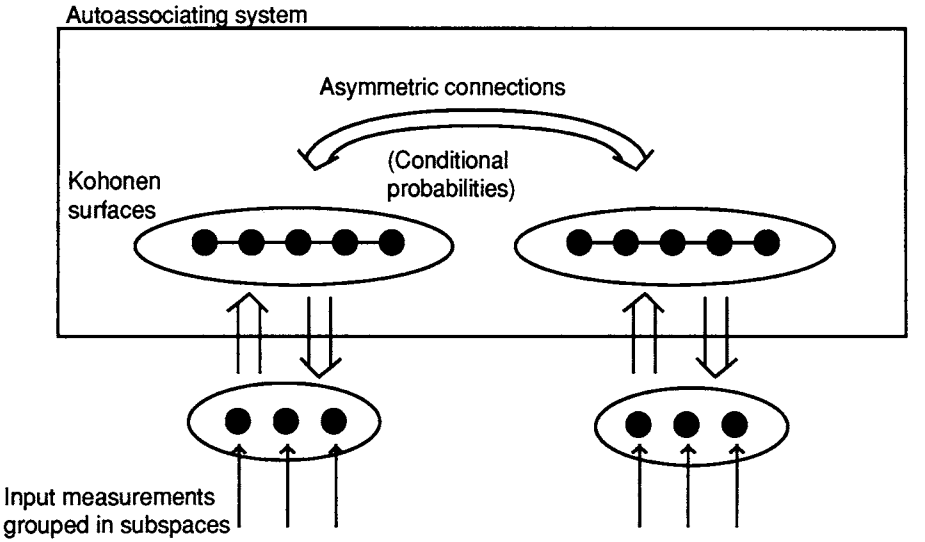


Figure 5: Architecture of inductive neural network

identified by the network.[13]

The Kohonen surfaces are then pairwise connected by asymmetric connections between the neurons in the two surfaces. The connections are total in the sense that each neuron on one surface is connected to all neurons on the other surface. The learning rule for these connections functions in such a way that the strength of the connection c_{ij} between a neuron x_i on one surface and a neuron y_j on another reflects the conditional probability (estimated from the learning examples) that y_j be activated given that x_i is activated.[14] The connections vary between -1 and +1 and obtain the extreme values only when x_i and y_j are never and always, respectively, activated together. In a sense, the network performs implicit computations of the inductive statistics.[15]

[13]The linguistic form of the names has, of course, to be provided by the user.

[14]The mathematical form of the connections are closely related to Hintikka's (1969, p. 328) measures of 'evidential support,' in particular the measure defined in his equation (27)*.

[15]The sense in which neural networks perform implicit statistic inferences can be made very precise. For example, see Lippman (1987) for a presentation of some of the results connecting least mean square and maximal likelihood solutions to the computations of

Once the learning phase has been completed, relating input receptors to Kohonen surfaces and these surfaces to each other, it is then possible to use the network to classify new objects. By feeding the system with a *partial* input vector, for example, the values for one of the subspaces, the network can then compute the *expected* values for all the other receptors and the expected locations for all the Kohonen surfaces. In this way the network *guesses* the unknown properties of the object it has only received partial information about. The network is thus able to *generalize* from earlier experience and make inductive inferences using the connections between the different Kohonen spaces.

Christian Balkenius and I have done some preliminary experiments using a network with the architecture that has been outlined here. So far, the results seem very promising. One example concerns a classification of 44 individual parasitical wasps. For each individual the values of twelve variables are supplied together with the species name it was assigned by an entomologist. These variables represent different kinds of data, some binary, some ordinal, and some real-valued. After discussions with the entomologist, we divided the input variables into four groups: One consisting of five variables on proportions of sizes of various body parts, the second consisting of four other morphological variables, the third consisting of three ecological variables, and the fourth simply a coding for the species name. Each of these four variable groups was mapped onto a one-dimensional Kohonen surface (i.e., a line), and the four surfaces were pairwise connected by asymmetric connections as described above.

After training the network by showing it the individual input vectors a number of times, it can be tested by feeding in all input variables for a particular individual, except for its species categorization and compare the output with that of the entomologist. In our tests, the network makes very few errors in classifying the 44 wasps. The results are, to some extent, dependent on the number of neurons on each Kohonen surface. If we allow as many as 50 neurons, which is more than the number of wasps, then the network can learn to correctly classify every individual in the sample. However, if there are only 20 neurons on the Kohonen surfaces, then it correctly classifies 43 out of 44. One can, of course, also feed in data (except for names of the species) about individuals that were not present in the learning sample. The species names that are produced as outputs from the network seem to have a high degree of validity. However, the performance of the network still awaits more detailed testing against empirical material.

The methodology of founding a classification on a large number of numerical data is similar to so called numerical (or phenetic) taxonomy (Sokal and

---

neural networks.

Sneath 1963, Ruse 1973) In my opinion, however, the mechanisms behind the classifications obtained by the neural network model and their biological validity are superior to what is achieved in numerical taxonomy. One essential difference is that, as a matter of methodology, all variables are treated as having equal value in the process of computing the numerical taxonomy. However, even if it may seem that the same holds of the inputs to the neural network described above, the variables are *implicitly* assigned different importance via the influence they have on the topology of the Kohonen surfaces that emerge during the learning period.

What are then the drawbacks of using neural networks of the type described here for inductive processes? A fundamental epistemological problem is that even if we know that the network will generate Kohonen surfaces that perform the right kind of job, we may not be able to 'describe' *what* the emerging dimensions represent. Even if we, for example, know that a system consisting of four one-dimensional Kohonen surfaces provides a perfect classification of a population of parasitical wasps, this may not help us in *interpreting* the 'meaning' of the surfaces, i.e., what overall features of the wasps that they *represent.* In other words, we may not be able to make the transition between the subconceptual level and the conceptual level. This kind of level problem is ubiquitous in applications of neural networks for learning purposes. The upshot is that a future theory of neural networks must somehow bridge the gap of going from the subconceptual level to the conceptual level. We may account for the information provided at the subconceptual level in term of a dimensional space with some topological structure, but there is no general recipe for determining what is the *conceptual* meaning of the dimensions of the space.

Other problems concern more methodological issues. How should the input variables be grouped before they are mapped onto different Kohonen surfaces? How does one decide how many dimensions to use in the target surface of the mapping from the input variables? These kinds of problems are found everywhere in science and they are not particular to using neural networks for the classifications.[16] In fact, the methodological problems involved in the procedure presented here seem to be smaller than the problems one encounters for other classification methods.

5. Conclusion: What is induction?

Where on the three levels that have been described here is real induction to be found? The answer is: nowhere and everywhere. The main thesis of

[16]For instance, very similar problems would be encountered when applying the multidimensional scaling methods that were outlined above.

this article is that there are several kinds of inductive processes. Depending on what perspective one takes on observations, different ways of generalizing the observations become relevant. Traditional philosophy of science has concealed these distinctions by neglecting the conceptual and subconceptual levels. For a complete account of induction, all three levels must be mustered.

What is the relation between the three levels? I hope it has become clear from my presentation that I do not view the three levels as being in conflict with each other. They should rather be regarded as three *perspectives* on observations that complement each other. Different aspects of inductive processes need to be explained on different levels. By disregarding some level one restricts the possibilities for understanding the mechanisms of inductive reasoning.

A three-level theory of cognitive representation that is related to the one proposed in this paper has been suggested by Harnad (1987) as a way of analysing problems in categorical perception.[17] He calls his lowest level the *iconic* representation (IR), "being an analog of the sensory input (more specifically, of the proximal projection of the distal stimulus object on the device's transducer surfaces)" (Harnad 1987, p. 551). The IRs are analog mappings which "faithfully preserve the iconic character of the input for such purposes as same-different judgements, stimulus-matching, and copying" (p. 552). It is obvious that this form of representation corresponds to what I have here called the sub-conceptual level.

The middle level Harnad calls *categorical* representation (CR). This representation eliminates most of the raw input structure and retains what is invariant in the produced categorization: "Whereas IRs preserve analog structure relatively indiscriminately, CRs selectively reduce input structure to those invariant features that are sufficient to subserve successful categorization (in a given context)" (p. 553).

Again, it is clear that this level corresponds to the conceptual level of this article. Unfortunately, Harnad says very little about how the categorization is acheived, except that it is some kind of filtering process. Furthermore, he provides no account of the structure of the categorical representation, with the exception that he presumes that categorization is to a certain extent context dependent. I believe that it is a strength of the theory of conceptual spaces outlined in Section 3 that it has strong, and to a large extent testable, implications for categorization and concept formation.

The highest level in Harnad's triad is *symbolic* representation (SR), which naturally corresponds to the linguistic level of this paper. He introduces

[17]This theory was brought to my attention by Paul Hemeren when the present article was almost finished.

a "description system" (p. 554), the expressions of which assign category membership to experiences. The description system presumes that the CRs are already *labeled*:

> Instead of constructing an invariance filter of the basis of direct experience with instances, it operates on *existing* labels, and constructs categories by manipulating these labels, in particular, assigning membership on the basis of stipulated rules rather than perceptual invariants derived from direct experience (p. 554).

Here it seems to me that Harnad is partly falling back on the Aristotelean tradition of concept formation. The upshot seems to be a hybrid theory:

> Descriptions spare us the need for laborious learning by direct acquaintance; however, they depend on the prior existence of a repertoire of labeled categories on which the combinatory descriptions can draw. Hence *symbolic* representations (SRs), which are encoded as mental sentences, define new categories, but they must be grounded in old ones; the descriptive system as a whole must accordingly be *grounded* in the acquaintance system (p. 556).

The use of the metaphor "grounded" indicates that Harnad views the three representation forms as separate systems. In contrast, the three levels presented here are three *perspectives* on one and the same system. Nevertheless, the similarities between mine and Harnad's are indisputable. Since Harnad proposes his three kinds of representations as a tool for understanding phenomena of categorical perception, these similarities strengthen the links between concept formation and the present analysis of induction.

It is also worthwhile comparing the three levels of observation and induction discussed in this article with Smolensky's (1988) distinction between the subsymbolic and symbolic levels in the context of connectionist models. In my opinion, his 'subsymbolic level' corresponds closely enough to what has here been called the subconceptual level. However, Smolensky confounds the symbolic and conceptual levels.[18] The reason why is simple: he is commiting himself to 'High Church Computationalism' by "limiting consideration to the Newell/Simon/Fodor/Pylyshyn view of cognition" (p. 3). One of the central tenets of the symbolic approach is what Smolensky formulates as 'hypothesis 4b':

[18]He even uses the two names: "I will call the preferred level of the symbolic paradigm the conceptual level and that of the subsymbolic paradigm the *subconceptual* level" (Smolensky 1988:3).

> The programs running on the intuitive processor are composed of elements, that is, symbols, referring to essentially the same concepts as the ones used to consciously conceptualize the task domain (p. 5).

He then gives the following reason for calling the symbolic level 'conceptual':

> Cognitive models of both conscious rule application and intuitive processing have been programs constructed of entities which are *symbols* both in the syntactic sense of being operated on by symbol manipulation and in the semantic sense of (4b). Because these symbols have the conceptual semantics of (4b), I am calling the level of analysis at which these programs provide cognitive models the *conceptual level* (*ibid.*).

However, there is a different tradition within cognitive science where the conceptual level of this paper is given independent standing. For example, I believe the theory of conceptual spaces presented in Section 3.1 can be seen as a generalization of the *state space* approach, advocated among others by P. M. Churchland (1986a,b), and of the *vector function theories* of Foss (1988). The theory of conceptual spaces is *a theory for representing information*, not a theory about symbol manipulation. (The symbol paradigm that Smolensky is referring to is called the 'sentential paradigm' by the Churchlands.[19])

Even though he fails to identify it as a separate level, Smolensky is well aware of this 'vectorial' approach, as can be seen from the following quotation:

> Substantive progress in subsymbolic cognitive science requires that systematic commitments be made to vectorial representations for individual cognitive domains. ... Unlike symbolic tokens, these vectors lie in a topological space in which some are close together and others far apart (Smolensky 1988, p. 8).

He even recognizes the importance of establishing a connection between the subconceptual level and the conceptual level:

> Powerful mathematical tools are needed for relating the overall behavior of the network to the choice of representational vectors; ideally, these tools should allow us to *invert* the mapping from representations to behavior so that by starting with a mass of data on human performance we can turn the mathematical crank

[19]Cf. P. S. Churchland (1986) and Gärdenfors (1992) for a discussion of the conflict between the two approaches.

> and have the representational vectors pop up. An example of this general type of tool is the technique of *multidimensional scaling* (Shepard 1962), which allows data on human judgments of similarity between pairs of items in some set to be tuned to vectors for representing those items (in a sense). The subsymbolic paradigm needs tools such as a version of multidimensional scaling based on a connectionist model of the process of producing similarity judgments (*ibid.*)

In conclusion, Smolensky's binary distinction between the symbolic and the subsymbolic level is insufficient. We need all three levels of representing information that have been presented in this paper to give an adequate description of the various inductive processes that are encountered in the human realm as well as in the artificial.

Acknowledgements

I wish to thank Christian Balkenius, Paul Davidsson, Jens Erik Fenstad, Paul Hemeren, Patrick Suppes, Peter Williams, and the Cognitive Science group in Lund for helpful discussions. My work with this article has been supported by the Swedish Council for Research in the Humanities and Social Sciences.

References

CARNAP, R. (1950), *Logical Foundations of Probability*, Chicago, IL: Chicago University Press.

CHURCHLAND, P. M. (1986a), *Cognitive neurobiology: A computational hypothesis for laminar cortex*, Biology & Philosophy, 1, 25-51.

CHURCHLAND, P. M. (1986b), *Some reductive strategies in cognitive neurobiology*, Mind, 95, no. 379, 279-309.

CHURCHLAND, P. S. (1986), *Neurophilosophy: Toward a Unified Science of the Mind/Brain*, Cambridge, MA: Bradford Books, MIT Press.

FAIRBANKS, G. and GRUBB, P. (1961), *A psychophysical investigation of vowel formants*, Journal of Speech and Hearing Research, 4, 203-219.

FOSS, J. (1988), *The percept and vector function theories of the brain*, Philosophy of Science, 55, 511-537.

GÄRDENFORS, P. (1990), *Induction, conceptual spaces and AI*, Philosophy of Science, 57, 78-95.

GÄRDENFORS, P. (1991), *Frameworks for properties: Possible worlds vs. conceptual spaces*, Acta Philosophica Fennica 49, 383-407.

GÄRDENFORS, P. (1992), *Mental representation, conceptual spaces, and metaphors*, forthcoming in Synthese.

GENESERETH, M. and NILSSON, N. J. (1987), *Logical Foundations of Artificial Intelligence*, Los Altos, CA: Morgan Kaufmann.

GOODMAN, N. (1955), *Fact, Fiction, and Forecast*, Cambridge, MA: Harvard University Press.

HARNAD, S. (1987): *Category induction and representation*, in Categorical Perception, ed. by S. Harnad, Cambridge: Cambridge University Press, 535-565.

HEMPEL, C. G. (1965), *Aspects of Scientific Explanation, and Other Essays in the Philosophy of Science*, New York, NY: Free Press.

HINTIKKA, J. (1969), *The varieties of information and scientific explanation*, in Logic, Methodology, and Philosophy of Science III, ed. by B. van Rootselaar and J. F. Staal, Amsterdam: North-Holland, 311-331.

KOHONEN, T. (1988), *Self-Organization and Associative Memory*, Second Edition, Berlin: Springer-Verlag.

LAKOFF G. (1987), *Women, Fire and Dangerous Things*, Chicago, IL: University of Chicago Press.

LANGACKER, R. (1986), *Foundations of Cognitive Grammar (Vol. 1)*. Stanford, CA: Stanford University Press.

LIPPMAN, R. P. (1987), *An introduction to computing with neural nets*, IEEE ASSP Magazine, April 1987, 4-22.

MICHALSKI, R. S. and STEPP, R. E. (1983), *Learning from observation: Conceptual clustering*, in Machine Learning, An Artificial Intelligence Approach, Los Altos, CA: Morgan Kaufmann, 331-363.

QUINE, W.V.O. (1969), *Natural kinds*, in Ontological Relativity and Other Essays, New York, NY: Columbia University Press, 114-138.

ROSCH, E. (1975), *Cognitive representations of semantic categories*, Journal of Experimental Psychology: General, 104, 192-233.

ROSCH, E. (1978), *Prototype classification and logical classification: The two systems*, New Trends in Cognitive Representation: Challenges to Piaget's Theory, ed. E. Scholnik, Hillsdale, NJ: Lawrence Erlbaum Associates, 73-86.

RUSE, M. (1973), *The Philosophy of Biology*, London: Hutchinson and Co.

SHAPERE, D. (1982), *The concept of observation in science and philosophy*, Philosophy of Science, 49, 485-525.

SHEPARD, R. N. (1962a), *The analysis of proximities: Multidimensional scaling with an unknown distance function. I.*, Psychometrika, 27, 125-140.

SHEPARD, R. N. (1962b), *The analysis of proximities: Multidimensional scaling with an unknown distance function. II.*, Psychometrika, 27, 219-246.

SMITH, E. and MEDIN, D. L. (1981), *Categories and Concepts*, Cambridge, MA: Harvard University Press.

SMOLENSKY, P. (1988), *On the proper treatment of connectionism*, Behavioral and Brain Sciences, 11, 1-23.

SOKAL, R. R. and SNEATH, P. H. A. (1963), *Principles of Numerical Taxonomy*, San Francisco: W.H. Freeman and Co.

Logic, Methodology and Philosophy of Science IX
D. Prawitz, B. Skyrms and D. Westerståhl (Editors)

WHEN NORMAL AND EXTENSIVE FORM DECISIONS DIFFER

TEDDY SEIDENFELD

Carnegie Mellon University

0. Introduction and outline.

The "traditional" view of normative decision theory, as reported (for example) in chapter 2 of Luce and Raiffa's [1957] classic work, *Games and Decisions*, proposes a reduction of sequential decisions problems to non-sequential decisions: a reduction of extensive forms to normal forms. Nonetheless, this reduction is not without its critics, both from inside and outside expected utility theory.[1] It is my purpose in this essay to join with those critics by advocating the following thesis.

THESIS: *Sequential decisions, in* extensive form, *may lead to different outcomes than their non-sequential,* normal form *versions, in a variety of problems where the normal form fails to eliminate some "future" options that will not be chosen.*

My plan for this paper is to review the non-equivalence of extensive and normal forms in the following contexts and show how the thesis applies in each one:

In section 1, I rehearse the Harsanyi-Selten (1988) argument, applied to Game Theory. They use this thesis to distinguish "perfect" from "imperfect" equilibria in extensive forms and show that this distinction is lost in the reduction to normal forms. They appeal to a "trembling hands" model of players' options to salvage a modified version of the reduction.

In section 2, I address an ingenious argument, due to M. Goldstein (in his [1983] "Prevision of a Prevision") which uses the extensive-normal form reduction to constrain a coherent (Bayesian) agent's current beliefs about his/her future degrees of belief. In particular, I point out (§ 2.1)

[1] See LaValle and Fishburn [1987] for a useful review of the issue for problems involving one decision maker.

where Goldstein's result leads to excessive use of Bayes' rule for updating: Temporal Conditionalization.[2] And I point out (§ 2.2) where it precludes the use of Bayes' rule in updating finitely additive probabilities.

Last, in section 3, I report on some relevant consequences of using sets of probabilities: Robust Bayesian analysis. In collaborated work with L. Wasserman (Statistics, CMU) we investigate a phenomenon we call "dilation" of sets of probabilities. This occurs when the set of unconditional probabilities for an event are (properly) smaller than the set of conditional probabilities for that event (given each outcome of a partition). I illustrate how "dilation" leads to a violation of the reduction of extensive to normal forms. In § 3.1 and § 3.2 I report some of our work-in-progress indicating necessary and sufficient conditions for "dilation".

1. Harsanyi & Selten's "trembling hands"

John Harsanyi and Reinhard Selten (1988) question the adequacy of Nash's concept applied to the normal-form version of an extensive form game. They deny the equivalence of normal and extensive game forms. Instead, they advocate a refined equilibrium concept for extensive form games, based on a "trembling hands" model of choice.

An equilibrium for extensive forms is acceptable, according to their account, provided it is robust over small perturbations in choice. One of their examples from (1992) beautifully illustrates the difference between the two kinds of equilibria. Each player has two pure strategies: In the extensive form, player-1 had choice set $\{a, b\}$ and, provided his/her information set is reached (provided player-1 chooses a), player-2 has a choice set $\{c, d\}$. In the corresponding normal form, the strategies are $\{A, B\}$ for player-1 and $\{C, D\}$ for player-2. Payoffs are displayed in the next two figures.

[2]The analysis of § 2.1 addresses Goldstein's reasons. I. Levi [1987] successfully responds to a variety of arguments purporting to show that Bayes' rule is mandatory for updating beliefs.

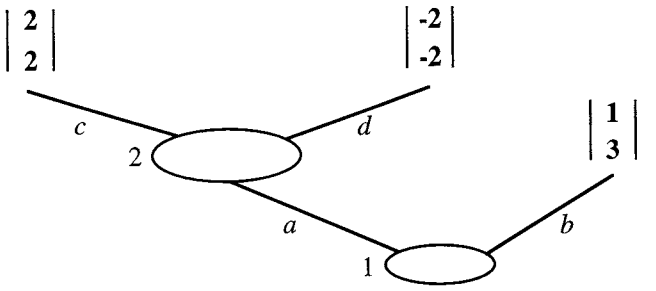


Figure 1.1 — the extensive form game
Player-1's payoff's are listed above Player-2's

	C	D
A	**2** **2**	**-2** **-2**
B	**1** **3**	**1** **3**

Figure 1.2 — Normal form of the game, above
Player-1's payoffs appear in the top-left corners.

Observe that, corresponding to the normal form Figure 1.2, there are two equilibria: the pairs $\{A, C\}$ and $\{B, D\}$. However, the latter in "imperfect" in the extensive form of Figure 1.1, as that requires player-2 to (threaten to) play option d in case choice node **2** is reached. Of course, at node 2, player-2 maximizes by playing option c instead of d, and player-1 knows this fact. Thus, the normal form equilibrium, $\{B, D\}$, depends, in the extensive form, upon ignoring that option D will not be chosen by player-2 if player-1 chooses B. To put the point another way, the normal form fails to distinguish between the extensive form of figure 1.1 and a different game where both play simultaneously, i.e., where player-2's information set does not reflect whether or not player-1 chooses a or b.

In order to avoid "imperfect equilibria", Harsanyi and Selten alter the basic moves in a game so that an agent selects one from a set of distributions (on pure options). A player chooses a mixed strategy rather than a

pure option. Figure 1.3 gives the normal form for the "trembling hands" perturbated game, where players may choose one of two mixed strategies in a perturbed extensive form game (not pictured).

In the perturbed game, the normal form options given in Figure 1.3 arise by using a two point distribution, with probabilities $(1 - \varepsilon)$ and ε assigned to each pure option in the corresponding perturbed extensive form.

	C*	D*
A*	$2-5\varepsilon+4\varepsilon^2$ $2+3\varepsilon+4\varepsilon^2$	$-2+7\varepsilon-4\varepsilon^2$ $-2+9\varepsilon-4\varepsilon^2$
B*	$1+\varepsilon-4\varepsilon^2$ $3-\varepsilon-4\varepsilon^2$	$1-3\varepsilon+4\varepsilon^2$ $3-5\varepsilon+4\varepsilon^2$

Figure 1.3

In the perturbed versions of the game, this difference between the two solutions pairs (which are in equilibrium in game form 1.2) is made evident. In the normal form 1.3, only the pair $\{A^*, C^*\}$ is in equilibrium. The $\{B^*, D^*\}$ pair is not in equilibrium since, when player-1 chooses B^*, player-2 improves his/her (expected) payoff by shifting from D^* to C^*, i.e., D^* is not player-2's best response to B^*.

The Harsanyi-Selten point is that "imperfect equilibria" are deficient because, in extensive game forms, they require a player to choose an outcome which fails to maximize his/her utility. Nonetheless, the suspect choice is justified by Nash's criterion of equilibrium in the corresponding normal form. In the extensive form of their game, player-2 does not maximize utility by choosing option d (if node 2 arises) — choice d is an *idle* threat. That move is inconsistent with the assumption that the players are utility maximizers and model each other that way. "Trembling hands", using sets of "ε-mixtures", is Harsanyi and Selten's ingenious way of reconstituting the reduction of extensive to normal forms in game theory. In section 3, I shall use sets of "ε-mixtures" of probabilities to defeat the extensive-to-normal form reduction!

2. The "prevision of a prevision" (M. Goldstein, 1983)

Goldstein's result concerns a coherent agent's currents beliefs about his/her future beliefs. It rests on the following, simple (yet suspect), lemma concerning sequential decisions.

LEMMA (Goldstein). *Let (terminal) decision D_1 lead to the "penalty" A. Suppose, also, there is a (sequential) option O to defer the choice between "penalties" A and B. Then, on pain of a sure loss, you may not now prefer D_1 over O.*

His proof (as summarized below), pivots on the extensive-to-normal form reduction.

" PROOF" (reductio): Suppose, now, you prefer D_1 to O by an amount greater than C. Then you are willing to pay amount C to receive D_1 over O. But then you suffer the sure loss C as you might just as well have only penalty A: first choice O (now), then A (later), rather than the larger penalty $A + C$.

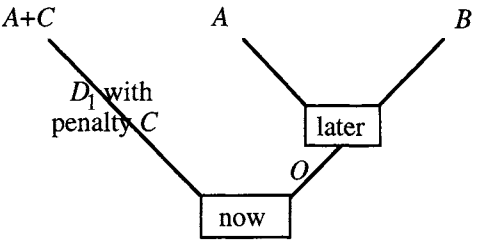


Figure 2.1 — the extensive version of Goldstein's argument

Goldstein's proof uses the reduction of the sequential option O to its normal form: a choice between penalties A and B. Goldstein compares $A + C$ and the better outcome A, without concern about what you know (now) you will choose "later". The "counterexamples" involve problems where you know (now) that were you to opt for O, then later you would choose B, which you now find inferior to $A + C$.

Next, let $P_t(E)$ denote your (currently unknown) probability for event E at the future time t. Let $P_{\text{now}}(E)$ be your current probability for E. And let $P_{\text{now}}(P_t(E))$ be your current expectation for the random variable $P_t(E)$. The result about your prevision of your (future) previsions is as follows.

THEOREM (Goldstein). $\quad P_{\text{now}}(P_t(E)) = P_{\text{now}}(E)$.[3]

PROOF: By the previous lemma on the value of deferred options.

Let us explore circumstances when this "theorem" fails, when the "lemma" fails, because extensive forms do not reduce.

[3]A related condition, called the "Principle of Reflection", is reported in van Fraassen's [1984] "Belief and the Will". See, e.g., Levi [1987] and Talbott [1991] for discussions.

2.1 Bayes' rule for updating — temporal conditionalization.

The dynamic version of Bayes' rule is this.

Suppose B summarizes the evidence acquired between (later) time t and now, then

$$P_t(\bullet) = P_{\text{now}}(\bullet \mid B).$$

If this temporal rule were mandatory then, as an extreme case: when you don't learn new evidence, you can't just change from one (coherent) distribution to another. Or, in a slightly different form using Goldstein's result, you aren't coherent if you now know that you are about to change your previsions from P to $P' \neq P$, though you will acquire no new evidence. However, in either of these cases the "lemma" does not apply as you are not prepared to equate the extensive and normal forms. The "lemma" fails to take into account that you know (now) certain choices will be rejected, yet you are asked to contrast such rejected (future) options with live current options.

The sequential argument offered on behalf of temporal conditionalization requires a questionable reduction to a normal form decision. The reduction is invalid because, by the agent's current lights, non-options are used in the normal form decision in order to show that violating the proposed dynamic rule leads to incoherent choices in the guise of a sure loss.

2.2 Non-conglomerability and the extensive to normal form reduction.

Next, I investigate where Goldstein's theorem precludes the use of Bayes' rule for updating. The case involves the use of probabilities which are finitely, but not countably additive. Let P be a f.a. probability defined on a σ-field of subsets of X. Let $E_p[\,]$ be the P-expectations for bounded, measurable functions f. And let $\pi = \{h_1, h_2, \dots\}$ be a countable partition of X.

DEFINITION (Dubins/de Finetti): Say that P is conglomerable in π provided that for each bounded, measurable function f, $\inf_\pi E_p[f \mid h] \leq E_p[f] \leq \sup_\pi E_p[f \mid h]$.

However, each P which is not σ-additive suffers a failure of conglomerability for some event E. (See Schervish et al, [1984].) That is, there exists an event E, a partition π and $\varepsilon > 0$ such that

$$P(E \mid h_i) < P(E) - \varepsilon \qquad (i = 1, \dots)$$

HEURISTIC EXAMPLE (Dubins, 1975). Figure 2.2 displays the finitely additive probabilities for "atoms". To help interpret P, assume that given

E, an integer 'i' is chosen "at random", $P(h = i \mid E) = 0$. Given E^c, a fair coin is flipped until a head appears and the number of flips determines i, $P(h = i \mid E^c) = 2^{-i}$. Also assume $P(E) = P(E^c) = 1/2$, leading to the values in Figure 2.2.

	h_1	h_2			h_i	
E	0	0			0	
E^c	2^{-2}	2^{-3}			$2^{-(i+1)}$	

Figure 2.2 — Dubins' example
$P(E) = P(E^c) = 1/2$, $P(h_i \mid E) = 0$ and $P(h_i \mid E^c) = 2^{-i}$ $(i = 1, \ldots)$.

Thus, $P(h_i) = 2^{-(i+1)}$ and $P(E \mid h_i) = 0$. So, $0 = P(E \mid h_i) \ll P(E) = 1/2$ $(i = 1, \ldots)$ and we see that P is not conglomerable in π.

Suppose the agent has P for his/her current personal probability, will learn which element of π obtains at t, and plans to use temporal conditionalization to update at t. Then, $P_{\text{now}}(E) = 1/2$ and $P_t(E) = 0$. Thus, $P_{\text{now}}(P_t(E)) = 0 \neq P_{\text{now}}(E) = .5$, and the "prevision of a prevision" theorem fails. Once again, Goldstein's "lemma" is false as the extensive form does not reduce to the normal form for decisions involving the random variable $P_t(E)$.

3. Dilation of sets of probabilities (work with Larry Wasserman)[4]

In this section, I report on a phenomenon we call "dilation", which leads in a different way to a non-equivalence of extensive and normal form decisions.

Let $\mathcal{P}$ be a (convex) set of probabilities on algebra $\mathcal{A}$. For an event E, denote by $P_*(E)$ the "lower" probability of E: $\inf_{\mathcal{P}} \{P(E)\} = P_*(E)$ and denote by $P^*(E)$ the "upper" probability of E: $\sup_{\mathcal{P}} \{P(E)\} = P^*(E)$. Let $\pi = (B_1, \ldots, B_n)$ be a (finite) partition.

DEFINITION: The set of conditional probabilities $\{P(E \mid B_i)\}$ *dilate* if

$$P_*(E \mid B_i) < P_*(E) < P^*(E \mid B_i) \qquad (i = 1, \ldots, n).$$

[4] An illustration of what we here call "dilation" was reported by Levi and Seidenfeld to I. J. Good in connection with Good's [1966] argument about the value of new evidence. That communication prompted Good's [1974] reply. Additional rebuttal is found in my [1981], where I link "dilation" with randomization in experimental design. A recently published example of dilation, relating to the worth of new evidence, appears on pp. 298–299 of P. Walley's [1991].

That is, dilation occurs provided that, for each event (B_i) in a partition π, the conditional probabilities for an event E, given B_i, properly include the unconditional probabilities for E. Dilation of conditional probabilities is the opposite phenomenon to the more familiar "shrinking" of sets of options with increasing shared evidence.[5]

HEURISTIC EXAMPLE OF DILATION

Suppose A is a highly "uncertain" event. That is $P^*(A) - P_*(A) \approx 1$. Let $\{H, T\}$ indicate the flip of a fair coin whose outcomes are independent of A. That is, $P(A, H) = P(A)/2$ for each $P \in \mathcal{P}$. Define the event E by, $E = \{(A, H), (A^c, T)\}$. It follows, simply, that $P(E) = .5$ for each $P \in \mathcal{P}$.

Then $0 \approx P_*(E \mid H) < P_*(E) = P^*(E) < P^*(E \mid H) \approx 1$

and $0 \approx P_*(E \mid T) < P_*(E) = P^*(E) < P^*(E \mid T) \approx 1$.

Thus, regardless how the coin lands, the conditional probability for event E dilates to a large interval, increasing from a "determinate" value .5.

This example mimics Ellsberg's (1961) "paradox", where the mixture of two "uncertain" events has a "determinate" probability. In different terms, event E is "pivotal" over the set $\mathcal{P}$.

Next, I indicate by example, that extensive forms do not reduce to normal forms when dilation is present.

HEURISTIC EXAMPLE (continued): Consider a sequential (that is, extensive form) choice between:

terminal option D_1 — Win \$.75 if E; Lose \$1.25 if E^c.

Or, sequential option O — observe the coin flip and choose between

D_2 — an even money \$1 bet on E

and D_3 — a "fee" of \$.50.

Thus, option $D_1 = D_2$ (an even money \$1 bet on E)+\$.25 "fee". Figure 3.1 illustrates the extensive form problem. [For convenience, hereafter, assume dollars are linear in utility.]

[5]For discussion of different senses in which a set of conditional probabilities may "shrink" with increasing evidence, see Schervish and Seidenfeld [1990].

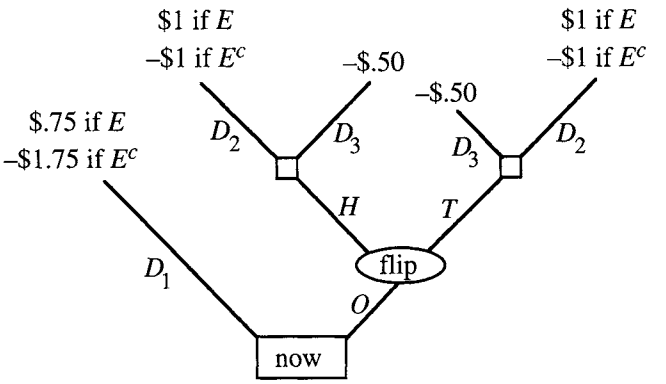


Figure 3.1 — Sequential Decision associated with the heuristic example of dilation.

Observe that in a pairwise choice between D_1 and D_2, option D_2 (simply) dominates option D_1. Therefore, in the normal version of this problem D_1 is not admissible. (D_1 fails to maximize expected utility for each $P \in \mathcal{P}$.) However, in the sequential (extensive form) problem above, after having seen the coin flip, conditional upon either H or T, both choices D_2 and D_3 are (pairwise) admissible according to expected utility considerations. That is, for some $P \in \mathcal{P}$ D_2 has higher expected utility than D_3 and for other probabilities this inequality reverses. But D_3 maximizes "security": D_3 has a better "worst" payoff, (-$.50 versus -$1.00)
or D_3 has a higher, minimum expected value (Γ-minimax).[6]

Thus, anticipating choices that will be made if the sequential option is taken, D_3 is the result of choosing O "now". Then, to complete the analysis, compare the two "live" options available "now": a choice between D_1 and D_3. But, between these two options D_3 fails to maximize expected utility for each $P \in \mathcal{P}$. Hence, D_1, which is inadmissible in the normal form, is the (sole) admissible option in the extensive form decision.[7]

[6] I allude, here, to decision theories like Levi's [1980] where an option is admissible from a choice set provided (i) it maximizes expected utility for some probability/utility in the agent's set of probabilities and utilities, and (ii) it maximizes a "security" index among those options passing the first condition. In the example here, "security" may be indicated by a maximum value or by a Γ-minimax value.

As an aside, I note that Γ-minimax requires an extraneous stipulation when sets of utilities are used. Specifically, depending upon how a set of utilities is standardized, i.e., depending upon which consequences are assigned 0 and 1, different options may be declared Γ-minimax.

[7] Of course, even when extensive forms do not reduce to normal forms, "backward

Using this example as a template, non-equivalence of extensive and normal forms can be manufactured whenever dilation occurs. In the following two sub-sections, I report on necessary and sufficient conditions for dilation.

3.1 Independence and dilation.

Independence is sufficient for dilation.

Let $\mathbb{Q}$ be a convex set of probabilities on an algebra $\mathcal{A}$ and suppose we have access to a "fair" coin which may be flipped repeatedly: coin-flip events are confined to algebra $\mathcal{C}$. Assume the coin flips are independent and, with respect to $\mathbb{Q}$, also independent of events in $\mathcal{A}$. Let $\mathcal{P}$ be the resulting convex set of probabilities on $\mathcal{A} \times \mathcal{C}$.[8]

THEOREM. *If $\mathbb{Q}$ is not a singleton, there is a 2×2 table of the form $(E, E^c) \times (H, T)$ where both:*

$$P_*(E \mid H) < P_*(E) = .5 = P^*(E) < P^*(E \mid H)$$
$$P_*(E \mid T) < P_*(E) = .5 = P^*(E) < P^*(E \mid T)$$

That is, dilation occurs.

PROOF (sketch): Let $A \in \mathcal{A}$ be "uncertain" with respect to $\mathbb{Q}$. Use the "fair" coin to form event F where $P_*(F) < .5 < P^*(F)$. Then mimic the "Ellsberg" heuristic example, above. □

Independece is necessary for dilation.

Let P be a convex set of probabilities on an algebra $\mathcal{A}$. the next result is formulated for subalgebras of 4 atoms: (p_1, p_2, p_3, p_4)

	B_1	B_2
A_1	p_1	p_2
A_2	p_3	p_4

Figure 3.2 — the case of 2×2 tables.

Define the quantity $S_P(A_1, B_1) = p_1 / (p_1 + p_2)(p_1 + p_3) = P(A_1, B_1) / P(A_1) P(B_1)$. Thus, $S_P(A_1, B_1) = 1$ iff A and B are independent under P.

induction" remains a valid sequential decision rule! See my [1988] discussion of this issue in connection with decision rules that abandon the "independence" postulate.

[8]The condition involving $\mathcal{C}$ is similar to, e.g., DeGroot's [1970] assumption of an extraneous continuous random variable, and is similar to the "fineness" assumptions in the theories of Savage [1954], Ramsey [1926], Jeffrey [1965], etc.

LEMMA. *If $\mathcal{P}$ displays dilation in this sub-algebra, then*

$$\inf_{\mathcal{P}} \{S_P(A_1, B_1)\} < 1 < \sup_{\mathcal{P}} \{S_P(A_1, B_1)\}.$$

PROOF: Direct calculation.

THEOREM. *If $\mathcal{P}$ displays dilation in this subalgebra, then there exists $P^{\#} \in \mathcal{P}$ such that*

$$S_P\#(A_1, B_1) = 1.$$

PROOF: By the lemma, there exists P_1 and P_2 such that $S_{P_1}(A_1, B_1) < 1 < S_{P_2}(A_1, B_1)$.

Write $P_x = xP_1 + (1-x)P_2$ and note that $S_{P_x}(A_1, B_1)$ is a continuous (quadratic) function of x, with coefficients involving $P_1(A_1)$, $P_1(B_1)$, $P_2(A_1)$ and $P_2(B_1)$. By the mean value theorem, for some $0 < x < 1$, $S_{P_x}(A_1, B_1) = 1$.

3.2 Dilation and ε-contamined models.

In this subsection, I report additional details about dilation for a particular (convex) set of distributions, known as the ε-contaminated model.

Given a probability P and $1 > \varepsilon > 0$, define the convex set

$$\mathcal{P}_\varepsilon = \{(1-\varepsilon)P + \varepsilon Q : Q \text{ an arbitrary probability}\}.$$

This "model" is popular in studies of Bayesian Robustness. (See, e.g., Huber, 1981.) As before, the following result applies to sub-algebras of 4 atoms.

THEOREM $\mathcal{P}_\varepsilon$ experiences dilation iff

case 1: if $S_P(A_1, B_1) > 1$,
$$\varepsilon > [S_P(A_1, B_1) - 1] \bullet \max\{P(A_1)/P(A_2); P(B_1)/P(B_2)\}$$

or

case 2: if $S_P(A_1, B_1) < 1$,
$$\varepsilon > [1 - S_P(A_1, B_1)] \bullet \max\{1; P(A_1)\,P(B_1)/P(A_2)\,P(B_2)\}$$

or

case 3: if $S_P(A_1, B_1) = 1$,
P is internal to $\mathcal{P}$.

(I omit the proof of this theorem.)

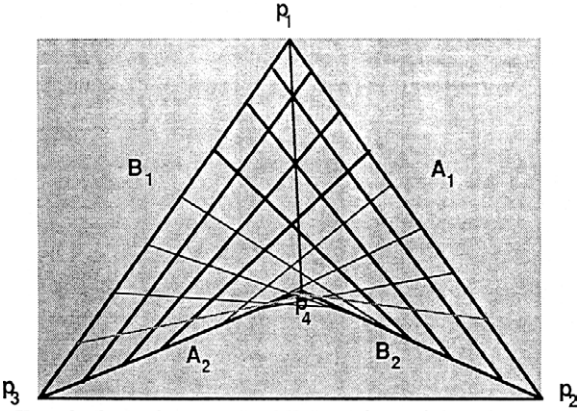
Thus, dilation occurs in the ε-contaminated model if and only if P is close enough (in the tetrahedron of distributions) to the saddle-shaped surface of distributions which make A and B independent.[9] The Figure 3.3 illustrates the "saddle" of probabilities satisfying $P(A, B) = P(A)P(B)$.

4. Summary

I have discussed three decision contexts where extensive forms do not reduce to normal forms:

1. Game theory — The Harsanyi-Selten argument about "imperfect" equilibria.
2. Denying "The Prevision of a Prevision" (M. Goldstein's argument)
   2a — involving failures of temporal conditionalization
   2b — involving non-conglomerability of finitely additive probability
3. Dilation of Sets of Probabilities.

The common reason why there is no reduction for these cases is that particular "future" options, which the agent knows (in advance) will *not* be chosen in the sequential decision are, nonetheless, used as though they were feasible options in the normal form. That is, an option which is inadmissible in the normal form may be admissible in an extensive form (generating that normal form). Rival choices which defeat that choice in the normal form turn out to be not feasible in the sequential form.



Tetrahedron showing "saddle" surface of distributions which make events A and B independent

Figure 3.3

[9] As a contrast, the Density Ratio model is immune to dilation. Let P be a fixed probability defined on the atomic algebra $\mathcal{A}$, with "atomic" probabilities denoted p_i. The Density Ratio model on $\mathcal{A}$, for $k \geq 1, \mathcal{R}(P, k) = \{Q : q_i/q_j \leq k \bullet p_i/p_j\}$.

REFERENCES

DEFINETTI, B. (1972) *Probability, Induction and Statistics.* New York: Wiley.

DEGROOT, M. (1970) *Optimal Statistical Decisions.* New York: Wiley.

DUBINS, L. (1975) "Finitely Additive Conditional Probabilities, Conglomerability and Disintegrations", *Annals of Prob.* 3. 89–99.

ELLSBERG, D. (1961) "Risk, Ambiguity, and the Savage axioms", *Q. J. Economics* 75, 643–669.

GOLDSTEIN, M. (1983) "The Prevision of a Prevision", *J. A. S. A.* 78, 817–819.

GOOD, I. J. (1966) "On the Principle of Total Evidence", *B. J. P. S.* 17, 319–321.

GOOD, I. J. (1974) "A Little Learning Can Be Dangerous", *B. J. P. S.* 25, 340–342.

HARSANYI, J. (1992) "Game Solutions and the Normal Form", in *Knowledge, Belief, and Strategic Interaction*, C. Biccieri and M. Della Chiara (eds.), Cambridge Univ. Press, 355–376.

HARSANYI, J. and SELTEN, R. (1988) *A General Theory of Equilibrium Selection in Games.* Cambridge: MIT Press.

HUBER, P. (1981) *Robust Statistics.* New York: Wiley.

JEFFREY, R. C. (1965) *The Logic of Decision.* New York: McGraw-Hill.

LAVALLE, I. and FISHBURN, P. (1987) "Equivalent Decision Trees and their Associated Strategy Sets", *Theory and Decision* 23, 37–63.

LEVI, I. (1980) *The Enterprise of Knowledge.* Cambridge: MIT Press.

LEVI, I. (1987) "The Demons of Decision", *The Monist* 70, 193–211.

LUCE, R. D. and RAIFFA, H. (1957) *Games and Decisions.* New York: Wiley.

RAMSEY, F. P. (1926) "Truth and Probability" in his Philosophical Papers [1990] Cambridge: Cambridge University Press (H. Mellor, ed.), pp. 52–109.

SAVAGE, L. J. (1954) *The Foundations of Statistics.* New York: Wiley.

SCHERVISH, M. and SEIDENFELD, T. (1990) "An approach to consensus and certainty with increasing evidence", *J. Stat. Planning and Inference* 25, 401–414.

SCHERVISH, M., SEIDENFELD, T., and KADANE, J. (1984) "The Extent of Non-conglomerability of Finitely Additive Probabilities", *Z. Wahr.* 66, 205–226.

SEIDENFELD, T. (1981) "Levi on the Dogma of Randomization", in R. Bogdan (ed.) *Henry E. Kyburg, Jr. & Isaac Levi.* Dordrecht: D. Reidel, pp. 263–291.

SEIDENFELD, T. (1988) "Decision Theory Without 'Independence' or Without 'Ordering': What is the Difference", *Economics and Philosophy* 4, 267–290.

TALBOTT, W.J. (1991) "Two Principles of Bayesian Epistemology", *Phil. Studies* 62, 135–150.

VAN FRAASSEN, B. C. (1984) "Belief and the Will", *J. Phil.* 81, 235–256.

WALLEY, P. (1991) *Statistical Reasoning with Imprecise Probabilities.* St.Edmunds, Suffolk: St. Edmundsbury Press.

Logic, Methodology and Philosophy of Science IX
D. Prawitz, B. Skyrms and D. Westerståhl (Editors)

ANDREI MARKOV AND MATHEMATICAL CONSTRUCTIVISM

N.M. NAGORNY

Computing Centre of the Academy of Sciences of the USSR 117333, Moscow

Andrei Andreevich Markov Jr. (22.9.1903 - 11.10.1979) whose contribution to mathematical constructivism is the subject of my paper is undoubtedly one of the most outstanding and original mathematicians and logicians of our time. He traversed a long and thorny path[1] whose twists and turns we shall retrace at least cursorily. His research in the theory of algorithms, mathematical logic and constructive mathematics that spanned thirty years of his life was the height of his scientific career. That period was marked by a clean break with the traditional fundamentals of mathematics—the set theory which he had for a long time adhered to, but later vehemently opposed. For the sake of brevity I shall call this period the "constructivist" period in Markov's career.

Not only did he obtain over this period a number of first-class concrete results (among them the solution of two famous algorithmic problems—the identity problem for semigroups and the homeomorphy problem; the results related to the recognition of invariant properties of associative calculi; the elaboration of the complexity approach in the theory of algorithms, etc.); and not only did he set up a scientific school whose representatives are to this day working in different countries.[2] He did over this period conceive a new view on mathematics (or at least on its foundations). There is hardly any need to explain that this is a rare achievement for a scientist.

In our country, on Markov's initiative, the trend of mathematics he inaugurated is called "constructivist". In the West this trend is called Soviet constructivism. I think it would be fairer to call it Markovian constructivism.

[1]Some details concerning Markov's scientific biography can be found in the articles [1-3] and in the Preface to the monograph [4].

[2]Belonging to Markov's school are such prominent mathematicians as O.Demuth, A.Dragalin, M.Kanovich, B.Kushner, S.Maslov, Yu.Matiyasevich, G.Mints, N.Nepeivoda, V.Orevkov, N.Petri, Phan Dinh Dieu, N.Shanin, D.Skordev, G.Tseitin, V.Yankov, I.Zaslavskii, and many others.

Today, Markovian constructivism holds a legitimate place in mathematics. It has opened up a wide field of research. Markov's own views, which were an extension and continuation (on a new basis) of the ideas of such brilliant thinkers as L.E.J. Brouwer and H.Weyl, raised the discussion on the degree of constructivity of the means used in mathematics to a new and higher stage. These views also characterize him as a profound and original philosopher.

Another outstanding Soviet mathematician, Andrei Kolmogorov, a man of Markov's age, was Markov's uncompromising opponent in many fundamental principles of mathematics. Shortly before Markov's death he spoke at the celebrations of the 20th anniversary of the chair of mathematical logic at Moscow University. That chair was founded by Markov and after his death was taken over by Kolmogorov. Speaking of Markov's place in mathematics, Kolmogorov put his name next to those of G.Cantor, L.E.J. Brouwer and D.Hilbert who, in his words, had felt great responsibility for the state of affairs in mathematics as a whole. It would be hard to disagree with such a highly competent opinion.

It must be noted that Markovian constructivism largely took shape in the pre-computer era. Now that science and society are getting more and more computerized, Markov's work reveals new qualities as a possible source of precise formulation of problems which would take into account actual computational practice and as a basis for solving such problems.

Of Markov's achievements in the "constructivist" period I shall dwell only on Markovian constructivism. I shall not enlarge on any of his specific achievements, however important they may be.

As a pupil of Markov's and an exponent of his school, I want to express my deep gratitude to the Programme Committee of the Eighth Section for giving me the opportunity of presenting this paper at the Congress. This, no doubt, is an authoritative recognition of Markov's services to science. Incidentally, he took an active part in the work of the International Union of History and Philosophy of Science as vice-president of the Division of Logic, Methodology and Philosophy of Science from 1967 to 1971 and from 1975 to 1979.

Before getting down to the main part of my report I shall touch upon the difficulties that Markovian constructivism encountered soon after its birth. Otherwise it would be hard to appreciate Markov's personal courage in those far from simple conditions.

Markov's switch to constructivism and his first steps in this new field coincided with one of the unhappiest periods in Soviet history. The war had just ended. Much of the country lay in ruins, and famine stalked the land. And yet, contrary to all common sense, public attention was, at the will of the Communist Party, focused on ideological matters with their multiple

problems. These problems had to be considered from the point of view of what was officially dubbed "the genuinely scientific and the only correct" doctrine—Marxist-Leninist materialism. This means that every scientist of the Soviet Union was required, even if only in words, to profess his adherence, or rather his allegiance, to this "philosophy". Many sciences were crushed by the juggernaut of the so called struggle against idealism. These included cybernetics (which was called a "reactionary bourgeois pseudoscience"). Mathematical logic also became suspect. By a twist of fate the set theory, with its undisputed platonism, was regarded as a materialist, and not "idealist", direction in mathematics. This is why, already after his first public statements spearheaded against the set theory, Markov was accused of "idealism and formalism". In those years such accusations suggested political heresy which might have the most dire consequences. Markov was attacked not only by Marxist philosophers and ideologists, but even by his colleagues the mathematicians who adhered to the traditional principles and who, after hearing him, felt distinctly uncomfortable. The heated debates that followed were, at least at the beginning, a one-man campaign against everybody. It was not only his prestige as a scientist and his temperament of a fighter that helped Markov to hold his ground, but also, I think, a stroke of good luck.

With time Markov consolidated his position, but neither he nor his school of thought were given the credit and honour they so richly deserved. He was never awarded distinctions that are usually accorded to scientists of his calibre (in particular he was not elected a full member of the Academy of Sciences of the USSR). His university chair was discriminated against, and his pupils had a much smaller chance of making the grade than their peers. None of them is presently a member of Markov's university chair or a fellow of Steklov Mathematical Institute of the Academy of Sciences. Some of his pupils trained for new professions, whereas others were forced to leave the Soviet Union. It would be no exaggeration to say that in its own country Markovian constructivism was among the sciences that fell victim to persecution.

And now I would like to give the reader an idea of the way Markov arrived at his views on mathematics, and tell him about some of the more typical features of Markov's personality which were reflected in these views and which made them so natural, consistent, dynamic and convincing.

Markov was born into the family of the celebrated Russian mathematician Andrei Andreevich Markov (Sr.; 1856 - 1922). He spent his early years in close contact with his father, a man of strong character, with a keen sense of justice, and great civic courage (in Russia he was known for his protest against the excommunication of Lev Tolstoy. Markov Sr. demanded that he

also be excommunicated together with Tolstoy). Markov Sr. sought to have an influence on his son outside the home as well. For example, for some time he taught mathematics in the class his son attended.

The young Markov received an excellent education and absorbed the best traditions of the highly cultured environment he grew up in: trust in the lofty mission of culture, love of and a sense of responsibility for the advancement of science, and high moral principles with a clear set of priorities.

Markov had faultless artistic taste, a keen sense of humour and a feel for language. He was a remarkable stylist and in almost every one of his works tackled a different stylistic problem. I think that Markov's urge towards a severe and sometimes ascetic style must, for purely aesthetic motives, have led him to reject the romantically unbridled freedom of set theory. His spirit and his temperament accorded more with the restrained, "syntactic", as it were, character of his constructivism and with the economy of the logical means involved.

At sixteen, Markov enrolled at the university, but at the chemistry department. Soon after his first scientific publication appeared. However, by the end of the second year at the university he realized that he would not make a good chemist: he was too temperamental and too impatient to just stand by and wait for the completion of a chemical reaction. In his third year at the university Markov switched to the physics and mathematics department. He majored in physics and, after graduating, worked for a short time at the State Physico-Technical Institute, wrote essays on theoretical physics (among them one of the first Soviet work [5] on quantum mechanics), and on applied geophysics. In 1925, Markov began his post-graduate studies later to become a Fellow at the Astronomy Institute. He published a succession of papers on celestial mechanics which retain their significance to this day.

His research in celestial mechanics gradually moved Markov on to purely mathematical problems and here he later tried himself in different spheres ranging from the applied (such as plasticity theory [6]) to the abstract (such as axiomatic set theory [7]). He made a signal contribution in almost every field that attracted his attention. Among the greatest achievements of Markov's "pre-constructivist" period, which lasted until about 1947, are his works on the dynamic systems theory, topology, and on the theory of topological groups.

Markov's formation as a mathematician must have taken place at the time of the triumph of the Cantorian set theory. He must have used that theory in his own research. It must have been the Cantorian theory that largely determined the theme of his work. And yet, his mathematical career began not just because he had received an education in the field. It was rather the career of a man who had studied nature and later embraced mathematics,

which was the result of the evolution of his creative interests. In this sense he was *a naturalist.* This also had a certain impact on his approach to every problem he worked on. If a theory caught his fancy he sought to give its development a certain slant so that its results would have a clear meaning.

One of the central mathematical ideas which actually forced mathematics to be an abstract, speculative science is the idea of infinity. This idea occurs in mathematics in two varieties: in the shape of so-called *"actual" infinity* and in its more modest version called *"potential" infinity.* These two varieties bring forth characteristic methods of abstracting: *the abstraction of actual infinity* and *the abstraction of potential feasibility.* We shall discuss them in connection with one of Markov's key concepts—the concept of *constructive process* (see [8, p.4] or [4, p.1]).

A constructive process, according to Markov, is a step-by-step process whereby complexes of symbols are generated, a process that leads from complexes which are adopted as initial complexes to new complexes which are formed in accordance with previously formulated rules. The complexes generated in the course of constructive processes are called *constructive objects.* Markov considers only the simplest constructive objects, namely words. Words are strings of symbols (called *letters*) picked out of a previously chosen collection (called *the alphabet*). This is a description of this process: initially, one letter of the alphabet is taken, after which another letter is added to the right of it, then a third letter, etc.

I would like to emphasize the particular tangibility of the constructive objects: we come to understand them through our senses, and not only by reason as we do it in the case of sets.

By unfolding a concrete constructive process we may eventually come up, at a certain stage, against a shortage of space, time or material. Abstraction of potential feasibility in this case consists in ignoring these obstacles, i.e. in allowing us to treat this process as unboundedly extendable. Thus another letter can always be added to any word. This means that by making use of this abstraction we pass from objects already generated in the process to objects eventually arising in this process (i.e. to *possible* objects). In a comment on the Russian translation of A. Heyting's book [9] which came out in 1965 under his editorship, Markov speaks about Brouwer's *mental constructions*, saying that these constructions are "usually just casts of material constructions that we can see everywhere .... On the other hand mental constructions often turn out to be blue prints for material constructions" (p.162). He could have said the same thing about constructive processes.

Having adopted the abstraction of potential feasibility, we make it possible to provide a satisfactory, from a constructivist point of view, answer to the question of what *an assertion of the existence* of a constructive object with

given properties really means. "In constructive mathematics the existence of an object with given properties is considered proven if one has provided a way for a potentially feasible construction of an object with these properties" ([10], p.9).

As for the abstraction of actual infinity in the given situation, it ignores the impossibility of completing the given process. This abstraction looks at this process as accomplished and integrates all the generated objects into one comprehensive set. The latter is considered an object of the same level as its components.

This abstraction is, undoubtedly, central to the Cantorian set theory. However, the validity of its utilization in mathematics has repeatedly been questioned. Brouwer and Weyl believed that this abstraction did not meet the requirements of intuitive clarity. Hilbert [11] criticized it as not corresponding to reality: "...das Unendliche ist in der Wirklichkeit nirgends zu finden, was für Erfahrungen und Beobachtungen und welcherlei Wissenschaft wir auch heranziehen. Sollte nun das Denken über die Dinge so unähnlich den Geschehnissen mit den Dingen sein und so anderartig vor sich gehen, so abseitig von aller Wirklichkeit?"

What was Markov's attitude to this abstraction? How did he treat the Cantorian set theory in general? As I said earlier on, Markov in his "pre-constructivist" period did use set theory. He also adhered to the then generally accepted standard of constructing mathematical theories which required defining all mathematical concepts strictly in set theoretical terms, and utilizing traditional Aristotelian logic as their logical apparatus.

However as time went on, Markov, who right from the start of his mathematical career strove for logical clarity and showed great interest in the foundations of mathematics, felt deeply dissatisfied with the groundwork upon which mathematics rested at that time.

Quite naturally, he was worried about the emergence in this groundwork of "sinister cracks" ([8], p.42) in the form of antinomies of set theory and the absence of any guarantees against the emergence of new antinomies after overcoming those already revealed. But he was also dissatisfied with the absence of any more or less precise definition of the central set theoretical concept, namely the concept of set. In fact, according to the standard just mentioned, the result of eliminating from whatever mathematical statement all definitions of concepts contained in it turns this into a statement about sets. But how is this proposition to be understood when even a definition of set is lacking?

Markov was also greatly disturbed by the fact that several important mathematical concepts (such as real numbers) are *non-constructively* defined. These non-constructive definitions speak about certain sets, and at

the same time ignore the possibility of constructing such sets. The non-constructivity of these definitions in turn determines the non-constructivity of the basic mathematical theories (such as mathematical analysis) which manifests itself in the so-called "*pure existence theorems*". These theorems assert the existence of certain objects with given properties without a clue as to how these objects can be found, and without so much as an attempt to make these objects tangible in any way (see [12], p.316). Pure existence theorems usually emerge out of *apagogical proofs*, which mainly result from the use of the so-called "*law of the excluded middle*". But the latter, as is known, is grounded on the abstraction of actual infinity.

Markov continues the criticism of his predecessors with regard to this abstraction. However, his whole approach to criticism is different from that of Brouwer, Weyl and Hilbert. He notes that it is possible to justify the law of the excluded middle as a general logical principle on this basis, and says: "The trouble is that this very idea is too fantastic. After all, one is unable to think of an infinite, i.e. never ending process as a finished one, without exercising brute force over our reason which rejects such contradictory fantasies. Actually, we wanted to consider infinite processes as finite processes, i.e. to abstract ourselves from their infinity" ([8], p.41). Speaking about "pure existence theorems" in the study [12] Markov points out that "abstractions are necessary in mathematics. However, they must not be exercised for their own sake, which would lead to a point where there is no way of getting back down to the ground. We must always remember the transition from the abstract thinking to praxis as the necessary phase of cognition of objective reality by man. When the possibility of such a tradition is very much in doubt, we should reconsider the use of old abstractions and replace them with new ones" (p.315-316).

In their sharp criticism of the set theory L.E.J.Brouwer [13,14] and later H.Weyl [15] looked for a positive way out of the crisis. They saw it in the building up of mathematics *without using* the abstraction of actual infinity. Brouwer's programme (his intuitionism) consisted in the building up of mathematics on the basis of "mental mathematical constructions". Brouwer showed that the study of such constructions requires a special logic which would be different from Aristotelian logic. He outlined the contours of this *intuitionistic logic* which, among other things, rejected the law of the excluded middle as a general logical principle.

Undoubtedly, Brouwer's intuitionism turned out to be a major event in mathematics, philosophy and logic.

Markov fully accepted the critical part of the intuitionistic programme. His view on the law of the excluded middle, as he himself put it "came very close to Brouwer's point of view" ([8], p.44). However, he was not satisfied

with the positive part of the programme: "Choice sequences (in Brouwer. - N.N.) are not constructive objects and they could hardly be 'looked at' without using the abstraction of actual infinity" ([8], p.45).

Markov closely followed major developments in foundations of mathematics. Quick to respond to anything new, he readily appreciated the significance of the first steps made by the theory of algorithms. He discerned not only the technical tools it offered for solving a number of famous algorithmic problems, but also the general logical and architectural perspectives which they opened in the foundations of mathematics. Markov thought highly of Church's Thesis which asserted the possibility of making more precise the general, somewhat vaguely formulated idea of algorithm. He appreciated the significance of that thesis which foresaw the part that it would play in constructive mathematics. The largely empirical justification of Church's thesis seemed quite convincing to Markov the naturalist (Markov compares the status of this thesis to that of the Law of the Conservation of Energy; see, for instance, [4, p.108]). Due to Church's Thesis, the study of algorithms boiled down to a consideration of constructive objects (and even just words) of a certain type.

Markov was greatly impressed by S.C.Kleene's work [16] which first came out in 1945. Later Markov said on many occasions that this article had exercised great influence on the evolution of his own views on the problems of foundations of mathematics. It was at that time that Markov turned to constructive mathematics once and for all. In 1947 - 1948 only two papers by him were published that pertained to set theoretical mathematics. Those were his last publications within the framework of this mathematics.

In later years Markov seldom and reluctantly spoke about his activities in the "pre-constructivist" period and said that there were things in them that he could no longer understand (which was his way of saying that he judged them fallacious). This explains why young colleagues thought him an expert par excellence in mathematical logic.

So, what does Markov's programme for mathematics consist in?
Markov suggests that:

1. Constructive mathematics should deal exclusively with constructive objects.

2. These objects should be treated only within the framework of potential feasibility abstraction, the abstraction of actual infinity being banned.

3. The term "algorithm" should be understood in a precise way (in the light of Church's thesis).

4. Due to the nature of the allowed objects (item 1) and abstractions (item 2), the understanding of mathematical statements should be based upon specially conceived constructive logical principles. For instance, the existence of an object with a given property should signify the possession of a way yielding a potentially feasible construction of an object with this property.

5. The used apparatus of logical deduction should be based upon a constructive logic that would exclude the provability of pure existence theorems. Such a logic should be free, in particular, from the so-called "Law of the Excluded Middle" (viewed as a general logical principle).

6. One should accept apagogical proofs of statements of the form: "algorithm f halts when applied to input x" (the so-called "Markov's Principle"; see, for instance, [4, p.348 - 350]).

Remarks:

To item 1.

1.1. The restriction set by this item means that the concepts of constructive mathematics are to be defined, as they arise, in terms of constructive objects. Let us observe that the concepts of set theoretical mathematics were defined in terms of sets.

1.2. Markov considers but the simplest constructive objects, namely words.

To item 3. The precise notion of an algorithm currently used in Markov's school is his own concept of Normal Algorithm, very attractive as, a research as well as a didactic, tool. Relatively few are those who are aware that this notion raised from research on the identity problem for semigroups, precisely, from the search for a satisfactory rendering of its solution.

To items 4 and 5. At the beginning of his "constructivist" period Markov adopted a semantics inspired by Kleene's realizability (the already mentioned paper [16]). However, if one tried to rely upon such a semantics for a systematic building up of constructive mathematics which should comprehend the theory of algorithms, one would fall into a vicious circle, as realizability is itself based upon the precise notion of algorithm. To overcome this difficulty as well as those aroused by the urge to understand the meaning of implication, Markov devised

the so-called *graded semantic system* (see for example, [17]). This was intended to provide a framework for the developement of constructive logic, theory of algorithms and specific mathematical theories in their interconnection. Having no space to dwell any longer upon this subject, I shall simply observe that while working on his system Markov developed a slant towards intuitionism which he had kept away from before, at the beginning of his constructivist period.

To item 6. Markov convincingly shows that his principle can be justified without using the abstraction of actual infinity. So, this principle is compatible with item 2 of Markov's programme.

In spite of its sobriety the sketched Markov's programme provides a framework in which many important parts of mathematics can be built up constructively. Mathematical analysis happens to be the most thoroughly developed area in Markov's school. Markov laid down its foundations in 1954 in the inaugural paper [18]. A suggestive result from the latter will be mentioned below. Basic concepts and a few chosen facts concerning constructive analysis can be found in Markov and Nagorny's monograph [4]; a systematic treatment in B.Kushner's book [19]; a detailed survey of results in this area being Kushner's paper [20]. Here I shall confine myself only to a few illustrations.

(a). We may introduce *natural numbers* as words built up from the symbols 0 and 1 which have the form

0	01	011	0111	01111
(zero)	(one)	(two)	(three)	(four)

and so on.

It is obvious how to define a constructive process generating these words.

(b). By adding to the natural numbers all the words of the form $-N$, where "$-$" is a new letter and N is a natural number, we get *the integers.*

Examples: 0111 (three) and -0111 (minus three).

(c). By adding to the integers all the words of the form M/N, where "/" is a new letter, M is an integer, N is a non-zero natural number, we get *the rational numbers.*

Examples: 011 (two), -0111 (minus three), $-011/0111$ (minus two-thirds).

(d). A normal algorithm (encoded in a proper way by a word) is called *a constructive sequence of rational numbers* if, being applied to any natural number, it produces a rational number.

(e). A pair of algorithms (encoded in a proper way by a word) is called *a constructive real number* if the first algorithm is a constructive sequence of rational numbers and the second effectively estimates the rate of convergence of this sequence.

For such constructive real numbers one can define in some natural way the relations of order and equality as well as arithmetical operations.

(f). A normal algorithm (encoded in a proper way by a word) is called a constructive real function if it satisfies the following properties:

(i): If it halts when applied to a constructive real number, then it produces a constructive real number;

(ii): If it halts when applied to a constructive number, then it halts when applied to every constructive real number equal to this;

(iii): If it halts when applied to two equal constructive real numbers, then it produces equal constructive real numbers.

In Markov's paper [12] (see also [18]) it is shown that *no constructive real function can be discontinuous*, i.e. that no real function can have a constructive discontinuity at any point.

Later in 1959 G.S.Tseitin [21] obtained a final result in this direction by showing that every constructive real function is in fact *continuous*.

(g). Cauchy's medium value theorem. It plays an important role in set theoretical mathematical analysis. This theorem implies that, given any continuous function assuming values of the opposite sign on the extremities of an interval, there exists an x in this interval such that $f(x) = 0$. This is a typical "pure existence theorem" which does not yield any method of finding the required x. It is worth noting in this connection that there are no satisfactory numerical methods of solving the equation $f(x) = 0$ for sufficiently large classes of sign changing functions f. Constructive analysis throws light upon this state of affairs. It happens that the following three theorems hold (G.S.Tseitin [22]):

(I). Given any continuous constructive real function f assuming values of opposite sign on the extremities of an interval, it is not true

that $f(x)$ is different from 0 for all constructive points x in this interval.

(II). There is no normal algorithm such that, being applied to any f of this sort, would produce a constructive real number x such that $f(x) = 0$.

However, the following theorem holds:

(III). There exists a normal algorithm which, being applied to any pair f, ε, where f is a function of the considered sort and ε is a positive rational number, produces a constructive point x such that $|f(x)| \leq \varepsilon$.

So, no uniform method can extract enough "information" from any continuous constructive real function f changing sign in order to find a solution to the equation $f(x) = 0$. It is possible, however, to find uniformly by f and ε a constructive point x such that $|f(x)| \leq \varepsilon$.

Let us observe that this is just the problem solved in practice.

It is worth noting that the proof of (III) requires the use of Markov's Principle (the 6th item of Markov's programme).

(h). Specker's Example. This is essentially a "constructive counterexample" to the well-known Weierstrass theorem, asserting that any increasing bounded sequence of real numbers converges to a limit.

Inspired by Weierstrass' "pure existence theorem" a mathematician could eventually succeed in finding the limit of this or that specific increasing bounded sequence. However, as follows from Specker's construction [23], he will fail to achieve this result every time. In fact, E.Specker exhibited an increasing bounded constructive sequence of rational numbers which does not converge to any constructive real number. Specker's example sets up an essential obstacle: no ingenuity would secure the computing of such a number.

An experienced specialist in numerical methods should, of course, have a strong feeling for the examples described here. However, we believe that the possibility of expressing this feeling in the form of precise mathematical assertions is undoubtedly an outstanding achivement of Markovian constructivism.

References

[1] LINNIK, YU.V., and SHANIN, N.A. *Andrei Andreevich Markov. On his fiftieth birthday* (Russian). Usp Mat Nauk, 1954, v.9, p.145 - 149.

[2] NAGORNY, N.M., and SHANIN, N.A. *Andrei Andreevich Markov. On his sixtieth birthday* (Russian). Usp Mat Nauk, 1964, v.19, p.207 - 223.

[3] DRAGALIN, A.G., NAGORNY, N.M., PETRI, N.V., and SHANIN, N.A. *Andrei Andreevich Markov. On his seventieth birthday* (Russian). Usp Mat Nauk, 1974, v.29, p.187 - 191.

[4] MARKOV, A.A., and NAGORNY, N.M. *The Theory of Algorithms.* Kluwer Acad Publishers, Dordrecht/Boston/London, 1988.

[5] MARKOV, A.A. *Ueber eine Minimumeigenschaft der Schrödingerschen Wellengruppen.* Zschr für Phys, 1927, v.42, p.637 - 640.

[6] MARKOV, A.A. *On variational principles in plasticity theory* (Russian). Prikl Mat & Mech, 1947, v.11, p.335 - 350.

[7] MARKOV, A.A. *On the non-independence of the axiom B6 from the other axioms of the Bernays - Gödel system* (Russian). Izv Akad Nauk SSSR, Ser Mat, 1947, v.12, p.569 - 570.

[8] MARKOV, A.A. *On the logic of constructive mathematics* (Russian). Znanie, Moscow, 1972.

[9] HEYTING, A. *Intuitionism. An introduction.* North Holland Publ Comp, Amsterdam, 1956. (Russian transl.: Ed. Markov A.A.; Mir, Moscow, 1965.)

[10] MARKOV, A.A. *On constructive mathematics* (Russian). Tr Mat Inst Steklov, 1958, v.67, p.8 - 14. (Engl. transl.: Amer Math Soc, Transl, Ser 2, 1971, v.98, p.1-9.)

[11] HILBERT, D. *Ueber das Unendliche.* Math Ann, 1926, Bd.95, S.161 - 190.

[12] MARKOV, A.A. *On constructive functions* (Russian). Tr Mat Inst Steklov, 1958, v.52, p.315 - 348. (Engl. transl.: Amer Math Soc, Transl, Ser 2, 1963, v.29, p.163-195.)

[13] BROUWER, L.E.J. *Over de grondslagen der wiskunde.* Thesis, Amsterdam, 1907.

[14] BROUWER, L.E.J. *De onbetrouwbaarheid der logische principes.* Tijdsch Wijsbegeerte, 1908, v.2, p.152-158.

[15] WEYL, H. *Das Kontinuum.* Veit, Leipzig, 1918.

[16] KLEENE, S.C. *On the intepretation of intuitionistic number theory.* J Symb Logic, 1945, v.10, p.109 - 124.

[17] MARKOV, A.A. *Essai de construction d'une logique de la mathématique constructive.* Rev Int Philos, 1971, v.98, p.477 - 507.

[18] MARKOV, A.A. *On the continuity of constructive functions* (Russian). Usp Mat Nauk, 1954, v.9, p.226 - 230.

[19] KUSHNER, B.A. *Lectures on constructive mathematical analysis.* Amer Mat Soc, v.80, 1984.

[20] KUSHNER, B.A. *Markov's constructive analysis: the expectations and the results.* In: Mathematical logic (Ed. P.P.Petkov), Plenum Press, New York/London, 1990.

[21] TSEITIN, G.S. *Algorithmic operators in constructive metric spaces.* (Russian). Tr Mat Inst Steklov, 1962, v.67, p.295 - 361. (Engl. transl.: Amer Math Soc, Transl, Ser 2, 1967, v.64, p.1-80.)

[22] TSEITIN, G.S. *Mean value theorems in constructive analysis* (Russian). Tr Mat Inst Steklov, 1962, v.67, p.362 - 384. (Engl. transl.: Amer Math Soc, Transl, Ser 2, 1971, v.98, p.11-40.)

[23] SPECKER, E. *Nicht konstruktiv beweisbare Sätze der Analysis.* J Symb Logic, 1949, v.14, p.145 - 158.

Logic, Methodology and Philosophy of Science IX
D. Prawitz, B. Skyrms and D. Westerståhl (Editors)

CONTRIBUTIONS TO THE HISTORY OF THE CLASSICAL TRUTH-DEFINITION

JAN WOLEŃSKI

Jagiellonian University, Institute of Philosophy, Kraków, Poland

1. Introduction

Although truth belongs to the family of crucial philosophical categories, writing its general history still remains a serious challenge for historians of philosophy. Also historical accounts of particular truth-theories are rather fragmentaric. Since the classical (also called "the correspondence") theory of truth has become the most popular and influential among all hitherto proposed answers to the philosophical problem of truth, a lack of its written history is especially strange, more than in the case of its various rivals; this theory maintains, roughly speaking, that truth consists in a relation of correspondence (agreement, adequacy or conformity) which holds between so called bearers of truth (judgements, ideas, thoughts, propositions, statements or sentences) and reality.

This paper presents a sketch of how the gap could be filled with respect to the classical concept of truth (CCT for brevity). It is just a sketch which by no means pretends to any completeness. The history of the classical (as well as every other) theory of truth requires taking into account at least four points, namely

(A) statements which have been explicitely intended as definitions (or other explications) of CCT;

(B) formulations which could be interpreted as definitions (or other explications) of CCT, independently of the intentions of their authors;

(C) the philosophical environment of formulations collected under (A) and (B); it is especially important for cases falling under (B);

(D) criticism of CCT and its defences against raised objections.

I would like to touch each of (A)–(D) but my principal goal is to contribute to (A) and (B).

Although the theory which is the subject of this paper goes back to the ancient Greeks, its presently used labels are rather new. The term 'correspondence' in the context of truth theory was introduced by Russell (see Russell 1910, 1912). However, Russell himself did not use (at least in his earlier works) the term 'the correspondence theory of truth'; in his book from 1984 (written in 1913), he distinguishes (see p. 149): theories which define truth by a correspondence, pragmatism and the coherence-theory. Certainly, the label 'the correspondence theory of truth' was invented under Russell's influence but it is difficult to say who employed it for the first time. The same concerns its German counterpart, namely 'Adäquationtheorie der Wahrheit' which became popular in the 1930s. Also there are difficulties as far as the matter concerns where and when the expression 'the classical theory of truth' has appeared in philosophy. Anyway, this name is very common among Polish philosophers (see (27) below).

2. *Aletheia* in old Greek (see Boeder 1958)

Leaving out the full etymology of *aletheia* (which for instance has led Martin Heidegger to far reaching claims concerning the concept of truth — "truth as openness"), let me note that this word was used in old Greek (especially in early Greek poems) in dialogical situations which involved knowing and asking persons. This use was neither predicative nor attributive; the word occurs together with so called (in Latin terminology) *verba dicendi*. Then, *aletheia* referred neither to abstract statements nor to things in itself but rather to locutions asserting something about concrete cases. To produce an *aletheia* (that is, to say "something true") meant to tell someone "how it is" with reference to a concrete object.

3. *Aletheia* in the Pre-Socratics

There are only very few fragments of the Pre-Socratics in which something is said on truth. Most of them are metaphorical or of a secondary importance. This is probably a reason why historians of philosophy are normally not attracted very much by the theory of truth in the Pre-Socratics; for instance, the index of subjects in G. S. Kirk, J. E. Raven and M. Schoefield 1957 does not contain the word 'truth'. Some philosophers try to derive (e.g. Herbertz 1913) certain consequences for the Pre-Socratics' account of truth from their more general epistemolological views, like direct or naive realism. So interpreted the Pre-Socratics, or rather some of them, especially Democritus, are presented as seeing the nature of truth in 'an agreement of thought and being'. A very similar view is also attributed

to Parmenides for his famous statement "[...] for the same thing is for conceiving as is for being" (cf. Coxon 1986, p. 54). Some authors (see for instance Krapiec 1959) regard this statement as the first strict account of the idea of an intentional relation between thought and its object.

I think that we are not able to derive any substantial theory of truth from the fragmentaric and cryptic texts of the Pre-Socratics. These reconstructions which appeal to their general standpoints have no confirmation in more concrete statements. In particular, no fragment on truth occuring in preserved texts of pre-Socratic philosophers might be literally translated with the help of such words as 'agreement', 'adequacy' or 'correspondence'.

Fortunately, grammarians (see Boeder 1958) have established several important facts for our problem. Namely, the Pre-Socratics extended the use of *aletheia* in such a way that it was no longer limited only to concrete dialogical situations. *Aletheia* (as referring to statements of a sort) for pre-Socratic philosophers is primarily an amount of a knowledge (conceived much more abstractly than in the Homeric era) consisting in a relation of a knowing person to a related object of knowledge. Thus, the statement 'snow is white' belongs to *aletheias* just because snow is white. A more sophisticated description of this usage of *aletheia* might consist in an appeal to a relation of correspondence between a statement and what is stated in it. However, the point is that no such appeal is involved in pre-Socratic "semiotics" concerning *aletheia*.

The observations made by grammarians show at least two things. Firstly, the Pre-Socratics used *aletheia* in a more depersonalized way than their pre-philosophical precedessors. Secondly, this more abstract treatment of *aletheia* must be considered as an essential step toward its predicative use.

4. Plato

Two principal fragments by Plato on truth are these (cf. Jovett 1953):

(1) Socrates: Come now, tell me this. Do you call anything "speaking truths" and "speaking falsehoods"?
Hermogenes: I do.
Socrates: So there would be such things as true and false speech?
Hermogenes: Certainly.
Socrates: So that which speaks of things that are, as they are, would be true speech? And that which speaks of them as they are not, would be false speech?
Hermogenes: Yes (*Cratylus 385 b*).

(2) Stranger: And the true one states about you the things which are

(or the facts) as they are.
Theatheus: Certainly.
Stranger: Whereas the false statement states about you things *different* from the things that are.
Theatheus: Yes.
Stranger: And accordingly states *things that are not* as being.
Theatheus: No doubt.
Stranger: Yes, but things that *exist*, different from things that exist in your case. For we said that in the case of everything there are many things that are and also many that are not.
Theatheus: Quite so (*Sophist 263 b*).

There are many points in both quoted fragments which require comments. Especially, we can ask how Plato sees relations between being and existence. However, without entering into this very difficult problem, we clearly observe that Plato links truth, existence (being) and predication. His account of truth is abstract — personal parameters play only a secondary role in the explanations offered by Socrates and the Stranger.

5. Aristotle

Almost everybody knows that it was Aristotle who proposed the classical (or correspondence) theory of truth for the first time. However, the fact that his writings contain different and often mutually non-equivalent statements on truth is less recognized. This is a sample of Aristotelian explanations concerning the concept of truth (cf. Ross 1924, Acrill 1963):

(3) To say of what is that it is not, or of what is not that it is, is false, while to say of what is that it is, and of what is not that it is not, is true (*Metaphysics* 1011 b);

(4) The fact of the being of a man carries with it the truth of the proposition that he is; and the implication is reciprocal: for if a man is, the proposition wherein we allege that he is, is true, and conversely, if the proposition wherein we allege that he is is true, then he is. The true proposition, however, is in no way the cause of the being of the man, but the fact of the man's being does seem somehow to be the cause of the proposition, for the truth or falsity of the proposition depends on the fact of the man's being or nor being (*Categories* 14 b);

(5) But since that which *is* in the sense of being true or *is not* in the sense of being false, depends on combination and separation, and truth and falsity together depend on the allocation of a pair of contradictory judgements; for the true judgement affirms where

the subject and predicate really are combined, and denies where they are separated, while the false judgement has the opposite of this allocation (*Metaphysics* 1027 b);

(6) [...] he who thinks the separated to be separated and the combined to be combined has the truth, while he whose thought is in a state contrary to that of the objects is in error (*Metaphysics* 1051 b);

(7) It is not because we think truly that you are pale, that you *are* pale, but because you are pale we who say this have the truth (*Metaphysics* 1051 b);

(8) Propositions correspond with facts (*Hermeneutics* 19 b).

The formulation (3) is usually taken as Aristotle's official definition of truth. Now (4) repeats the content of (3) but adds that being is in a sense more basic for truth than an assertion which is qualified as true. The two statements are not equivalent because neither does (4) follow from (3) nor does the reverse entailment hold. Statements (5) and (6) introduce an explicit ontological parameter, namely combination and separation; these statements seem to be equivalent (or at least "nearly" equivalent). On the other hand, there is no direct entailment from (5) (or (6)) to (3) or (4), and back.

Perhaps one might say that 'a is b' is true if and only if the relation which holds between referents of a and b is mapped by the relation holding between a and b, and false if the mapping is not the case. If we decide to label mapping as 'combination' and not-mapping as 'separation', we obtain something very close to (5) and (6). And if we look at combination as correspondence and separation as non-correspondence, (5) and (6) become popular formulations of the classical definitions of truth.

The statement (7) seems to exemplify previous explanations, particularly (3). Finally, (8) explicitly speaks about facts and correspondence but it is only a marginal remark made by Aristotle when he considered the celebrated sea-battle problem. Hence, there are no sufficient reasons to treat (8) as a serious proposal to define the concept of truth.

If we take (3) as Aristotle's official truth-definition (and, *a fortiori*, as the first mature explanation of CCT), then other Aristotelian formulations should be understood rather as more or less auxiliary comments than proper definitions of truth. The point is very important because no idea of correspondence is directly involved in (3). Although, as my previous remarks show, 'combination' can be replaced by 'correspondence' but nothing forces us to dress Aristotle's truth-theory into "correspondence talk". In fact, (3)–(7) may be explained without any reference to such ideas as correspondence, agreement, adequacy or conformity; recall that

(8) is only a marginal remark. I think that the best understanding of what is going on in Aristotle's theory of truth consists in looking at (3) as something which is very closely related to (1) and (2). Then if we think of Plato's philosophy of truth as a further step in the tradition beginning with old Greek poems and continued by the Pre-Socratics, Aristotle should also be considered in the same way. Under this assumption, (3) schematically says how to answer the question: how is it? Although Aristotle supplements (3) with considerable ontological equipment, his main intuition concerning the concept of truth seems very simple.

6. Schoolmen

Various explanations by Peter Abelard of the concept of truth offered in his Logica Ingrediendibus lead to (see De Rijk 1956, p. LIV)

(9) the sentence p is equivalent with 'p is true' if and only if p is the case.

Clearly, (9) anticipates the semantic definition of truth but it was not properly understood in the Middle Ages (nor later).

The most famous medieval explication of the concept of truth comes from Thomas Aquinas. His formulation is this:

(10) Veritas est adequatio intellectus et rei, secundum quod intellectus dicit esse quod est vel non esse quod non est (*De Veritate 1,2*).

The passage which begins with the word *secundum*, is simply a repetition of Aristotle's main formulation (see (3) above). But the first part of (10) — *veritas est adequatio intellectus et rei* — is an obvious addition to Aristotle, actually related to (5) or (6). Usually, (10) is quoted in its simplified version limited to its first part: *veritas est adequatio intellectus et rei*; in fact, this shortened formula is the most popular wording of the classical truth-definition. However, everybody who employs this simplified record of CCT as "Aristotelian", must remember that it is certainly not Aristotelian to the letter. The question whether and to which extent it is Aristotelian in spirit requires special investigations that exceed the scope of this paper. So I restrict myself to some remarks on *adequatio intellectus et rei*.

One can link the meaning of *adequatio* in (10) with the second (Aristotelian) part of this formula. However, Thomas Aquinas also uses such terms as *conformitas*, *correspondentia* and *convenientia* to explain his understanding of CCT. It suggests his *adequatio* expresses (or at least might express) contents which is not quite reducible to Aristotelian intuitions.

What is going on in the first part of (10)? There are several possible answers. Let me indicate three. Firstly, *veritas est adequatio intellectus*

et rei may be regarded as a counterpart of (5) or (6). Secondly, the fact that the *adequatio*-formula opens Thomas' definition seems to suggest that he changed the centre of gravity in the Aristotelian truth-theory in such a way that *adequatio, correspondentia, conformitas* or *convenientia* became crucial ideas in defining truth. Thirdly, the *adequatio*-formula was invented by the Schoolmen to capture intuitions concerning truth in a simple way; the Schoolmen very much liked brief formulations. It is very difficult to decide today which interpretation (I am very far from claiming that my three cases exhaust all possible interpretations of (10)) is correct with respect to Aquinas' original intentions. However, the next development of Thomism rather followed the second interpretation. For instance, Suarez says that *veritas transcendentialis significat entitatem rei, connotondo cognitionem seu conceptum intellectus, cui talis entitas conformatur vel in quo talis res representatur* (*Disputationes metaphysicae*, 8, 2.9). The content of (3) is completely absent in Suarez. He proposes instead an analysis of truth with the help of the concept of *representatio* and seems to assume that a *conformitas* (*adequatio, correspondentia*) holds between thoughts and their objects. That is what I mean by "changing the centre of gravity". Most post-medieval thinkers adopted this route in their thinking on truth and tried to explain how *adequatio* should be understood.

It is now proper to introduce an important distinction (see Woleński and Simons 1989), namely that of weak and strong concept of correspondence. If the concept of correspondence is goverened by (3) (or similar statements), we are dealing with correspondence in the weak sense. On the other hand, Suarez's approach employs correspondence in the strong sense. I am inclined to regard the distinction of the two concepts of correspondence as fairly crucial for the history of CCT. Thus, we must always ask which concept of correspondence is used in particular truth-theories because many difficulties with interpreting philosophers' view on truth are rooted in their view of the distinction in question. As far as the matter concerns the concept of correspondence, it has been explained by notions like sameness, similarity, model, picture, co-ordination, isomorphism or homomorphism (see some definitions listed in section 9 below).

Let me finish this section with some historical remarks (see Gilson 1955). Thomas Aquinas notes that his definition of truth is derived from *Liber de definitionibus* by Isaac Israeli; Aquinas also refers to Avicenna in this context. However, *adequatio* does not occur in Israeli's truth-definition which (in Latin version) is this: *Et sermo quidem dicentis: veritas est quod est, enuntiativus est natura veritatis et essentiae ejus, quonian illud sciendum quod est res, vera est; est veritas nonnisi quod est*; this formula is

fairly Aristotelian. Avicenna in his *Metaphysics* says (in Latin translation) that *veritas [...] intelligitur dispositio in re exteriore cum est ei aequalitas*; the last word suggests the strong sense of 'correspondence'. It was William of Auvergne who introduced the term *adequatio* in philosophy for the first time. He refers (in *De universo*) to Avicenna in the following way: *[...] et hoc [intentio veritas] ait Avicenna, est adequatio orationis et rerum.* Then William adds that the truth is *adequatio intellectus ad rem.* In Albertus Magnus' treatise *De bono* we find that truth is *adequatio rei cum intellectu.* Then comes (10).

7. Modern philosophy from the Renaissance to Kant

(11) Veritas auterm enunciationis seu iudicii nihil aliud est quam conformitas ore factae aut iudicii mente peracto cum ipsa enuntiata seu iudicata (Gassendi, *Syntagma philosophiae Epicuri* I, 1);

(12) [...] mot verité, en sa propre signification, denote la conformité de la pensée avec l'objet (Descartes, A letter to Mersenne, 1639).

(13) Truth is the marking down in words the agreement or disagreement of ideas as it is [...] [Signs] [...] contain *real truth* when [...] are joined, as our ideas agree, and when our ideas are such as we know are capable of having an existence in nature but by knowing that such (Locke, *An Essay Concerning Human Understanding*, IV, V, § 9).

(14) Those propositions are true which express things as they are; or truth is conformity of those words or signs, by which things are expressed, to the things themselves (Wollaston, *The Religion of Nature Delineated*, I).

(15) Idea vera debet convenire cum suo ideato (Spinoza, *Ethica*, axiom VI);

(16) Contentons nous de chercher la verité dans le correspondence des propositions qui sont dans l'esprit, avec les choses dont il s'agit (Leibniz, *Nouveaux Essays*, IV.5, § 11).

(17) Veritas est consensus iudicii nostri cum objecto seu re representata (Wolff, *Philosophiae rationalis sive logica*, § 505);

(18) Die Namenklärung der Wahrheit, dass sie nämlich die Übereinstimmung der Erkenntnis mit ihren Gegenstande sei, wird hier geschenkt und vorausgesetzt (Kant, *Kritik der reinen Vernuft*, A 58).

These samples show that philosophers who represented radically different epistemological views used the formula "truth consists in confor-

mity (agreement) of thought with its object" to express their own truth-theories. This is a surprise because we know that they did not share the same views on truth. The Cartesian account of truth is much better captured by his statement that *verum est quod clarae ac distinctae percipio* which expresses the main tenet of the evidence theory. Spinoza and Leibniz belong to the family of coherentists; Wolff is a fairly Leibnizian philosopher who defends his master against various objections. Kant is famous for his strong attack on CCT. Only Gassendi, Locke and Wollaston are genuine correspondists in this company. Thus, the correspondence formula was used in the 16th and 17th centuries as a convenient scheme for recording very different, often mutually conflicting, intuitions on truth. However, independently of differences in particular cases, the concept of correspondence has a constant element in all formulas (11)–(18), namely it occurs in its strong meaning. So the distance between (11)–(18) and (3) is rather far.

8. The Nineteenth century

Bernard Bolzano's semantic approach to the concept of truth is perhaps from the contemporary point of view the most interesting contribution to CCT in the 19th century. Although interesting, it was not influential because Bolzano's work was not appreciated in a proper way at that time; to some extent, Bolzano's fate resembles that of Petrus Abelard.

Several important criticisms of the classical theory of truth appeared in the 19th century. Jacob Friedrich Fries advanced Kantian objections in this way: "We cannot, as is usually done, speak of truth as opposed to error by saying that truth is the correspondence of a representation with its object. We can only say that the truth of a judgement is its correspondence with the immediate cognition of reason in which it is grounded. [...] The general meaning of truth is only the internal agreement of mediate cognition with the immediate. This immediate recognition possesses its truth from its sheer presence of reason" (Fries, 1989, p. 31; the German original was published in 1805). This passage contributes to how Kant understood correspondence and, moreover refreshes some traditional objections against CCT (stated as far back as by ancient sceptics) by pointing out that there is no truth-criterion if truth is conceived as conformity of our knowledge with transcendental reality.

Franz Brentano (who himself defended a kind of evidence theory of truth) raised other objections against the classical theory of truth (see Brentano 1930). For him, the *adequatio*-formula leads to a fundamental misinterpretation of Aristotle's conception of truth. Moreover, Brentano

argued that this formula raises serious difficulties of its own, independently of its historical relation to Aristotle or any other author. The difficulties are these:

(a) Let A be a sentence and F^A a fact corresponding to A. To assert that A corresponds to F^A one must use a sentence B which says that A corresponds to F^A. However, it raises the question of correspondence of B to F^B and so *ad infinitum*. For Brentano, the outlined argument shows that the correspondence theory of truth is inevitably burdened by *regressus ad infinitum*.

(b) If truth consists in correspondence with existing reality, we must ask what negative existentials, for instance, the statement 'Pegasus does not exist' correspond to.

(c) For Brentano, every logical tautology may be translated into a negative existential statement. So we encounter the problem of truth for tautologies.

Independently of Brentano, also Gottlob Frege (see Frege 1892, 1918) and Francis Bradley (see Bradley 1914) raised the *regressum* objection. Moreover, for both Frege (truth is not definable for him) and Bradley (he defended a coherence-theory), each theory of truth based on the concept of correspondence must admit what has been called a Great Fact to which all true propositions correspond. However, Frege and Bradley maintained that this is an obvious absurdity because the correspondence theory requires that if a proposition is true, it corresponds not to the whole reality but to a particular fact.

Nevertheless, the correspondence theory of truth was fairly popular among philosophers in the 19th century. Let me mention three German definitions (though the respective books were published after 1900, they expressed thoughts "belonging" to the 19th century):

(19) Die Wahrheit unserer Erkenntnis ist die, Übereinstimmung unserer Urteile mit der Wirklichkeitswelt; da unsere Urteile rückschreitend bis auf Sinneseindrücke zurückführen so ist die Wahrheit unserer Erkenntnis schliesslich auch die übereinstimmung unserer Vorstellungen und Sinneseindrücke mit der "Wirklichkeit" (Mauthner 1902, p. 360).

(20) Ungesucht bietet sich die alte aristotelische Antwort dar, die bis in die gegenwart herein ihr ansehen behauptet hat: das Urteil misst sich, indem es wahr sein will, an der Wirklichkeit übereinstimmen. Die Unhaltbarkeit dieser Definition fällt indessem in die Augen, sobald man ihr nun ihre genaue Fassung, gibt. Nicht von einer Übereinstimmung des Urteils, sondern nur von einer Übereinstim-

mung des Urteilsgegenstands mit der Wirklichkeit kann die Rede sein. In der Tat ist dies der genuine Sinn der aristotelischen Wahrheitstheorie (Maier 1926, p. 223).

(21) *Materiale* [Wahrheit] ist, ganz allgemein, "Übereinstimmung" (Konformität) des Denkens mit den Sein. Es gibt aber zwei Arten der Materialen [Wahrheiten]: a) *Empirisch-immanente* [...]. Hier bedeutet die "Übereinstimmung" von Denken und Sein [...] *nicht* die *Abbildung* u. dgl. des Sienden im und durch das Denken, sondern *Übereinstimmung* des Einzelurteils mit der methodisch gesetzen Realität, die in einem System von Wahrnemungs- und Urteilsnotwendigkeiten sich darstellt [...]. b) *Metaphysische* [Wahrheit] ist die Übereinstimmung des Denkens mit der absoluten Wirklichkeit [...]. Auch hier kann von einen "Abbilden" keine Rede sein, sondern die "Übereinstimmung" bedeutet hier ein mehr oder weniger treffendes "Nachkonstruieren" der transzendenten Wirklichkeits-Verhältnisse *in immanenten, begrifflichen Symbolen* (Eisler 1930, pp. 450/451).

In fact, the definitions (19)–(21) are attempts to adjust the correspondence theory (in the strong meaning of correspondence) to Kantian objections; this tendency is especially evident in Maier's case who attributes correspondence in its strong sense to Aristotle. Eisler's views are particularly interesting in this context. His *Dictionary* summarizes German philosophical experience at the end of the 19th century. Reading his explanations, we can clearly see how difficult it was to explain words like 'Konformität' or 'Übereinstimmung'. These key words are put in quotes or surrounded by phrases like 'mehr oder weniger'.

9. The Twentieth century

(22) Every judgement is a relation of mind to several objects, one of which is a relation; the judgement is true if the relation which is one of the objects relates to the other objects, otherwise it is false (Russell 1910, p. 156).

(23) The belief is true when the objects are related as the belief asserts that they are. Thus the belief is *true* when there is a certain complex which must be a definable function of the belief, and which we shall call the *corresponding* complex, or the *corresponding fact* (Russell 1984, p. 144).

(24) A judgement that *uniquely designates* a set of facts is called *true* [...] the concept of truth was almost always defined as an agree-

ment between thought and its object — or, better, between judgement and what is judged [...] here is no doubt that this definition expresses a correct conception. But which conception? [...] the notion of agreement, in so far as it is to mean sameness or similarity, melts away under the rays of analysis, and what is left is unique coordination. It is in the latter that the relationship of true judgements consists, and all those naive theories according to which our judgements and concepts are able in some fashion to "picture" reality are completely demolished. No other sense remains for the word "agreement" than that of unique coordination or correspondence (Schlick 1974, p. 61; the German original was published in 1918).

(25) 4.011 A proposition is a picture of reality [...]. A proposition is a model of reality [...].
4.022 [...] A proposition *shows* how things stand *if* it is true [...].
4.05 Reality is compared with proposition.
4.06 Propositions can be true or false only by being pictures of reality (Wittgenstein 1922).

(26) The propositional function p is true is simply the same as p (Ramsey 1978, p. 45; the first edition of Ramsey's papers was published in 1931).

(27) We should like our definition to do justice to the intuitions which adhere to the *classical Aristotelian conception of truth* (see (3) above — J. W.). If we wish to adapt ourselves to modern philosophical terminology, we could perhaps express this conception by means of the familiar formula:

The truth of a sentence consists in its agreement with (or correspondence to) reality.

(For a theory of truth which is to be based upon the latter formulation the term "correspondence theory" has been suggested.)
[...] we could possibly use for the same purpose the following phrase:

A sentence is true if it designates an existing state of affairs.

However, all these formulations can lead to various misunderstandings, for none of them is sufficiently precise and clear (though this applies much less to the original Aristotelian formulation than to either of the others; at any rate, none of them can be considered a satisfactory definition of truth. It is up to us to look for a more precise expression of our intuitions [...].
Thus, if the definition of truth is to conform to our conceptions, it

must imply the following equivalence:

The sentence "snow is white" is true, if and only if snow is white (Tarski 1944, pp. 342/343).

[...] we arrive at a definition of truth and falsehood simply by saying *a sentence is true if it is satisfied by all objects, and false otherwise* (Tarski 1944, p. 353).

(28) Reverting to the analysis of truth, we find that in all sentences of the form '*p* is true', the phrase 'is true' is logically superfluous. When, for example, one says that the proposition 'Queen Anne is dead' is true, all that one is saying is that Queen Anne is dead. Thus, to say that a proposition is true is just to assert it, and to say that it is false is just to assert its contradictory. And this indicates that the terms 'true' and 'false' connote nothing, but function in the sentence simply as marks of assertion and denial" (Ayer 1946, pp. 117/118).

(29) An atomic sentence [...] consisting of a predicate followed by an individual constant is true if and only if the individual to which the individual constant refers possesses the property to which the predicate refers (Carnap 1947, p. 5).

(30) I accept the commonsense theory (defended and refined by Alfred Tarski) that truth is correspondence with facts (or with reality); or, more precisely, that a theory is true if and only if it corresponds to the facts (Popper 1972, p. 44).

(31) The combination 'it is a fact that' is vacuous [...] 'It is a fact that snow is white' reduces to 'Snow is white'. Our account of the truth of 'Snow is white' in terms of facts has now come down to this: 'Snow is white' if and only if snow is white. [...] Here, as Tarski has urged, is the significant residue of the correspondence theory of truth. To attribute truth to the sentence is to attribute whiteness to the snow. Attribution of truth to 'Snow is white' just cancels the quotation marks and says that snow is white. Truth is disquotation (Quine 1987, p. 213).

The formulations (22)–(31) present a considerable variety of definitions intending to capture the classical intuitions. We can preliminary divide these proposals into three groups:

(a) strong correspondence definitions (Russell, Wittgenstein, Schlick, perhaps Popper);

(b) semantic definitions (Tarski, Carnap);

(c) redundancy and disquotational definitions (Ramsey, Ayer, Quine).

It is interesting that in (a) and (c) we find a reference to Tarski — Popper does it in the group (a) and Quine in (c). There is an irony here because Popper and Quine defend with help of Tarski those formulations which he regarded as wrong. For Tarski, (30) is simply obscure but disquotational and redundancy theories have difficulties with a proper analysis of the following statement: logical consequences of true sentences are true.

Both Schlick and Tarski criticize traditional versions of the classical truth-definition but they do it in radically different ways: Schlick tries to strenghten the concept of correspondence, Tarski entirely abandons the concept of strong correspondence in favour of something that perhaps could be called 'semantic correspondence' (satisfaction by all objects).

I think that the concept of semantic correspondence is a very good explicatum for the concept of weak correspondence. Now, if (3) is to be interpreted *via* weak correspondence, the semantic theory of truth has an obvious philosophical import as a modern realization of Aristotelian intuitions. This view is opposite to Max Black's very often quoted statement: "[...] the neutrality of Tarski's definition with respect to the competing philosophical theories of truth is sufficient to demonstrate its lack of *philosophical relevance*" (Black 1948, p. 63). Let me remind you that the formula *veritas est adequatio intellectus et rei* has been employed (see section 7. above) by competing philosophical theories of truth but, as far as I know, nobody has considered it as devoid of "philosophical relevance".

REFERENCES

ACRILL, J. L. 1963, *Aristotle's Categories and De Interpretatione*, tr. J. L. Acrill, Clarendon Press, Oxford.

AYER, A. J., 1946, *Language, truth and logic*, 2nd ed., Gollancz, London.

BLACK, M., 1948, *The semantic definition of truth*, Analysis, vol. 8, pp. 49–63.

BOEDER, H., 1959, *Der fruegriechische Wortgebrauch von Logos und Aletheia*, Archiv für Begriffsgeschichte, vol. 4, pp. 82–112.

BRADLEY, F., 1914, *Essays on truth and reality*, Clarendon Press, Oxford.

BRENTANO, F., 1930, *Wahrheit und Evidenz*, Meiner, Leipzig.

CARNAP, R., *Meaning and Necessity*, University of Chicago Press, Chicago.

COXON, A. H., 1986, *The fragments of Parmenides: A critical text with introduction, translation, ancient* testimonia *and a commentary*, tr. A. H. Coxon, Van Gorcum, Assen/Maastricht.

DE RIJK, L., 1956, *Petrus Abelardus, Dialectica*, ed. L. De Rijk, Van Gorcum, Assen/Maastricht.

EISLER, R., 1930, *Woerterbuch der philosophischen Begriffe*, vol. III, Mittler, Berlin.

FREGE, G., 1892, *Über Sinn und Bedeutung*, Zeitschrift für Philosophie und philosophische Kritik, vol. 100, pp. 25–50.

FREGE, G., 1918, *Die Gedanke: eine logische Untersuchung*, Beiträge zur Philosophie der deutschen Idealismus, vol. 1, pp. 58–77.

FRIES, J. F., 1989, *Knowledge, belief, and aesthetic sense*, tr. K. Richter, Dinter, Köln.

GILSON, E., 1955, *History of Christian philosophy in the Middle Ages*, Random Press, New York.

HERBERTZ, R., 1913, *Das Wahrheitsproblem in der griechischen Philosophie*, Reimer, Berlin.

JOVETT, B., 1953, *The Dialogues of Plato*, tr. B. Jovett, 4th ed., Clarendon Press, Oxford.

KIRK, G. S., RAVEN, J. E. and SCHOEFIELD, M., 1957, *The Presocratic Philosophers*, Cambridge University Press, Cambridge.

KRAPIEC, M., 1959, *Realizm ludzkiego poznania Realism of human knowledge*, Pallotinum, Poznán.

MAUTHNER, F., 1902, *Zur Grammatik und Logik* , 3rd vol. of *Beitraege zu einer Kritik der Sprache*, Cotta, Stuttgart.

MEIER, H., 1926, *Wahrheit und Wirklichkeit*, Mohr, Tübingen.

POPPER, K., 1972, *Objective knowledge*, Clarendon Press, Oxford.

QUINE, W. V., 1987, *Quiddities An intermittently philosophical dictionary*, Harvard University Press, Cambridge, Mass..

RAMSEY, F., 1978, *Foundations Essays in Philosophy, Logic, Mathematics and Economics*, Routledge and Kegan Paul, London.

ROSS, D. W., 1924, *Aristotle's Metaphysics*, tr. W. D. Ross, Clarendon Press, Oxford.

RUSSELL, B., 1912, *Philosophical Essays*, Longmans, London.

RUSSELL, B., 1912, *The Problems of Philosophy*, William & Norgate, London.

RUSSELL, B., 1984, *Theory of Knowledge The 1913 Manuscript*, Allen and Unvin, London.

SCHLICK, M., 1974, *General theory of knowledge*, tr. A. Blumberg, Springer, Wien.

TARSKI, A., 1944, *The semantic theory of truth and the foundations of semantics*, Philosophy and Phenomenological Review, vol. 4, pp. 341–374.

WOLEŃSKI, J. and SIMONS, P., 1989, *De veritate: Austro-Polish contributions to the theory of truth from Brentano to Tarski*, The Lvov-Warsaw School and the Vienna Circle, ed. K. Szaniawski, Kluwer Academic Publishers, Dordrecht.

Logic, Methodology and Philosophy of Science IX
D. Prawitz, B. Skyrms and D. Westerståhl (Editors)

NOTES ON THE VALUE OF SCIENCE

LARS BERGSTRÖM

Stockholm University

It is generally believed that science is a good thing. (I use the term "science", in this paper, to include not only the natural sciences, but also the social sciences and the humanities.) Many people—including most scientists—take it for granted that scientific knowledge is valuable for its own sake. In addition, scientific research has very important social effects, and I think the predominant view is that while some of these may be bad or neutral, the total impact of science on society is positive rather than negative. After all, we do spend a lot of money on science, and scientists have a lot of prestige in our society. This might be explained by the assumption that most people think that science is valuable. But is the belief true? Is science, on the whole, good or bad? This is the problem I want to discuss in the present paper.[1]

Everyone would agree that so far science has had some positive as well as some negative effects. For example, it has given us electricity, which may be used to make our lives more comfortable, but it has also given us terrible weapons, which may one day put an end to our very existence. Einstein once described the situation as follows:

> Penetrating research and keen scientific work have often had tragic implications for mankind, producing, on the one hand, inventions which liberated man from exhausting physical labor, making his life easier and richer; but on the other hand, introducing a grave restlessness into his life, making him a slave to his technological environment, and—most catastrophic of all—creating the means for his own mass destruction.[2]

[1]This paper partly derives from a talk given in January 1990 to a seminar on "Humanistic Aspects of Scientific and Technological Progress" at the Institute of Philosophy of the USSR Academy of Sciences in Moscow. I am grateful to the participants for many helpful comments. I also wish to thank Hans Mathlein, Torbjörn Tännsjö, and Jan Österberg of Stockholm University for comments on the first written version.

[2]Albert Einstein, "A message to intellectuals" (1948), p. 148, in *Ideas and Opinions*, New

Most people would accept this statement. However, there may be some disagreement over other alleged effects of science. For example, some people may claim that only certain natural sciences, like physics, chemistry, and biology have negative effects, and that other sciences, e.g. the humanities, have only good effects, in addition to being valuable for their own sake. Rousseau, on the other hand, makes no such distinctions when he claims, in his famous first *Discourse* of 1750, that "our minds have been corrupted in proportion as the arts and sciences have improved".[3] He says that the sciences "generate idleness" and contribute to "the destruction and defamation of all that men hold sacred".[4] Rousseau's ideal is Sparta, which is "as famous for the happy ignorance of its inhabitants, as for the wisdom of its laws", and which is also "eternal proof of the vanity of science".[5] This view is very far from being generally accepted.

In order to make an overall evaluation of science, it seems that one would have to do two things. First, one would have to decide what the positive and negative aspects of science are. Second, one would have to weigh these positive and negative aspects against one another and decide whether or not the positive aspects outweigh the negative ones.

Different people may come to different conclusions here. A majority believes that the positive aspects prevail, but there are also some dissidents. For example, in a note from 1947, Wittgenstein writes:

> It isn't absurd [...] to believe that the age of science and technology is the beginning of the end for humanity; that the idea of great progress is a delusion, along with the idea that the truth will ultimately be known; that there is nothing good or desirable about scientific knowledge and that mankind, in seeking it, is falling into a trap. It is by no means obvious that this is not how things are.[6]

A similar, but even stronger thesis has been advanced by Michael Dummett. He says that

> it seems to me indisputable, with hindsight, that we should be, on balance, far better off than we are if, in 1900 or in 1920, all scientific research had come to a permanent stop. With the

York: Crown Publishers, 1954, pp.147-51.

[3]Jean Jacques Rousseau, "A Discourse on the Moral Effects of the Arts and Sciences", in *The Social Contract and Discourses* , translated with introduction by G.D.H. Cole, Everymans's Library, London: Dent, 1968, p. 123.

[4]See ibid., pp. 131-2, 135-6.

[5]Ibid., p. 126.

[6]Ludwig Wittgenstein, *Culture and Value*, Oxford: Blackwell, 1980, p. 56.

> experience of what happened, we have little reason to doubt that the net practical result of future research will be increasingly disastrous.[7]

It seems to me that the pessimism expressed by Wittgenstein, Dummett, and others ought to be taken seriously. It is not obviously correct, but it is not obviously wrong either.

1. Science and knowledge.

The main argument for engaging in scientific activities is that this is a way—and perhaps the only way, or the best way—of gaining *knowledge*. And knowledge, in turn, is supposed to be valuable, either intrinsically or extrinsically (instrumentally) or both.

However, contrary to popular opinion it might be argued that scientific work does not really give us much knowledge. I am taking it for granted, then, that knowledge (in a strict sense) has to be true and justified. This is in accordance with the standard account of knowledge.[8] If I know that p, then p is true and I am justified in believing that p is true. Now, it seems clear that science produces theories. So, if science produces knowledge, the theories in question have to be both true and justified. But several influential positions in modern philosophy of science seem to imply that science cannot or may not produce theories which are both true and justified. Let us consider some of these positions.

(1) Popper. According to Karl Popper there is no criterion of truth.[9] Even observational statements may be mistaken—mainly because "all observation involves interpretation in the light of our theoretical knowledge"[10]— and there can be no inductively valid inference from observational statements to theories. A given scientific theory may happen to be true, but we never have any good reason to believe that it is true. We are never justified in believing that a theory is true. Our so-called "knowledge" merely consists of conjectures: "even if we hit upon a true theory, we shall as a rule be merely guessing, and it may well be impossible for us to know that it *is* true".[11]

---

[7]Michael Dummett, "Ought research to be unrestricted?", *Grazer Philosophische Studien*, vol. 12/13 (1981) p. 292.

[8]See e.g. the article "Knowledge and belief" in P. Edwards (ed.) *The Encyclopedia of Philosophy*, New York and London: Macmillan and The Free Press, 1967, vol. 4, pp. 345-52, esp. p. 345.

[9]See e.g. Karl R. Popper, *Conjectures and Refutations*, New York and London: Basic Books, 1962, p. 28.

[10]Ibid. p. 23.

[11]Ibid., p. 225.

Popper sometimes seems to imply that we have some knowledge, as e.g. when he says that "by far the most important source of our knowledge—apart from inborn knowledge—is tradition".[12] But in such cases he seems to be using the term "knowledge" in a loose or weak sense. What is referred to as "knowledge" in this sense may very well be false and unjustified. For all we know, it might still have some value, but it is certainly not knowledge in the strict, standard sense. Knowledge in the strict sense is impossible according to Popper.

(2) Kuhn. According to Thomas Kuhn, a theory cannot even be true. Kuhn seems to hold that the term "true" has only intra-theoretic applications, and that there is no sense in which one theory may be a better approximation to the truth than another.[13] Hence, on Kuhn's view, we cannot know that a theory as a whole is true or approximately true. It may still be possible to know some things, but it seems that Kuhn is rather pessimistic about the epistemological potential of science.

Of course, this does not mean that Kuhn is pessimistic about the value of science. He believes that some theories are better than others—according to criteria which are internal to science—and that, in general, later theories are better than earlier ones.[14] His position is also compatible with the view that the scientific enterprise is useful and/or valuable for its own sake. It is just that any value it may have must be independent of our coming to know the truth.

(3) Quine. One of W.V. Quine's most well-known theses is that theories are underdetermined by all possible observations.[15] But if theories are underdetermined in Quine's sense, it seems that we can have no real evidence for them. Whatever observation would be counted in favour of a given theory counts equally in favour of some completely different theory. There is always more than one "best explanation" of any given set of data. Hence, we can never know that a given theory is true. Our evidence can never single out our own theory from a set of rival theories.

Quine is still willing to say, of any theory that he himself accepts, that it is true. There is nothing objectionable about this, for by saying that a

[12]Ibid., p. 27.

[13]See e.g. Thomas S. Kuhn, "Reflections on my critics", in Imre Lakatos and Alan Musgrave (eds.) *Criticism and the Growth of Knowledge*, Cambridge: Cambridge University Press, 1970, pp. 321-78, esp. pp. 264-6.

[14]See e.g. ibid., p. 264.

[15]For a recent formulation of the underdetermination thesis, see W.V. Quine, *Pursuit of Truth*, Cambridge, Mass.: Harvard University Press, 1990, pp. 95-102. For a discussion of the thesis and further references, see Lars Bergström, "Quine on underdetermination", in R. Barrett and R. Gibson (eds.) *Perspectives on Quine*, Oxford: Blackwell, 1990, pp. 38-52.

theory is "true", Quine just expresses his acceptance of it; in his own words, "to call a sentence true is just to reaffirm it".[16] Maybe he would even be willing to say, of theories that he accepts, that he *knows* that they are true. By saying this he would merely express his acceptance of the theories in question and perhaps also his belief that this acceptance is "justified" in some sense.[17] However, the theories in question may still be false. And, more to the point, underdetermination still seems to rule out the possiblity of justification in the sense of a good reason for believing that the theories are true. If Quine is right, there seems to be no room for knowledge in the strict sense. There may be perfect coherence within one's total theory of the world, but underdetermination seems to guarantee that there would be equally perfect coherence within some completely different total theory.

(4) Instrumentalism. Another threat to scientific knowledge is instrumentalism. According to instrumentalism, scientific theories are neither true nor false; instead, they are tools which may be used to predict future occurrences. Tools may or may not be useful, but they do not tell us anything about the nature of reality. Tools do not tell us anything at all. We may be able to use a given theory for predicting what will happen under certain conditions, but the theory does not tell us why this will happen. Hence, if instrumentalism is right, the common belief that science gives us knowledge about the world is to a large extent mistaken.

A similar conclusion can be arrived at from the assumption that the acceptance of a theory consists in the belief that the theory is empirically adequate. This may be combined with the view that theories are true or false. The point is that the acceptance of a theory does not involve the belief that what the theory says about theoretical (non-empirical) states and events is true. An account of this kind has been developed by Bas van Fraassen.[18] In so far as scientific theories are only meant to be accepted in this way, science gives us no knowledge about the unobservable features of the world. Of course, it may still give us knowledge concerning the observable world. But I think most people believe that science tells us more than that.

(5) The pessimistic induction. If you reject instrumentalism and adopt a realist conception of scientific theories, you are faced with another argument against the possibility of scientific knowledge. For the history of

[16]W.V. Quine, "On empirically equivalent systems of the world", *Erkenntnis*, 9 (1975), pp. 313-28, esp. p. 327.

[17]See W.V. Quine, *Quiddities*, Cambridge, Mass.: Harvard University Press, 1987, pp. 108-10. Quine also points out that the concept of knowledge "does not meet scientific and philosophical standards of coherence and precision", ibid., p. 109.

[18]See e.g. Bas C. van Fraassen, *The Scientific Image*, Oxford: Clarendon Press, 1980, p. 12.

science seems to suggest that every theory which is accepted by the scientific community at some time will be rejected sooner or later. One possible explanation of this is that there are so many ways in which a theory may be wrong and only one way in which it can be right. This point is stressed by Rousseau in the following passage:

> What a number of wrong paths present themselves in the investigation of the sciences! Through how many errors, more perilous than truth itself is useful, must we pass to arrive at it? The disadvantages we lie under are evident; for falsehood is capable of an infinite variety of combinations; but the truth has only one manner of being.[19]

If we learn about the nature of reality by trial and error—as Popper suggests[20] — and if the nature of reality is very complicated, it is only to be expected that our theories are mostly wrong. It has even been held that no theory will be accepted for more than two hundred years.[21] A similar judgement has been expressed by a leading sociologist of science as follows:

> After all, the majority of all the theories which scientists have ever put forward have been rejected as false or misconceived, and the majority of the findings which they have reported have been forgotten. Scientific knowledge has an extremely short lifetime. The knowledge routinely accepted and used in any scientific field is on the whole extraordinarily recent: scarcely any fields make use of materials more than a few decades old, and such older material as is used is very rarely accepted just as it stands. Yet because we place such trust in it now, many people have difficulty in seeing that our present knowledge is likely to be treated in three or four generations much as we ourselves treat the knowledge of three or four generations ago.[22]

In other words, all scientific theories may well be false. If this is so, it seems that science produces (useful) delusions rather than knowledge.

(6) Interpretations. Instrumentalism is perhaps a plausible account of many theories in the natural sciences. It is less plausible in the humanities and the social sciences. Theories within these latter areas cannot so easily

[19]Rousseau, "Discourse", pp. 130-1.

[20]See e.g. Popper, *Conjectures and Refutations*, pp. vii and 312-3.

[21]See W.H. Newton-Smith, *The Rationality of Science*, Boston, London and Henley: Routledge, 1981, p. 14.

[22]Barry Barnes, *About Science*, Oxford: Blackwell, 1985, p. 66.

be regarded as tools for prediction. There is not much prediction within these sciences—and even less successful prediction. However, theories in the humanities and the social sciences are often regarded as "interpretations",[23] and such interpretations are sometimes held to be neither true nor false.[24] If this is right, interpretations do not express or contain knowledge.

The upshot of all these considerations, then, is that science may produce much less knowledge than is ordinarily assumed. From the point of view of those who believe that science is valuable because it produces knowledge, this is bad news.

2. Knowledge and ignorance.

However, for the sake of argument, let us now accept the more normal view that science has given us a lot of knowledge. In any case, most of us would agree that science has provided us with much low-level empirical knowledge and technical know-how which is amazingly reliable. Nevertheless, it is paradoxical that science has also increased our ignorance in many ways.

The point may be put this way. As science progresses, more things are known, but at the same time each person knows less of what there is to know. Scientific progress has led to an extreme specialization and fragmentation of the scientific enterprise.[25] Even so, the literature within any given special field is unsurveyable. I think it is fair to say that for almost every important scientific problem it is completely impossible to find out what has been written about it. Even before the Second World War the situation was desperate. J.D. Bernal reports that there were 33.000 different scientific periodicals in 1934.[26] Maybe there are ten times as many today? Or more? The rate of growth is truly terrible! For example, Bernal also tells us that the number of entries in the *Biological Abstracts* had grown from 14.506 in 1927 to 21.531 in 1934. That is approximately a 50% increase in only seven years. In general, it has been estimated (in the 1960s) that for at least two or three centuries, "the crude size of science in manpower or in publications tends

[23]See e.g. Charles Taylor, "Interpretation and the sciences of man", *The Review of Metaphysics*, vol. 25 (1971), pp. 3-51.

[24]See e.g. Joseph Margolis, *Art and Philosophy*, Atlantic Highlands, N.J.: Humanities Press, 1980, Ch. 6, and Lars Bergström, "Explanation and interpretation of action", *International Studies in the Philosophy of Science*, vol. 4 (1990) pp. 3-15, esp. pp. 13-4.

[25]For example, it has recently been reported that "A catalogue of fields of study at German universities at present lists more than 4000 fields", see Martin Carrier and Jürgen Mittelstrass, "The unity of science", *International Studies in the Philosophy of Science*, vol 4 (1990) pp. 17-31, esp. p. 17.

[26]J.D. Bernal, *The Social Function of Science*, London: Routledge, 1939, p. 117.

to double within a period of 10 to 15 years".[27] The result is frightening. Already in 1939, Bernal claims that

> it has become impossible for the average scientific worker, who does not wish to devote the major part of his time to reading, to keep up with the progress in his own field.[28]

This was over 50 years ago. Today the situation is certainly much worse. Of course, much of what is written is not worth reading, but this is a poor consolation, for in order to know what is worth reading, one has to read everything. This may not be strictly true, but it is at least sufficiently true to present us with a genuine problem. Quine puts the point as follows:

> The mass of professional journals is so indigestible and so little worth digesting that the good papers, though more numerous than ever, are increasingly in danger of being overlooked.[29]

Quine is referring primarily to philosophical journals, but I am inclined to believe that the same is true in most or all fields. For some years I have had the habit of asking professors from various disciplines, whom I happen to meet, whether they would agree that something like 75% of the research done in their own field is bad or uninteresting. So far everyone has agreed to this. It is also confirmed by Bernal, who claims that scientific publications are "of very unequal value; a large proportion of it, possibly as much as three-quarters, does not deserve to be published at all".[30] So, a large percentage of the scientific work which is published is bad or boring or both, and possibly the percentage of bad work is larger the more that is published.[31]

A related point is this. Partly because of the increased specialization, and partly because of scientific progress, scientific theories have become increasingly difficult to understand. Most people may be able to grasp the principles behind the steam engine, for example, but only a small minority understand the functioning of a laser or a computer. Quantum theory is certainly more difficult than Newtonian mechanics. Hence, it may be safely assumed that educated people have never before been as ignorant of

[27]Derek J. de Solla Price, *Little Science, Big Science ... and Beyond*, New York: Columbia University Press, 1986, p. 5. However, as Barry Barnes points out, "the rate of scientific growth has fallen off very markedly since the early 1960s", Barnes, *About Science*, p. 5.
[28]Bernal, *The Social Function of Science*, p. 117.
[29]W.V. Quine, *Theories and Things*, Cambridge, Mass.: Belknap Press, 1981, p. 197.
[30]Bernal, *The Social Function of Science*, p. 118.
[31]The last point is also supported by Quine. He says that new journals "were needed by authors or articles too poor to be accepted by existing journals", *Theories and Things*, p. 196.

the science of their time as they are today. This is also true of the scientists themselves. Something like this is suggested by Kuhn in the following passage:

> Is it not possible, or perhaps even likely, that contemporary scientists know less of what there is to know about their world than the scientists of the eighteenth century knew of theirs? Scientific theories, it must be remembered, attach to nature only here and there. Are the interstices between those points of attachment perhaps now larger and more numerous than ever before?[32]

In addition, there is also the mechanism that the more one knows about something, the more sceptical one becomes of various theories and ideas in the field in question, and the more one becomes aware that one really knows very little.

Hence, my general conclusion is that because of the scientific progress, there is more ignorance than before. If ignorance is bad, this is an unfortunate effect of science.

3. The intrinsic value of scientific knowledge.

If knowledge is good, then either it is good in itself, or it is good as a means to something else. Its value is either intrinsic or extrinsic—or both. I shall discuss the extrinsic value of knowledge in sections 4 and 5. In this section, I shall consider the question of whether knowledge is good for its own sake.

I guess most scientists would answer this question in the affirmative. The claim that knowledge is valuable for its own sake is perhaps especially popular among people who work in areas where economically or socially useful applications are rare or nonexistent. But it also seems to be accepted within more "useful" disciplines.

So, the claim is widely accepted. But is it true? It is hard to say with certainty, but it seems to me that we have no reason to believe that it is true and some reason to believe that it is false.

In order to support the claim that knowledge is intrinsically good, we might refer to the fact that people do desire knowledge for its own sake.[33] This is the argument that is usually given, in so far as any argument is

[32]Thomas S. Kuhn, "Logic of Discovery or Psychology of Research?", in Lakatos and Musgrave (eds.) *Criticism and the Growth of Knowledge*, pp. 1-23; the quotation is from pp. 20-1.

[33]Such an argument reminds one of Mill's "proof" of the Principle of Utility in Chapter 4 of his *Utilitarianism*, see J.B. Schneewind (ed.), *Mill's Ethical Writings*, New York and London: Collier, 1965, pp. 308-15. An argument of this kind is also suggested by Richard Brandt. He writes: "There is *prima facie* support in our attitudes for the intrinsic worth

given. But it is not a good argument. In the first place, it is doubtful whether anyone really desires *knowledge* for its own sake. Rather, what people desire is the state of affairs that they *themselves* know the answer to some *particular* question or questions. And what a scientist typically desires for its own sake (if anything) is probably something even more specific, viz. that he himself be the *discoverer* of the answer to some question. In general, we do not desire the state of affairs that someone at some time knows the answer, or that mankind dicovers it.

Secondly, once we realize that people desire very different and very specific knowledge-states, we can also appreciate the fact that most people are completely indifferent to most knowledge-states. In most cases, people do not even desire that they themselves know the answers to scientific questions. Barry Barnes puts the point this way:

> Most people see science, quite rightly, as an activity beyond their understanding. And very many have in any case not the slightest interest in understanding it: many of the most popular newspapers and magazines devote more space to astrology and horoscopes than they do to natural science and its results.[34]

Similarly, most people cannot care less about the latest scientific news about the use of adverbs in Shakespeare's plays or the causes of inflation in Yugoslavia in 1970-75. In fact, I want to suggest that *everyone* is completely indifferent to most knowledge-states, and that most knowledge-states are *not* desired for their own sake.

However, it might be claimed that there are exceptions to this general rule. For example, it has been suggested by Richard Brandt that there are certain pieces of knowledge that we desire for everybody. Brandt writes:

> Yet it does seem that there are certain kinds of knowledge we do wish everyone to have—not isolated bits, as if there were value in memorizing paragraphs from Keynes on economic theory, without understanding what they mean, but systems of knowledge: the understanding of the physical and social world, of man's nature, of science and the evidence for scientific theory, and so on. These we wish all to have. That we do so is doubtless part of the basis for advocating a "liberal" education and requiring acquaintance with certain fields of knowledge. Nor is the reason

of knowledge; we do seem to want at least some knowledge on its own account", see R.B. Brandt, *Ethical Theory*, Englewood Cliffs, N.J.: Prentice-Hall, 1959, p. 335. For counter-arguments and further references, see Lars Bergström, "On the value of scientific knowledge", *Grazer Philosophische Studien*, vol. 30 (1987) pp. 53-63.

[34]Barnes, *About Science*, p. 20.

> for this simply that we wish everybody to have some common areas about which he can *converse* with other people.[35]

Well, maybe this is something we do wish. However, I very much doubt that it is something we want for its own sake. I think that those of us who want it would agree, on reflection, that we want it as a means to some more hedonistic value, such as well-being. We believe that people need some kind of world-view, and that they are bound to be frustrated if their beliefs about the observable features of their surroundings are radically mistaken. But it is not essential that their beliefs are true. For example, Newton's mechanics would do fine for physics, even if it is not strictly true. In fact, it might not matter much if our high-level theoretical beliefs were completely false, as long as our low-level empirical beliefs are approximately true. Moreover, different systems of belief may be quite acceptable in different cultures. Knowledge in a strict sense is not necessary.

Thirdly, even if some knowledge-states really are desired for their own sake, it does not follow that these states are intrinsically good. (Neither, of course, does it follow that *all* knowledge-states are intrinsically good.) It does not follow logically, for according to Hume's Law evaluative conclusions do not follow logically from factual premisses. But it does not follow inductively either, for the intrinsic value of knowledge does not constitute the best explanation of the intrinsic desire for knowledge. Rather, one can assume that the desire is best explained in either of the following two ways: (1) the intrinsic desire for knowledge may have survival value,[36] or (2) knowledge may often be desired as a means to something else, and there is a psychological mechanism to the effect that what is often desired as a means will easily come to be desired also for its own sake.[37]

Moreover, we cannot argue that knowledge is intrinsically good in virtue of some axiological principle to the effect that what is desired for its own sake is good in itself. For there seem to be many counterexamples to such a principle: some people may desire money, or fame, or power, and so on for its own sake, but we would not like to conclude from this that money, fame, and power are intrinsically good.

At this point, someone might say that even if knowledge as such is not intrinsically good, the *pursuit* of knowledge is. In other words, if we make

[35]Brandt, *Ethical Theory*, pp. 337-8.

[36]Notice, though, that desiring knowledge for its own sake may not have survival value under all circumstances. (For example, curiosity killed the cat.) More importantly, even if curiosity has had survival value in human history so far, it may not have survival value in the future, since technological conditions have changed quite a lot.

[37]See e.g. Charles L. Stevenson, *Ethics and Language*, New Haven: Yale University Press, 1944, pp. 193-8.

the usual distinction between science as a *process* and science as a *product*, we can see that it is the former, rather than the latter, that has intrinsic value. Notice, that this version of the doctrine avoids the objection that the product of science is often ignorance or error rather than knowledge. Perhaps it is also more in accordance with Aristotle's conception of the intellectual virtues as the most important constitutents of *eudaimonia* (happiness or human flourishing), which in turn is the supreme good.[38] As before, the claim that scientific activities are intrinsically valuable cannot be supported by reference to our desires, but it is perhaps more plausible in itself than the corresponding claim for knowledge.

However, it seems to me that there are at least three considerations which tend to make one sceptical of both versions of the doctrine. In the first place, it seems somewhat *ethnocentric* to believe that scientific knowledge and/or the pursuit of such knowledge is intrinsically good. There are many cultures in which science is not regarded as important. Indeed, the majority of mankind is probably not at all interested in science. It may be that every culture needs some form of " intellectual" enterprise. (This may even be true by definition.) But one may think of various alternatives to science here, such as religion, music, story-telling, magic, gardening, poetry, painting, chess, astrology, Hermann Hesse's *Glasperlenspiel*, and so on. It is hard to see why science should be *intrinsically* better than any of these. On the other hand, if *all* these activities are intrinsically good to the same degree, then there is nothing special about science: its value, as compared to that of its alternatives, has to be judged exclusively by its external results.

Secondly, those who claim that knowledge or the pursuit of knowledge is intrinsically good are usually themselves scientists or intellectuals. It is obvious that they have a vested interest in this doctrine. They have something to gain from propagating it. Scientists are privileged in our society. Therefore, they need to justify their life-style, both to themselves and to those who pay for it. Moreover, scientists are usually the sort of people who are culturally influential. This is quite sufficient as an explanation of why the doctrine is so widely accepted. This explanation also seems to undermine the plausibility of the doctrine.

Thirdly, it is possible to construct plausible counterexamples to the idea that knowledge or the pursuit of knowledge is good in itself. For example, suppose that Ivan has a fatal disease that will kill him within a few weeks. He is in bed, and the only thing he can do is to watch television. There are two alternatives: on one channel there is a series of rather good movies, on the other channel there are good educational programmes. The movies will give him a lot of pleasure, the educational programmes will give him somewhat

[38]See Aristotle, *Nichomachean Ethics*, VI.

less pleasure but much more knowledge. Ivan prefers to watch the movies. However, if knowledge or the pursuit of knowledge were intrinsically good, he ought to watch the educational programmes, since this would produce more intrinsic value. Moreover, and for the same reason, if he does not watch the educational programmes of his own free will, his wife ought to persuade him to do so (other things being equal). But this seems quite absurd. It is certainly all right for Ivan to watch the movies, and to enjoy his last weeks as much as possible. Therefore, neither knowledge as such, nor the pursuit of knowledge, is intrinsically good.

4. The effects of science.

If my argument so far is correct, we may now disregard the idea that science or knowledge is valuable for its own sake. If it has any value at all, this must be purely extrinsic. The value of science must depend exclusively upon the value of its effects or consequences. Moreover, I shall assume that the only effects that are relevant here are those which somehow affect the welfare or happiness of sentient beings. In the present context, this assumption seems quite reasonable.

Science has many different effects which are relevant to the welfare of sentient beings. Notice that some relevant effects have already been touched upon in sections 1 and 2, for states of knowledge and ignorance may in turn affect people's welfare. The intended effects of science are described by J.D. Bernal as follows:

> Science as an occupation may be considered to have three aims which are not mutually exclusive: the entertainment of the scientist and the satisfaction of his native curiosity, the discovery and integrated understanding of the external world, and the application of such understanding to the problems of human welfare.[39]

This sounds reassuring, but it must be remembered that science also has effects which are not intended. Some of these may not even be predictable. And the value of science depends upon all its effects, whether they are intended (or predictable) or not.

This might be disputed. It is often suggested by scientists that they themselves are only responsible for the scientific quality of the theories that they produce, and, in particular, that they are not responsible for the effects of each practical application of those theories. Scientists take pride in good effects of applications of their theories, but they are much less willing to accept

[39]Bernal, *The Social Function of Science*, p. 94.

responsibility for bad effects which are unintended or unpredictable. Similarly, it might be held that the value of science is independent of unintended or unpredictable effects.

I myself do not accept the view that scientists are responsible only for the intended effects of what they do. But even if we were to accept this view, it seems clear that we may still insist that *the value of science* depends upon unintended and unpredictable effects as well. To some extent, this is also agreed to and even stressed by people who argue that science is valuable. It is often pointed out that the future applications of basic scientific research is always to a large extent unpredictable, but that in many cases such research turns out to be extremely useful. This is regarded as an argument for the positive value of basic research—particularly in cases where practical applications cannot be imagined. Similarly, I would say, negative effects of science and its applications are relevant to the overall value of science, even if they are unpredictable.

In any case, it would be completely arbitrary to claim that the instrumental value of science depends upon the good consequences of scientific activity (such as improved health and more efficient communications, and so on) but not upon bad consequences (such as nuclear and chemical war, pollution, bad TV programmes, and so on).

The effects of science are of course very varied. In order to approach an answer to the question of whether they are, on balance, good or bad, I suggest that we consider them under the following five headings (where we start with the first of the aims mentioned by Bernal): entertainment, power, health, security, and education.

(1) Entertainment. Scientific research can be quite entertaining. Scientists typically enjoy solving problems, and they can also derive satisfaction from studying the work of other scientists. In fact, Bernal considers the idea that the ultimate justification of science is that it is "quite an amusing pastime", and he goes on to say that this attitude, "though rarely admitted, is actually extremely widespread among scientists, particularly those in the safer and more comfortable positions".[40] There is probably some truth in this. Moreover, some parts of science are also entertaining to non-scientists and non-specialist. (In many cases, this presupposes popularization.)

On the other hand, I think nearly everyone would agree that a lot of science is extremely boring, and that a lot of it is in fact completely unintelligible to the non-specialist. Scientific work can also be rather tiresome and unrewarding. I would suggest that, on the whole, modern science has comparatively little value as entertainment. As far as I can see, it is quite possible that alternative activities like alchemy, literature, religion, music, gardening, and

[40]Ibid., p. 97.

the game of trivial pursuit might be equally or more entertaining.

It must not be forgotten, of course, that science also has more indirect effects which have to do with entertainment. The application of science has given us technological inventions like radio, television, aeroplanes, personal computers, motor cars, tape recorders, gramophone records, and so on, which can be used for our entertainment. The impact of science in this respect is indeed overwhelming. And we certainly do enjoy all these technological gadgets. However, in the last analysis it may very well be doubted whether they have made us happier than we would have been without them. Maybe they have just changed our social habits, and provided us with alternative means of enjoyment. It is not at all clear that the institution of science can be justified on the ground that it provides entertainment.

(2) Power. By increasing our knowledge of the world, we automatically increase our power over it. This is a classical idea, which goes back primarily to Francis Bacon. But there are different kinds of power. In fact, what Bacon seems to have had in mind here is just what Bernal refers to as the application of scientific understanding to the problems of human welfare. Bacon writes as follows:

> It will not be amiss to distinguish the three kinds and as it were grades of ambition in mankind. The first is of those who desire to extend their own power in their native country; which kind is vulgar and degenerate. The second is of those who labour to extend the power of their country and its dominion among men. This certainly has more dignity, though not less covetousness. But if a man endeavours to establish and extend the power and dominion of the human race itself over the universe, his ambition (if ambition it can be called) is without doubt both a more wholesome thing and a more noble than the other two. Now the empire of men over things depends wholly on the arts and sciences. For we cannot command nature except by obeying her.[41]

Unfortunately, the first two kinds of ambition are all too common, and I think Bacon would have had to agree that science can be used to satisfy those as well. But this is not what science is for, according to him. In another place he writes:

> For what is at stake is not merely a mental satisfaction but the very reality of man's wellbeing, and all his power of action. Man is the helper and interpreter of Nature. He can only act and

[41] Francis Bacon, *Novum Organon*, 129. Quoted from Benjamin Farrington, *Francis Bacon. Philosopher of Industrial Science*, London: Macmillan, 1973, p. 7.

> understand in so far as by working upon her he has come to perceive her order. Beyond this he has neither knowledge nor power. For there is no strength that can break the causal chain: Nature cannot be conquered but by obeying her. Accordingly these twin goals, human science and human power, come in the end to one. To be ignorant of causes is to be frustrated in action.[42]

This view of science more or less originates with Bacon. Before him, the pursuit of truth was not in general regarded as a means to the improvement of the conditions of life for mankind.[43] After him, of course, similar ideas were central to the Enlightenment.

No one would deny that science has in many ways increased our power in Bacon's sense. However, three further points should be noticed here. In the first place, there seem to be many scientific disciplines which have *not* been of much use to mankind in the way Bacon aimed at. Examples of such disciplines might be theology, astronomy, philology, political science, fundamental particle physics, archaeology, musicology, futurology, topology, philosophy, zoology, and the history of art and literature.[44]

Secondly, the power over nature that science has given us is used very selectively. It has often been pointed out that knowledge is more commonly used for the benefit of the few than for the benefit of all,[45] and Bernal goes even further when he says that "science is being used mainly for the enrichment of the few and the destruction of the many".[46] I will say more about this below.

Thirdly, it seems that there is something about scientific progress itself which may, at least in some important cases, *reduce* our power over nature. For scientific progress seems to lead to larger and more complex systems

---

[42]Francis Bacon, *The Great Instauration*, Part 6. Quoted from Farrington, *Francis Bacon*, p. 91.

[43]See e.g. Farrington, *Francis Bacon*, p. 5.

[44]This list of examples might be disputed. I shall not try to argue for it here. Let me just give the following quotation, which concerns one of the least obvious and most expensive items on the list: "The *cause célèbre* at present is the study of the fundamental particles of matter in high energy nuclear physics. This is of great interest academically— physicists are agreed on that. On the other hand, it is also a very expensive field of research, because enormous accelerators are required to bring particles to high enough energies. There are no signs of any useful applications emerging from knowledge of these fundamental particles. It is important to be quite clear that this really means exactly what it says: no use can even be envisaged", F. R. Jevons, *The Teaching of Science. Education, Science, and Society*, London, 1969, p. 75. For a similar, but more recent judgement, see Barnes, *About Science*, p. 27.

[45]Compare e.g. Harold D. Lasswell, "Must science serve political power?", *The American Psychologist*, vol. 25 (1970), pp. 117-123, esp. p. 117.

[46]Bernal, *The Social Function of Science*, p. 97.

of economy and technology, and it is far from clear that science can tell us how best to achieve our ends within such systems. The experts very often disagree when it comes to matters which are relevant to important decisions concerning economic policy and large-scale technology.[47] It seems reasonable to assume that the greater the socio-political impact of a given decision, the higher is the probability that the experts will disagree and that their views will be influenced by political considerations and by their personal and economic relations to various organizations in society. The debates concerning nuclear power, the greenhouse effect, and the transition from socialism to market economy in former socialist countries illustrate this.

So, the situation is not as simple as Bacon might have thought. As regards our power to improve our conditions of life it seems that science has been only partly beneficial. In particular, let us briefly consider two main dimensions or indicators of human well-being, viz. health and security.

(3) Health. For some people today, the health situation is of course very much better than it was for most people before the age of science. The progress of medical science has had the effect that many diseases have completely disappeared in certain areas, and that many of the remaining ones can be treated with excellent results. The infant mortality rate has decreased, and the average length of life has increased. In particular, this is true in the rich countries.

On the other hand, the situation is obviously much worse in the poor countries, where there is also a severe lack of medicine and effective health organizations. Thus, for example, around 40.000 children die every day in the world, and at least half of them could have been saved by quite simple means (polio vaccine, etc). Moreover, if some diseases have disappeared, others have replaced them, and some are even caused by the very technological progress which is in turn based on science. There is a shortage of food and clean water in many areas, and the environment is polluted almost everywhere. This affects people's health in a negative way. And while it is true that the average length of life has been increased in many countries, it is also true, even in the rich countries, that the quality of life is often rather bad for old people and for sick people who are kept alive by artificial means.

If we think about the state of health of sentient beings, we should also note the fact that a great many animals are made to suffer as a result of our technological progress. Whole species are extinguished or at least severely threatened and reduced by changes in the environment brought about by us, and every year hundreds of millions of animals are killed, often in very

[47] See e.g. Dorothy Nelkin (ed.), *Controversy. Politics of Technical Decisions*, Beverly Hills and London: Sage Publications, 1979.

painful ways, in scientific research.[48]

In short, it is not at all obvious that the average level of health of sentient beings has been improved as a result of the scientific development. It seems quite possible to me that it is rather the other way round. Moreover, there is really no indication that the situation can be improved by further scientific and technological progress. We already have the knowledge and the technological means to help the sick and the starving, but we do not use them. Perhaps it is, and will remain, politically impossible to do so.

(4) Security. It is possible that we feel more secure when we know more about the causes of events and about human nature, and when we do not believe that we are at the mercy of gods and evil spirits. Again, we are more secure when we can protect ourselves against wild animals, illness, and natural disasters like floods and thunderstorms. Science can be useful here. Thereby it contributes to a higher level of security, which in turn increases our welfare.

Science can also help us to defend ourselves against other people. As science has developed, the police have been provided with more efficient techniques, involving e.g. weapons, information storage systems, and a developed technology of surveillance. Similarly with the armed forces used for national defense.

On the other hand, obviously, the production and distribution of arms and other military technology also reduces security in many cases, and it has led to a lot of suffering and death. It seems that the number of wars in the world per year has been more or less constant during the rise of science, but that the average number of people killed in wars increases drastically with time. For example, 0.8 million people were killed in 92 wars in the years 1820-1859, 4.6 million were killed in 106 wars in the period 1860-1899, and 42.5 million were killed in 117 wars in the period 1900-1949. If the trends are extrapolated, it turns out that virtually 100 per cent of the world population will be killed in wars before the year 3000.[49] Of course, science has played an important role here. Without science, it would simply not be possible to kill

[48]See e.g. Richard D. Ryder, "Speciesism in the laboratory", in Peter Singer (ed.), *In Defence of Animals*, Oxford: Blackwell, 1985. Ryder says: "It has been estimated that between 100 million and 200 million animals die in laboratories around the world each year" (ibid., p. 79). Another commentator says that "the total number of laboratory animals now used throughout the world anually is 200 to 250 million. The United States accounts for about 100 million of these animals as follows: 50 million mice, 20 million rats, and about 30 million other animals, including 200,000 cats and 450,000 dogs"; see Bernard E. Rollin, *Animal Rights and Human Morality*, Buffalo, N.Y.: Prometheus Books, 1981, p. 91.

[49]See Robin Clarke, *The Science of War and Peace*, London: Jonathan Cape, 1971, pp. 10-12.

so many people. And a very large proportion of scientific research has indeed been directed towards the development of weapons systems. For example, in the United States most of the economic resources used for research and development in recent decades has been used for military purposes.[50]

Besides, the application of scientific theories has created new environmental problems. For example, there are dangerous emissions from the chemical industry, there is radioactive waste from nuclear power plants, and so on. This tends to make our existence less rather than more secure.

(5) Education. Scientific research and higher education go together. Each presupposes the other. The scientific community will die out if it does not reproduce itself and make itself respected in the rest of society, and higher education will become scholastic and boring if it is not intimately related to research. Moreover, it might be held that a high level of education is essential to the welfare of a population. Education will make people better equipped to solve problems, to communicate with others, and to learn from the experiences of earlier generations. It may be suggested that science has increased our freedom. It has provided new opportunities for action, and it has made it easier for people in general to choose the alternatives that they really want.

In this way, science may indeed have had beneficial effects. Moreover, this particular function of science is not restricted to natural science. Many of the human and social sciences may be even more useful in this particular respect. For example, to people in general, disciplines like economics and philosophy are probably more useful than physics and geology.

On the other hand, higher education may also generate new inequalities and preserve old ones. Scientists and educated people are privileged in our society. (Indeed, it seems quite likely that science as we know it would cease to exist if scientists were not privileged.) Besides, even if people's freedom has been increased in some respects, because of the development of science, it also seems to have been reduced in certain other ways. Harold Lasswell puts the point as follows:

> If the earlier promise was that knowledge would make men free, the contemporary reality seems to be that more men are manipu-

[50] "Between 1950 and 1985, 65-70% of federal research and development funds were channeled through the Department of Defense, only 1-3% through the NSF [National Science Foundation]. (If one includes the Department of Energy, whose major focus is nuclear weapons, and the National Aeronautics and Space Administration, which is under heavy contract to the military, the military-related totals go even higher.)" Carl Mitchum, "The Spectrum of Ethical Issues Associated with the Military Support of Science and Technology", in *Ethical Issues Associated with Scientific and Technological Research for the Military*, edited by Carl Mitcham and Philip Siekevitz, The New York Academy of Science, New York 1989, pp. 1-9, p. 4.

> lated without their consent for more purposes by more techniques by fewer men than at any time in history.[51]

In other words, even if a high level of education is desirable in many ways, there are certain aspects of it which are not desirable.

5. The overall extrinsic value of science.

What I have discussed above are the effects of *past* science. Past science may be defined as the totality of all scientific activities which have taken place in the world so far. (I have not yet been concerned with *future* science and its effects; this will be the subject of section 6 below.) Let us now consider the question of whether the overall *extrinsic value* of past science is positive or negative.

This question may be interpreted in different ways. For example, it may be taken as (1) the question of whether the total consequences of past science are on the whole good or bad or indifferent. Or it may be (2) the question of whether these total consequences are better or worse than the consequences of some alternative to past science. Finally, it may be (3) the question of whether the total consequences of past science are better or worse than the consequences of that particular alternative to past science that would in fact have taken place if none, or very few, of the activities within past science had occurred. Of these three interpretations, (3) is the most interesting one. Our attitude towards science should depend upon our answer to this question. Moreover, I shall stick to the idea above that value is determined by the relative welfare of sentient beings.

However, if we reflect upon this formulation of the problem, we can see that it is far beyond our power to answer it in an intersubjectively reliable way. Even if we had access to a normatively acceptable, quantitative, and operational definition of "welfare", the combined resources of all scientific disciplines would not be sufficient to provide a reliable answer. The question of whether the psychological well-being of human beings, or of sentient beings in general, is favourably affected by past science seems to be a factual question. But science cannot solve it. It involves interpersonal (and "inter-organism") utility comparisons, large-scale conterfactual conditionals, and completely unsurveyable initial conditions which science cannot handle.

Sixty years ago, John Dewey expressed a similar scepticism concerning the future impact of science on society as follows:

> Externally, science through its applications is manufacturing the conditions of our institutions at such a speed that we are too be-

[51] Lasswell, "Must science serve political power?", p. 119.

> wildered to know what sort of civilization is in process of making. Because of this confusion, we cannot even draw up a ledger account of social gains and losses due to the operation of science. But at least we know that the earlier optimism which thought that the advance of natural science was to dispel superstition, ignorance, and oppression, by placing reason on the throne, was unjustified. Some superstitions have given way, but the mechanical devices due to science have made it possible to spread new kinds of error and delusion among a larger multitude. The fact is that it is foolish to try to draw up a debit and credit account for science.[52]

What Dewey says here is, I think, correct in many ways. However, his conclusion should be resisted. We should at least try to "draw up a debit and credit account" for science. We must realize that this cannot be done in a scientifically reliable way, but this is no excuse for ignoring the question. Science is too important an element of our culture to be taken for granted without criticism. Our attitude towards science must be based upon personal judgement, and this judgement can at best be made in awareness of arguments and considerations of the kind exemplified in section 4 above, concerning the effects of science.

My own judgement, for what it is worth, is that past science is probably *not* extrinsically good. We would have been better off without it.[53] In general, when its effects are beneficial, they are beneficial only to small minorities which are already quite well off. For example, the products of military science are useful mainly to arms dealers and superpowers. Other technology based on natural science is useful to industrialists and shareholders, and it also yields economic profit to other citizens in the highly developed countries. Social science may be useful to political elites by helping them to control the masses and to legitimize the policies preferred by the elites. (Of course, much social science is critical of political elites, but this makes it even more useful as a harmless token of tolerance and freedom in the society.) The humanities, finally, are usually regarded as fairly useless, at least if we disregard the personal satisfaction which they may give to some of the very few people who are actually working within these disciplines.

The thesis that past science is not extrinsically good is also reinforced by the following considerations. Science has produced technology, which has to

[52] John Dewey, *Philosophy and Civilization*, Gloucester, Mass.: Peter Smith 1968, p. 319.

[53] Quite possibly, many people in the rich countries today would find life intolerable if they were moved, miraculously, back to the 17th or 18th century. But this does not show that humanity is happier now than it used to be. Nor does it show that humanity is happier now than it would have been without science.

some extent improved our material conditions of life (such as health, security, and so on). However, psychological well-being is not a simple function of such material conditions. It is more dependent upon the extent to which expectations are satisfied. And technological progress does not guarantee the satisfaction of expectations. Indeed, it may have the opposite effect. Nicholas Rescher puts the point as follows:

> There is what might be called the Fundamental Paradox of Progress: progress produces dissatisfaction because it inflates expectations faster than it can actually meet them. And this is virtually inevitable because the faster the expectations *are* met, the faster they escalate.[54]

Moreover, it seems clear that the extrinsic value of science should be taken to depend upon psychological well-being rather than upon material conditions of life. From a normative point of view, the latter are relevant only if they affect the former. Health, wealth, security, and power are not intrinsically valuable. They are only valuable as means to pleasure and happiness.

It is often pointed out that scientific and technological research is needed in order to neutralize or remove undesirable effects of scientific and technological research. More sophisticated weapons are needed to counteract the sophisticated weapons already in existence. New energy systems must be deviced in order to prevent or reduce pollution of the environment. And so on. Similarly, it might be held, higher education is required in order to avoid alienation and apathy among ordinary people in scientifically and technologically advanced societies. This means that a high level of education in a population is perhaps best thought of as an antidote against the bad effects of science. In short, it might turn out that the main use of science nowadays is to protect us from the bad effects of science.

This point has some relation to the much discussed question of the rationality of science. Some philosophers think that the development of science is governed by a series of rational choices on the part of the scientific community.[55] But even if each individual choice is rational, the enterprise as a whole may be irrational or non-optimal. Individual choices may be

[54]Nicholas Rescher, "Technological Progress and Human Happiness", p. 19, in *Unpopular Essays on Technological Progreess*, Pittsburgh: University of Pittsburgh Press, 1980, pp. 3-22.

[55]See e.g. Newton-Smith, *The Rationality of Science*. For the view that rationality is not very common and not very desirable in science, see e.g. Lars Bergström, "Some remarks concerning rationality in science", in Risto Hilpinen (ed.), *Rationality in Science*, Dordrecht, Boston, and London: Reidel, 1980, pp. 1-11.

rational relative to the internal aims of science, but if past science has left us worse off it seems irrational from the point of view of humanity.

6. Prospects for the future.

If I am right, science has not been a good thing so far. What about the future? The first point to be noted here is that future science is to a large extent unpredictable.[56] And if the content of future science is unpredictable, so are its effects. It is easy to suppose that the future will be like the past. But such an induction is extremely risky. Most of us may take it for granted that science will continue to grow as before, but there is really no justification for such a belief. There may be a saturation limit to the growth of science.[57] In fact, we may be close to such a limit right now. And if there is stagnation, there may also be decline. This possibility is recognized by Bernal:

> The continued existence of this institution of science is in general far too easily taken for granted; because science in its association with industry has in the past made such enormous progress, it is assumed that this progress will automatically continue. Intrinsically, however, there is no more justification for continued progress in science than for continued progress in industry. ... We have seen, in the course of history, institutions grow up, stagnate, and die away. How do we know that the same will not happen to science?[58]

In any case, even if past science has been extrinsically bad, this may not be true of future science. However, it is hard not to be pessimistic. Michael Dummett makes the following prediction:

> For it seems to me evident that, were the option a live one, there exist overwhelming grounds for bringing *all* scientific research to a halt. Of no reseach is it possible to foresee what applications will be made. Even so intelligent a man as Rutherford is reported to have thanked God that his research was practically useless; but we have no excuse for making a similar mistake.

[56] This point is stressed e.g. by Karl Popper in the "Preface" to *The Poverty of Historicism*, London: Routledge, 1961.

[57] "In its typical pattern, growth starts exponentially and maintains this pace to a point almost halfway between floor and ceiling, where it has an inflection. After this, the pace of growth declines so that the curve continues toward the ceiling in a manner symmetrical with the way in which it climbed from the floor to the midpoint", de Solla Price, *Little Science, Big Science*, p. 18.

[58] Bernal, *The Social Function of Science*, p. 11.

> All that we can say with confidence is that, of the scientific research carried out within any given future period, much of it will have applications, some of them quite unexpected, and that, of these applications, most of those that yield unqualified benefits for mankind will either be unexploited or, at best, used to enhance the lives only of people in the wealthy nations, while some will, for certain, be used to create as yet unimagined dangers and horrors.[59]

I think Dummett is wrong when he says that this prediction can be made "with confidence" and that we know this "for certain", but the content of his prediction may very well be right. It seems to me that if well-informed people disagree on the validity of Dummett's prediction—as they can be expected to do—their different views are caused mainly by different personality traits. Some people are optimists and some are pessimists, and this is all there is to the disagreement.

Dummett is probably right that we *cannot* bring all scientific research to a halt—except, of course, by starting the last world war. However, we might be able to discourage and reduce certain kinds of research by re-allocating the available economic resources to other disciplines or to non-scientific projects. As a rule of thumb, we might even assume that the more "useful" (in a conventional sense) a given field of research is considered to be, the more dangerous it is, and the less money should be invested in it. In particular, it may be a good thing to invest less money in disciplines which tend to generate technological applications. Some of the money spent in this way could instead be absorbed by the humanities, which are fairly harmless.

It is sometimes said that curiosity is part of being human, and hence that if we stop doing science, we stop being human. But this is mainly rethoric. In conclusion, three points may be noted. First, we can be curious even if we are not scientists, and even if there is nothing like modern science around. Second, even if we continue to do science, we can concentrate upon the more harmless disciplines. Third, many people are *not* curious in the sense of being interested in science, but we should not conclude from this that they are not human.

[59]Dummett, "Ought research to be unrestricted?", p. 291.

Logic, Methodology and Philosophy of Science IX
D. Prawitz, B. Skyrms and D. Westerståhl (Editors)

MORALITY AND HUMAN EVOLUTION[1]

ALLAN GIBBARD

Department of Philosophy, University of Michigan, Ann Arbor, MI 48109, U.S.A.

How might we find a good framework for thinking about the psychology of morals? Recent advances in evolutionary theory may help, I want to suggest. Over the last few decades, biologists have developed ways of thinking that were suggested by Darwin, but that needed much theoretical clarification. Darwin's own most pertinent writing, in his book *The Expression of Emotion in Man and Animals* (1872), by now reads strangely, as a mixture of brilliant observations, hypotheses, and speculations, on the one hand, and theoretical blunders on the other. George Williams (1966) helped get the current wave going by warning against facile "good of the species" arguments in explanations of animal behavior. William Hamilton (1964) pioneered in developing rigorous mathematical models of the evolution of behavior. John Maynard Smith (1974, 1983) showed how game-theory could be applied to the genetic evolution of behavioral propensities. As a result of this and much other work, the evolution of behavior has now become a medium-scale interdisciplinary field of study. And its successes, I think, suggest new ways of thinking about moral thought and motivation in human beings.

Evolutionary thinking about human behavior is immensely controversial, of course, and treating morality evolutionarily has been especially controversial. There is a lot of chaff to sort from the wheat on both sides of these debates. This threshing, though, will not be a business of mine here. Instead of jumping into these controversies, I shall be laying my own speculations on the table, and making a few comments as I go.

[1]This work was supported by a John Simon Guggenheim Memorial Foundation Fellowship.

1. Accepting norms

In 1990 I published a book called *Wise Choices, Apt Feelings: A Theory of Normative Judgment*. It is mostly a book of moral philosophy, not of moral psychology. It contained, though, speculations on moral psychology and the genetic evolution of human moral propensities. Let me review these speculations and add some thoughts.

My speculation centers on a special psychological state which I call "accepting a norm". This is a state I am in when I think, for instance, that it makes most sense to get lots of sleep before delivering a lecture, or when I think that it doesn't make sense to be angry at the critic who finds a major flaw in my argument. I accept a norm that says to get lots of sleep, or I accept a norm that says not to be angry at critics who make good points against my favorite theses. My speculation is, first, that there is such a state as accepting a norm, and that this state is an important one in the human psychic makeup.

When I speak of such a state, I suppose that genetically programmed psychic mechanisms underlie it. These mechanisms were shaped by natural selection in the course of human evolution. They were shaped by selection pressures to do certain things that promoted reproduction; these jobs are the *biological functions* of the mechanisms.[2]

The mechanisms at work in human norm acceptance, then, have biological functions. I speculate that their chief biological function is one of *coordination*, in a special, game-theoretic sense of the term. These mechanisms coordinate the actions of different people through two chief tendencies; I call them "normative discussion" and "normative governance". In normative discussion, people tend to avow the norms they accept, and tend to be influenced toward accepting the norms that others avow. The upshot is a tendency toward consensus in the norms discussants accept. By "normative governance", I mean a motivation to act in accordance with the norms a person accepts. Combine normative discussion with normative governance, and everyone will tend to act in accordance with the same norms. Normative discussion tends toward consensus on norms, and normative governance then tends toward everyone's acting on those norms. Thus actions are coordinated. Normative discussion and normative governance combined tend to coordinate actions.

Some commentators have mistaken this hypothesis for a group selection account: that we are evolved to coordinate our actions because doing so is good for the group. I should stress that I mean no such thing. Whether anything that should be called group selection is important in

[2]See Wright (1973) and Symons (1979), 10–14.

the evolution of behavior is a dicey question. The experts agree, though, that no facile move from the good of the group to the selection of a characteristic is going to work. My own explanation attempts no such move; mine is an individual selection account. It is meant to work within Dawkins' framework of thinking about advantages to a selfish gene (1976; 1982, Ch. 2).

It was Thomas Schelling (1960, Ch. 2) who developed the broad notion of coordination I am appealing to. He showed how coordinating one's actions with the actions of those around one is often greatly to an individual's advantage. Many examples are familiar enough; think of driving on the left or driving on the right. Coordinated people don't bump into each other—whether they coordinate to the left or to the right. Each, given the actions of others, does best for himself if he conforms to the system. This parable, Schelling showed, expands to encompass a great deal in human affairs.

Biologists have applied this pattern to the affairs of creatures on down, in order of simplicity, at least as far as spiders. Pecking orders are one example. Various arrangements are possible: I might defer to you over food, or you defer to me, or we might share the food peacefully. Suppose neither of us can overpower the other at low risk. Then if I am disposed to defer to you, and you are disposed to peck at me if I don't, then I am doing as well as I can, given your dispositions. The same goes if I am disposed to peck and you are disposed to defer. Either way, we avoid a ruinous fight over who gets the food. Game theorists call these two outcomes alternative *Nash equilibria:* A Nash equilibrium is a combination of propensities to action such that each, given the propensities of the others with whom he is interacting, does at least as well with that propensity as he could with any other.

John Maynard Smith (1974, 1983) called a variant of this—or its evolutionary analogue—an *evolutionarily stable strategy.* When strategies are evolutionarily stable, then each individual has genetic propensities that constitute the best available response to the propensities of the others. Each individual's propensities are individually advantageous, given the propensities of the others. And so explanations in terms of evolutionarily stable strategies invoke individual selection, not group selection. Coordination does benefit the group, it may be, but this is not what explains the propensities' being naturally selected. They are selected because of the benefit each individual derives, given the propensities of the others.

This talk of "advantage", "good", and "benefit" has a fairly precise, technical meaning. I am talking of increases in prospects for reproduction —we might say, one's expected number of great-grandchildren. This,

of course, is not what advantages and benefits are in the usual senses of these terms, and so we have to keep the figurative, technical senses of these terms straight from their ordinary senses. For increases in prospects for great-grandchildren, let me use the term "selective advantages". This term, then, I appropriate as a technical one. Selective advantages may not be advantages in any ordinary sense of the term. It might be to the selective advantage of the male praying mantis to be eaten by his mate after he impregnates her, since he thereby nourishes his young. That doesn't mean that he likes it, or that being eaten is good for him. The relation between selective advantages and the sorts of things we regard as advantages will be complex. The two will tend to be the same, but they will not be entirely coincident.

I have been running through evolutionary commonplaces; now let me return to my own evolutionary speculation. Human affairs are far more complex than pecking orders. They were far more complex, all indications are, even among hunting-gathering proto-human populations. Genetic selection in these populations left us with the human genetic propensities we have today, and our large brains may result from the complex social demands of proto-human life—demands that can make differences of life and death, and big diferences in reproductive opportunities. In all human groups we know today, people spend enormous quantities of thought and energy on courtship, sex, and marriage. They spend enormous quantities of thought and energy on various matters of property, such as division of the hunt, gifts, and establishing entitlements. Everywhere people tend their social relationships, and often they engage in rivalries that extend to feud and war. The complexities of human life and their connection to matters of life and death and reproductive opportunities are especially vivid in groups that live in anarchy, without strong, effective governmental control of killing, property, and the like. Then, conflicts and quarrels always threaten death. Even, though, when force and violence are rare, they stand in the background. Whom a man could mate with without risking severe consequences is almost always a highly social matter.

My speculation is that as proto-human social life became more complex, simple systems like pecking orders and territoriality became inadequate to the job of coordination. More powerful mechanisms of coordination were needed. Now human evolution included the evolution of a capacity for language, and language allowed for more complex and powerful coordinating mechanisms. Capacities for language no doubt conferred many different kinds of selective advantages on our ancestors, but some of the important ones, I suggest, must have been matters of the kinds of coordination language made possible.

Language allows a group to share representations of absent situations. Discussants can share responses to a situation they are not in right here and now—past situations, expected situations, hypothetical cases, and even fictions. Discussants can rehearse coordinated responses together, and so coordinate their actions when similar situations arise in the future.

One way to do this is by what I am calling normative avowal. People might discuss, say, the conflict between a younger and an older man, and say "Let the younger man defer." If they all accept this imperative, then in similar situations in the future, older men will tend to insist on deference, and younger men will tend to defer. Their actions will thus tend to be coordinated, and mutually threatening quarrels will be avoided.

Another way normative discussion can coordinate actions is by coordinating peoples' feelings about actions. Conversants may come to agree, say, that theft by stealth is shameful. They avow norms that say to feel ashamed of secret stealing, and to disdain anyone they learn has tried secretly to steal something. Then any of them, faced with an opportunity for secret theft, will be deterred by the feelings the norms he accepts tell him to have. And that may well be a good thing for him, because others accept norms that say to have disdainful feelings toward him if he practices secret theft, and so they will act badly toward him if they learn that he has done so.

I place morality in this second pattern: Actions being coordinated by norms for feelings. Moral norms coordinate feelings of indignation. We can think of indignation as a form of anger—anger governed by norms one accepts as impartial. Moral norms also coordinate feelings of anger with feelings of guilt. A person can feel guilty about his own actions, and others can feel angry at him over them; norms can coordinate these first-person feelings of guilt with others' feelings of anger. The prospect of guilt deters, and guilt can also motivate a person to placate anger and make amends. Coordinated guilt and anger can thus coordinate actions. To think an act morally reprehensible is to accept norms that tell the person who did it to feel guilty for having done it, and tell others to feel impartially angry at him for having done it.

2. Realism, morality, motivations

My sketch of the speculations in my book has been quick. I now want to ask how well grounded they are likely to be. I realize the drawbacks of armchair empiricism, but I think that at this point we can still get somewhere by asking ourselves questions. Given what we know, I want to ask, what is most likely to be right in this speculation, and what is most

likely to be wrong?

Some features of my story strike me as hard to doubt—at least once we start looking at human affairs with a story like this in mind. Human life is intricate, and at all stages of human evolution, life has been full of coordination problems.

Feelings are the work of genetically evolved emotional mechanisms, and these mechanisms have evolved to respond to the social environment with refined heuristics. By heuristics, I mean schemes of response that were selectively advantageous for our ancestors—not always, but often enough to promote their reproductive prospects on average. One crucial kind of selective advantage lay in various ways that feelings made for social coordination. Talk interacted with feelings, and important psychic mechanisms made for this interaction. They were shaped accordingly by natural selection. Scolding and criticism, complaint, praise, gossip, stories, ceremony, oratory—all of these are widespread in human life, perhaps universal. All of them combine language with feelings. It would be amazing if all this talk and feeling were idle in human affairs, or if it were not the work of psychic mechanisms intricately adapted to the range of social circumstances our ancestors faced.

These broad generalities I find hard to doubt. I find it hard to doubt that they are crucial to explaining human capacities for moral reasoning and moral motivation. On the other hand, my specific story of human morality is no doubt far too simple and schematic. The human brain is immensely complex, and no short story of an aspect of its workings is likely to be right. Natural selection is amazing in the devices it can design. It can optimize responses to complex problems without having to understand them. This optimization has its limits, to be sure. Think, though, of a bird's wing, and contrast it with what most of us could design as a flying device.

My own schema was very simple: People avow norms for action and for feelings. They are influenced by the avowals of others. They tend to act and feel according to the norms they accept. Feelings tend toward actions. Contrast this with real human life, or with the intricacies of dialog, feeling, and action in any piece of fiction that we would find worth hearing. We don't know how life was in hunting-gathering groups when they had the run of the rich portions of the earth, but we do have available to us elaborate accounts of somewhat primitive conditions. Icelandic sagas, for instance, tell of an anarchic agricultural society in harsh country (Miller, 1990). At least one ethnographer has studied oratory among horticultural headhunters (Rosaldo, 1973). These accounts amply confirm the intricacy of language, feeling, and action in human life, and the high stakes involved.

In short, then, on the one hand I am confident that in human beings, there are refined, genetically coded psychic mechanisms that make for significant interactions among language, feelings, and action. On the other hand, I can't claim to have good hypotheses about what these mechanisms are like. My talk of avowal and mutual influence may be a helpful first approximation, but clearly it will need refinement. My account of what gets avowed should be taken with even more caution—as I'll try to explain.

What gets avowed, I suggested, is norms—often norms for feelings. Norms are imperatives of a kind, so my schema would have us saying to each other, "Feel such-and-such a way about so-and-so." Obviously, very little of human talk sounds much like that. What I can seriously hypothesize, then, is not that we are adapted to use language that fits my schema directly. We can use such language, but mostly we don't, and our ancestors may not have used it at all.

The power I can claim for my schema is one of translation. We can say confidently, I claim, that one central aspect of human and proto-human life is language directing feelings and actions. I hypothesize further—and this the more risky speculation—that these connections of language to feeling and action can be represented by an austere language of norms, feelings, and actions. Much of the talk that coordinates human actions, I hypothesize, can be helpfully translated into talk of advisable actions and warranted feelings. The translation displays the ways these kinds of language are tied to motivation.[3]

Sometimes the translations will be quite direct. When we call something shameful, we mean pretty directly that it warrants shame on the part of the person who did it. When we call a state of affairs sad, we mean pretty directly that sad feelings toward it would be apt. Indeed Latin has a whole verb form that fits this pattern.

Sometimes the translation will be more distant. Moral philosophers have written much, in recent years, about "thick concepts", concepts that both respond to fairly specific features in the world, and direct feeling and action.[4] One question we should be asking about thick concepts is this: What kinds of mental mechanisms stand behind our capacities to use them? These mechanisms somehow connect features of a situation to feelings and actions in response. My serious hypothesis is that this interplay can be perspicuously represented by my austere language of norm, feeling, and action.

Crucial aspects of human language are not best interpreted simply as representing straightforward states of affairs, but as regulating and coordi-

[3]See Chapter 5 of my book (1990).

[4]See Williams, 1985, 140–52; Wiggins, 1987, Chs. 1, 3; McDowell, 1988.

nating feelings and actions. We need some way of displaying the patterns of regulation. A language of norms, feelings, and actions will do this job, I am hypothesizing. I doubt that it will do more ambitious psychological jobs, such as giving the true structure of a language of thought that is always involved in such regulation.

Let me now scrutinize my talk of mutual influence and consensus. It too is no doubt far too simplistic. Sometimes we are influenced, and sometimes we resist influence stubbornly. We can be influenced in some directions with ease, and in other directions only with great difficulty. I did say a little in my book about these matters, but we need to know far more. Psychic mechanisms assess whether to be influenced by a given group or person in a given direction. What heuristics do these mechanisms employ?

From the game-theoretic structure of bargaining situations, we might speculate about two kinds of mechanisms. One kind assesses what one can get away with. Another kind assesses whom to accept as models.

First, then, assessing what one can get away with. Much of human and animal life involves what game theorists call bargaining situations. Dominance hierarchies illustrate some of their simplest aspects, and so perhaps they can help us understand some of the design features of the psyche of a highly social animal. In a pecking order, how best to act depends partly on one's prowess, and partly on others' expectations. Dangers being equal, it is better to be higher in a pecking order than lower. On the other hand, in the short term at least, it is dangerous to try to advance in a pecking order. And if birds below one do seriously try to advance, it is even dangerous to defend one's place. This kind of situation lends itself well to mathematical modelling, using either analytic techniques or computer simulation. It illustrates some of the factors that must have been present in the more complex situations proto-humans must have faced. In short, we can say this: Coordination brings great advantages, and if those were the only advantages involved, no bird would ever challenge another for its place in a pecking order. These advantages, though, are not the only ones involved, and so at times it is advantageous to upset a scheme of peace and coordination, in hopes of a more advantageous position in a revised scheme of coordination.

A few comments about what I have been saying. "Advantage" as I am using the term should be understood as selective advantage, as genetic fitness, as expected reproduction. And the question to ask is not, directly, what course of action holds out most advantageous prospects. The question is what heuristics will tend to work advantageously, on average, in the run of situations the animals are in. Questions like these should sug-

gest hypotheses about human beings. What psychic mechanisms stand behind such things as noble bearing, dignified bearing, and menial bearing? What mechanisms stand behind gestures of dominance, equality, and submission, and the language of challenge and obedience? Understanding pecking orders helps us understand some of the demands these mechanisms are shaped to meet.[5] With human beings, though, things will be more complex than in pecking orders. Human bargaining is not only a matter of what one does, but also of what one says. One can claim a more advantageous position with words, one can seize it openly, or one can take advantages in secret. We need to think about what mechanisms could advantageously govern one's words—the norms one accepts, in my schema—and what mechanisms could advantageously govern the match or mismatch between words and actions.

Also with human beings, alliances are crucial. Two-person game theory can tell us something about human life, but most study of humanity needs many-person game theory—and that, we know, is a far messier subject. A person's individual powers have some bearing on the alliances he can make, but often what matters are such things as family ties and badges of affiliation. There is much to be explored, here, about why even distant kinship often matters in human life, and why human beings devote so much heed and effort to such things as clothing, styles of language, and rituals.

That brings me to the question of modelling one's style and behavior on others. Lore and experience suggest that people are very selective in whom they will imitate, whose actions, style, and words they will be influenced by, and whose they will dismiss. Boyd and Richerson (1985) cite a report of an agricultural assistance program in Pakistan. Selected farmers were encouraged to use fertilizer and high-yield grain, in hopes that neighbors would observe their success and emulate them. At first things went as hoped. Then the farmers with modern techniques grew so prosperous that they started wearing clothing that came below their knees. At that point, their bare-kneed neighbors stopped emulating their methods. Boyd and Richerson's suggestion is that we have mechanisms that assess who is relevant—on the basis, partly, of lifestyle—and then assess who, among relevant people, are successful and what they do. They offer a mathematical model of the selective advantages of having such mechanisms of discriminating imitation. It may be better to imitate those who are successful and who are like oneself, than to analyze why some ways of life lead to success and some do not. The analysis may be too complex and fallible; a somewhat crude heuristic of finding models and imitating

[5]See Nisbett and Ross (1980) on human heuristics in this sense.

them may work better.

I have been touching on some of the kinds of things that must complicate my simple pattern of avowal, influence, and acting on what one avows. What emerges, I hope, is a strategy for investigation. Suppose that human action is a matter of psychic mechanisms that respond to cues and that lead to action. Ask what kinds of mechanisms would do well in important game-theoretic structures. Then look for indications that such mechanisms might be at work in human affairs. These mechanisms, I am suggesting, will include tendencies to avow, to be influenced, and to act on what one avows. There will be refined mechanisms, though, for assessing whom to be influenced by, in what directions. There will be refined mechanisms for assessing how personally advantageous a set of norms one can get away with avowing. And there will be refined mechanisms for assessing whether to act on one's avowals. Even a sketchy understanding of the game-theoretic structure of recurrent human interactions will suggest kinds of psychic mechanisms we might expect to find in human beings.

3. Morality

How does all this bear on human morality? Morality in a narrow sense, I suggested, we could understand as consisting in norms for guilt and anger. This identification is somewhat arbitrary; there are various features of our own everyday morality that we could take as its defining features. But there are advantages to thinking of morality as centering on blame. A person is to blame for something he did, it seems fair to say, if he and others can reasonably blame him for it. We can think of guilt as the feeling of self-blame, and anger as the second or third party feeling of blame.

Now if we delineate narrow morality as I have proposed, then a moral psychology will be a psychology of norms for guilt and anger. Or rather, we should speak of a moral social psychology, since people work out what norms to accept in interaction with each other. A social psychology of morality will be successful if it explains the psycho-social dynamics of normative discussion and normative governance. In particular, it will deal with discussion that can be well translated as developing norms for guilt and anger. It will explain what sorts of norms, if any, people are likely to accept for guilt and anger in various different kinds of circumstances. It will discover what psychic mechanisms are involved, and how they interact to produce moral convictions. It will also explain why, often, people avoid the kinds of actions they think would warrant guilt and anger, and under what circumstances they are most likely nevertheless to do things they would condemn.

It is open to question, though, how significant morality is in the narrow sense I have proposed delineating. Even if my proposal does point to crucial features of morality as we know it, it still might be that morality as we know it is peculiar to modern Europe and its cultural offshoots. That would mean, among other things, that whatever the genetically coded psychic mechanisms are that make for moral thinking and motivation, these mechanisms are adapted to doing something else.

Science provides a good parallel. Modern science started as a peculiarly European development, though like many features of European culture it has now spread globally. It would not make sense, then, to look for genetically coded psychic mechanisms adapted to doing science. It would make sense, though, to ask what psychic mechanisms are involved in doing science, and what those mechanisms were shaped by evolution to be able to do. It might be likewise with morality.

Still, even if morality narrowly conceived has been rare among human groups, morality in broader senses is clearly widespread. An interplay of language, action, and feelings about actions is found everywhere I have heard anything about. It may be universal that such talk, and the motivations that attach to it, play a crucial role in adjusting human interactions. If we suppose that morality, in the narrow sense I have proposed, is peculiarly modern European, then the picture we should accept is this. What is peculiar to Europe is a reliance on guilt and anger as the feelings called on to play a central regulative role. Europeans have identified these feelings, cultivated them, and elaborated norms for them. This involves the workings of psychic mechanisms biologically adapted to coordinate human actions. In special cultural circumstances, though, these mechanisms work in special ways.

Is narrow morality, then, really peculiar to modern Europe and its cultural offshoots? This is a big question, and I am in no position to begin to tackle it. Much needs to be done to investigate what features of European-influenced morality are special, and what features are widespread or universal. I have not been trying to answer these questions, but to suggest one way of formulating them.

Let me glance, though, at a few scattered indications. Sometimes the rules of other cultures will strike us as strange and bizarre, but they are accepted as moral requirements—or so many observers are inclined to say. In anarchic tenth century Iceland, open killings and raids on property were assessed as glorious at times and imprudent at times, but there was no general moral condemnation of these acts of affront or domination. Secret killings and thefts, in contrast, were regarded as ignominious (Miller, 1990). The African Ik whom Colin Turnbull studied shocked him

in many ways, but stuck to the rule that food must be shared with all those present. That did not mean that the Ik were generous to each other, but that they went to great lengths to be alone when they had acquired meat (Turnbull, 1972).

Are these rules, strictly speaking, moral rules? I don't know. It would be marvelous if ethnographic psychologists could tell us more about them. How are rules like these taught or shared? What is the developmental psychology of their uptake? What are the emotional flavors of criticism, blame, and the like to be found among exotic peoples? To what lengths are standards of conduct elaborated in discussion, criticism, quarrels, orations, and the like? More information on these scores could begin to indicate whether the criticisms ethnographers identify as moral are tied to feelings like guilt and anger, or have some other kind of sanction.

We can be confident that some aspects of modern European morality will turn out to be universal among humanity, and that others will be more or less peculiar. I have proposed a schematization of European morality that might help us pose the right questions. Like any schematization, it is bound to oversimplify European ethical life. It may, though, point to some of the chief features of the thought, discussion, and motivation of modern Europeans on which European moral terminology and theories have fastened. If so, it may give us a way of asking what role, if any, those features play in human life in general.

4. Concluding remarks

From a social psychology of judgments of right and wrong, nothing follows directly about what things really are right or wrong. The same goes for evolutionary hypotheses. Judgments may be correct or incorrect, for all psychological or evolutionary theories of their origin tell us. Such theories may have some legitimate bearing on questions of right and wrong, but the bearing will be complex and indirect. All this would be a long story.

Return, then, to evolutionary moral psychology, and let me finish with a few methodological remarks: I have been working within a scientific world picture, but I have not been claiming anything like good scientific levels of proof for my speculations. I don't think, though, that we should confine ourselves to thoughts that could be proved or refuted. Science does not jump from a void in thought and discussion to well confirmed theories. It does not progress from one clearly testable hypothesis to another. We need to alternate looking at evidence and thinking about where the evidence might be leading us.

Sometimes evolutionary hypotheses do attain a high degree of credibil-

ity. Eventually, we might hope, hypotheses about the evolution of human capacities for moral judgment will attain this status. Already, I think, we have good evolutionary hypotheses about some other aspects of human psychic capacities. I have in mind, for instance, some of Donald Symons' work in his book *The Evolution of Human Sexuality*.[6] I have not attempted the kind of search of the literature that could begin to make my hypotheses credibly scientific, and to refine them in light of the evidence. But my hope is that against the background of the rough hypotheses I have been setting out, such a survey would be rewarding.

Success in this realm is not a matter of decisive tests. We have it to a high degree when the evidence fits a coherent, plausible evolutionary picture, and does so much better than we could expect it to fit any equally coherent and plausible alternative. This ambition, I think, will sometimes be attainable. We can observe some of the outcomes of past evolutionary histories. Often we can have a good enough idea of what kinds of selection pressures must have been at work. All this can sometimes let us infer with fair confidence what must have happened in the distant past.

The degree of justified confidence we can attain will vary from topic to topic. Making unreasonable demands for standards of proof can stifle promising investigations. We should make sure that we demand no standards of proof that would forbid our saying that fish gotta swim and birds gotta fly, and that wings and fins are adaptations. In addition, though, we should not disdain thought about human evolution that is far less certain than these things are, as long as we recognize the degree of uncertainty. We should, of course, subject evolutionary speculations to vigorous analysis and criticism. We should distinguish loosely supported speculation from hypotheses that fit extensive evidence, and distinguish these from hypotheses where a careful work has been done to eliminate alternative hypotheses. Our goal, though, should not be to weed out all uncertainty or to close down speculation. It should be, rather, to identify the most promising lines of investigation, and to assess our degrees of knowledge and ignorance.

When we think about human beings, to be sure, we have to recognize the dangers. Evolutionary pseudo-science has been exploited to horrible ends. I don't think the cure, though, is to demand complete certainty in all permissible evolutionary thinking about humanity. We aren't going to go through life as if we thought we knew nothing of what human beings are like or could be like—and other kinds of thinking can lead us astray too. Most of the cure for human brutality isn't a matter of rules of evidence at all. Part of the cure, though, is to be skeptical of everything to the

[6]Symons (1979); see also Daly and Wilson (1988).

degree that it warrants skepticism.

REFERENCES

BOYD, ROBERT and RICHERSON, PETER (1985), *Culture and the Evolutionary Process*, (Chicago: University of Chicago Press).

DALY, MARTIN and WILSON, MARGOT (1988), *Homicide*, (New York: Aldine de Gruyter).

DARWIN, CHARLES (1872), *The Expression of Emotion in Man and Animals.*

DAWKINS, RICHARD (1976), *The Selfish Gene*, (Oxford: Oxford University Press).

DAWKINS, RICHARD (1982), *The Extended Phenotype*, (Oxford: Oxford University Press).

GIBBARD, ALLAN (1990), *Wise Choices, Apt Feelings: A Theory of Normative Judgment*, (Oxford: Oxford University Press).

HAMILTON, WILLIAM D. (1964), *"The Genetic Evolution of Social Behavior"*, Journal of Theoretical Biology Vol. 7: 1–52.

MAYNARD SMITH, JOHN (1974), *"The Theory of Games and the Evolution of Animal Conflicts"*, Journal of Theoretical Biology Vol. 47: 209–21.

MAYNARD SMITH, JOHN (1983), *Evolution and the Theory of Games*, (Cambridge, England: Cambridge University Press).

MCDOWELL, JOHN (1988), *"Projection and Truth in Ethics"*, (Lawrence, KA: Department of Philosophy, University of Kansas).

MILLER, WILLIAM (1990), *Bloodtaking and Peacemaking: Feud, Law, and Society in Saga Iceland*, (Chicago: University of Chicago Press).

NISBETT, RICHARD and ROSS, LEE (1980), *Human Inference: Strategies and Shortcomings of Social Judgment*, (Englewood Cliffs, NJ: Prentice-Hall).

ROSALDO, MICHELLE (1973), *Language in Society Vol. 2: 193–223.*

SCHELLING, THOMAS (1960), *The Strategy of Conflict*, (Cambridge, MA: Harvard University Press).

SYMONS, DONALD (1979), *The Evolution of Human Sexuality*, (Oxford: Oxford University Press).

TURNBULL, COLIN (1972), *The Mountain People*, (New York: Simon and Schuster).

WILLIAMS, BERNARD (1985), *Ethics and the Limits of Philosophy*, (Cambridge, MA: Harvard University Press).

WILLIAMS, GEORGE C. (1966), *Adaptation and Natural Selection*, (Princeton: Princeton University Press).

WIGGINS, DAVID (1987), *Needs, Values, Truth: Essays in the Philosophy of Value*, (Oxford: Basil Blackwell).

WRIGHT, LARRY (1973), *"Functions"*, Philosophical Review Vol. 82: 139–68.

Logic, Methodology and Philosophy of Science IX
D. Prawitz, B. Skyrms and D. Westerståhl (Editors)

CONCEPTUAL ISSUES IN ETHICS OF SCIENCE AND TECHNOLOGY

QIU RENZONG

Institute of Philosophy, Chinese Academy of Social Sciences, Beijing, China

In this paper I will deal with some conceptual issues in arguing for an ethics of science and technology as a species of professional ethics. My argument will start with the crisis of conscience scientists have experienced since the explosion of the first atomic bomb in Hiroshima. These experiences, in my view, refuted the established model of science as a value-free pursuit of truth or self-governing activity independent from its social and cultural context. In the second, third and fourth sections I will argue that obligation, virtue and just distribution of resources should be the essential aspects of the ethics of science and technology. In the conclusion, I will argue that the ethics of science and technology as a species of professional ethics is the only way out of conflict between professional autonomy and social control by a scientific community's self-regulation of its own conduct. In this paper technology is classified as applied science, science mainly refers to natural sciences but does not exclude human and social sciences.

Value conflicts and model of science

Since 1945 there has been three major events by which many scientists were so shocked that they experienced a crisis of conscience and felt it necessary to reflect on their own activities.

The first is the explosion of the first atomic bomb in Hiroshima. It shows that, as B. Baumrin [3] points out, "a scientific theory which had begun in philosophical nursings and scratchings on paper had culminated in the terrible deaths of thousands".

The second is the trial in Neuremberg in 1945. This event revealed that scientific research which aims at the discovery of the impersonal truth of universe could be proceeded in such an inhumane way that it violated

fundamental human rights and killed innocent people.

The third is the sudden discovery of the silence in spring — the worldwide environment pollution which threatens human existence on this planet and even puts its very existence at stake.

These major events, among others, caused many scientists and lay people to be seriously concerned with the social consequences of their research results, the impact of their application on society, humanity and ecology, as well as the scientific conduct. Those concerns that are laden with value conflicts or ethical dilemmas could be listed as follows:

(1) Concerns about risks

There are some research activities that are themselves dangerous because of the materials employed or produced which are risky to human health, such as radioactivity in the research of high-energy physics, or pathogenic microorganisms in the research of molecular biology. Biochemists at Ashilomar adopted a self-imposed moratorium on recombinant DNA research in fear of unknown pathogenic microorganisms escaping from laboratories. The Three Mile Islands and Chernobyl accidents revealed the possible catastrophic consequences of nuclear generators. Let alone such products as PCBs, freon, and even excessive release of CO_2 which have proved harmful to human health or human environment.

(2) Concerns about misuse

Research results can be misused to cause harm to some part of a population or to society as a whole. Scientists worried that the results of research on the relationship between race and IQ would be misused to support racism, the discovery of centers in brain controlling behaviour and the results of genome mapping be misused to control human behaviour. Many technological applications that seemed a blessing to mankind when first introduced became threats when their use became widespread. For a well-known example, DDT was employed to eliminate disease-carrying pests and raise agricultural productivity, but it also threatened ecological systems, including the food chains of fishes, birds and eventually human beings. And the use of psychotropic drugs to treat children with learning disabilities or hyperactive children has already led to misuse in many cases.

(3) Moral concerns

Scientists are sometimes challenged that the knowledge they acquire might create a danger by undermining human values or threaten the foundations of public morality. For example, does Darwinism undermine the

religious values cherished by many Americans? Does Mondal-Morgan's biology, Virchow's cellular pathology and Wiener's cybernetics undermine the ideological values of dialectical materialism in a socialist country? It was feared that Frankenstein-like monsters would be created from science and technology and get rid of human control or even enslave their own creators — human beings violating the beliefs about free will and self-determination. Let alone that the practice of abortion, the use of fetal tissues, the research on embryos, and even the use of RU 486 (abortion pill) etc. are thought to violate moral and legal principles by 'killing' an innocent 'human being' if the upper limit of human life was set at the beginning of conception. People have worried that reproductive technology such as contraception, artificial insemination, in vitro fertilization, surrogate mothers, and embryo transfers etc. would undermine the traditional structure of marriage and family. And the use of prenatal sex selection has already jeopardized the balance of the sex ratio in China which could in turn cause a series of social problems.

(4) Concerns about equity

The research results and the application of science and technology may benefit some part of a population and burden or harm others. Prenatal research led to a technology which can save the life of a very low birth weight (say, 500 g.) newborn at the expense of half a million US Dollars. Sophisticated life-supporting technology can prolong the life of comatose patients, those in a persistent vegetative state, and terminally ill at the expense of unbearable burden on others and society without improving their quality of life. And all researchers need grants from governments or private foundations. For each researcher it seems that the more the better. Thus arises the problem of just allocation of resources.

(5) Concerns about individual rights

In biomedical, psychological and social research some experiments have to be done on human subjects. In these experiments, subjects will be exposed to harm in order to benefit society as a whole, or their confidential information will be revealed to the researcher. So voluntary informed consent has to be obtained from human subjects. Otherwise, the rights to privacy and self-determination of these subjects will be violated and their interest be infringed upon [5, 15]. Scientists worried that the knowledge of reproductive bio-medicine would be applied to construct such a world as that described by A. Huxley in his *Brave New World* in which not only the traditional structure of marriage and family is destroyed, but also the individual rights are systematically violated.

Value conflicts are involved in all these concerns; the interest of researchers, the progress of knowledge in the given discipline made by the research, benefit to people or society from the research, burden or cost of the research on society, harm caused by the research itself or the application of its results to part of the people, interest and rights of human subjects, impact of its application on other social, economical, political or cultural factors of society, and on human environment and ecology. Scientists gradually became aware that they have to balance all these values concerned to make decisions about their actions.

Value conflicts led scientists to moral dilemmas in many cases. Moral dilemmas arise when there is a conflict between two obligations — both ought to be fulfilled, but only one can be fulfilled. When the progress of some knowledge is in many people's interest but the acquisition or application of this knowledge will cause harm to some people — what ought the researcher to do? There is a conflict between the obligation of benefiting people by the research and the obligation of doing no harm to people.

The fact that scientists are increasingly aware of the social implication of scientific research and the potential misuse of scientific knowledge has refuted the established model of science, according to which science is characterized as a value-free pursuit of truth or a self-governing activity independent of its social and cultural context.

It was argued in [16] that this model of science could be derived from R. Merton's [13] values or rules of the scientific game, which make up the ethos of modern science and serve as a guide for its practitioners: universalism, communalism, disinterestedness, and organized scepticism. The point is open to discussion. However, these values are mainly concerned with the internal aspect of scientific knowledge. Indeed, some of them suggest that science is operative in a context of individual sovereignty. An example is the value 'disinterestedness'. A colleague [9] interpreted it as 'selfishlessness'. I doubt this interpretation to be closer to what is meant by Merton. Even so it is hard to define what 'selfishlessness' means in the given context. Does it require a scientist to do research without any personal concerns or only without harm to others? So this interpretation seems to be irrelevant here. Rather, 'disinterestedness' is to be interpreted as arguing for a science for the sake of science, a science without concerns about its utility. A science for the sake of science, or a science in the ivory tower is a myth in modern society. And it is this value which has affinity with the established model of science.

This model ignored the fact that the interaction between science and society, and the process of integration of science into society at large

have been more explicit and intensified since the end of the Second World War. On the one hand, as D. Nelkin [16] points out, the technological sophistification of research requires considerable dependence on external funds, which in turn leads to a social process of collectivization and instrumentalization of science, inevitably implying less disinterested research in a context of personal autonomy or individual sovereignty. Institutional and societal pressures force scientists in a tightened economy to select research questions based less on disinterested judgements of intrinsic scientific merit than their institutional and social needs. They also have to take into account the public concern about the ethical implication of research [16].

On the other hand, the interval between the result obtained in research and its application is considerably shorter than before, and it may be more probable than before that scientists can anticipate, even if they cannot precisely predict, the potential application of their research through many links between them. These factors make it more explicit that scientific research is not a neutral, value-free, merely cognitive activity within a self-sufficient system.

It is plausible when B. Vitale claimed the demystification both of the concept of science as the neutral pursuit of truth and the image of scientists as priests of truth [21]. According to the established model of science, scientists are described as people who are members of an elite in society, capable of possessing some sort of omniscience, and it is believed that what is good for them is good for society. As B. Vitale argues, at the end a beautiful picture of scientists is constructed as priests of truth and guardians of people's welfare. This image does not take into account the fact that scientists have to be dependent on external support, so they and the ruling class need each other. As B. Vitales points out: "The power of the ruling classes is strong, and scientists oscillate between their interest in knowledge and their interest in power, prestige and privileges, leaning rather heavily towards the latter" [21].

The refutation of the established model of science will have some important implications for the ethics of science and technology. First, we have to answer the following questions: As far as the end-product of science is concerned, whether or to what extent the scientist ought to be held responsible for the consequences of her/his discovery? As far as scientific conduct and method are concerned, how ought he/she to proceed with human experimentation? Those questions lead us to a theory of obligation for scientists as an essential part of the ethics of science and technology.

Obligation

In conformity with the established model of science, the traditional theory of scientists' obligation is that scientists' job is discovering and inventing, and so they would bear no responsibility for the application of their work. For example, P. N. Bridgeman argued that a scientist should never bear the responsibility for any application of his work relying on the fact that the harm resulting from scientific work arises not at the point of research but at that of industrial manufacture [3]. This argument can be countered as follows: First, risks can appear in the process of research itself, e.g. high-energy physical research and biological research. Secondly, in some cases the cost of research is very high, but the cost of manufacture is rather low, so control of research is often more practical than is control of industry [3].

Another argument for the traditional theory of scientists' obligation is that science is good, the pursuit of scientific knowledge is a good activity in its own right, and even better since scientific knowledge is an absolute good apart from its consequences.

In the counter-argument some sociologists went too far in arguing that science seems a vice in the sense that they discovered that individual scientists are 'all-too-human', that the life of the scientific community can be portrayed as a career oriented power struggle, and scientific research is used by the members as a main tool in their ruthless self-promotion and careerism. It is hard for me to agree with them. But what does it mean that an activity is good? Does it mean that if an activity is good, its products are always good? Can a good activity give rise to a thoroughly undesirable product? And it is hardly possible to absolutely separate the acquisition of knowledge from the application of knowledge. Knowledge is to be tested in practical applications as well as in laboratory experiments. Moreover, resources available to scientists in a society is limited, and the limited resources ought to be given priority to be allocated to the solution of pressing social problems and the promotion of the common good. Even if the resources are sufficient, not all that can be done ought to be done. What can be done is a scientific or technological judgement. What ought to be done is a moral or ethical judgement. As B. Baumrin [3] argues, "if scientists choose to justify their choice of engaging in disinterested science on basis of its being pleasant for them, or its beauty or its self-fulfillment, then to ask for social support for their activity is facetious". Instead, as J. Neilands [14] points out, "the research scientists have the responsibility for misdirected technology, simply because it is we who preside over the wellsprings of knowledge in this area". For example, both arms race and

environmental pollution arise from technology, which in turn, is rooted in scientific research. "It is part of the inevitable 'peril and promise' of science" [14] that researchers of basic science cannot escape responsibility for these two evils.

So, as J. Ravetz [19] argued, the goal of science is not only one, but two, by which it is made legitimate: the attainment of the True and the production of the Good. The production of the Good includes acquiring knowledge in a moral way, and applying knowledge to benefit people and society. The realms in which these goals are attained are Cosmos and Humanity. Or, in F. Bacon's phraseology, 'light' and 'fruit' are two goals of science. It was F. Bacon's claim that the moral value of the pursuit of scientific knowledge came from the value of the general social utility promoted by its achievements. The crisis of conscience which was experienced by scientists during the event of Hiroshima or others implies the refutation of the traditional theory on scientists' obligation. At least, they thought this theory was inadequate.

This inadequacy was reflected also in the argument for the transition from the philosophy of knowledge to the philosophy of wisdom made by N. Maxwell [12]. He argues that according to the philosophy of knowledge, the basic aim of science is to produce 'reliable, objective, factual knowledge' (p. 16). Most scientists seem to believe that this can be achieved only if that aim is pursued independently of psychological, sociological, economic, political, moral and ideological factors and pressures. But Maxwell believes that many members of the scientific community see the internal moral order in relation to large humanitarian enterprises, the general betterment of the quality of human life. On this humanitarian or Baconian view, the connection is simple, by providing mankind with objective truths scientific community contributes to the humanitarian aim. And Maxwell tries to subsume under the one 'philosophy of wisdom' both the morality of a science directed to the resolution of human problems, and the attitude to science as a road to a deep understanding.

If we accept the thesis that scientists have the obligation not only of attainment of the True, but also of production of the Good, we are forced to make judgements about what kinds of knowledge should be pursued or given priority to be pursued, and how knowledge should be applied. These are moral or ethical judgements. Science is what is, ethics is what ought to be. One cannot derive an 'ought' from an 'is', or vice versa. What can be done scientifically and technologically does not amount to what ought to be done morally or ethically.

The conscience of scientists is more strongly expressed in the movement to accept greater responsibility for the use of their knowledge in which

scientists are playing multiple roles, such as advisors to policy-makers, consultants to governments and private enterprises, expert witnesses in courts, social critics, popularizers of science, advocates for community groups, etc. In recent years Chinese scientists stepped forward to argue against the proposed 'Three Gorges Project', in which the upper reaches of the Yangtze River will be dammed to build a huge hydraulic power plant. Many Chinese scientists expressed their great concerns about the potential disastrous social and ecological consequences of this project, and forced the policy-makers to postpone their decision [4]. These roles, intrinsic to the scientific enterprise, contradict the expectation of disinterested research and the image of neutrality. It was plausible to argue as B. Baumrin does [3], that doing disinterested science or science for the sake of science and justifying that activity in terms of its benefit to mankind are both immoral in the sense that they break the duty to be honest and give rise to avoidable deleterious consequences.

Virtue

Knowledge is power. Modern science and technology have the great and awesome power of changing the social, economical and cultural pattern and the structure of life including food, clothing, shelter and transportation, and even birth, age, illness and death. To provide the Good with science it is necessary to have norms to control and optimize the social uses of scientific knowledge and resources. It was argued that scientific activities were best characterized in Merton's values mentioned above in which cognitive norms as well as ethical norms are implied [9]. The norms which are implied as ethical are inadequate. Morality cannot naturally flow from cognitive norms. And ethical norms need to be made from these values. But even if we did have ethical norms, it is not sufficient.

Now fraud in science has become a pandemic spread all over the world. Even reasearchers in esteemed institutes or famous scientists commit fraud. In our country, a favourite scientist, who is at the top of the power structure, played a leading role in the fraud of using falsified data to advocate parapsychology or justify a stupid policy which led to catastrophic consequences. The cases of fraud in science are absolutely not few, but we do not know if they were simply episodes that will drift into the history of science as footnotes, or the top of an iceberg. Anyway, fraud gravely damages the image of science and scientists. Some scientists blame the problem on the external pressures on science, and others on the moral deterioration of the society at large.

The epidemic of fraud in science can be explained only by the combi-

native effect of internal and external factors. The current reward system are external reward oriented. The incentives to scientists are money, prestige, power and fame. As C. J. List [10] points out, internal rewards are self-caused, and external rewards are other-caused. External rewards are limited and socially arranged, contingent upon the existence of institutions that control and distribute them. As for internal factors one is the weakness of cultivating virtues in scientists since the traditional virtue-oriented ethics has been replaced by modern norm-oriented ethics.

As A. MacIntyre [11] pointed out, "virtue is an acquired human quality, the possession and excercise of which tend to enable us to achieve those goods which are internal to practices and the lack of which effectively prevents us from achieving any such goods". Virtues, such as justice, courage and honesty have to be accepted as necessary components of scientific practices. For example, a scientist should have courage enough to propose a bold, novel theory, should be honest in testing her/his theory, and should be just in comparatively assessing her/his theory and others' competitive theories. Faust who sold his soul to the devil is no longer the model of a good scientist. It is these virtues in which the norms of science must be grounded. The mere existence of norms, cognitive or ethical, is insufficient to deter fraud in science [10]. As Confucius once argued: If a man has no virtue, what has he to do with norms? [1]

There are some arguments against virtue-oriented ethics [20]. The proper virtue set is not obvious, the proper set of virtues for a particular role is not obvious, virtue theory can lead to wrong acts, virtue theory is unnecessary in science or technology that is practiced among people who are essentially strangers. These counter-arguments seem plausible. In a pluralist society it is not easy, if not impossible, for people to agree on which virtues should be possessed. However, as Confucius claims, human natures are similar, only nurture makes them apart. Otherwise, how can we sit here to discuss the topic we are all interested in? And it does not follow from the fact that the proper set of virtues is not obvious that we cannot get nearer and nearer to the proper set of virtues on the basis of practical experiences.

As for the relationship between a virtuous character and right acts, it is true that there is no direct, necessary relation between them, and there is no guarantee that any act of a virtuous man is necessarily or unavoidably right. However, we can say that it is more probable for a person with a virtuous character than for a person with a vicious character to do right acts. And it is not the case that virtue is unnecessary in science and technology in a world of "strangers". Our planet is becoming smaller and smaller. The beneficiaries or victims of an application of your discovery

or invention may be strangers you never met before, may be your relatives or even yourself. In any practice if we want to achieve higher standards of excellence, we must possess and exercise some virtuous character traits. Last but not least, virtue ethics is not designed to replace normative ethics. They are complementary to each other. It is implausible to suppose that they are incompatible so that if we accept one we have to reject the other. Both of them are needed to achieve higher standards of excellence.

As for the countermeasure to deter fraud in science, together with the cultivation of scientists' virtuous character, the institutional external reward-oriented system must be reformed. Scientists should be promoted to attain the self-caused internal rewards obtained only by scientists who possess virtues and observe norms.

Distributive justice

Scientists who are engaged in scientific practice as an institutional activity have to deal with problems of just distribution of benefits and burdens on the parties concerned. Among others, it is uncontroversial that decisions have to be made on the allocation of resources at macro-, meso- and micro-levels on the basis of the balance of different or even incompatible values.

The ethical problem of resource allocation at macro-level is that among the resources available to a society or country, how much is proper to be allocated to the department of science and technology, and among the resources available to the department of science and technology, how much is proper to be allocated to each group of disciplines or basic research, theoretical research and applied research etc.

What is 'proper'? 'Proper' does not mean that the resources allocated to science are 'the more, the better'. Science and tecnology as a department of the national economy are interrelated and interdependent with other departments such as agriculture, industry and especially education. So at macro-level we have to balance the needs of different departments of the national economy to set a proper proportion of resources to be allocated to science and technology for optimizing their development in a given country at a given time. It is the same with the relationship between basic research, theoretical research, applied research and development within science and technology.

The ethical problem of resource allocation at meso-level is that among the resources available to an institution, how much is proper to be allocated to each section or unit that is affiliated with the institution.

The ethical problem of resource allocation at micro-level is how to properly use the resources available to a scientist.

It is admitted that there is an interdependence between different disciplines, or different departments in an institute or different groups managed by a scientist. Over-allocation of resources to a discipline is not necessarily good to the discipline itself, as well as mal-allocation to a discipline is not necessarily good to other disciplines, because all disciplines depend on a coordinate development with one another. In Confucian phraseology, 'Perfect is the Mean', or going too far is as bad as not going far enough [2]. There are many cases in China in which much resources were allocated to an incompetent scientist in one department, who, however, was very competent to boast her/his achievement to the decision-maker in resource allocation and got excessive funds for her/his research at the expense of her/his colleagues' interest. This is immoral in two senses, the first is to violate the duty to be honest by deliberately exaggerating her/his achievement beyond normal error. The second is that her/his conduct led to unjust allocation of resources.

Science as profession

There has been very little talk about science as a profession. It has been argued that science is not a profession because of its unique characteristics distinctive from typical professions like medicine and law. Now let us look for a while at the characteristics of typical professions, and examine whether there is any essential difference between science and them [6] [2].

First, members of typical professions provide personal services to individuals with whom they have a special professional relationship, namely clients, patients, parishoners or students in law, medicine, divinity or university faculty, and this relationship is, in some respect, very close and intimate. It seems that for scientists there are no such personal services to individuals and no such intimate relationship with them. However, nor do some lawyers or doctors provide personal services and have intimate relationship, such as the lawyer employed by a corporation, or the epidemiologist working in public health administration. On the other hand, some scientists, such as psychologists or geneticists who provide councelling services could have personal and intimate relations with their clients. Because of institutionalization, professionals are often in a tripartite relationship with their clients and employers, such as client-lawyer-office or court, patient-physician-hospital, parishioner-minister-church, student-university faculty member-administration. It is this tripartite relationship

which sometimes puts the professional into moral dilemmas when there is an incompatible conflict between the obligation to her/his client and the obligation to her/his employer. Scientists who are employed by government or private enterprises share with other professionals the tripartite relationship between their employers and bearers of the consequences of their research results. The only difference between them is that for scientists there is no personal relationship with the bearers of consequences. The question is whether this difference is wide enough to make science a non-profession. If scientists provide counselling services they will enter in personal relationship with the bearers of consequences. On the basis of what has been said above the difference between science and a so-called typical profession is rather in degree than in kind. Science may be at the one end of the profession continuum.

Second, these professions traditionally control the services they offer, the standards of evaluation of these services, and the qualification for membership in the profession. This seems to hold for science and technology. However, all professions are now undergoing the pressure of external control. There is no exception for science. In China scientists complained of the intervention in evaluation of research results from 'omnipotent' officials, journalists or even writers, and in the United States physicians complained of the control of their services by insurance companies. The sophistication of equipment in science, medicine and university faculty makes these professions vulnerable to intervention and control by external powers.

Third, the members of these professions often assume the role of moral arbiters of what is morally good for their clients and even of what is morally good for society. Scientists assume the same role. As we argued in the previous section, scientists have the obligation to tell their employers, people and society which kind of knowledge should be pursued with limited resources and how to apply the knowledge. However, for scientists and other professionals this role is limited, because of complexities it is often not clear what is morally good for their clients or society. This moral role should be played in cooperation with colleagues from other disciplines, especially those from human and social sciences.

Fourth, all these professions provide a body of theory which affects practice and forms a professional culture. When lay people enter into a relationship with a professional in an institutional framework, they go to a new and strange place, where people speak a different language and have strange manners and values. This puts lay people at a disadvantage, they are outsiders, weak and unequal, and therefore necessarily dependent on the good will of the professionals. This holds completely for science.

Fifth, these professionals represent the élite and belong to the upper social class, superior in education, intelligence, wisdom and authority, and they are moral superiors too, and expected to have a sense of honour, and to devote themselves to the welfare of others. Scientists are the most superior among them. In China, even if the salary of a scientist may be lower than that of a factory worker, and far much lower than that of a private merchant, her/his social status is the highest. In a survey to young people recently made in China, among the responses to the question of 'what is your most preferable profession?' the first is 'scientist' and the second is 'physician'.

Sixth, many professions possess an ethical code. But science has none. That does not mean that science is not a profession. Possessing an ethical code may not be the characteristic of a profession, but the result of being classified as a profession. If a scientific community feels the need or is determined to possess an ethical code, they will have one.

I should like to add one more argument. The relationship between clients and professionals is characterized as a fiduciary one in all these typical professions in which trust or trustworthiness is the unique feature [17]. As early as in the 1960's M. Polanyi [18] claimed that in understanding scientists' statements we must make the 'fiduciary transposition'. Assertions of fact must be transposed into the 'fiduciary mode'. They are to be read not as 'This is true' but as 'Trust me when I say that ...'. To adopt the fiduciary mode is itself a fiduciary act. Recently R. Harré argued [7] that science is not just an epistemological but also a moral achievement, and the scientific community exhibits a model or ideal of rational co-operation set within a strict moral order (p. 1). He claims that science is a cluster of material and cognitive practices, carried on within a distinctive moral order, whose main characteristic is the trust that obtains among its members and should obtain between that community and the large community with which it is interdependent, and at the heart of the morality with which the scientific community practices is the commitment that the products of this community shall be trustworthy (p. 6). Science shares this characteristic with other professions.

In any case there is no adequate reason to preclude science from the professions.

Scientists may refuse to accept the notion of science as a profession, because this notion implies codes and certificate procedures. They are accustomed to informal collegial control and concerned about individual sovereignty. Scientific societies have traditionally been more concerned with promoting 'good science' than with controlling the actions of their

members. They thought that the norms of scientific honesty are sufficient without codification. However, it follows from what has been argued above that the norms of scientific honesty are not sufficient. The problem is that if scientists and their community do not come to regulate their activities by themselves, they will not be able to resist the external pressures intruding on scientific autonomy. Self-regulation is the only path avilable to scientists for the retention of autonomy. The voluntary self-imposed moratorium on recombinant DNA research adopted by biochemists at Ashilomar, though it might be overcautious in retrospective, can be seen as an attempt to avoid outside intervention. So ethics of science and technology as a kind of social control can be reconciled with professional autonomy by a scientific community's self-regulation of its own conduct.

REFERENCES

[1] *Analects*, in W.-T. Chan (ed.): A Source Book in Chinese Philosophy, Princeton University Press, Princeton, NJ, 1963, pp. 18–49.

[2] *The Doctrine of the Mean*, op.cit., pp. 97–114.

[3] BAUMRIN, BERNARD, *The Immorality of Irrelevance: The Social Role of Science*, in: B. Baumrin & B. Freelman (eds.), Moral Responsibility and the Professions, Haven Publications, New York, 1963, pp. 230–250.

[4] Dai, Jing (ed.): *Yangzi, Yangzi: The Debate on the Three Gorge Project*, Guizhou People's Press, 1989.

[5] DAVIS, BERNARD: *Limits in the Regulation of Scientific Research*, in T. Segerstedt (ed.): Ethics for Science Policy, Pergamen Press, Oxford, 1979.

[6] GREENWOOD, ERNEST: *Attributes of a Profession*, in: Moral Responsibility and the professions, pp. 20–32.

[7] HARRÉ, ROM, *Varieties of Realism: A Rationale for the Natural Sciences*, Basil Blackwell, Oxford, 1986.

[8] LADD, JOHN *Philosophy and the Moral Professions*, Oelgeschlager, Grun & Main, Boston, MA, 1985, pp. 11–30.

[9] LIU, JUNJUN, *Sociology of Science*, Shanghai People's Press, 1990, pp. 183–194.

[10] LIST, C. J., *Scientific Fraud: Social Deviance or the Failure of Virtue?*, Science, Technology and Human Values I, No. 4, 1985, pp. 25–36.

[11] MACINTYRE, A., *After Virtue: A Study in Moral theory*, University of Notre Dame Press, Notre Dame, IN, 1981, p. 178.

[12] MAXWELL, NICHOLAS, *From Knowledge to Wisdom*, Blackwell, Oxford, 1984.

[13] MERTON, ROBERT, *Science and Technology in a Democratic Order*, Journal of Legal and Political Sociology I, 1942, pp. 115–126.

[14] NEILANDS, J. B., *Communication with Others — the SCIENTISTS" Responsibility*, in: Ethics for Science Policy, pp. 177–185.

[15] NELKIN, DOROTHY, *Science as a Source of Political Conflict*, in: Ethics for Science Policy, Pergamen Press, Oxford, 1979, pp. 9–24.

[16] NELKIN, DOROTHY, *Social Controls in the Changing Context of Science*, in: Social Controls and the Medical Profession, pp. 83–94.

[17] PELLEGRINO, E., VEATCH, R. and LANGAN, J. (eds.), *Ethics, Trust, and the Professionals: Philosophical and Cultural Aspects*, Georgetown University Press, Washington, DC, 1991.

[18] POLAYI, MICHAEL, *Personal Knowledge: Towards a Post-Critical Philosophy*, Roulledge & Kegan Paul, London, 1962.

[19] RAVETZ, JEROME, *A Critical Awareness of Science*, in: Ethics for Science Policy, pp. 49–56.

[20] VEATCH, ROBERT, *Against Virtue: A Deontological Critique of Virtue Theory in Medical Ethics*, in: E. E. Shelp (ed.), Virtue and Medicine: Explorations in the Character of Medicine, Reidel, Dordrecht, 1985, pp. 329–346.

[21] VITALE, BRUNO, *What becomes of a Scientist?*, in: Ethics for Science Policy, pp. 145–152.

Logic, Methodology and Philosophy of Science IX
D. Prawitz, B. Skyrms and D. Westerståhl (Editors)

A NEW PARADOX IN TYPE THEORY

THIERRY COQUAND

Department of Computer Science, University of Göteborg/Chalmers
S-412 96 Göteborg, Sweden
coquand@cs.chalmers.se

Introduction

The aim of this paper is to present a new paradox for Type Theory, which is a type-theoretic refinement of Reynolds' result [24] that there is no set-theoretic model of polymorphism. We discuss then one application of this paradox, which shows unexpected connections between the principle of excluded middle and the axiom of description in impredicative Type Theories.

1. Minimal and polymorphic higher-order logic

1.1. Minimal higher-order logic

1.1.1. A presentation of the system

We assume known simply typed lambda calculus (see for instance [3].) The lambda-terms will always be considered up to β-conversion. The *types* of minimal higher-order logic consist of one basic type o and function types of the form $\alpha \to \beta$. The *terms* of minimal higher-order logic are those of simply typed-lambda terms - constants, variables, abstractions - with the usual type constraints. Write $a : \tau$ to mean "a is of type τ." The only constants are the constant $\Rightarrow$ of type $o \to o \to o$, and for each type τ, the constant $\forall_\tau : (\tau \to o) \to o$. A *proposition* is a term of type o.

We write $\phi \Rightarrow \psi$ for $\Rightarrow (\phi, \psi)$, and $\forall x^\alpha.\phi$ for $\forall_\alpha(\lambda x.\phi)$. The application of a to the successive arguments $b_1, \ldots, b_n$ is written $a(b_1, \ldots, b_n)$. The notation $[a/x]b$ stands for the substitution of the term a for the variable x in b.

We define inductively when a proposition ϕ is *entailed* by a finite set Γ of propositions, notation $\Gamma \vdash \phi$. A proposition is *provable* or *true* iff it is

entailed by the empty set of propositions. This is given by the rules.

$$\frac{\phi \in \Gamma}{\Gamma \vdash \phi} \qquad (HYP)$$

$$\frac{\Gamma \cup \{\phi\} \vdash \psi}{\Gamma \vdash \phi \Rightarrow \psi} \qquad (ABS)$$

$$\frac{\Gamma \vdash \phi \Rightarrow \psi \quad \Gamma \vdash \phi}{\Gamma \vdash \psi} \qquad (MP)$$

$$\frac{\Gamma \vdash \forall x^{\alpha}.\phi}{\Gamma \vdash [x/t]\phi} \qquad (INST)$$

$$\frac{\Gamma \vdash \phi}{\Gamma \vdash \forall x^{\alpha}.\phi} \qquad (GEN)$$

In the rule $(INST)$, t is a term of type α, and in the rule (GEN), it has to be assumed that x^{α} does not appear free in any proposition of Γ.

1.1.2. Definition of other logical connectives

It is possible to define other logical connectives. This fact was in essence already known to Russell [25], at least for negation and conjunction.

$$\begin{array}{lcll}
\bot & = & \forall \phi^{o}.\phi & : o \\
\neg & = & \lambda \phi^{o}.\phi \Rightarrow \bot & : o \rightarrow o \\
\top & = & \forall \phi^{o}.\phi \Rightarrow \phi & : o \\
\wedge & = & \lambda \phi^{o}, \psi^{o}.\forall \delta^{o}.(\phi \Rightarrow \psi \Rightarrow \delta) \Rightarrow \delta & : o \rightarrow o \rightarrow o \\
\vee & = & \lambda \phi^{o}, \psi^{o}.\forall \delta^{o}.(\phi \Rightarrow \delta) \Rightarrow (\psi \Rightarrow \delta) \Rightarrow \delta & : o \rightarrow o \rightarrow o \\
\exists_{\tau} & = & \lambda P^{\tau \rightarrow o}.\forall \delta^{o}.(\forall x^{\tau}.P(x) \Rightarrow \delta) \Rightarrow \delta & : (\tau \rightarrow o) \rightarrow o
\end{array}$$

Often, we shall not write explicitely the type of a bound variable when it can be infered. For instance, the definition of $\exists_{\tau}$ will also be written $\lambda P.\forall \delta.(\forall x.P(x) \Rightarrow \delta) \Rightarrow \delta : (\tau \rightarrow o) \rightarrow o$.

1.1.3. Church's higher-order logic

The original logic of Church [3] was formulated for classical logic and had a ground type of individuals. Yet another difference was the introduction of a description operator (another version contains an extensionality axiom and the axiom of choice). It is possible to interpret classical higher-order propositional logic in minimal higher-order logic (see [11]).

1.1.4. Semantics

Minimal higher-order logic has a direct set-theoretic semantics. Each type denotes a finite set: the type o a set with two elements $\{\mathrm{T}, \mathrm{F}\}$, and the function type operator is interpreted as set-theoretic exponentiation. The constants $\Rightarrow$, $\forall_\tau$ are then interpreted following the usual truth-table laws of boolean logic. By induction, it is seen that a provable proposition gets the value T under this semantics. This insures the *consistency* of minimal higher-order logic, that is, there are propositions that are not provable. For instance, the proposition $\bot = \forall \phi^o.\phi$ gets the value F, and hence cannot get a proof.

Such a semantics is not faithful to the intuitionistic character of minimal higher-order logic. Topos theory provides various intuitionistic interpretations [13], which fail however to reflect the definitional equality on propositions.

1.2. Polymorphic higher-order logic

1.2.1. Second-order lambda calculus

Second-order lambda-calculus has been introduced independently by Girard [10] and Reynolds [22]. One motivation is to provide a syntax for polymorphic (or generic, or uniform) procedure. Typically, the identity operation is of type $\alpha \to \alpha$, where α is arbitrary, and such an operation behaves "uniformly" in α. It is quite difficult however to describe precisely this notion of uniformity, as it is shown by the paradox we will present.

The types of second-order calculus are either type variables, written α, $\beta, \ldots$, or function types $\sigma \to \tau$, or product types $\Pi\alpha.\tau$. For instance, the type of the polymorphic identity is $\Pi\alpha.\alpha \to \alpha$.

A *closed* type is a type without free type variables.

The syntax of the terms of second-order lambda calculus is the one of simply typed lambda-calculus, extended with *type instantiation* $a\{\tau\} : [\alpha/\tau]\sigma$, where a is a term of type $\Pi\alpha.\sigma$, and *type abstraction* $\Lambda\alpha.a : \Pi\alpha.\sigma$, where a is a term of type σ. For this rule, the type variable α should not appear free in the type of the term variables of a.

For instance, the polymorphic identity $\mathsf{id} = \Lambda\alpha.\lambda x^\alpha.x$ is a term of type $\mathsf{A} = \Pi\alpha.\alpha \to \alpha$. Notice that it is possible to intantiate id on its own type $\mathsf{id}\{\mathsf{A}\} : \mathsf{A} \to \mathsf{A}$, and to apply the result to id, getting $\mathsf{id}\{\mathsf{A}\}(\mathsf{id}) : \mathsf{A}$.

The β-conversion of typed lambda-calculus is extended with *type β-conversion*

$$(\Lambda\alpha.a)\{\tau\} = [\alpha/\tau]a : [\alpha/\tau]\sigma.$$

For instance, the term $\mathsf{id}\{\mathsf{A}\}$ is convertible to the term $\lambda x^{\mathsf{A}}.x$, and hence id is convertible to $\mathsf{id}\{\mathsf{A}\}(\mathsf{id})$.

1.2.2. A presentation of the system

We consider second-order lambda calculus with one ground type o, one constant $\Rightarrow : o \to o \to o$, and, for each closed type τ, one constant $\forall_\tau : (\tau \to o) \to o$. Terms are always considered up to conversion. As above, we write $\phi \Rightarrow \psi$ for $\Rightarrow(\phi, \psi)$, and $\forall x^\alpha.\phi$ for $\forall_\alpha(\lambda x.\phi)$. A *proposition* is a term of type o.

We define exactly as in minimal higher-order logic when a proposition is entailed from a finite set of proposition, with the inductive clauses (HYP), (ABS), (MP), $(INST)$ and (GEN), and when a proposition is provable. We get an extension of minimal higher-order logic, called *polymorphic higher-order logic.*

1.2.3. An example of a derivation

Here is a simple example that shows the expressive power of polymorphic higher-order logic. We define:

$$\begin{array}{lcll}
\mathsf{N} & = & \Pi\alpha.\alpha \to (\alpha \to \alpha) \to \alpha & \\
\mathsf{O} & = & \Lambda\alpha.\lambda x^\alpha.\lambda f^{\alpha\to\alpha}.x & : \mathsf{N} \\
\mathsf{S} & = & \lambda n^{\mathsf{N}}.\Lambda\alpha.\lambda x^\alpha.\lambda f^{\alpha\to\alpha}.f(n\{\alpha\}(x,f)) & : \mathsf{N} \to \mathsf{N} \\
\mathsf{E}_\tau & = & \lambda x^\tau, y^\tau.\forall P^{\tau\to o}.P(y) \Rightarrow P(x) & : \tau \to \tau \to o
\end{array}$$

The term E_τ is called *Leibniz's equality* over the closed type τ. The propositions expressing that E_τ is an equivalence relation over τ are directly provable [26]. Notice next that if we define

$$\mathsf{P} = \lambda n^{\mathsf{N}}.n\{o\}(\bot, \lambda\phi^o.\mathsf{T}) : \mathsf{N} \to o$$

then we have the conversion

$$\begin{array}{lcll}
\mathsf{P}(\mathsf{O}) & = & \bot & : o \\
\mathsf{P}(\mathsf{S}(x)) & = & \mathsf{T} & : o
\end{array}$$

and, from this, it follows that the proposition $\forall x^{\mathsf{N}}.\neg \mathsf{E}_{\mathsf{N}}(\mathsf{O}, \mathsf{S}(x))$ is provable. This proposition expresses the fourth Peano axiom for arithmetic.

2. A type-theoretic refinement of Reynolds' theorem

2.1. A heuristic presentation of Reynolds' theorem

Reynolds' theorem [24, 21] states that there is no set-theoretic model of second-order lambda-calculus. We do not need here to detail the notion of "set-theoretic model" required in order to make this statement precise. But we will however give some comments in order to motivate the argument of the next section.

Since there is no set of all sets, there is a problem in interpreting set-theoretically second-order lambda-calculus. However, in [23], Reynolds conjectured that there is a nontrivial set-theoretic model where the function operator is interpreted as set-theoretic exponentiation. The idea was that, in interpreting a product of a family of sets (A_X), indexed over all sets, we consider only families $a_X \in A_X$ that are "uniform", with a strong enough notion of uniformity so that the collection of uniform families is small enough to be considered as a set.

Let us motivate this conjecture by some concrete examples. For the type $\mathsf{A} = \Pi\alpha.\alpha \to \alpha$, the idea is to consider only "parametric" families (t_X), with $t_X \in X^X$. One definition of parametricity expresses the notion of "representation independence" (cf. [23]): for all sets X and Y, if $R \subseteq X \times Y$, and $R(x, y)$ then we shall have $R(t_X(x), t_Y(y))$. It is then the case that there is only one "parametric" family, which corresponds to the polymorphic identity. Indeed, given a set Y, and $y \in Y$, we can always take for X the singleton set $\{0\}$, and R the relation holding only between 0 and y. If (t_X) is parametric, we shall have $R(0, t_Y(y))$ which implies $t_Y(y) = y$. In this way, the type $\Pi\alpha.\alpha \to \alpha$ gets interpreted by a singleton.

For the type $\mathsf{N} = \Pi\alpha.\alpha \to (\alpha \to \alpha) \to \alpha$, the condition of uniformity of a family (t_X) becomes: for all sets X and Y, if $R \subseteq X \times Y$, if $R(a, b)$ and for all $x \in X$, $y \in Y$, $R(x, y)$ implies $R(f(x), g(y))$, then we shall have $R(t_X(a, f), t_Y(b, g))$. It is then the case that if (t_X) is parametric, there exists a fixed integer n_0 such that $t_X(x, f) = f^{n_0}(x)$ for all set X, $x \in X$ and $f \in X^X$. This integer n_0 is $t_\omega(0, S)$ where S is the successor function. Indeed, given a set Y, and $b \in Y$, $g \in Y^Y$, we let $R \subseteq \omega \times Y$ be the relation holding between n and $y \in Y$ only if $y = g^n(b)$. If (t_X) is parametric, we shall have $R(n_0, t_Y(b, g))$ which implies $t_Y(b, g) = g^{n_0}(y)$. In this way, N gets interpreted essentially by ω.

Such an argument is directly generalised to any type of second-order lambda-calculus determined by any algebraic signature: the type gets interpreted essentially by the initial algebra of this signature. This is shown in [23].

It is then natural to look for the case of a signature that has no set-theoretic initial algebra, and the simplest example is the signature with only one constructor that maps elements of $B^{(B^X)}$ into elements of X, where B is a fixed set. This leads to the consideration of parametric families for the type $\Pi\alpha.(((\alpha\to\tau)\to\tau)\to\alpha)\to\alpha$, where τ is a fixed type, and to Reynolds' proof in [24].

2.2. A type-theoretic formulation

The intuitive arguments of the previous section cannot be formulated in polymorphic higher-order logic. Indeed, the uniformity condition involves in general a quantification over all sets, and we have no quantification over type variables in polymorphic higher-order logic. Instead, we will express it as a kind of "induction principle" over a given type.

We first consider the following expressions. Given a type expression α, we let $\Phi(\alpha)$ be $(\alpha\to o)\to o$.

$$
\begin{array}{lcll}
\mathsf{A}_0 & = & \Pi\alpha.(\Phi(\alpha)\to\alpha)\to\alpha & \\
\phi & = & \Lambda\alpha,\beta.\lambda f.\lambda z.\lambda u.z(\lambda x.u(f(x))) & : \Pi\alpha,\beta.(\alpha\to\beta)\to\Phi(\alpha)\to\Phi(\beta) \\
\mathsf{iter} & = & \Lambda\alpha.\lambda f.\lambda u.u\{\alpha\}(f) & : \Pi\alpha.(\Phi(\alpha)\to\alpha)\to\mathsf{A}_0\to\alpha \\
\mathsf{intro} & = & \lambda z.\Lambda\alpha.\lambda f.f(\phi\{\mathsf{A}_0,\alpha\}(\mathsf{iter}\{\alpha\}(f),z)) & : \Phi(\mathsf{A}_0)\to\mathsf{A}_0 \\
\mathsf{match} & = & \mathsf{iter}\{\Phi(\mathsf{A}_0)\}(\phi\{\Phi(\mathsf{A}_0),\mathsf{A}_0\}(\mathsf{intro})) & : \mathsf{A}_0\to\Phi(\mathsf{A}_0)
\end{array}
$$

All these definitions can be done in second-order lambda-calculus. The term ϕ expresses that the map $\alpha\longmapsto\Phi(\alpha)$ can be seen as a "functor", and the term iter expresses some kind of "weak initiality" of A_0 w.r.t. this "functor". This corresponds to the functor $T(X)=2^{(2^X)}$ in set theory, and we are going to build in polymorphic higher-order logic what would be an initial T-algebra (see [21]).

If α is a type, we write $\mathsf{Rel}(\alpha)$ the type $\alpha\to\alpha\to o$. If $E:\mathsf{Rel}(\alpha)$, we say that E is a *relation* on α. Let us say that a relation is a *partial equivalence relation* iff it is provably symmetric and transitive.

If we have $f:\alpha\to\beta$, E relation on α and F relation on β, let us write $\mathsf{morphism}(E,F,f)$ the proposition $\forall x,y^{\alpha}.E(x,y)\Rightarrow F(f(x),f(y))$. We say that f is a *morphism* between E and F if, and only if, the proposition $\mathsf{morphism}(E,F,f)$ is provable. If furthermore $g:\beta\to\alpha$ is a morphism between F and E, we say that the pair (f,g) is an *isomorphism* between E

and F if, and only if, both propositions $\forall x, y.E(x,y) \Rightarrow E(x, g(f(y)))$ and $\forall x, y.F(x,y) \Rightarrow F(x, f(g(y)))$ are provable.

The next definitions associate to the types o and A_0 a relation that is provably a partial equivalence relation.

$$
\begin{array}{lcll}
\equiv & = & \lambda\phi, \psi.(\phi \Rightarrow \psi) \wedge (\psi \Rightarrow \phi) & : \mathsf{Rel}(o) \\
\mathsf{sym} & = & \Lambda\alpha.\lambda E.\forall x, y.E(x,y) \Rightarrow E(y,x) & : \Pi\alpha.\mathsf{Rel}(\alpha) \rightarrow o \\
\mathsf{trans} & = & \Lambda\alpha.\lambda E.\forall x, y, z.E(x,y) \Rightarrow E(y,z) \Rightarrow E(x,z) & : \Pi\alpha.\mathsf{Rel}(\alpha) \rightarrow o \\
\mathsf{per} & = & \Lambda\alpha.\lambda E.\mathsf{sym}\{\alpha\}(E) \wedge \mathsf{trans}\{\alpha\}(E) & : \Pi\alpha.\mathsf{Rel}(\alpha) \rightarrow o \\
\mathsf{power} & = & \Lambda\alpha.\lambda E.\lambda f, g.\forall x, y.E(x,y) \Rightarrow f(x) \equiv g(y) & \\
 & & & : \Pi\alpha.\mathsf{Rel}(\alpha) \rightarrow \mathsf{Rel}(\alpha \rightarrow o) \\
\phi_2 & = & \Lambda\alpha.\lambda E.\mathsf{power}\{\alpha \rightarrow o\}(\mathsf{power}\{\alpha\}(E)) & : \Pi\alpha.\mathsf{Rel}(\alpha) \rightarrow \mathsf{Rel}(\Phi(\alpha))
\end{array}
$$

The term ϕ_2 extends the action of $\alpha \longmapsto \Phi(\alpha)$ to relations over types. The term $\mathsf{per}\{\alpha\}(E)$ represents the fact that E is a relation symmetric and transitive on the type α.

It is direct to show:

LEMMA: *if E is a partial equivalence on the type τ, then $\mathsf{power}\{\tau\}(E)$ is a partial equivalence relation on the type $\tau \rightarrow o$.*

We introduce next a term that represents the intersection of all relations E on A_0 that are partial equivalence relations and such that intro is a morphism between $\phi_2\{\mathsf{A}_0\}(E)$ and E.

$$
\mathsf{E}_0 = \lambda x, y.\forall E.\mathsf{per}\{\mathsf{A}_0\}(E) \Rightarrow \mathsf{morphism}(\phi_2\{\mathsf{A}_0\}(E), E, \mathsf{intro}) \Rightarrow E(x,y) : \mathsf{Rel}(\mathsf{A}_0)
$$

Since these two properties of a relation on A_0 are closed under intersection, we have:

LEMMA: *the relation E_0 is a partial equivalence relation on A_0, and intro is a morphism between $\phi_2\{\mathsf{A}_0\}(E_0)$ and E_0.*

The relation E_0 can be seen as a construction in polymorphic higher-order logic of the initial T-algebra of the functor $T(X) = 2^{(2^X)}$.

LEMMA: *the term match is a morphism between $\phi_2\{\mathsf{A}_0\}(E_0)$ and E_0; furthermore $(\mathsf{intro}, \mathsf{match})$ is an isomorphism between E_0 and $\phi_2\{\mathsf{A}_0\}(E_0)$.*

For this, we essentially follow the usual argument that the morphism parts of initial T-algebra are isomorphisms (see [21] and the references given there).

THEOREM: *Polymorphic higher-order logic is inconsistent.*

That is, all propositions are provable, or alternatively, $\bot$ is provable. This follows directly from the lemmas, and the usual intuitionistic proof of Cantor's theorem, that there cannot be onto maps from a set to its power set (see for instance [19]).

This argument has been checked and found using a computer, and the formal proof is presented in [7].

2.3. Connection with Girard's paradox

In [11], Girard considers essentially the extension of polymorphic higher-order logic with quantification over type variables (called "system U") and proves that a form of Burali-Forti paradox holds for this extension. The question of the consistency of polymorphic higher-order logic (called "system U^-") is then raised and left open. The theorem above solves this question negatively.

Reynolds' argument, as it is presented in [24] can be directly formulated in the system U, but not in polymorphic higher-order logic, because the notion of parametricity used there is defined with a quantification over set variables. The idea of replacing this quantification by an "induction principle" appears also, independently, in the framework of topos theory in a paper of A. Pitts [19].

In [5], a slight simplification of Girard's argument is presented. We have not been able however to formulate a "Burali-Forti" like paradox in polymorphic higher-order logic, that is, we have not seen if it was possible to avoid the quantification over type variables used in [11, 5].

3. Application to impredicative type theory

3.1. Impredicative type theory

Impredicative type theory has been introduced in [4] and is analysed in [6]. We will not present in detail this type theory, but limit ourselves to a short description.

Impredicative type theory is a direct expression of the principle of "propositions-as-types" and "proof-as-objects" for minimal higher-order logic. In order to stress this aspect, we represent by Set the type of propositions, that are now thought of as intuitionistic sets (the set of their proofs). The objets of type Set are themselves considered as types. We let a "small type", or "set" be a type that is also an object of type Set. The basic operation is the dependent product, written $(\Pi x : A)B(x)$ of a dependent family of types $B(x)$ $(x : A)$ over a type A. The basic feature of impredicative Type

Theory is that small types are closed by product. If $B(x) : \mathsf{Set}\ (x : A)$, then $(\Pi x : A)B(x) : \mathsf{Set}$. The theorem of Reynolds shows that it is impossible to think of the present sets as sets in the sense of Zermelo-Skolem-Fraenkel.

This basic feature is the main difference with Martin-Löf's logical framework, as presented in [18]. Otherwise, these systems are quite similar. In particular, a fundamental role is played by the notion of *context*, which is a finite set of typed variables declaration. This notion is also a basic notion of Automath, and we refer to the article [2] for an intuitive description of contexts.

If A and B are types, we let $A \rightarrow B$ be the product of the constant type family B over the type A. In the case where A and B are small types or sets, we write it also $A \Rightarrow B$. Minimal higher-order logic has a direct interpretation in impredicative type theory: o gets interpreted by Set, and a proposition gets interpreted by a small type, which represents the type of its proofs. For instance, the proposition $\mathsf{T} = \forall\phi.\phi \Rightarrow \phi$ is interpreted by $(\Pi X : \mathsf{Set})X \Rightarrow X$. The rules of inference (HYP), (ABS), (MP), $(INST)$ and (GEN) are then a consequence of the general principle that a proposition is true if, and only if, its corresponding type of proofs is inhabited. For instance, the usual proof of T is the polymorphic identity $(\lambda X : \mathsf{Set})(\lambda x : X)x$ over intuitionistic sets. We will use the same notations for logical connectives introduced in minimal higher-order logic, suitably reinterpreted in the framework of impredicative Type Theory.

The "truth table" semantics of minimal higher-order logic described above is directly extended to a model of impredicative Type Theory where a type is interpreted as a finite set, and a small type as a set that has at most one element. Let this model be the *proof irrelevance model*, so called because it forgets proof objects. This terminology is inspired by [2].

3.2. Definite descriptions and excluded middle

3.2.1. Proof irrelevance

The principle of *proof irrelevance* is

$$(\Pi A : \mathsf{Set})(\Pi x, y : A)\mathsf{E}_A(x, y).$$

It states that any set (or intuitionistic proposition) has at most one element w.r.t. Leibniz's equality. Since Leibniz's equality is the weakest possible notion of equality, in the sense that if E is an equivalence relation on A, then $\mathsf{E}_A(x, y)$ implies $E(x, y)$, the principle of proof irrelevance implies that any set has at most one element w.r.t. any notion of equality over this set.

3.2.2. The principle of definite description

Let A be a set, and $\phi(x) : \mathsf{Set}$ $(x : A)$. As in minimal higher-order logic, we let $(\exists x : A)\phi(x)$ be $(\Pi X : \mathsf{Set})[(\Pi x : A)[\phi(x) \Rightarrow X]] \Rightarrow X$ and we let $(\exists! x : A)\ \phi(x) : \mathsf{Set}$ be the term

$$(\exists x : A)[\phi(x) \wedge (\Pi y : A)[\phi(y) \Rightarrow \mathsf{E}_A(x, y)]].$$

This expresses that there exists one and only one element satisfying ϕ, where the equality on A is Leibniz's equality. The principle of *definite descriptions* is

$$(\Pi A, B : \mathsf{Set})(\Pi R : A \to B \to \mathsf{Set})[(\Pi x : A)(\exists! y : B)R(x, y)] \\ \Rightarrow [(\exists f : A \to B)(\Pi x : A)R(x, f(x))]$$

This principle appears in the system of Church [3], in the form of a description operator ι. The motivation comes from Russell's work on denoting (see [27, 26]).

3.2.3. Excluded middle

The last principle we shall consider is the principle of *excluded middle.*

$$(\Pi A : \mathsf{Set})A \vee \neg(A).$$

The extension of Martin-Löf's set theory with this principle has been considered by J. Smith in [29]. It is direct to check that this principle is equivalent to

$$(\Pi A : \mathsf{Set})\neg(\neg(A)) \Rightarrow A.$$

3.3. An application

We can now state the application of the inconsistency of polymorphic higher-order logic.

LEMMA: *The set*

$$(\exists f : o \to o \to o)(\Pi x, y : \mathsf{B})\ T(f(x, y)) \equiv [T(x) \Rightarrow T(y)] \qquad (IMP).$$

and, for each set A, the set

$$(\exists f : (A \to \mathsf{B}) \to \mathsf{B})(\Pi P : A \to \mathsf{B})\ T(f(P)) \equiv [(\Pi x : A)T(P(x))]\ \ (UNIV_A).$$

are inhabited.

PROOF: We show only how to build a proof of (IMP); the case of $(UNIV_A)$ can be solved in a similar way.

If $x : \mathsf{B}$, we let $T(x) : \mathsf{Set}$ be $\mathsf{E}_{\mathsf{B}}(x, \mathsf{true})$, $F(x) : \mathsf{Set}$ be $\mathsf{E}_{\mathsf{B}}(x, \mathsf{false})$, and $B(x) : \mathsf{Set}$ be $T(x) \vee F(x)$.

Notice that if we have $B(z_1)$, $B(z_2)$ and $T(z_1) \equiv T(z_2)$, then we have also $\mathsf{E}_{\mathsf{B}}(z_1, z_2)$. Indeed, the axiom $\neg(\mathsf{E}_{\mathsf{B}}(\mathsf{true}, \mathsf{false}))$ rules out the case $T(z_1)$, $F(z_2)$ and the case $F(z_1)$, $T(z_2)$. If $T(z_1)$ and $T(z_2)$, then $\mathsf{E}_{\mathsf{B}}(z_1, z_2)$, because Leibniz's equality is symmetric and transitive. Similarly, if $F(z_1)$ and $F(z_2)$, then $\mathsf{E}_{\mathsf{B}}(z_1, z_2)$.

This can be expressed in intuitive terms as the fact that the operator $T(z)$ $(z : \mathsf{B})$ is "one-to-one" on elements of B that satisfy the predicate B.

For getting a proof of (IMP), we build a proof of a stronger statement

$$(\exists f : o \to o \to o)(\Pi x, y : \mathsf{B})\ B(f(x, y)) \wedge [T(f(x, y)) \equiv [T(x) \Rightarrow T(y)]].$$

This follows from the principle of definite description and

$$(\Pi x, y : \mathsf{B})(\exists! z : \mathsf{B})\ B(z) \wedge [T(z) \equiv [T(x) \Rightarrow T(y)]].$$

This is a direct consequence of the axiom $\neg(\mathsf{E}_{\mathsf{B}}(\mathsf{true}, \mathsf{false}))$, and of the principle of excluded middle. Indeed, by the principle of excluded middle, we have $T(x) \Rightarrow T(y)$ or $\neg(T(x) \Rightarrow T(y))$. In the first case, we can choose $z = \mathsf{true}$, and in the second case $z = \mathsf{false}$. Furthermore, we have seen that the axiom $\neg(\mathsf{E}_{\mathsf{B}}(\mathsf{true}, \mathsf{false}))$ implies that the operator $T(z)$ $(z : \mathsf{B})$ is one-to-one on elements of type B that satisfies the predicate B. □

THEOREM: *In impredicative type theory extended with excluded middle, the principle of definite description implies the principle of proof irrelevance.*

PROOF: We place ourselves in the context

$$\Gamma = \mathsf{B} : \mathsf{Set}, \mathsf{true} : \mathsf{B}, \mathsf{false} : \mathsf{B}, h : \neg(\mathsf{E}_{\mathsf{B}}(\mathsf{true}, \mathsf{false})),$$

and we build a proof of $\bot$ in this context.

It will then follow from the principle of excluded middle that $\mathsf{E}_{\mathsf{B}}(\mathsf{true}, \mathsf{false})$ is derivable in the context

$$\mathsf{B} : \mathsf{Set}, \mathsf{true} : \mathsf{B}, \mathsf{false} : \mathsf{B}.$$

Hence, the principle of proof irrelevance is derivable in the empty context.

First, we give a way to interpret each closed type of polymorphic higher-order logic by a set of impredicative Type Theory. We interpret the type of propositions o by the set B, and in general a type of polymorphic higher-order logic will be interpreted as a set, interpreting the function type operator as exponentiation on sets and the product over type variables as the product over set variables. For instance, the type $\Pi\alpha.\alpha \to \alpha$ is interpreted as the set $(\Pi X : \mathsf{Set})X \to X$.

Next, we consider a fixed derivation of the absurd proposition in polymorphic higher-order logic. In this derivation, we have used only a finite number of universal quantification over a finite number of closed types. Let $A_1, \ldots, A_n : \mathsf{Set}$ be an enumeration of the translation of those types in impredicative Type Theory. Consider then the context Γ extended by

$$f_0 : \mathsf{B} \to \mathsf{B} \to \mathsf{B}, h_0 : (\Pi x, y : \mathsf{B})\ T(f_0(x, y)) \equiv [T(x) \Rightarrow T(y)],$$

and for each set A_i,

$$f_i : (A_i \to \mathsf{B}) \to \mathsf{B}, h_i : (\Pi P : A_i \to \mathsf{B})\ T(f_i(P)) \equiv [(\Pi x : A_i)T(P(x))].$$

In this extended context Δ, We can translate the given proof of the absurd proposition into a construction of a term of type $\bot$. For this, we interpret $\Rightarrow$ as f_0, and each universal quantification by one of the term f_i.

By this way, we get a construction of type $\bot$ in the extended context Δ.

Using the lemma, we get a proof of (IMP), and of $(UNIV_{A_1})$, $\ldots$,$(UNIV_{A_n})$. This allows us to transform this derivation of $\bot$ in the extented context Δ into a derivation of $\bot$ in the context Γ. □

4. Related results and problems

4.1. Looping combinators

The inconsistency of polymorphic higher-order logic, or even of the system U of [11], entails, by direct translation, the existence of a non normalisable term in a type system with a type of all types (see [16, 5, 12]). The existence of a fixed-point combinator in such a type system is an open problem since [16]. The article [12] contains a proof, using computers in an essential way, that shows the existence of a family of looping combinators, that is, a family of terms $Y_n : (X : \mathsf{Type})X \to X$ such that $Y_n(X, f) = f(Y_{n+1}(X, f))$ $(X : \mathsf{Type}, f : X \to X)$. The fact that we get a family of looping combinators, and not a fixed point combinator seems to be closely connected to the well-known "mismatch" in the representation of destructors for recursively defined types

represented in second-order lambda-calculus (as presented for instance in [17]). But the author has not been able to make this connection precise.

The existence of a family of looping combinators entails the undecidability of type checking for a type system with a type of all types. In [9], the existence of a family of looping combinators is derived from A-translation in polymorphic higher-order logic.

In [8], it is shown that it is possible to build such a fixed-point operator in the presence of a the well-founded type operator of Martin-Löf [15].

4.2. Strong existence

The results about excluded middle in impredicative theory were first expressed as consequence of the inconsistency of the system U of Girard, which extends polymorphic higher-order logic with quantification over type variables [6]. It was then shown that it is possible to interpret system U in the context

$$\mathsf{B} : \mathsf{Set},\ E : \mathsf{B} \to \mathsf{Set},\ \epsilon : \mathsf{Set} \to \mathsf{B},\ H : (X : \mathsf{Set})\ X \equiv E(\epsilon(X)).$$

Hence, it is possible to derive $\bot$ in this context. The author does not know any "direct" derivation of $\bot$ in this context.

A consequence of this is the fact that, in presence of a strong existence operator [15] added to impredicative Type Theory, the principle of excluded middle implies the principle of proof irrelevance. A different proof, somewhat more direct and based in a different idea than Reynolds', has been given by S. Berardi, and checked in the proof checker LEGO of R. Pollack.

The present result, which concerns the principle of definite descriptions, generalises and was motivated by a result of G. Pottinger [20].

4.3. Consistency and independence results

S. Berardi has shown by a model theoretic argument that the axiom of description, and hence the axiom of choice, is not provable in impredicative Type Theory (personal communication.) A "syntactic" version of this model is described in [1]. It is similar to the proof irrelevance model, but the inhabited sets are interpreted instead by the set of all untyped lambda terms. This also models the principle of excluded middle, but not the principle of proof irrelevance. It shows that the principle of proof irrelevance is independent of excluded middle.

In [28], a purely proof theoretic argument shows the consistency of a context implying classical arithmetic, where the set of integers is interpreted as a small type.

4.4. Related results in category theory

The results about excluded middle seem to have some connections with the two following results in category theory. Both are described in Lambek and Scott's book on categorical logic [13].

The first one is Diaconescu's theorem, that in a topos, the axiom of choice implies the principle of excluded middle. The analysis of the proof given in [13] reveals an essential use of the extensional equality, and this result does not seem to be easily interpretable in Type Theory, where the equality between propositions is definitional [18].

The second one is Joyal's result, that says that any "boolean category" is trivial (see [13], page 67). In this case also, this result does not seem to be easily interpretable in Type Theory, because the equality on proofs is definitional [18]. For instance, it is not the case in general that a set $\bot \to A$ has only one element w.r.t. definitional equality, but the fact that it has only one element for the extensional equality is used in an essential way in the proof presented in [13].

Conclusion

We hope to have shown that the study of paradoxes in Type Theory is a rich topic. Quite characteristic is the use of computers in the process of checking, and analysing such paradoxes [1, 5, 12]. However, the feeling of the author is that we have only superficially yet explored this question, and that a more basic understanding of the nature of paradoxes connected to impredicativity is still missing. Nontrivial results in this direction may bring a new light on the status of the "reducibility axiom", a question raised almost one century ago [26], and left essentially open since then.

Acknowledgement

The inconsistency of polymorphic higher-order logic was suggested first by a discussion with Ch. Paulin-Mohring about her representation of inductive types in impredicative Type Theory (see [17]), and then by a discussion with H. Barendregt about the representation of Girard's paradox in polymorphic higher-order logic. Studying the status of excluded middle in impredicative Type Theory was suggested by a discussion with S. Yoccoz. S. Berardi and G. Pottinger provided helpful comments on a draft version of this paper. I want also to thank G. Huet, who urged me to analyse paradoxes in type theories.

References

[1] BERARDI S. *Type dependence and constructive mathematics.* Ph. D. thesis, Dipartimento Matematica, Università di Torino, 1990.

[2] DE BRUIJN N.G. *A survey of the project Automath.* In: To H.B. Curry: Essays in combinatory logic, lambda calculus and formalism, ed. J.P. Seldin and J.R. Hindley, Academic Press 1980, pp. 579 - 606.

[3] CHURCH A. *A formulation of the simple theory of types.* Journal of Symbolic Logic (1940), 56–68.

[4] COQUAND TH. AND HUET G. *Constructions: A Higher Order Proof System for Mechanizing Mathematics.* EUROCAL85, Linz, Springer-Verlag LNCS 203 (1985).

[5] COQUAND TH. *An Analysis of Girard's Paradox.* Proceedings of the first Logic in Computer Science, Boston, 1986, pp. 227 - 236.

[6] COQUAND TH. *Metamathematical Investigations of a Calculus of Constructions.* Logic and Computer Science, P. Odifreddi editor, Academic Press 1990.

[7] COQUAND TH. *Reynolds' paradox, with the Type:Type axiom.* in The Calculus of Constructions, Documentation and user's guide, G. Huet ed., projet FORMEL, 1989.

[8] COQUAND TH. *The paradox of trees in Type Theory.* to appear in B.I.T., 1991.

[9] COQUAND TH. and HERBELIN H. *An application of A-translation to the existence of families of looping combinators.* Submitted to the Journal of Functional Programing, 1991.

[10] GIRARD J.Y. *Une extension de l'interprétation de Gödel à l'analyse, et son application à l'élimination des coupures dans l'analyse et dans la théorie des types.* Proc. 2nd. Scand. Log. Symp., North Holland, 1971.

[11] GIRARD J.Y. *Interpretation fonctionnelle et elimination des coupures de l'arithmetique d'ordre superieur.* These d'Etat, Paris VII (1972).

[12] HOWE D.J. *The Computational Behaviour of Girard's Paradox.* Proceedings of the 2nd Logic in Computer Science 87, Ithaca, 1987.

[13] LAMBEK J. and SCOTT P.J. (1986) *Introduction to Higher Order Categorical Logic.* Cambridge University Press.

[14] MARTIN-LÖF P. *An intuitionistic theory of types: predicative part.* Logic Colloquium, North-Holland (1975).

[15] MARTIN-LÖF P. (1984) *Intuitionistic Type Theory.* Bibliopolis.

[16] MEYER A. R. and REINHOLD M. B. *"type" is not a type.* In Conference record of the thirteenth annual ACM symposium on principles of programming languages, Association for Computing Machinery, SIGACT, SIGPLAN, 1986.

[17] CH. PAULIN-MOHRING. *Inductive definitions in the Calculus of Constructions.* in The Calculus of Constructions, Documentation and user's guide, G. Huet ed., projet FORMEL, 1989.

[18] NORDSTRÖM B., PETERSSON K., SMITH. J. M. (1990), *Programming in Martin-Löf Type Theory.* Oxford Science Publications, Clarendon Press, Oxford.

[19] PITTS A. *Non-trivial Power Types can't be Subtypes of Polymorphic Types.* Proceedings of the fourth Logic in Computer Science, 1989, pp. 6 - 13.

[20] POTTINGER G. *Definite Descriptions and Excluded Middle in the Theory of Constructions.* Communications in the TYPES electronic forum. October 1, 1989.

[21] REYNOLDS J.C. and PLOTKIN G.D. *On Functors Expressible in the Polymorphic Typed Lambda Calculus.* Logical Foundations of Functional Programming, edited by G. Huet, Addison Wesley, 1989.

[22] REYNOLDS J. C. *Towards a Theory of Type Structure.* Programming Symposium, Paris. Springer Verlag LNCS 19 (1974) 408–425.

[23] REYNOLDS J.C. *Types, abstraction and parametric polymorphism.* In: Information Processing 83, edired by R.E.A. Mason. Elsevier Science Publishers B.V. (North-Holland), Amsterdam, 1983, pp. 513 - 523.

[24] REYNOLDS J.C. *Polymorphism is not Set-Theoretic.* Semantics of Data Types, edited by G. Kahn, D.B. MacQueen, and G.D. Plotkin, LNCS 173, Springer-Verlag, 1984, pp. 145 - 156.

[25] RUSSELL B. (1903) *The Principles of Mathematics.* Cambridge University Press.

[26] RUSSELL B. and WHITEHEAD A.N. (1912) *Principia Mathematica.* Volume 1,2,3 Cambridge University Press.

[27] RUSSELL B. *On Denoting.* Mind, 1905. Reprinted in Robert C. Marsh, Logic and Knowledge, George Allen & Unwin, London, 1956, pp. 41-56.

[28] SELDIN J.P. *Excluded Middle without Definite Description in the Theory of Constructions.* MWPLT91 Spring, Proceedings of The First Montreal Workshop on Programming Language Theory, M. Okada and P.J. Scot eds, 1991, pages 74-83.

[29] SMITH J. *On a Nonconstructive Type Theory and Program Derivation.* The Proceedings of Conference on Logic and its Applications, Bulgaria, Plenum Press, 1986.

Logic, Methodology and Philosophy of Science IX
D. Prawitz, B. Skyrms and D. Westerståhl (Editors)

TAKING FORMALISM SERIOUSLY

EDWARD NELSON

Department of Mathematics, Princeton University

Almost all mathematicians believe that the natural numbers exist. How would the world be different if numbers suddenly ceased to exist? Share with me a fantasy: we open our morning newspaper to find a report with a banner headline, "Numbers Vanish!—Early last night the natural numbers, so called because they have always been found in nature, suddenly disappeared. Mathematicians have expressed stunned despair. Without numbers, they say, they can no longer prove theorems. Numbers are the foundation of science and technology, and without them humanity will soon revert to barbar- (*continued on an inside page*)."

This is nonsense. The newspaper could still put marks 2, 3, etc., on the inside pages for ease of reference. Machines could still function as before. And we mathematicians could continue to put marks on paper, just as before, and hopefully submit them to editors of mathematical journals. We do not need the natural numbers.

Some would maintain that the natural numbers exist of necessity and could not disappear. This is a religious belief that I do not share.

Formalism is that view of the foundations of mathematics which maintains that the natural numbers and other mathematical entities do not exist, that mathematics is the manipulation of marks according to specified rules. Formalism is associated with the name of David Hilbert, but Hilbert did not take formalism seriously. For him, it was a tactical device in his skirmishes with the intuitionists, designed to enable mathematicians to continue to dwell in Cantor's paradise.

In this talk I shall outline what I believe are the consequences of taking formalism seriously.

Writing correct proofs

There is a gap between the professions of formalists and mathematical practice. Few mathematical works contain full proofs that obey in detail explic-

itly formulated rules. This has been too tedious, both for the author and the reader. But since the dawn of mathematics in Greece it has been a worthy ideal. Mathematics is a highly social endeavor, and although we may dispute with each other questions of priority or questions of value, through the centuries we have striven—with a remarkable measure of success—for common standards of what constitutes a correct argument. More than the practitioners of any other discipline, mathematicians judge work by commonly held criteria. But it is a frequent occurrence for incorrect results to be published, and it is even more frequent for results to be published with a serious gap in the argument. And when a truly important result is claimed, several mathematicians may spend weeks checking the proof.

With the advent of digital computers, this will soon begin to change, and it will change radically during the lifetime of many of those present. There will be a central data bank of theorems, arranged hierarchically to facilitate search by mathematicians attempting to prove new theorems. Interactive programs will be developed to help us construct fully formalized proofs, and when these are submitted they will be verified and entered into the data bank with the name of the inventor and date of construction.

For my own purposes, I am constructing a proof checking program called *qed*, written in Larry Wall's *perl* language. It is a primitive program, but it has some features that I believe more sophisticated programs will share. I am speaking about programs for mathematical reasoning as distinguished from mathematical computation.

Such a program should have two levels: a very rapid verification program for fully formalized proofs and a hierarchically structured interactive search program.

The verification should exploit duality. Each logical operator should be encoded by a single byte, and there should be a byte for affirmation (the dual of negation) so that the negation of any formula can be achieved simply by translating each logical operator byte into its dual.

Every step of every proof should be an argument by contradiction. Suppose that $A_1, \ldots, A_n$ is a list of formulas that have been proved, or are assumed as hypotheses in a deduction, and we want to deduce A. Then we adjoin $\neg A$ to the list and for each formula on the list we look to see whether it or its negation is a subformula of some other formula on the list; if so, we make the appropriate reduction by the laws of the propositional calculus with equality. This can be done very rapidly. If we find a contradiction, then A has been deduced. But if we do not find a contradiction, our work may not have been wasted: each formula on the reduced list is a consequence of assuming $\neg A$ as a hypothesis, and the argument can be continued.

There are many levels of search involving an exponentially large world

of possibilities. By necessity, such searches are time consuming. The most primitive level of search is for the appropriate terms to substitute for the free variables in a theorem. This is usually routine, as in applying the commutative law $x \cdot y = y \cdot x$ to conclude that $a+b \cdot c = a+c \cdot b$, but sometimes it involves insight, as in applying the Liouville theorem "f is a bounded entire function implies f is a constant" to the function $1/p$ where p is a polynomial without roots. As the years pass, more and more sophisticated searches will be programmed into computers. But good mathematics is the fruit of deep personal and cultural experience. Digital computers and people have different search skills, and a fruitful program will involve an interactive partnership.

Questioning Church's thesis

0 is a numeral; if n is a numeral, then Sn is a numeral. Thus the numerals are 0, S0, SS0, SSS0, and so forth. Numerals are used to count things.

An effectively computable function is a program that for any numerals as arguments terminates in a finite number of steps and yields a numeral as value. This is a somewhat vague concept, but the concept of a recursive function is precisely formalizable; Church's thesis is that the two concepts are equivalent.

For example,

$$
\begin{aligned}
x + 0 &= x, && (1)\\
x + \mathrm{S}y &= \mathrm{S}(x + y), && (2)\\
x \cdot 0 &= 0, && (3)\\
x \cdot \mathrm{S}y &= (x \cdot y) + x, && (4)\\
x^0 &= \mathrm{S}0, && (5)\\
x^{\mathrm{S}y} &= x^y \cdot x, && (6)\\
x \Uparrow 0 &= \mathrm{S}0, && (7)\\
x \Uparrow \mathrm{S}y &= x^{x \Uparrow y} && (8)
\end{aligned}
$$

give constructions of addition, multiplication, exponentiation, and superexponentiation as recursive functions.

No attempt to construct an effectively computable function that is not recursive has been successful; Church's thesis has withstood challenges from above. But we can challenge it from below: how do we know that every recursive function is effectively computable?

Computer scientists agree that the dividing line between functions computable in practice and those not computable in practice lies roughly between

multiplication and exponentiation. But I want to discuss computability in principle, as nearly as I can recall the meaning that I used to think I understood by the phrase "in principle".

The *finitary dogma* asserts that every recursive function is effectively computable; that is, it asserts that for any variable-free term made up from 0 and function symbols representing recursive functions, such as

$$\mathrm{SS0} \Uparrow (\mathrm{SS0} \Uparrow (\mathrm{SSSSS0})),$$

if we apply the rules of construction sufficiently often we end up with a numeral.

Finitists have verified this for simple cases. In other cases they may give up, but they know that if only they persisted long enough the computation would terminate. If asked how long that is, their answer would be: roughly the number of the numeral I am computing.

Spoiled children throw temper tantrums; they have verified in simple cases that this gets them what they want. In other cases they may give up, but they know that if only they persisted long enough they would get what they want. If asked how long that is, their answer would be: until I get what I want.

In saying this, I am casting no aspersions on those who hold opinions different from mine; I am simply expressing my opinion that finitism is a self-validating belief system for which there is no evidence, and which may very well be incorrect.

Let us examine the finitary dogma more closely. In general, the logical terminology and notation used in this talk are those of Shoenfield 1967. Let T be a strong theory, one that formalizes contemporary mathematics. Assume that T is consistent. Let $\hat{\mathrm{T}}$ be the theory obtained by adjoining to T a new unary predicate symbol ϕ and the axioms

$$\phi(0), \tag{9}$$

$$\phi(x) \;\rightarrow\; \phi(\mathrm{S}x). \tag{10}$$

These axioms express the idea of counting, parallel to the metamathematical definition of a numeral, and the intended meaning of $\phi(x)$ is "x is equal to a numeral". Let F be a recursive function, say of two variables, formalized in T. Can we prove in $\hat{\mathrm{T}}$ a theorem expressing the idea that if x and y are numerals, then $\mathrm{F}(x, y)$ is equal to a numeral?

Even for addition, we cannot prove in $\hat{\mathrm{T}}$

$$\phi(x) \;\&\; \phi(y) \rightarrow \phi(x + y). \tag{11}$$

To see this, take a non-standard model of T (i.e., one in which the natural numbers contain a non-standard ν) and interpret $\phi(x)$ as signifying that x

is less than ν plus some standard natural number. Then we have (9), (10), and $\phi(\nu)$ but not $\phi(\nu + \nu)$, so (11) is not provable in $\hat{\mathrm{T}}$.

Let us try another approach. Can we find an inductive formula A of $\hat{\mathrm{T}}$ (i.e., a formula such that A(0) and $\mathrm{A}(x) \to \mathrm{A}(\mathrm{S}x)$, and consequently A(n) for every numeral n, are theorems of $\hat{\mathrm{T}}$) such that

$$\mathrm{A}(x) \;\&\; \mathrm{A}(y) \to \phi(\mathrm{F}(x, y)) \tag{12}$$

is a theorem of $\hat{\mathrm{T}}$? If so, then we can say that we have proved that if x and y are numerals then $\mathrm{F}(x, y)$ is equal to a numeral.

For addition we can let A be the formula $\phi^2(x)$ given by

$$\forall y[\phi(y) \to \phi(y + x)]$$

and for multiplication we can let A be the formula $\phi^3(x)$ given by

$$\forall y[\phi^2(y) \to \phi^2(y \cdot x)];$$

see Chapter 5 of Nelson 1986. But the proofs of (12) for addition and multiplication use the associativity of these operations, and exponentiation is not associative.

In fact, the following metatheorem is proved in Chapter 18 of Nelson 1986 (though not formulated in quite this way):

If F *is the exponential function, there does not exist an inductive formula* A *of* $\hat{\mathrm{T}}$ *such that* (12) *is a theorem of* $\hat{\mathrm{T}}$.

In other words, the only way to prove that the exponential of one numeral by another is equal to a numeral is to beg the question, by postulating for the formalization of the concept "x is equal to a numeral" some property going beyond the original concept of counting. It is interesting that the dividing line for demonstrable computability in principle is the same as that for computability in practice.

This result should give finitists pause. Finitism is the last refuge of the Platonist.

Seeking a contradiction

Robinson's theory Q (see R. M. Robinson 1950) is the theory whose nonlogical axioms are

$$\mathrm{S}x \neq 0, \tag{13}$$

$$\mathrm{S}x = \mathrm{S}y \to x = y, \tag{14}$$

(1)–(4), and

$$x \neq 0 \to \exists y[\mathrm{S}y = x]. \tag{15}$$

This may be reformulated as an open theory Q_0 by introducing a function symbol P (for predecessor) and replacing (15) by

$$x \neq 0 \rightarrow \mathrm{SP}x = x. \tag{16}$$

Robinson's theory is the simplest theory in which one can do nontrivial mathematics. It is essentially Peano Arithmetic without induction.

There are several consistency proofs for Q. The most familiar is the infinitary argument that the natural numbers are a model for Q. But in a world from which numbers have vanished this carries no conviction.

There is also a finitary consistency proof. The first point is that Q_0 is quasi-tautologically consistent. This means that it is impossible to derive a contradiction in the theory without using quantifiers. More precisely, let Q_0' be the formal system with the same language and nonlogical axioms as Q_0 but with no quantifiers and with instance and quasi-tautological consequence as rules of inference; then Q_0' is consistent. There is certainly a finitary demonstration of this. I believe that this is in fact true, and that a demonstration of this assertion can be given that will convince the most skeptical of formalists. The second point is that the consistency of Q follows from this by the Hilbert-Ackermann theorem; see Shoenfield 1967. This is a finitary algorithm for eliminating quantifiers from proofs. But it relies on the finitary dogma; specifically, that superexponentiation is effectively computable. For one who has put aside credence in the finitary dogma, this proof also carries no conviction.

Taking formalism seriously entails regarding the consistency of Q as an open problem. It would be solved if one could derive a contradiction in a theory interpretable in Q. This is what I am working on now.

Seeking a demonstrably consistent mathematics

Who would have believed ten years ago that it is possible, in a theory for which there is a finitary consistency proof, to do modern mathematics? Yet that is what Nelson 1986 and Nelson 1987 together accomplish; see the description of Q* in the last chapter of Nelson 1986. In Nelson 1987, modern ideas of stochastic processes find direct expression, liberated from the heavy weight of Cantorian set theory. This is made possible by using a small portion of Abraham Robinson's nonstandard analysis (Robinson 1974).

It is an open problem to develop a formal system that is demonstrably—without appeal to the finitary dogma—consistent, and yet is powerful enough for modern mathematics. A candidate for such a system is Q* but with only bounded quantifiers permitted.

The phrase "foundations of mathematics" is a static architectural image. I prefer to think of mathematics as a growing tree. Rather than construct foundations, the formalist can study and nourish the roots of this tree. By taking a fresh look at the way good mathematics is actually done, we may find to our surprise that it is demonstrably consistent.

A personal note

The talk I actually gave at the Congress was entitled "A formalism for developing interactive programs". The chairman of the session was Per Martin-Löf. He was quite surprised to hear me present, with enthusiasm and in total ignorance, a primitive version of his own intuitionistic type theory which he had invented at least twelve years previously. My subsequent embarassment was assuaged by Martin-Löf's graciousness, and we agreed that the event had a strongly comical apect. I hope at some point to make a contribution to this beautiful field, but only after I have done my homework by studying the literature.

Bibliography

EDWARD NELSON, 1986 *Predicative Arithmetic,* Mathematical Notes No. 32, Princeton University Press, Princeton, NJ.

EDWARD NELSON, 1987 *Radically Elementary Probability Theory,* Annals of Mathematics Studies 117, Princeton University Press, Princeton, NJ.

ABRAHAM ROBINSON, 1974 *Non-standard Analysis,* Rev. ed., American Elsevier, New York.

R. M. ROBINSON, 1950 *An essentially undecidable axiom system,* Proc. Int. Cong. Math., Cambridge, MA, Vol. I, 729–730.

JOSEPH R. SHOENFIELD, 1967 *Mathematical Logic,* Addison-Wesley, New York.

Logic, Methodology and Philosophy of Science IX
D. Prawitz, B. Skyrms and D. Westerståhl (Editors)

WHAT IS THE PHILOSOPHICAL BASIS OF INTUITIONISTIC MATHEMATICS?

RICHARD TIESZEN

Department of Philosophy, San José State University, San José, USA

How should we understand the philosophical basis of intuitionistic mathematics late in the Twentieth Century, some 25 years after the death of Brouwer? I believe this is an important question because there are insights in intuitionism that are found nowhere else in the philosophy of mathematics, insights that ought to be preserved, clarified and extended. Chief among these is the idea that a proof is a mental construction. The idea that a proof is a mental construction already distinguishes intuitionism from other philosophical views of mathematics like (ontological) platonism, nominalism, and formalism. There are also many philosophical problems that can be raised for intuitionism and I intend to discuss some of them below. I shall first briefly consider some views on the question that are found in the literature on intuitionism. I shall then argue for what I take to be a good working answer to the question, an answer which I think is in the tradition of Brouwer and Heyting but which can be used to clarify their views and to defend some of the key philosophical insights of intuitionism.

1. A brief survey of views on the question

For Brouwer the philosophical basis of intuitionist mathematics was to be found in the concept of intuition. In particular, Brouwer portrayed intuitionism as abandoning Kant's apriority of space but adhering all the more resolutely to Kant's idea of time as an *a priori* form of intuition. Brouwer describes the basic intuition upon which mathematics is founded in a number of places in his writings. In "Intuitionism and Formalism" [Brouwer 1912], for example, he describes it as follows:

> This neo-intuitionism considers the falling apart of moments of life into qualitatively different parts, to be reunited only while re-

> maining separated by time, as the fundamental phenomenon of the human intellect, passing by abstracting from its emotional content into the fundamental phenomenon of mathematical thinking, the intuition of bare two-oneness.

Brouwer often notes the role of memory in retaining earlier life moments while the succession of life moments continues. He says that it is introspectively realized how this basic operation of mathematical construction, this intuition of two-oneness, successively generates the finite ordinal numbers, inasmuch as one of the elements of the two-oneness may be thought of as a new two-oneness, and the process may then be repeated indefinitely. Also important for Brouwer's view of the role of this basic intuition is the claim that (i) what has meaning in mathematics is derived from the basic intuition and (ii) that mathematics is a languageless activity of mind. Many of Brouwer's comments suggest a very strong separation of thought from language. This figures into Brouwer's conception of how the basic intuition of mathematics can provide a foundation for mathematics that is exact and free from error and misunderstanding. In "Weten, willen, and spreken" [Brouwer 1933], for example, Brouwer says that

> ... the languageless constructions originating by the self-unfolding of the primordial intuition are, by virtue of their presence in memory alone, exact and correct; ... the human power of memory, however, which has to survey these constructions, even when it summons the assistance of linguistic signs, by its very nature is limited and fallible. For a human mind equipped with an unlimited memory, a pure mathematics which is practiced in solitude and without the use of linguistic signs would be exact; this exactness, however, would again be lost in an exchange *between* human beings with unlimited memory, since they remain committed to language as a means of communication.

On the basis of comments like these it appears that the certainty that is supposed to be guaranteed by founding intuitionism on intuition is certainty for the *ideal* mathematician only. The actual, practicing mathematician does not have such certainty, nor does there appear to be any intersubjective certainty for Brouwer, since the expression of mathematical ideas needed for communication is always imperfect. The separation of thought from language leads to the charge that Brouwer's notion of intuition and the concept of meaning it supports is thoroughly solipsistic. Insofar as Brouwer's view is solipsistic I think it clearly deviates from (or is inconsistent with) the Kantian view of intuition that he invokes elsewhere, and to deleterious effect. I shall come back to this later and argue that Brouwerian solipsism is philosophically untenable. In any case, I think it

is fair to say that Brouwer did not have a philosophically sophisticated conception of intuition.

Heyting's work adds an interesting and important new dimension to the discussion of the foundations of intuitionism. His 1931 address on the intuitionistic foundations of mathematics is especially rich in philosophical content [Heyting 1931]. In the address Heyting explained and defended the intuitionistic viewpoint by suggesting that we view mathematical propositions as expressions of intentions, in the sense of Husserl's theory of intentionality. "Intentions" in this sense not only refer to states of affairs thought to exist independently of us but also to experiences thought to be possible. Heyting then identifies proofs, as mental constructions, with fulfillments of intentions. In the 1931 address he goes on to describe the meaning of the logical constants of the intuitionistic propositional calculus in these terms. Martin-Löf, in his lectures on the meanings of the logical constants [Martin-Löf 1983–84], has said that Heyting did not just borrow these terms from Husserl but that he also applied them appropriately. I agree, but in agreeing with Martin-Löf's remark I think we are at the same time committing ourselves to the need for the kind of clarification and extension of the philosophical views of Brouwer and Heyting that is called for by Heyting's identification. I discuss this below. Note, by the way, that Heyting does not use the term "intuition" in his description but anyone who knows Husserl's philosophy knows that the concept of intuition is defined in terms of the fulfillment of intentions. So in identifying proofs with fulfilled (or fulfillable) intentions Heyting too holds that intuitionism is founded on the evidence provided by intuition, only now Heyting has invoked a much more sophisticated and philosophically developed conception of intuition than had Brouwer. This concept of intuition also forms part of an elaborate theory of meaning, but one that is different from Brouwer's in several important respects. In particular, I shall argue for a theory of meaning that is not solipsistic.

Martin-Löf's views on the philosophical basis of intuitionistic mathematics are similar to Heyting's. In his 1983–84 lectures on the theory of meaning, for example, Martin-Löf says that

> ... the proof of a judgment is the evidence for it ... thus proof is the same as evidence ... the proof of a judgment is the very act of grasping, comprehending, understanding or seeing it. Thus a proof is, not an object, but an act. This is what Brouwer wanted to stress by saying that a proof is a mental construction, because what is mental, or psychic, is precisely our acts ... and the act is primarily the act as it is being performed, only secondarily, and irrevocably, does it become the act that has been performed.

Martin-Löf's comments emphasize the intuitionistic view that a proof is a cognitive act or process before it is an object, an act or process in which we come to see or intuit something. In his discussions of Heyting's Husserlian interpretation [Martin-Löf 1983–84, 1987, 199?] of the logical constants Martin-Löf has been more careful than some writers to distinguish between proof as an act or process and proof as an object. We might say that proof as an object is constituted, in its most primitive form, in an act of proof by virtue of the kind of retention in memory that Brouwer emphasizes in his descriptions (see [Tieszen 1989], pp. 99–111). Also, Martin-Löf's system of intuitionistic type theory in [Martin-Löf 1984] use four basic forms of judgment, among which are the two that "S is a proposition" and "a is a proof (construction) of the proposition S". Martin-Löf notes that one can read these, equivalently, as "S is an intention (expectation)" and "a is a method of fulfilling (realizing) the intention (expectation) S", respectively. Thus, one can understand his system as a formalization of features of the informal concepts of intentionality, intuition and evidence.

In a somewhat different vein, Troelstra and van Dalen, in their two-volume book *Constructivism in Mathematics* [Troelstra and van Dalen 1988], have argued that since the perfect introspection that Brouwer postulates for the ideal mathematician is simply not accessible to us we must look elsewhere for the philosophical foundations of intuitionism. They suggest that 'informal rigor' (in Kreisel's sense) is the main source of mathematical knowledge in intuitionistic mathematics. As Kreisel described it [Kreisel 1967], the idea of informal rigor is that we obtain definitions, axioms or rules by analyzing intuitive notions as precisely as possible and putting down their properties. Kreisel thought the general idea applied equally well to realist or idealist conceptions of mathematics. In idealist conceptions one supposes that the intuitive notions are related to thought or cognition instead of to a mind-independent, external world. The general form of the view of Troelstra and van Dalen that intuitionistic mathematics is based on informal rigor therefore amounts to the idea that we attempt to rigorously analyze intuitive concepts concerning various cognitive acts, structures and abilities instead of analyzing intuitive concepts concerning a mind-independent, external world. Their view is, I think, clearly consistent with the views of Heyting and Martin-Löf. In Heyting's work we already have the idea that in intuitionism we are to focus on intentions insofar as they refer to experiences thought to be possible, and this is distinguished from focusing on the reference of intentions to states of affairs which are thought to exist independently of us. The view of Troelstra and van Dalen also has the same form that is involved in

understanding Martin-Löf's intuitionistic type theory as a formalization of features of the informal concepts of evidence, intuition and intentionality. Thus, I would argue that founding intuitionistic mathematics on the idea of informal, rigorous concept analysis in the sense of Troelstra and van Dalen is not at all incompatible with the above-mentioned views of Heyting and Martin-Löf. In fact, the general form of all these views is remarkably similar to what is called for in parts of phenomenological analysis.

Let us now turn briefly to Dummett's views. Dummett has written more on the philosophical basis of intuitionism than anyone in recent times [Dummett 1975, 1976, 1977, 1991]. Prawitz, Sundholm and many others have discussed and elaborated on Dummett's arguments [see especially Prawitz 1977, 1978, 1980, and Sundholm 1983, 1986]. In his argument for rejecting classical logic in favor of intuitionistic logic Dummet takes the philosophical basis of intuitionism to lie in considerations involving the philosophy of language and, in particular, the theory of meaning. Indeed, Dummet argues that there is no way to approach these questions independently of or prior to investigations in the philosophy of language. Dummett's view is that the theory of meaning underlying intuitionism is, roughly, Wittgenstein's theory that meaning is determined by use. To say the meaning of a mathematical statement is exhaustively determined by its use is to say that the meaning of a statement cannot contain anything which is not fully manifest in the use of the statement. If two people agree completely about the use to be made of a statement they agree about its meaning. 'Undertanding' consists of knowledge of meaning in this sense. Prawitz, it should be noted, has modified the claim that meaning must be fully determined by observable uses of sentences. He suggests instead that "the samples of use with which we are presented never completely determine the meaning but only enable us to form some theories or hypotheses about the meaning" [Prawitz 1977]. This leads him to formulate an adequacy condition on meaning theories that is weaker than Dummett's requirement that implicit knowledge is to be fully manifest in behaviour.

Dummett contrasts his view with the view that the meaning of a proposition is determined by its truth conditions. The problem with the latter view of meaning, which is essentially the view embodied in classical two-valued logic and also in platonism, is that it gives us a notion of meaning which is not recognizable by us, or which transcends our knowledge or understanding. It cannot be a view on which meaning is fully determined by use. The argument for this claim is based on the fact that, in general, truth is not decidable. In particular, a platonist theory of meaning re-

quires us to have an understanding of quantification over infinite domains but this transcends our capacity to recognize statements which quantify over infinite domains as true. Suppose the meaning of an undecidable statement Φ is given by its truth-conditions and we know the meaning of Φ. How could this knowledge be manifested? Not by giving a proof or disproof of Φ. The best we could do is to paraphrase or restate Φ, but that does not give the meaning of Φ except to someone who already knows it. Knowledge of the meaning of a mathematical statement could not, on pain of an infinite regress, consist solely of *explicit* verbalizable knowledge, of the ability to state or to paraphrase the meaning of a statement, for then it would be impossible for anyone to learn a language who was not already equipped with a fairly extensive language. Knowledge of meaning must ultimately be *implicit*, and the ascription of implicit knowledge requires saying in what the full manifestation of the knowledge consists. There must be observable differences between the behaviour or capacities of someone who is said to have such knowledge and someone who is said to lack it. Thus, the truth-conditional view of meaning underlying classical logic and platonism cannot give substance to the idea of having implicit knowledge of what the condition for the truth of a mathematical statement is, since there can be nothing which constitutes a manifestation of such knowledge.

Dummett's argument is framed by views about how language would not be learnable if meaning were not fully determined by use, for if there were some kind of meaning that transcended the use made of an expression then we would have to say that someone might have learned a language and behaves in every way as if he had learned and yet does not understand, or understands incorrectly. Such a view would make meaning private, ineffable. It would be inconsistent with the idea that meaning is communicable and with mathematics as a social enterprise. Dummett also does not want to be understood as a radical conventionalist, as was Wittgenstein, about what counts as *correct use* of mathematical statements. Another important component of Dummett's argument is his rejection of meaning holism, i.e., of the view that nothing less than the total use of language determines the meaning of an individual statement. Dummett argues that the theory that meaning is determined by use would rule out revisionism in logic and mathematics if meaning holism were correct, because in that case the question of justifying deductive practices cannot really arise. Much more could be said about these matters than we have space for here.

2. Putting intuition back into intuitionism

The view of the philosophical basis of intuitionism that I shall argue for is different from Dummett's view in some important respects, although it can incorporate parts of Dummett's argument. It is, I believe, consistent with the views of Heyting, Martin-Löf, and Troelstra and van Dalen. I would like to say that it is also different from Brouwer's views in some ways, although it is still in the Kantian spirit of founding intuitionism on intuition.

Dummett of course points out that in his argument for rejecting classical logic in favor of intuitionistic logic he is not concerned with the exegesis of intuitionistic writings or with how well his account jibes with the views of the intuitionists themselves. I think this goes without saying since elements of the Wittgensteinian theory of meaning that Dummett takes as the philosophical basis of intuitionism are really quite alien to intuitionism as it has traditionally been expounded. What happens on Dummett's account, for example, to the distinctive idea in intuitionism that a proof is a *mental act* in which we can come to *see* something, or to have *evidence* for a judgment? Not too surprisingly, it disappears. While Dummett distances his position from classical behaviorism it is nonetheless true that on his account of intuitionism the entire vocabulary of cognitive acts, processes and abilities in fact disappears, or in some cases is reinterpreted, after Wittgenstein, in terms of observable practices and abilities. The distinction between inner and outer phenomena vanishes, along with the very basic distinction between act and object. So much then for Martin-Löf's description, cited above, of a proof of a judgment as the mental act of 'grasping', 'comprehending', 'understanding' or 'seeing' the judgment. I think this shift is also a source of the objection raised by Troelstra and van Dalen that they do not see how the formulation of axioms based on the process of informal rigor, such as their own formulation of axioms for lawless sequences, fits into Dummett's theory ([Troelstra and van Dalen 1988], p. 851).

Of course one of the things that disappears along with the idea of mental acts and processes on Dummett's approach is any philosophical objection to intuitionism based on solipsism or subjective idealism. But I shall argue that Dummett goes too far here, that we can perfectly well keep the distinction between inner and outer phenomena without succumbing to the pitfalls of solipsism. Dummett's account of intuitionism, on the other hand, contains no theory of intentionality, fulfilled intentions and evidence of the sort that Heyting appeals to. It has no theory of mathematicians as cognitive information processors, of the structure of cognition, of mental

acts and meaning, of mental representation, of the content of mental acts, of implicit or qualitative content, of consciousness, and the like. I think, however, that if proof is really to be understood as either a cognitive act or as an object of an act then these notions must figure into our understanding of the philosophical basis of intuitionism. Thus, I will argue that the philosophical basis of intuitionistic mathematics is best understood along the lines suggested by Heyting's 1931 address. I shall distinguish several key components of this view and then indicate how they can be used to enrich the philosophical understanding of intuitionism and also to defend intuitionism as a philosophy of mathematics.

The first component of the view is that mathematical propositions are to be understood as expressions of intentions, in the sense of a theory of intentionality [Tieszen 1989, 1991]. Intentionality is the characteristic of "aboutness" or "directedness" possessed by various kinds of mental acts. The intentions of acts can be determined by "that"-clauses in attributions of beliefs and other cognitive acts to persons, as in propositions of the form "M believes that Φ". M's intention here is expressed by Φ. Note that this view of expression entails a philosophy of language and also, as we shall see in a moment, a theory of meaning. Thus, I do not wish to suggest, as against Dummett, that a philosophy of cognition is independent of and anterior to a philosophy of language. The two are intertwined and may stand or fall together. However, there is not only one kind of philosophy of language or meaning. Witness, for example, the subtle interactions between the study of language, meaning and cognition in treatments of transformational grammars, in the semantics of propositional attitudes, and in what has come to be called propositional attitude psychology.

The second component of the view is that mathematical intentions can either be fulfilled or not, or even partially fulfilled, by additional acts carried out through time. Intentions can also be understood as expectations that can either be realized or not. When an (empty) intention is fulfilled then the object intended in an act is actually seen. That is, we have direct *evidence* for it. If the intention is fulfillable then it is possible to find the object intended. So, following Heyting, a *proof* or a *construction* in intuitionistic mathematics is a fulfilled (or fulfillable) mathematical intention. A fulfilled intention is an intention for which we possess evidence. The intention/fulfillment relation can also be understood in terms of Kolmogorov's interpretation [Kolmogorov 1932] of propositions as problems or tasks and proofs as solutions, and in terms of Martin-Löf's suggestion that we view empty intentions as specifications and fulfillments as programs that satisfy those specifications [Martin-Löf 1982]. I want to argue, however, that something crucial to the intuitionistic view that mathemat-

ics is the precise part of human thinking would be lost if these alternative explications were not understood in terms of a theory of intentionality. This means, for example, that we ought not to immediately identify fulfillments, understood as programs, with programs computable by Turing machines. There might be a difference between machine computability and human computability and we do not want to ignore the problematic status of Church's Thesis in intuitionistic mathematics.

The third component is that intentional acts are responsible for meaning or interpretation in the sense that strings of signs, noises, and so on would not be taken to have meaning, value, or significance if there were not intentional systems in the universe. This does not mean, as its detractors sometimes claim, that a person must always consciously, as it were, perform some mental act in order to understand a string of signs. It is a crude caricature of the view to suppose that there is first some completely uninterpreted sign configuration, and then someone performs a mental act which bestows sense upon it, whereupon it is understood. Rather, we normally understand the meaning of expressions quite automatically and prereflectively. The point is rather that it is a condition for the possibility of meaningfulness that there be individuals in the universe that have cognitive states that are characterized by intentionality. That is all that is meant by saying that mental acts are involved in meaning or understanding.

The fourth component I would like to mention is that mathematical statements can be meaningful even if they are not fulfilled or are not fulfillable. We have constructions for some mathematical intentions but not for others. But surely we can understand the meaning of a statement independenntly of knowing its truth value, for as Frege and Husserl remind us, and as Brouwer may have failed to recognize, we must not confuse lack of (intuition of) reference with meaninglessness, nor even logical inconsistency with meaninglessness. But the kind of distinction between meaning and reference implied by these remarks is not part of the Wittgensteinian theory of meaning that Dummett takes as the basis of intuitionism. In saying this I do not of course wish to deny that fulfilled mathematical intentions have a more determinate or explicit meaning than empty intentions. Fulfilled mathematical intentions provide more information about the object or state of affairs in question, including specific numerical or computational meaning that is otherwise lacking.

Now let us fill in somewhat the view of the philosophical basis of intuitionism that is associated with these points. Of particular concern, vis-à-vis Dummet, is whether the intuitionistic emphasis on proof as mental construction can be preserved and defended. Dummett has argued that

if meaning were not fully determined by observable uses of sentences we would have to say that someone might have learned a language and behaved as if she had learned and yet does not understand. Language would not be learnable if meaning were not fully determined by observable use. Meaning would be private, and ineffable, which is inconsistent with the possibility of communication and with the social character of mathematics. These arguments, however, are far from being decisive. First, I do not see any problem with saying that someone can appear from observable uses of sentences to have learned a language but in fact does not understand the language. Perhaps the simplest way to see this in recent times is through the type of argument John Searle gives about Chinese syntax manipulators [Searle 1980]. The person or machine in Searle's argument interacts with others in Chinese and passes the test for understanding Chinese based on the criterion of considering all possible observable uses of sentences, but does not understand a word of Chinese. What is missing? Intrinsic intentionality. Hilary Putnam has made arguments about the evolution of "perfect actors" to also show that observable linguistic behavior does not suffice to determine understanding or meaning [Putnam 1965]. It can be argued that observable behavior or practice generally underdetermines what we know or understand. We see this in linguistics, perception, mathematics, and elsewhere. Compare, for example, our linguistic performance and our linguistic competence.

Now, because observable practice in using sentences underdetermines our knowledge or understanding, and does not suffice to explain it, we must make an inference to unobservable, inner processes or structures to fill in the explanation. This is a pattern of reasoning that goes back to Kant and is now used widely in linguistics, cognitive science and artificial intelligence studies. The role of informal rigor in intuitionism, if it is to be the source of mathematical knowledge, must evidently be to unfold and clarify our knowledge of these cognitive processes or structures. On this view there is no reason to expect a direct correlation between a set of observable linguistic behaviors and the structures of a semantic theory, where these structures may be cognitively real. This is not, of course, to say that there is no correlation of any kind. Surely there is some relationship between our internal cognitive states and our observable linguistic behavior, but it would not do to suppose that we know enough about the relationship at this point in time to simply substitute the latter for the former. This seems to be especially true in the case of mathematics, where our thinking appears to have much more complexity, subtlety or nuance than it does in some of our other cognitive or practical endeavors. Observable linguistic behavior in mathematics, one might believe, is just

too coarse to do justice to this complexity, except perhaps at the level of pebble arithmetic. The pattern of reasoning we are invoking establishes a distinction between inner and outer phenomena, but not in a way that makes meaning private, non-learnable, non-communicable or non-implicit. Why not? The answer is straightforward: because human beings are so constituted as to have at least some isomorphic cognitive structure, which is what makes learnability and communicability possible.

In other words, the intuitionistic idea that a proof is a *mental* act of construction can be defended against the charge of incoherence on the following grounds, which are basically Kantian. We *start* by taking the science of mathematics as a social enterprise as given and then attempt to deduce the kinds of cognitive structures that are necessary to make it possible. On such a view there can be no philosophical defense of a solipsistic notion of proof, if solipsism is the position that there could be proofs that are in principle understandable to only one person. The rejection of solipsism does not, however, entail that there has to be intersubjective agreement at all times on all mathematical statements. Nonetheless, we have an explanation of how it is possible for there to be intersubjective agreement in at least elementary parts of mathematics and the explanation implies that the concept of a proof, as a fulfilled intention, is not solipsistic, and that it need not involve introspection. Thus, I argue that speaking of a proof as the fulfillment of an intention for a particular mathematician depends on the possibility of fulfillment of the same intention for other mathematicians.

I freely admit that this view of proof contrasts sharply with some of Brouwer's remarks. In the early *Leven, Kunst en Mystiek*, for example, Brouwer says that even in logic and mathematics "no two persons will think the same thing in the case of the fundamental notions" [Brouwer 1905]. Brouwer's viewpoint, however, fails to do justice to the fact that the science of mathematics, and intuitionism itself, exists as a social enterprise, that different people make contributions to it at different times and places. So I agree with Dummett insofar as he is pointing out that Brouwer's solipsism, so construed, is philosophically indefensible but, unlike Dummet, I do not jettison the idea of proof as an act of mental construction. I do not want to dispose of the act/object distinction, nor of some version of the meaning/reference distinction and of the view of epistemology that goes with it.

Several logicians have suggested that perhaps Dummett's view of the philosophical basis of intuitionism is, after all, consistent with the view I have presented thus far, or that Dummett's is a complementary view, for Dummett is simply emphasizing the external or observable aspects

of acts while I am emphasizing the internal aspects. On the other hand, some other logicians have felt that the views expressed above are definitely inconsistent with Dummett's views. I have been inclined to agree with the latter position, but perhaps further philosophical analysis is needed.

The other main point I want to discuss is one which bothers many logicians and mathematicians when the subject of intuitionism arises, and that is the question of the relation of intuitionistic to classical mathematics. In particular, I wish to ask how we should understand classical mathematics from an intuitionistic standpoint. In *The Elements of Intuitionism* and "The Philosophical Basis of Intuitionistic Logic" Dummett is interested in developing an argument for showing that the classical way of construing mathematics is "incoherent and illegitimate", that it is "unintelligible". He is concerned to find grounds for the revision of mathematical practice, and his argument is not favorably disposed toward an eclectic position on this issue.

Now I shall argue that there is a sense in which classical mathematics need not be construed as incoherent, illegitimate or unintelligible for an intuitionist, although it may be so construed if it is taken to do justice to mathematical knowledge. In order to grasp this let us first recall a response Dummett has made to a defense of platonism. Dummett has argued that human practice is simply limited and there is no extension of it, by analogy, that will give us an understanding of the capacity to run through an infinite totality. Meaning must be derived from *our* capacities. It cannot be derived from a hypothetical conception of capacities we do not have. To think otherwise only shows the extent to which illusions are involved in understanding our own language. It has of course been pointed out by Crispin Wright and others that one of the problems with this argument is that intuitionism is committed to some of its own rather strong idealizations of human practice, so that someone who took the limitations of our capacities seriously, like a strict finitist, could direct a similar line of reasoning to Dummett's own position [Wright 1982]. Thus, a strict finitist might argue that there is no extension of our practice which, by analogy, will give us an understanding of an effective procedure which is not feasible, according to some measure of computational complexity. Meaning cannot be derived from a conception of hypothetical capacities that transcends feasibility. Does this show that intuitionism is incoherent and unintelligible? I believe it no more shows this than Dummett's argument shows that classical mathematics is incoherent and unintelligible. But it does show us that something is amiss in Dummett's conception of how meaning is connected with idealizations of practice, especially as this is supposed to figure into the difference between basing meaning on use

and basing meaning on truth conditions.

On the Husserlian view of intentionality and meaning invoked by Heyting, we are to view mathematical statements as expressions of intentions, as expectations, or as problems. It is just that, as intuitionistic (weak) counterexamples show us, we have reason to believe that some of our expectations, understood in their full generality, will never be realized. But I do not see why intuitionists or even strict finitists need to deny that general logical principles like $P \vee \neg P$, or universal quantifications over infinite domains, can function in our experience as regulative ideals in a Kantian sense. That is, $P \vee \neg P$ can be viewed as an expectation of what should be the case at a research point lying at infinity, a kind of postulation of reason that reflects a natural tendency of human cognition, no matter how much we may try to suppress it. Then we can think of intuitionistic mathematics as an expression of the view that we *know* far less about objects than we can *reason* about on a classical model of reasoning. We are inclined in our reasoning to postulate certain closure conditions, forms of completeness or of "perfection" which cannot be verified in intuition. We try to complete the incomplete. But, at the same time, this can be useful because in the process we come to grasp and measure the degree and defects of the incomplete. If Kant's view is correct then regulative ideals drive scientific research and problem solving. For example, they induce mathematicians, including intuitionists, to work toward the solution of open problems with the expectation that a solution is to be found, although the source of the expectation is now taken to be immanent to cognition and is not derived from the idea of a mind-independent realm of truths, as it might be for an ontological platonist. For $P \vee \neg P$ to have this kind of meaning is of course not the same thing as having an intuitionistic proof that $P \vee \neg P$. On the other hand, it does not follow that classical mathematics, with its attendant notion of "meaning as determined by truth conditions", is unintelligible or incoherent, provided we now view it as postulating 'truth' as an absolute or regulative ideal, analogous to the abstraction from a finite bound on computation involved in the intuitionist's own conception of acceptable mathematical reasoning.

Intuitionists owe us at least an explanation of the origins of classical mathematics, if not of its status and significance, and on the view just described we have such an explanation. An intuitionist can ask about the conditions for the possibility of classical mathematics, and the answer will come in terms of some aspect of our cognitive makeup, some function involving the effort to complete the incomplete, to attain a kind of "cognitive closure". Parts of mathematical practice will be a product of this cognitive makeup, and in those parts where our idealizations are especially

far-flung lie the possibilities of antinomies, paradoxes, or illusions. Just as traditional rationalistic metaphysics existed, so parts of mathematical practice that cannot be constructively justified actually exist. There is, nonetheless, a foundation in our cognitive structure for classical mathematics and classical mathematics cannot be meaningless to us. In this way we can explain classical mathematics as a part of human practice to which different mathematicians in different times and places make contributions. The non-constructive meaning of mathematical statements also need not be construed as private, non-learnable, non-communicable, or non-implicit because, as I am construing intuitionism, humans are so constituted as to have at least some isomorphic cognitive structure, which is what makes learnability and communicability possible. Humans are bound to project their knowledge beyond their actual, even possible experience, but there is intersubjective agreement in doing this, even if the specific views that result from doing it are sometimes different.

Thus, intuitionists can say of classical mathematics that it constitutes an illegitimate and perhaps even a dangerous extension of what we can be said to *know* about objects, but that it is cognitively inevitable and does serve some purpose in human affairs. It is just that intuitionism calls for a kind of experiential verifiability not found in classical mathematics. This boils down to the fact that the kind of "grounding in experience" called for in constructive mathematics generally gives us a foothold in reality, a standard, and a common, "objective" basis for mathematics. There is a core of elementary mathematics on which the views of mathematicians of quite different philosophical persuasions overlap, and this core is constructivist. Intuitionism loses none of its substance in making the point that we need to be careful in saying that we "know" classical mathematics, in making the point that we do not really "know" something that results from striving for cognitive closure when doing so could lead to illusions.[1]

[1]I would like to thank the many LMPS IX participants with whom I discussed this paper for their comments, and especially Susan Hale, Geoffrey Hellman, Per Martin-Löf, Dag Prawitz, Hilary Putnam, Michael Resnik, Sören Stenlund, Göran Sundholm, and Dirk van Dalen.

REFERENCES

BENACERRAF, P. and PUTNAM, H. 1983, *Philosophy of Mathematics: Selected Readings*, 2nd edition, Cambridge University Press, Cambridge.

BROUWER, L. E. J. 1905, *Leven, Kunst en Mystiek*, Waltman, Delft.

BROUWER, L. E. J., 1912, *Intuitionisme en formalisme*, Nordman, Groningen. The English translation by A. Dresden, "Intuitionism and Formalism", has been reprinted in Benacerraf and Putnam, 77–89.

BROUWER, L. E. J., 1933, "Weten, willen, and spreken", *Euclides* 9, 177–193.

DUMMETT, M. 1975, "The Philosophical Basis of Intuitionistic Logic", in *Logic Colloquium '73*, Rose, H., and Sheperdson, J. (eds.), North-Holland, Amsterdam, 5–40.

DUMMETT, M. 1976, "What is a Theory of Meaning? (II)" in *Truth and Meaning*, Evans, G. and McDowell, J. (eds.), Oxford University Press, Oxford, 67–137.

DUMMETT, M. 1977, *Elements of Intuitionism*, Oxford University Press, Oxford.

DUMMETT, M. 1991, *The Logical Basis of Metaphysics*, Harvard University Press, Cambridge, Mass.

HEYTING, A. 1931, "Die intuitionistische Grundlegung der Mathematik", *Erkenntnis* 2, 106–115. The English translation, by E. Putnam and G. Massey, is reprinted in Benacerraf and Putnam, 52–61.

HEYTING, A. 1956, *Intuitionism*, North-Holland, Amsterdam.

KOLMOGOROV, A. N. 1932, "Zur Deutung der Intuitionistischen Logik", *Mathematische Zeitschrift* 35, 58–65.

KREISEL, G. 1967, "Informal Rigor and Completeness Proofs" in *Problems in the Philosophy of Mathematics*, Lakatos, I. (ed.), North-Holland, Amsterdam, 138–186.

MARTIN-LÖF, P. 1982, "Constructive Mathematics and Computer Programming", in Rose, H. and Sheperdson, J. (eds.), 1982, *Logic, Methodology and Philosophy of Science VI*, North-Holland, Amsterdam, 73–118.

MARTIN-LÖF, P. 1983–84, "On the Meanings of the Logical Constants and the Justifications of the Logical Laws", *Atti Degli Incontri di Logica Matematica*, Vol. 2, Università di Siena, Italia, 203–281.

MARTIN-LÖF, P., *Intuitionistic Type Theory*, Bibliopolis, Napoli.

MARTIN-LÖF, P. 1987, "Truth of a Proposition, Evidence of a Judgment, Validity of a Proof", *Synthese* 73, 407–420.

MARTIN-LÖF, P. 199?, "A Path from Logic to Metaphysics", an unpublished talk given at the congress Nuovi problemi della logica e della filosofia della scienza, Viareggio, Italia, January 1990.

PRAWITZ, D. 1977, "Meaning and Proofs: On the Conflict Between Classical and Intuitionistic Logic", *Theoria* XLIII, 2–40.

PRAWITZ, D. 1978, "Proofs and the Meaning and Completeness of the Logical Constants", in *Essays on Mathematical and Philosophical Logic*, Hintikka, J., et al. (eds.), D. Reidel, Dordrecht, 25–40.

PRAWITZ, D. 1980, "Intuitionistic Logic: A Philosophical Challenge", in von Wright, G.H. (ed.), *Logic and Philosophy*, Nijhoff, The Hague, 1–10.

PUTNAM, H. 1965, "Brains and Behavior", in Butler, R.J. (ed.), *Analytical Philosophy*, vol. 2, Blackwell, Oxford.

SEARLE, J. 1980, "Minds, Brains and Programs", *The Behavioral and Brain Sciences* 3, 417–457.

SUNDHOLM, G. 1983, "Constructions, Proofs and the Meanings of the Logical Constants", *Journal of Philosophical Logic* 12, 151–172.

SUNDHOLM, G. 1986, "Proof Theory and Meaning", in Gabbay, D., and Guenthner, F. (eds.), *Handbook of Philosophical Logic*, Vol. III, Reidel, Dordrecht, 471–506.

TIESZEN, R. 1989, *Mathematical Intuition,*, Kluwer, Dordrect.

TIESZEN, R. 1991, 1991, "What is a Proof?", in *Proof, Logic and Formalization*, Michael Detlefsen (ed.), Routledge, London.

TROELSTRA, A., and VAN DALEN, D. 1988, *Constructivism in Mathematics* (2 vols.), North-Holland, Amsterdam.

WRIGHT, C. 1982, "Strict Finitism", *Synthese* 51, 203–282.

Logic, Methodology and Philosophy of Science IX
D. Prawitz, B. Skyrms and D. Westerståhl (Editors)

ASYMPTOTICS, SINGULARITIES AND THE REDUCTION OF THEORIES

MICHAEL BERRY

H H Wills Physics Laboratory, Tyndall Avenue, Bristol BS8 1TL, UK

1. Introduction

In science we strive to integrate our experiences, observations, and experiments into a single explanatory framework - 'a theory of everything'. Of course this goal has not been achieved, and probably never will be. What we have instead are the partial descriptions provided by biology, chemistry, physics, etc., and, within these, the various subfields such as fluid mechanics and quantum mechanics. The different areas of study do not fit tidily together. Particular difficulties arise when a more general description is supposed to encompass an older, less general, one, usually by providing a microscopic explanation of its principles. It is hoped that a less general theory can thus be 'reduced' to a more general one. But this comfortable picture is often spoilt by certain classes of higher-level, or 'emergent', phenomena which are well described by the older theory but obstinately refuse to emerge from the supposedly encompassing one.

To illustrate the point with a familiar example, consider life. Is it contained in, or implied by, Schrödinger's equation for the 10^{23} electrons and nuclei in an organism, plus rules for incorporating the environment? I suspect that most scientists, especially physicists, would, if pressed, answer yes, but be uncomfortable. The discomfort stems from a dilemma. We know that writing down the Schrödinger equation and gazing at it is not a promising strategy for finding a cure for AIDS, or learning why we do not live for ever. But we feel that invoking something else, outside physics, at a fundamental level, is mysticism. Somehow, life might emerge from physics in some limit (possibly involving increasing complexity), but we have no clear idea how to convert this dream into science.

Of course this problem of reduction has been studied a great deal by philosophers. Sometimes the discussion centres on the conflict between the

two views summed up by the terms 'correspondence' and 'incommensurability': in brief, two theories correspond if one can be deduced as a special case of the other, and are incommensurate if their foundations are logically incompatible. My intention here is to present an idea which seems to capture an essential aspect of the problem of reduction of emergent phenomena and which goes some way towards dissolving the antinomy between incommensurability and correspondence, but which has not to my knowledge been considered by philosophers. I will confine myself to reductions of theories within physics, but of course hope that the idea could eventually prove useful in grander contexts such as the reduction of biology to physics (or chemistry).

To begin, realise that theories in physics are mathematical; they are formal systems, embodied in equations. Therefore we can expect questions of reduction to be questions of mathematics: how are the equations, or solutions of equations, of one theory, related to those of another? The less general theory must appear as a particular case of the encompassing one, as some dimensionless parameter - call it δ - takes a particular limiting value. A general way of writing this scheme is

$$\text{encompassing theory} \rightarrow \text{less general theory} \quad \text{as} \quad \delta \rightarrow 0 \tag{1}$$

Thus reduction must involve the study of limits, that is asymptotics. The crucial question will be: what is the nature of the limit $\delta \rightarrow 0$? We shall see that very often reduction is obstructed by the fact that the limit is *highly singular*. Moreover, the type of singularity is important, and the singularities are not only directly connected to the existence of emergent phenomena but underlie some of the most difficult and intensively-studied problems in physics today.

There is one aspect of the study of limits in physics which has attracted the attention of philosophers, beginning with Berkeley, that I will not be considering here, even though there are interesting and subtle points still to be brought out. This centres on the fact that the limit $\delta = 0$ is always an idealization; in any actual situation, δ is always finite. Instead of discussing this important matter, which involves the relation between the world and our models of it, I shall remain firmly in the realm of theory.

Before proceeding to examples, I must disambiguate an irritating terminological orthogonality. Philosophers consider the less general theory as being 'reduced by' the encompassing theory, because the latter employs principles that are more elementary to explain more phenomena [1]. Physicists, however, find it more natural to think of the reduction as occurring the other way, that is by the more general theory 'reducing to' the less general one as $\delta \rightarrow 0$ because the less general one is a special case (thus the function $\cos\theta$ 'reduces to' 1 as $\theta \rightarrow 0$).

2. Singular limits and emergent phenomena

Here are six examples of the scheme (1) in physics, together with the meaning of the dimensionless parameter δ.

- special relativity $\rightarrow$ Newtonian mechanics, $\delta = v/c$.
- general relativity $\rightarrow$ special relativity, $\delta = Gm/c^2a$.
- statistical mechanics $\rightarrow$ thermodynamics, $\delta = 1/N$.
- viscous (Navier-Stokes) flow $\rightarrow$ inviscid (Euler) flow, $\delta = 1/Re = \eta/\rho a v$.
- wave optics $\rightarrow$ ray optics, $\delta = \lambda/a$.
- quantum mechanics $\rightarrow$ classical mechanics, $\delta = \hbar/S$.

Here the meaning of the symbols is as follows. v : speed of body; c: light speed; G: Newton's gravitational constant; m: mass of body; a: typical linear dimension of body; N: number of particles; Re: Reynolds' number; η : viscosity; ρ: density; λ :wavelength; $\hbar$: Planck's constant; S: typical classical action.

Reduction in its simplest form is well illustrated by the first example. Every physics student learns that one form of the connection between the encompassing theory of special relativity and the less general theory of Newtonian mechanics is contained in the 'low speed' series expansion

$$\sqrt{1-\delta^2} = 1 - \frac{1}{2}\delta^2 - \frac{1}{8}\delta^2 + \cdots \tag{2}$$

The left side represents special relativity, and the right side is a convergent Taylor series whose first term represents Newtonian mechanics. Mathematically, special relativity is analytic in δ at $\delta = 0$, so that the limit is unproblematic (the hyper-relativistic limit $\delta = 1$ is singular, but that is a different matter).

My main point will be that this simple state of affairs is an exceptional situation. Usually, limits of physical theories are not analytic: they are singular, and the emergent phenomena associated with reduction are contained in the singularity. Often, these emergent phenomena inhabit the borderland between theories.

To begin, consider the third example, namely the reduction of thermodynamics by statistical mechanics as the number of particles ($N = 1/\delta$) increases to infinity (the 'thermodynamic limit'). Standard arguments [2] involving large-N asymptotics show that for a fluid the thermodynamic

equation of state, e.g. the pressure $P(V,T)$ as a function of volume and temperature, can (in principle and to a large extent in practice) be derived from the principles of statistical mechanics and a knowledge of the forces between the atoms. But the reduction runs into difficulty near the *critical point* P_c, V_c, T_c , where the compressibility $\kappa \equiv [-V(\partial P/\partial V)_T]^{-1}$ is infinite. The problem is to find the form of the divergence of κ as $T \to T_c$. This is a power-law, whose exponent is wrongly given by otherwise useful models such as the Van der Waals theory.

The reason for the difficulty is fundamental, and only after a decade of concentrated effort was it clarified, and techniques developed for the correct calculation of 'critical exponents'. Thermodynamics is a continuum theory, so reduction has to show that density fluctuations arising from interatomic forces have a finite (and microscopic) range. This is true everywhere except at the critical point, where there are fluctuations on all scales up to the sample size. Thus at criticality the continuum limit does not exist, corresponding to a new state of matter [3]. In terms of our general picture, the critical state is a singularity of thermodynamics, at which its smooth reduction to statistical mechanics breaks down; nevertheless, out of this singularity emerges a large class of new 'critical phenomena', which can be understood by careful study of the large-N asymptotics.

A particularly vicious example, at the cutting edge of applied mathematics nowadays, is the fourth on the above list, namely the mechanics of a fluid as its viscosity is decreased or its speed is increased (so that δ gets smaller). Exact solution of the Navier-Stokes equation for smooth flow down a pipe, driven by a pressure difference ΔP , predicts that the mass flow rate is proportional to ΔP . For small δ, however, experiment shows a rate close to $\sqrt{\Delta P}$. The reason is that the predicted flow is unstable, and the true flow is not smooth but disorderly, that is, *turbulent*. In turbulence [4-6], instead of viscous dissipation vanishing smoothly as $\delta \to 0$, the dissipation concentrates onto a set of zero measure which is fractal in form. Again the limit $\delta \to 0$ is singular, and out of the singularity emerges an important phenomenon, namely turbulence, whose mathematical nature is still far from understood.

3. Quantum and classical mechanics

Now we come to the examples I shall discuss in most detail - not because they are more fundamental than the others but because they lie closest to my own research interests [7] - namely the reduction of ray theory (e.g. geometrical optics) to wave theory, and (closely related) of classical to quantum mechanics. Here, singular limits abound, even in the simplest problems, as

the following example shows.

A wave (of light, sound or water, for example) travelling along the x-axis with speed ν can be represented by

$$\psi = \cos\left\{\frac{2\pi}{\lambda}(x - \nu t)\right\} \tag{3}$$

In the ray limit (where for example geometrical optics provides a consistent and serviceable description of, for example the operation of telescopes and cameras), we have $\lambda \to 0$. But this limit is singular! ψ is non-analytic at $\lambda = 0$, so that it cannot be expanded in powers of λ ; instead, this wavefunction oscillates infinitely fast and takes all values between -1 and $+1$ infinitely often in any finite range of x or t. Only if we consider the wave *intensity*, corresponding to ψ^2, and average over a small interval corresponding to the finite resolution of a detector, do we get the finite and smooth result corresponding to the intensity of the system of parallel rays corresponding to (3); often it is convenient to average over time (reflecting the fact that for light or sound the wave frequency is too high to measure directly):

$$\left\langle \psi^2 \right\rangle_t = \left\langle \cos^2\left\{\frac{2\pi}{\lambda}(x - \nu t)\right\}\right\rangle_t = \frac{1}{2} \tag{4}$$

Now consider the superposition of two such waves, with speeds ν and $-\nu$, giving

$$\psi = \cos\left\{\frac{2\pi}{\lambda}(x - \nu t)\right\} + \cos\left\{\frac{2\pi}{\lambda}(x + \nu t)\right\} = 2\cos\left\{\frac{2\pi x}{\lambda}\right\}\cos\left\{\frac{2\pi\nu}{\lambda}\right\} \tag{5}$$

and the time average

$$\left\langle \psi^2 \right\rangle_t = 2\cos^2\left\{\frac{2\pi x}{\lambda}\right\} \tag{6}$$

This describes a spatially fixed interference pattern such as that produced by a double slit. Again there is a powerful singularity at $\lambda = 0$. To eliminate it requires an extra average, this time spatial, and then we obtain

$$\left\langle \psi^2 \right\rangle_{t,x} = 1 \tag{7}$$

Thus to obtain from wave theory the simple fact that in ray theory two beams of intensity $1/2$ add to give intensity 1, with no interference, requires a double average over a mathematically pathological function.

Having seen that interference is associated with a $\cos^2(1/\lambda)$ singularity in the ray limit, we now examine the anatomy of other sorts of wave singularity. An interesting case occurs when waves reach places that rays do not.

Examples are the outside of a glass-air interface within which total internal reflection occurs, the dark side of a rainbow, and the thin layer of air near a hot road in which mirage reflections are seen. In the ray limit, the wave and its intensity are zero, but there are nevertheless waves present, whose amplitude is typically

$$\psi \propto \exp\left\{\frac{-\text{function of } x}{\lambda}\right\} \tag{8}$$

Again this is singular, and cannot be expanded in a power series in λ (all terms are zero).

An important role is played by the *transition* between these two sorts of singularity ($\cos^2\{1/\lambda\}$ and $\exp\{-1/\lambda\}$). (This is somewhat analogous to the transition T through T_c in thermodynamics, for large N.) The transition happens across a *caustic*, which is an envelope of a family of rays (a generalized focal surface in space, or line in the plane), marking the boundary between regions with different numbers of rays. In the simplest case, the regions have two rays and no rays, corresponding to the 'interference' and 'penetration' regimes represented by (6) and (8). A caustic is a collective phenomenon, a property of a family of rays that is not present in any individual ray. Probably the most familiar example is the rainbow. The singularity across a caustic must interpolate between (6) and (8). How this happens was first elucidated by Airy in 1838 as part of an attempt to understand supernumerary rainbows, that is oscillations on the lit side of the bow, in the intensity of light of a given colour. It was necessary for him to invent a new function $Ai(z)$, oscillatory for $z < 0$ and decaying for $z > 0$. In terms of $Ai(z)$, the wave across a caustic has the form

$$\psi = \frac{1}{\lambda^{1/6}} Ai\left\{\frac{Kx}{\lambda^{2/3}}\right\} \tag{9}$$

In this transition, the emergent phenomenon is the fringe pattern associated with a caustic: in the ray limit $\lambda \to 0$, its intensity grows as $\lambda^{-1/3}$, and the spacing of the fringes shrinks as $\lambda^{2/3}$.

Caustics can themselves have singularities, whose classification is the province of catastrophe theory [8]. At such places, the envelope of rays is itself singular. These singular envelopes are decorated with wave patterns ψ whose $\lambda \to 0$ singularities (shrinking fringe spacings and diverging intensities) depend on the geometry of the catastrophe. Such 'diffraction catastrophes' have intricate and beautiful structures [9,10], and constitute a hierarchy of nonanalyticities, of emergent phenomena par excellence. The patterns inhabit the borderland between the wave and ray theories, because when λ is zero the fringes are too small to see, whereas when λ is too large the

overall structure of the pattern cannot be discerned: they are *wave* fringes decorating *ray* singularities.

Quantum mechanics is a particular wave theory, whose corresponding ray theory is classical mechanics, and where Planck's constant $\hbar$ plays the role of wavelength λ (through De Broglie's relation $\lambda = 2\pi\hbar/p$ where p is momentum). Its relation to classical mechanics should be through the *semiclassical limit* $\hbar \to 0$. When the limit is not singular, we have the correspondence principle: quantum observables tend to their classical counterparts as $\hbar \to 0$. Usually, though, the limit is singular, and then the correspondence principle, while often a useful guide [7], is too crude to be a substitute for mathematical asymptotics. From the analogy with other sorts of waves we expect that the nonanalyticities and emergent caustic phenomena described above will occur in quantum mechanics, and these have indeed been seen in the scattering of electrons, nuclei and atoms. In addition, the $\hbar \to 0$ limit is enriched by another limit, which is fundamental, namely the *long-time limit* $t \to \infty$.

There are several reasons to study the long-time limit in conjunction with the semiclassical limit:

- Spectra of atoms and molecules involve the quantized energies of these systems when in stationary states. These are states that persist over infinite time, so their semiclassical study – spectra near the classical limit – inescapably involves $t \to \infty$ too.
- Experiments on atoms traversing strong oscillating fields begin to probe the combined $\hbar \to 0, t \to \infty$ limit.
- It is only after infinite time that chaos may occur in the classical orbits. Chaos [11, 12] is unpredictability arising from exponential sensitivity to initial conditions in a bounded region. Therefore any attempt to study how classical chaos is reflected in the semiclassical limit of quantum mechanics ('quantum chaology' [13, 14]) must evidently involve $t \to \infty$ as well.

The essential point is that *the two limits do not commute*: taking the classical limit first, and the long-time limit second, leads to a different result from taking the limits in the reverse order. Such a clash of limits implies a singularity at the origin of the plane with coordinates $\hbar, 1/t$. One way to try to resolve the clash is to take both limits at once, in a controlled way, i.e.

$$\hbar \to 0, \ t \to \infty, \ \hbar t \equiv \tau = \text{constant} \tag{10}$$

In the one case where it has been possible to take the combined limit explicitly [15], for a system whose classical dynamics is trivial, analysis shows that the point $\hbar = 1/t = 0$ is truly a 'dragon's lair' , so singular that the behaviour exhibits a fantastic complexity which depends on the *arithmetic nature* of τ.

When the classical orbits are chaotic, the clash of limits generates some remarkable emergent phenomena. I will briefly describe just one: *the statistics of spectral fluctuations.* Consider a bound quantum system, that is one with a discrete spectrum of energy levels, and ask about the distribution of these levels in the semiclassical limit. The simplest fact about the levels is that as $\hbar \to 0$ they get closer together – their mean spacing is proportional to $\hbar^N$, where N is the number of freedoms. This must happen, because in the classical limit the levels form a continuum. (It is worth pausing to remark that this particular passage to the limit provides a nice illustration of the 'incommensurability' and 'correspondence' approaches to reduction. In the first, it is emphasized that for any finite $\hbar$, however small, the spectrum is always discrete: the classical continuum is never reached, and so cannot be said to be logically contained in the semiclassical limit. On the other hand, when $\hbar$ is sufficiently small the inevitably finite resolution of any spectroscopic measuring device means that the results of all observations will be the same as if the spectrum were continuous, and the correspondence principle can be said to apply.)

Now imagine looking at the set of levels with a microscope [14] whose power is proportional to the mean level density, thus generating a rescaled spectrum consisting of a set of numbers whose mean density remains constant as $\hbar \to 0$. What is the statistical nature of the fluctuations of this set of numbers about its (unit) mean density? The answer is remarkable: apart from trivial exceptions, the fluctuations are *universal* [14, 16], that is, independent of the details of the system and dependent only on whether the orbits of its classical counterpart are regular or chaotic. Paradoxically, the spectral fluctuations are those of a sequence of random numbers (Poisson distribution) when the classical motion is regular, and are more regularly distributed (exhibiting the level repulsion characteristic of the eigenvalues of random matrices) when the classical motion is chaotic. We are beginning to understand this quantum universality [7] in terms of semiclassical asymptotics: it arises from a similar universality in the distribution of long-period classical orbits.

Universality of the spectral fluctuations is a novel qualitative phenomenon emerging from quantum mechanics in the combined semiclassical long-time limit. It was not predicted by analysis of the Schrödinger equation, but was discovered in numerical experiments (and later seen in real experiments) motivated by some physical arguments. Nevertheless Schrödinger's equation does contain it, albeit well concealed behind some very tricky (and incompletely explored) asymptotics.

4. Divergent series

So far, we have considered only the leading-order behaviour in the parameter δ whose vanishing describes how the encompassing theory reduces to the less general one. In those cases where any sort of mathematical treatment was possible, the leading-order behaviour was quite complicated (cf. equation (9)), and this of course reflects the singular nature of the limit. But determination of the leading order is only the first step: a complete treatment requires understanding the series consisting of all the correction terms – usually involving powers of δ. The determination of such series is still in its infancy, but it has been carried out for certain of the simpler problems of wave physics described in §3.

The most important characteristic of such series, and one which almost certainly extends to all series associated with singular reductions, is that they *diverge*. This was one of the factors prompting a re-examination [17] of the mathematics and physics of divergent series. The main results reinforce earlier indications [18] that the divergent tail, conventionally discarded as mathematically meaningless, contains important information in coded form. When decoded, these tails not only enable the function being expanded to be approximated to previously unequalled levels of accuracy but also describe physical effects, associated with the reduction that the asymptotics is attempting to describe, which are qualitatively different from those contained in the leading terms. Examples are the exponentially weak births of rays beyond caustics [19] and the generation of transitions between quantum states [20] in the adiabatic limit of slow driving.

It seems clear that these ideas, and further developments of them, must be involved in any complete description of how the less general theory is embedded in the structure of the encompassing theory.

5. Concluding remarks

Even in what philosophers might regard as the simplest reductions, between different areas within physics, the detailed working-out of how one theory can contain another has been achieved in only a few cases and involves sophisticated ideas on the forefront of physics and mathematics today. This is because in all nontrivial reductions the encompassing theory is a singular perturbation (parameterised by δ) of the less general one. The singularities are reflected in the quantities of the encompassing theory being nonanalytic at $\delta = 0$, and the nonanalyticities describe emergent phenomena in the borderland between the theories. As examples of these phenomena I described thermodynamic critical behaviour in fluids, fluid turbulence, interference

patterns decorating optical caustics, and the chaology-dependent statistics of energy-level fluctuations in quantum mechanics.

It should be clear from the foregoing that a subtle and sophisticated understanding of the relation between theories within physics requires real mathematics, and not only verbal, conceptual and logical analysis as currently employed by philosophers. One can hope that these ideas generalize beyond physics (for example to the reduction of biology or chemistry). This would mean that the problem of theory reduction would itself have been 'reduced', to the mathematical asymptotics of singularities. From the evidence so far, the task will be far from easy, and will require the development of new physical ideas and new mathematical concepts and techniques.

Finally, I would be the first to admit that the ideas explored here lack precision in several respects, and have not been presented in their final form. I hope they will benefit from the attention of philosophers.

References

[1] BULLOCK, A. and STALLYBRASS, O. 1977, *The Fontana Dictionary of Modern Thought.* (Collins, London). "Reduction: ... In philosophy ... the process whereby concepts ... that apply to one type of entity are redefined in terms of concepts ... of another kind, normally one regarded as more elementary ...".

[2] TOLMAN, R.C. 1938, *The Principles of Statistical Mechanics* (Oxford, University Press).

[3] WILSON, K.G. 1975, Rev. Mod. Phys. 47, 773-840.

[4] MANDELBROT, B.B. 1982, *The Fractal Geometry of Nature* (San Francisco: Freeman).

[5] FRISCH, U, 1983, in *Chaotic Behaviour in Deterministic Systems*, eds G. Iooss, R.H.G. Helleman and R. Stora, Les Houches Lecture Series XXXVI (Amsterdam: North-Holland) pp665-704.

[6] SREENIVASAN, K.R. and PRASAD, R. 1989, Physica D38, 332-339.

[7] BERRY, M.V. 1991, in *Chaos and Quantum Physics*, eds M-J Giannoni, A. Voros and J. Zinn-Justin, Les Houches Lecture Series LII (Amsterdam: North-Holland).

[8] POSTON, T. and STEWART, I.N. 1978, *Catastrophe Theory and its Applications* (London: Pitman).

[9] BERRY, M.V. and UPSTILL,C.1980, *Progress in Optics* 18, 257-346.

[10] BERRY, M.V. 1990, *Current Science* (India), 59, 1175-1191.

[11] STEWART, I.N. 1989 *Does God Play Dice? The Mathematics of Chaos* (Oxford: Blackwell).

[12] BERRY, M.V. 1990, Proc.Roy. Institution of Gt. Britain, 61, 189-204.

[13] BERRY,M.V., 1989, Physica Scripta 40 335-336.

[14] BERRY, M.V., 1987, Proc.Roy.Soc. A 413, 183-198.

[15] BERRY, M.V., 1988, Physica D, 33, 26-33.

[16] BOHIGAS, O. and GIANNONI, M-J.1984 in *Mathematical and Computational Methods in Nuclear Physics,* eds., J.S.Dehesa, J.M.G.Gomez and A.Polls, Lecture Notes in Physics 209 (N.Y.:Springer-Verlag) pp1-99.

[17] BERRY, M.V. 1991 *Asymptotics, superasymptotics, hyperasymptotics...* in Asymptotics beyond all orders, ed. S. Tanveer (New York: Plenum) in press.

[18] DINGLE, R.B. 1973, *Asymptotic Expansions: their Derivation and Interpretation* (London: Academic Press).
[19] BERRY, M.V., 1989, Publ. Math.of the Institut des Hautes Études scientifique, 68 211-221.
[20] BERRY, M.V. 1990 Proc.Roy.Soc.Lond, A429, 61-72.

Logic, Methodology and Philosophy of Science IX
D. Prawitz, B. Skyrms and D. Westerståhl (Editors)

REALISM AND QUANTUM MECHANICS

HANS PRIMAS

Laboratory of Physical Chemistry, ETH-Zentrum, CH-8092 Zürich, Switzerland

A realistic interpretation of quantum mechanics is imperative

By and large, working scientists are unabashed realists, they stubbornly believe that there is a real external world. For many theoreticians, this belief is the only raison d'être of physical theories. They would like to have a description of how, fundamentally, the world *is*. But many popular presentations tell us that quantum mechanics is not compatible with realism. If this view would be true we would be in real trouble. Scientists take no thought of abandoning quantum mechanics since it is probably the empirically best confirmed scientific theory. In spite of many counter-intuitive quantum-theoretical predictions, there is not a single well-performed experiment which contradicts quantum mechanics. Certainly, there are open questions, but no flagrant contradictions between theory and experiment. On the other hand, we cannot abandon realism since the very confirmation of quantum mechanics is based on the acceptance of everyday realism. In the early days, quantum mechanics has been considered as a theory of the microworld, and most scientists did not realize that they cannot consistently adopt different ontologies for the microworld and the everyday world of laboratory instruments. Nowadays we cannot any longer take this position because we know that quantum mechanics is valid also for mesoscopic systems–like DNA-molecules with biochemically important quantum properties *and* genetically important classical properties. Since no scientist is willing to give up some kind of realism in the domain of laboratory experience, we really have to care for a *realistic interpretation of quantum mechanics.*

The philosophical notions about quantum mechanics held by many philosophers and theoretical physicists are incompatible with the actual practice of the working scientist. The lack of a well-founded philosophical discourse on quantum mechanics has harmful consequences in research and in teaching. Nevertheless, quantum theory is by no means in a state of crisis. The problem is only that many scientists and most philosophers are not familiar

with the modern technical developments of quantum mechanics, and therefore they still try to solve conceptual problems of quantum theory– like the theory of the measuring process– in terms of old-fashioned Hilbert-space quantum mechanics *which is valid only for finite closed systems.* Strictly speaking, such systems do not exist. Since all material systems are inextricably coupled to the electromagnetic and to the gravitational field, even "reasonably isolated" finite systems do not exist. This does not mean that it is not instructive to study the fiction of closed systems, but one should not confuse tentative investigations and the full-grown theory. In this sense, no exegesis of the writings of Niels Bohr, Werner Heisenberg and other pioneers will lead to a satisfactory solution of the conceptual problems of contemporary quantum mechanics.

I would like to advocate to investigate carefully

(i) *what we mean by realism, and whether we should expurgate objectionable ideas taken over from realism as understood in classical physics.*

(ii) *what we mean by quantum mechanics from a contemporary point of view, and whether philosophically important features in our understanding of quantum physics have changed in the last sixty years.*

The Cartesian split, the death of atomism and the limitations of contemporary science

Classical physics and a large part of contemporary science rest on Descartes' idea that nature is intrinsically divided into two parts: mind (*res cogitans*) and matter (*res extensa*). In addition, it is a tacit assumption of all engineering and experimental sciences that nature can be *manipulated* and that the initial conditions required by experiments can be created by interventions using means *external* to the object under investigation. That is, *we take it for granted that the experimenter has a certain freedom of action which is not accounted for by first principles of physics.* Man's free will implies the ability to carry out actions, it constitutes his essence as an actor. Without this freedom of choice, experiments would be impossible. *The framework of experimental science requires this freedom of action as a constitutive though tacit presupposition.* Traditionally, free will is understood as something belonging to the spiritual world, therefore contemporary science cannot dispense lightly with Cartesian dualism.

Many scientists and philosophers praise quantum mechanics as the *fun-*

dament of modern physics, molecular chemistry and molecular biology, but rarely it is stressed that quantum mechanics also put an *end to atomism.* The historical idea that the material world is already structured by some kind of interacting 'elementary systems' is in sharp contradiction to the structure suggested by quantum mechanics. According to quantum mechanics, the material world is a whole, *a whole which is not made out of parts.* If one agrees that quantum mechanics is a serious theory of matter, then one cannot adopt the classical picture of physical reality with its traditional metaphysical presuppositions. In particular, the nonseparability and nonlocality of the material world and its holistic features are not compatible with the ontology usually adopted in classical physics.

The experimentally well-confirmed holistic character of the material world casts severe doubts upon the consistency of the Cartesian separation of the *material* reality from the *spiritual* one—this idea may well be radically in error. Nevertheless, *present-day experimental science* still requires an *epistemological* dualism of subject vs. object. It is true that quantum theory has clearly put in evidence the limitations and the narrowness of today's scientific conception of reality, but *the often heard statement that quantum mechanics has already given up Cartesian dualism is unfounded.* In every experimental investigation of a quantum system, the measuring apparatus is described positively in terms of classical or engineering physics. *In quantum physics man's consciousness does not enter the physical discourse in any other way than in classical physics.* In the words of Wolfgang Pauli: "Die alte Frage, ob unter Umständen der psychische Zustand des Beobachters den äusseren materiellen Naturverlauf beeinflussen kann, findet in der heutigen Physik keinen Platz" [1]. In fact, contemporary quantum mechanics–as it is used by all experimentalists–is still in a kind of "peaceful coexistence" with Cartesian dualism. That does not mean that the Cartesian separation is not misconceived and that we should not try to create a non-Cartesian science. However, today's physics is ill-disposed and technically incapable to start such a project. At present, it would be science fiction to link quantum events to conscious events, or trying to incorporate a representation of conscious processes into physical representations of brain processes. Since there is no sound theory which includes consciousness in the realm of physics, I prefer to acknowledge that there is a gap in the reasonings of present-day science. In this sense, all physical theories at our disposal are *essentially incomplete theories:* they are incapable to deal with the *complementarity* of matter and spirit.

Contemporary quantum mechanics requires an engineering approach with a division into a part "which sees" and a part "which is seen". According to the formalism of quantum mechanics, this cut is *context-dependent* and not

identical with the Cartesian cut. The Cartesian separation would require an *intrinsic* separation of the whole reality into *res extensa* and *res cogitans*, while engineering quantum mechanics requires a contextual subject–object tensor-product decomposition of the whole reality such that there are no Einstein–Podolsky–Rosen-correlations between the *observed object* and the *observing tools*. This requirement is a precondition of experimental science. In the formalism of algebraic quantum mechanics, it implies that the observing tools have a representation as *classical* quantum systems[1]. In all engineering applications of quantum mechanics, the conscious human observer is a *part* of the "observing tools" so that the experimenter can be regarded as a "*detached observer*" in the sense of Bohr [2]. Inasmuch as the Cartesian cut is put within the classical domain, a direct conflict between quantum theory and the Cartesian ontology is avoided. This is in accordance with the modern experimental techniques where the observing and recording devices are often completely automated to the extent that the role of the human observer is reduced to simple acts of cognition of the numeric displays of classical measuring instruments. Hence the free will and the awareness of the observing scientist play exactly the same role they have in classical physics and engineering science. Also, in the cosmological or biological evolution there are objective happenings, encodings and registrations which are independent of the existence of beings having a consciousness. For these reasons, we conclude that in general the irreversible transmutation from possibilities to facts cannot depend on anthropogenic preparation and registration procedures, or on the consciousness of a human observer.

Realism

The historical Copenhagen view does not present quantum mechanics as a universal theory, it presupposes observational tools, but does not describe them quantum-mechanically. According to the Copenhagen view, quantum mechanics gives just the rules to calculate the probability of quantum events, but does not describe the events themselves. This attitude was reasonable in the pioneer years of quantum mechanics since at that time the mathematical tools for describing open systems and their interactions with the environment were not available. In order to analyze modern experiments of molecular science and the phenomena of molecular biology from a quantum-

[1] A quantum system is said to be classical if its algebra of observables is commutative. Note that every classical quantum system depends on Planck's constant. The existence of contextual classical quantum systems has not to be postulated, but is a consequence of a proper mathematical codification of quantum mechanics.

mechanical point of view, *the Copenhagen view is not sufficient.* In molecular and mesoscopic science we need a theory which is *universally valid in the whole molecular domain*, including systems of mesoscopic and macroscopic dimension, having both quantum and classical properties, and which can describe *individual dynamical processes* in an objective way.

Since the atomistic view of classical physics is very different from the holistic view of quantum mechanics, it is plain that the traditional notion of reality used in classical physics and the notion of reality required in quantum theory clash. But these notions only clash because philosophers were not careful enough in their attempts to give an explication what we could mean by realism. A number of views of traditional realist philosophy is incompatible with the results of modern science.

Many formulations of what realism asserts are so vague that it is difficult to evaluate their claims in the domain of science. Often such formulations are unnecessarily coupled with unfounded assumptions about the structure of the material world. For example, it has been said that in a realistic interpretation the theoretical terms genuinely refer (maybe in some approximate way) to objects existing in the world. Such a characterization is inadmissible since it makes a specific assumption about the *physical* structure of the world, namely that the world consists or is built out of well-defined and independently existing objects. From the viewpoint of modern quantum theory, any *a priori* identification of "material objects" (presumably tacitly supposed to be well-localized in physical space) with "material reality" is unacceptable, since–whatever the precise meaning of "material objects" may be–we have to expect that such systems are entangled by Einstein–Podolsky–Rosen-correlations, so that they have no individuality. Quantum mechanics does not describe 'things as they really are' since, according to this theory, there are no things in an absolute sense. Even macroscopic objects are correlated by Einstein–Podolsky–Rosen–correlations. A description corresponding to our inborn pattern-recognition mechanism and common-sense conceptions is possible only if such Einstein–Podolsky–Rosen–correlations are declared as irrelevant. Such a demand is not unreasonable because *without abstractions there is no science.* Every scientific description depends on the decision which effects we consider to be relevant and which effects we decide to ignore. Nevertheless, quantum mechanics allows a *contextual* realistic interpretation, provided we do not claim that matter is *made out* of elementary particles (like electrons). We have to use the more judicious formulation that the material reality can be described–under appropriate circumstances–in terms of elementary systems. Yet it is an *objective property of the material reality* that it can manifest itself under pertinent experimental conditions in a way that is best *described* in terms of elementary systems.

In scientific theories, the *problem of realism* is the question of the ontological status of the material reality while it is not observed. Since the existence of an external reality is not provable with the means available to science, we have to consider *realism as a purely metaphysical regulative principle*, free from any experimentally testable physical content, and without presupposing a particular compartmentalization of the material world. Furthermore, the investigation of the role of potential and actualized properties of physical objects is the business of physics, not of philosophy. In classical physics we are allowed to posit that all potential properties are always actualized but a priori there are no reasons to assume that such a convention is always logically possible.

The *scientific problem* is *not* to prove the existence of an independent reality, but to show that an appropriate regulative principle concerning a reality existing independently of human experience is *useful and compatible with the formalism of a fundamental scientific theory like quantum mechanics, together with all experimental results.* Moreover, the concept of realism should not be combined with structures taken over from classical physics or with specific physical ideas like atomism, localizability, separability, or determinism. I will adopt the following characterization of realism:

(i) *There exists a material world which is independent of our awareness of it.*

(ii) *Our knowledge of the material reality depends also on occurrences external to our consciousness.*

(iii) *Physical theories refer to some intrinsic aspects of the material reality.*

Note that in this characterization, realism does neither assert nor deny the existence of any kind of objects. Furthermore, it is not denied that *some* features of the observable aspects of the material reality may be due to our mental organization. In fact, we have to expect that common-sense descriptions of the outer world always depend on the psychic properties of the observer.

In the framework of theories which include the engineering domain, it is most reasonable to add the following regulative principle:

> *If a universally valid physical theory is restricted to the domain of engineering science, then the adopted realistic interpretation should cum grano salis give the every-day realism of the engineering world.*

Quantum mechanics of mesoscopic, macroscopic and open systems

Practical quantum mechanics–as used by the working scientists–is *not* based on a rigorously specified axiomatization but on some not too well defined 'first principles' and a bunch of working rules. The historical Hilbert-space formalism–as introduced by von Neumann [3] in his book of 1932–is limited to locally compact phase spaces. That is, this theory is restricted to strictly closed systems, and does not, for example, allow a mathematically proper description of the interaction of a charged particle with its electromagnetic field (which is a system having infinitely many degrees of freedom). As a consequence, the axiomatic Hilbert-space formalism does not include genuinely irreversible processes or the possibility of symmetry breakings. An important instance of the breakdown of a fundamental physical symmetry is the emergence of classical observables, that is, observables which commute with all observables and behave like observables in classical mechanics[2]. Von Neumann's Hilbert-space codification is based on the Stone–von Neumann uniqueness theorem for the representations of the canonical commutation relations. It is a simple corollary of this theorem that for finitely many degrees of freedom there exist no spontaneous symmetry breakings and no classical observables. No philosophical conclusions can be drawn from the fact that the traditional Hilbert-space codification cannot explain these features. The resolution is almost trivial: *The uniqueness theorem by Stone and von Neumann says that symmetry breakings and classical observables are impossible in this unnecessarily restricted codification.* That is, von Neumann's Hilbert-space formalism is not an adequate codification of quantum mechanics considered as a universally valid theory. Its straightforward generalization–the Fock-space quantum field theory–is theoretically inconsistent. Clearly, one should not try to conceive a realistic interpretation of quantum mechanics on the basis of a codification which is unable to explain mesoscopic and macroscopic physics. Fortunately, *there is no reason to identify quantum mechanics with the historical Hilbert-space or Fock-space codifications.*

If we consider quantum mechanics as *universally valid* in the atomic, molecular, mesoscopic and engineering domain, then we have to require that a proper mathematical codification of this theory must be capable to describe all phenomena of molecular and engineering science. Already rather small molecules can have classical properties, so that a classical behavior is *not* a characteristic property of large systems. The existence of molecular superselection rules and of molecular classical observables is an

[2]Note that the classical aspects of quantum systems have nothing to do with the limiting behavior when Planck's constant $\hbar$ can be regarded as "small". Classical quantum systems depend in an essential way on Planck's constant but nevertheless obey the laws of classical mechanics.

empirically well-known fact in chemistry and molecular biology. The chirality of some molecules, the knot type of circular DNA-molecules, and the temperature of chemical substances are three rather different examples of molecular classical observables. Such empirical facts can be described in an ad hoc phenomenological manner, but it is not so easy to explain these phenomena from the first principles of quantum mechanics. A universally valid theory of matter has not only to describe but also to *explain* why the chirality of biomolecules (like the L-amino acids, the D-sugars, lipids, or steroids) is a *classical* observable. The reality of this breakdown of the superposition principle of traditional quantum mechanics on the molecular level is dramatically demonstrated by the terrible Contergan tragedy which caused many severe birth defects. Contergan was the trade name of the drug thalidomide (3–phtalimido–2,6–dioxopiperidin, $C_{13}H_{10}N_2O_4$) which exists in two enantiomeric forms. The left-handed stereoisomer of thalidomide is a powerful and maybe safe tranquilizer, but the right-handed isomer is a teratogenic agent, causing disastrous physiological deformities in the developing embryo and foetus [4].

In the *engineering domain*, quantum theory must in principle be able to provide a description of measuring instruments and of our general experimental laboratory equipment. Therefore, a full-grown codification of quantum mechanics must include the successful engineering theories like classical point mechanics, chaotic nonlinear dynamical systems, continuum mechanics, hydrodynamics, classical stochastic processes, thermostatics including phase transitions, Maxwell's electrodynamics, Newton's gravitation. In the *mesoscopic domain* manifestations of both quantal and classical properties at one and the same object are nothing out of the ordinary, but they cannot be understood by some "correspondence rules"; their description requires a full-blooded theory which includes both traditional quantum mechanics *and* classical mechanics as special cases. For example, DNA-molecules–the material carrier of genetic information– possess important properties which definitely require a quantum-mechanical description, e.g. its photochemical reactivity. On the other hand, every DNA-molecule has a tertiary structure which is manifestly classical, and biologically important for the mechanism of genetic recombination. Moreover, circular DNA-molecules may be knotted, and there are enzymes which can change their knot-type. The knot-type of a DNA-molecule is an example for a classical property which cannot be explained by any variant of a "correspondence principle". Molecular biology is a rich source for such mixed quantal-classical systems. *Enzymes* act as molecular measuring devices and require a classical behavior for their function. The *immune system* is a molecular quantum system with an only classically describable memory, warranting the individual molecular iden-

tity. In order to understand such systems, one needs a theory of matter which can describe both the quantal and the classical properties of single individual objects. Since the cross-over from quantum to classical behavior is not given by Bohr's correspondence principle, one of the most important theoretical problems in molecular quantum mechanics is the correct analysis of the interaction of an individual, small, non-isolated quantum object with its environment and with classical degrees of freedom.

In every scientific investigation we divide the universe into an *object system* and its *environment*–which is all the rest. The environment acts as *background* which is indefensibly neglected in historical quantum mechanics. The idea of a physical object without an environment is an outrageous and incongruous abstraction. Eddington, in his posthumous book *Fundamental Theory*, called attention to the inevitability of considering the background: "The environment must never be left out of consideration. It would be idle to develop formulae for the behaviour of an atom in conditions which imply that the rest of matter of the universe has been annihilated. In relativity theory we do not recognise the concept of an atom as a thing complete in itself. We can no more contemplate an atom without a physical universe to put it in than we can contemplate a mountain without a planet to stand on" [5,p.13]. Therefore, the abstract structure of a tough-minded theory must be rich and complex enough to describe the essential features of the environment of an object under study.

A complete, mathematically rigorous and empirically correct theory of *open quantum systems* and of *mesoscopic* and *macroscopic quantum systems* is still a great desideratum, but it seems that most mathematical tools are available in terms of *algebraic quantum mechanics*. Algebraic quantum mechanics is *not a new*, but just a physically and mathematically correct formulation of quantum theory; it is nothing else but a proper codification of the basic principles of quantum mechanics. No ad hoc modifications, no hidden variables, and no quantization procedures are necessary. Algebraic quantum mechanics encompasses all kinds of physical systems, e.g. finite systems (with a locally compact phase space) and infinite systems (whose phase space is not locally compact). There is a dramatic difference between the behavior of finite and infinite systems. According to the uniqueness theorem by Stone and von Neumann, finite systems have a unique Hilbert-space representation while infinite systems have infinitely many *physically inequivalent* W^*-representations which account for the stupendous complexity of observable phenomena in nature.

In the framework of algebraic quantum mechanics, it can be proven that, in general, open quantum systems undergo symmetry breakings and possess *classical observables*. Contextual classical observables are *emergent* in the

sense that they are generated by the algebra of intrinsic observables together with a new contextual topology, but they are not functions of the intrinsic observables [6–10]. A typical example for an emergent classical observable is the *temperature* of systems in thermal equilibrium. It turns out that the contextual classical part of a dynamical quantum system is always a *stochastic* dynamical system, it depends in an essential way on Planck's constant but nevertheless obeys the laws of classical mechanics. In addition, the emergence of classical observables does *not* depend on the macroscopic character of the system under investigation, already rather small molecules can have classical properties.

Endophysical and exophysical descriptions

Certainly, present-day quantum mechanics is not the ultimate theory of matter. But even if we had a truly universal ultimate theory it would not give us all the information we need to describe an observed phenomenon. That is, the statement "universally valid" cannot be literally correct since a language which encompasses everything would have to be semantically closed, and hence engender antinomies. The impossibility of a complete description is not a flaw of the theory but a logical necessity. Every theory which attempts to describe its own means of verification is necessarily self-referential. In order to avoid paradoxes of self-reference, we need an at least two-leveled theory where the second level represents the metatheory which must be formulated in another language, a so-called metalanguage. This metalanguage has to be essentially richer than the language of the basic physical theory. If the two languages would be identical (or translatable into each other) we would have a semantically closed language with self-referential sentences [11, 12]. So we have to split the world into two parts, the observed part and the observing part. Our description depends on this cut but *this cut cannot be derived from any kind of an ultimate theory.* Hence the language of a hypothetically posited universal theory can at most describe a *part* of the full reality, perhaps even only a tiny area. Traditionally, the physical sciences exclude the subject of cognizance from their enquiry. No known physical theory deals with the reality of man in his freedom.

In the following I adopt the working hypothesis that quantum mechanics in its algebraic codification is *universally valid* in the atomic, molecular, mesoscopic and non-cosmological macroscopic domain. Our confidence in the trustworthiness of quantum mechanics as a fundamental physical theory is in an essential way based on its confirmation by laboratory experiments. That is, both the validity of engineering physics and the feasibility of experimenters having free will is *presupposed*, and not derived from the first

principles of quantum theory. I do not assume that consciousness or free will can be reduced to physical properties of the organism such as brain states. All ideas of choice and purpose must be included in the relevant *regulative principles* which are not derivable from physical first principles. Clearly, such regulative principles play a central role in our picture of the world. These postulates lead to the necessity to distinguish between endophysics and exophysics. This helpful distinction has been made by Otto Rössler [13] and David Finkelstein [14, 15]. Probably misusing their ideas, I adopt nevertheless their way of speaking:

> *A strictly closed physical system without any concept of an observer is called an endosystem.*
>
> *If the endoworld is divided into an observing and an observed part, we speak of an exophysical description.*
>
> *The world of the observers with their communication tools is called an exosystem.*

Note that endophysics is different from exophysics. *All fundamental universally valid first principles we know refer to strictly closed systems, hence belong to endophysics.* They are supposed to be universally valid, but they are not operational. Strictly speaking, there is nothing outside an endosystem. The endophysical description is a view without perspective, it is God's panorama, a *"view from nowhere"*.

Already the *formalism* of quantum mechanics predicts that quantum systems like electrons, atoms or molecules are always *entangled* with the rest of the world, so they cannot be possible candidates for individual entities which "really exist". Provided we accept quantum theory as a holistic theory, *a consistent variant of scientific realism cannot postulate an independent existence of building blocks like strings, quarks, electrons, atoms or molecules. We construct* building blocks to describe matter from a particular point of view, but the world is *not made out of* some building blocks. This insight is not in contradiction with the view that quantum mechanics is a story about what there really is. Objectivity does not reside in transcendental entities like molecules, atoms, electrons or quarks, these are just *manifestations* of the material reality. On a fundamental level, we have to emphasize different aspects like symmetries.

First principles are not natural laws but fundamental ideas. To a certain extent it is a matter of taste what we consider as first principles and what as pragmatic working rules. As far as possible and appropriate, first principles should be context-independent. For that reason, first principles are

always extravagantly remote from our every-day experience. All popular first principles refer to situations with high intrinsic symmetry. Experience tells us that symmetry is an effective criterion for selecting first principles so we adopt the view that *maximal symmetry* is a typical characteristic of an *endophysical first principle.* Such fundamental symmetries are, as a rule, not manifest in the everyday domain. So it is necessary to break these symmetries, as clearly recognized by Pierre Curie [16]: "C'est la dissymétrie qui crée le phénomène". That is, genuine endophysical symmetries are directly inaccessible by experience, they can empirically be found only by exophysical symmetry breakings. On that account we consider all laws or rules showing *broken symmetries* to be contextual and belonging to a particular *exophysical description.* For example, for endophysics we posit a bidirectional deterministic time evolution distinguished by a time-inversion symmetry, while the *arrow of time* of most exophysical descriptions manifests a broken time-inversion symmetry.

Quantum *endo*physics cannot predict what happens in a physical experiment, since in an endoworld there is not yet any concept of observing tools or observers. It is a strictly deterministic theory, set up to describe the reality existing independently of human observations. Note that the fact that quantum endophysics is deterministic does not imply that it is *determinable* by an internal or an external observer. The *endophysical description refers to an immanent ontology,* it pictures an *independent reality* in a non-operational way. Every operational description of the world requires the transition from the *endophysical* to an *exophysical* description by introducing a cut between the observed and the observing part. The *exophysical description* refers to the *empirical reality* in the sense of d'Espagnat [17, 18]. Yet, *the endophysical first principles are not sufficient for a characterization of exosystems* since every exophysical description depends not only on first endophysical principles but also on the choice of the cut. This fact does not imply that we cannot go from endophysics to an exophysical description, but that for such an enterprise we need additional *regulative principles.* Every exophysical description is therefore *contextual* and at most *weakly objective* (in the sense of an intersubjective agreement of observers choosing the same cut).

The inverse problem is building up a picture of the world independent of the perceiving subject from experimental data, or in our terminology, a logically consistent reconstruction of conjectured endophysics from the operationally accessible exophysical descriptions. The theoretical construction of an endophysically immanent ontology can be considered as a *realization problem.* That is, we are asking for an ontically interpreted theoretical structure which, together with appropriate regulative principles, allows us to *derive* all legitimate exophysical description of all aspects of the material reality

encompassed by the basic theory. This realization problem is, in the main, a *consistency problem.* If it has a solution, it has many solutions. We can reduce this nonuniqueness by some *minimality requirements* (Ockham's razor) and by adopting an ontology whose restriction to the engineering domain gives the realism almost universally adopted in classical physics. Therefore, endophysics never can be a literally true story of what the world is like. An endophysical conception of reality must be compatible but cannot be derived from empirical data. In the words of Albert Einstein: " 'Being' is always something which is mentally constructed by us, that is, something which we freely posit (in the logical sense). The justification of such constructs does not lie in their derivation from what is given by the senses. Such a type of derivation (in the sense of logical deducibility) is nowhere to be had, not even in the domain of pre-scientific thinking. The justification of the constructs which represent 'reality' for us, lies alone in their quality of making intelligible what is sensorily given ..." [19, p. 669].

On interpretations

An interpretation always refers to a logically consistent and empirically well-confirmed theoretical formalism. That is, we assume that we have a mathematically rigorous codification of a physical theory (the 'formalism'), a minimal interpretation of the theory which allows an operationalization and an empirical verification of the theoretical predictions. We adopt the following definition:

> *An interpretation of a physical theory is characterized by a set of normative regulative principles which can neither be deduced nor be refuted on the basis of the mathematical codification and the minimal interpretation.*

Since theories are not determined by their empirical consequences, we have some freedom for choosing an interpretation. First of all, we distinguish between epistemic and ontic interpretations. *Epistemic interpretations* refer to our knowledge of the properties or modes of reactions of systems "as we perceive them", while *ontic interpretations* refer to the properties of the "object in itself", regardless of whether we know them or not, and independently of any perturbations by observing acts. An ontic interpretation of quantum mechanics makes assertions about values *possessed* by observables. A realistic world view demands an individual ontic interpretation of *quantum endophysics*, it is intrinsically objective but not operational. The operationalistic view requires an exophysical epistemic interpretation,

and usually works with a statistical description. By a proper choice of the regulative principles, one can get a contextually objective and operational *exophysical* description of quantum reality.

To be sure, an ontic interpretation of quantum mechanics does refer only to a fictitious *theoretically immanent reality*, and not to the *ultimate reality*. But under the *working hypothesis*–which nobody really believes–*that quantum mechanics is a universally valid theory*, an ontic interpretation allows us a consistent way of speaking *as if* we would refer to reality.

Individual and statistical descriptions of quantum systems

Both individual and statistical descriptions of material reality are possible, but the appropriate mathematical formulations are fundamentally different. Moreover, a coherent statistical interpretation requires an individual interpretation as a backing. In classical theories this requirement is automatically fulfilled since the convex set of all statistical states is a simplex so that a *unique* decomposition of every mixed state into pure states is warranted. In quantum theories, a mixed state has many feasible realizations in terms of pure states so that it is not at all clear what the *conceptual* meaning of a statistical state is. On the other hand, a complete individual interpretation is always in terms of *ontic* states, mathematically described by pure states. The solution of the equation of motion for this pure state requires a knowledge of the initial conditions of all degrees of freedom of the whole environment. From an experimental point of view, this information is never available so that we are forced to introduce an *epistemic* state by some kind of optimal estimate of the initial conditions of the environment. This procedure leads to a well-defined mixture in terms of ontic states, hence to a conceptually well-defined statistical state. These statistical states are epistemic states, *they refer to our knowledge of the ontic state.*

The usual mathematical formalism of quantum mechanics refers to a *statistical* description, and one would be ill-advised to use this mathematical formalism also for the individual description. *The mathematical formalism required for an individual description is different from the formalism required for a statistical description.* In classical point mechanics, the usual individual description is given in terms of a symplectic phase space Ω, where the individual state of the system at time t is given by a point ω_t of Ω. According to Gelfand's representation [20, p.16], there is a one-to-one correspondence to the algebraic description in terms of the C^*-algebra $\mathcal{C}^\infty(\Omega)$ of continuous functions on Ω which vanish at infinity. In this algebraic description the individual states are given by the extremal elements of the dual of $\mathcal{C}^\infty(\Omega)$. The statistical description of the same mechanical system can be formulated

in terms of probability densities, that is of positive and normalized elements of the Banach space $L^1(\Omega)$. The dual of this Banach space is the W^*-algebra $L^\infty(\Omega)$ of bounded Borel-measurable functions on Ω, and is called the algebra of bounded observables. Just as in classical point mechanics, the individual description of an arbitrary quantum system can be given in terms of an appropriate separable *C^*-algebra* $\mathcal{A}$, where the individual states are represented by the extremal elements of the dual $\mathcal{A}^*$ of $\mathcal{A}$. The statistical description of a quantum system has to be given in terms of an appropriate *W^*-algebra* $\mathcal{M}$ with a separable predual $\mathcal{M}_*$. In quantum mechanics the algebras $\mathcal{A}$ and $\mathcal{M}$ are in general noncommutative. In the special case of commutative algebras we speak of classical quantum systems, and we can represent these algebras as in historical classical mechanics by $\mathcal{A} = \mathcal{C}^\infty(\Omega)$ and $\mathcal{M} = L^\infty(\Omega)$, where $\mathcal{M}_* = L^1(\Omega)$.

Ontic interpretation of endo–quantum mechanics

While quantum phenomena require a radical revision of our ideas about physical reality, they do not prevent us from accepting a reasonable realistic individual interpretation. For this we do not require any kind of hidden variables, faster-than light influences, or an exotic continuously splitting many-worlds description. Quantum mechanics does not force us to give up realism, but it forces us to distinguish carefully between *potential* and *actualized* properties. It is a misconception (though one surprisingly widespread among philosophers and scientists) that physical quantities have to be truth-definite. A popular working rule of pragmatic quantum mechanics says that "an observable has *no value* before a measurement"[3]. This is in contrast to the usual metaphysical commitment of classical mechanics that every observable *has* a value at all times. This commitment cannot be transferred to quantum mechanics since there is a theorem saying that for a full set[4] of states of a C^*-algebra $\mathcal{A}$, a hypothetical attribution of definite truth values to *all* elements of $\mathcal{A}$ requires that $\mathcal{A}$ is commutative[5]. However, instead of a positivistic renouncement we can adopt the intrinsic, internally consistent

[3]Of course, a positivist would not say so much. For example, Reichenbach adopts the following definition: "In a physical state not preceded by a measurement of an entity u, any statement about a value of the entity u is meaningless" [21].

[4]A set $\mathcal{S}$ of states on a C^*-algebra $\mathcal{A}$ is said to be *full* if an element A of $\mathcal{A}$ satisfies $A \geq 0$ if and only if $\rho(A) \geq 0$ for all $\rho \in \mathcal{S}$.

[5]The relevant basic theorem is due to Misra [22]: A C^*-algebra $\mathcal{A}$ (different from the complex numbers) admits a dispersion-free state if and only if it has a nontrivial norm-closed two-sided ideal $\mathcal{I}$ such that the quotient algebra $\mathcal{A}/\mathcal{I}$ is commutative. This theorem implies that in traditional quantum mechanics there are no states which are dispersion-free for *all* observables.

ontic interpretation that at every instant there is a maximal set of truth-definite observables. A truth-definite observable possesses a value–whether we know this value or not, is at this stage of the theoretical discussion entirely irrelevant. This point of view corresponds exactly to the usual interpretation of classical point mechanics, where the ontological question of 'having a value' is clearly separated from the entirely different question how to get empirically some information about this value.

The natural referent for quantum endophysics is a *single system.* A *statistical* interpretation of quantum mechanics presupposes the existence of an *external* measuring system with a *classical irreversible dissipative* behavior, so that it is a topic of quantum *exo*physics. Therefore, a statistical interpretation of quantum endophysics makes no sense, but a non-operational and *intrinsically nonprobabilistic individual ontic interpretation* is possible in a logically consistent way. Algebraic quantum mechanics allows to give a precise definition of an ontic interpretation which is free of inner contradictions. In algebraic quantum mechanics, quantum endophysics is characterized by a C^*-algebra $\mathcal{A}$ of intrinsic observables. The referent of an endophysical ontic interpretation of quantum mechanics is the whole universe of discourse. The *intrinsic potential properties* describe independently of any observation what is physically real, they are represented by the selfadjoint elements of the C^*-algebra $\mathcal{A}$ of intrinsic observables. The *intrinsic ontic state* of an object at time t is characterized by the set of all intrinsic potential properties which are actualized at the instant t. That is, *the intrinsic potential properties characterize the object, while the actualized intrinsic properties characterize the ontic state of the object.* An ontic state can be represented by a positive linear functional and is characterized by the fact that there are no other linear functionals with the same collection of actualized observables. It can be proved that there is a one-to-one correspondence between the ontic states of an object and the *extremal*, normalized positive linear functionals on $\mathcal{A}$ (the so-called 'pure states').

Mathematical supplement

A selfadjoint operator $A \in \mathcal{A}$ is said to be *dispersion-free* with respect to a state $\rho \in \mathcal{A}^*$ if $\rho(A^2) = \rho(A)^2$. In this case, the observable A is said to *possess* the value $\rho(A)$ with respect to a state ρ. The set of all observables on which a state $\rho \in \mathcal{A}^*$ is dispersion-free, is called the *definite set* $\mathcal{D}_\rho$ of ρ [23],

$$\mathcal{D}_\rho := \{A \in \mathcal{A} \mid A = A^*,\ \rho(A^2) = \rho(A)^2\}.$$

The complex span $\mathcal{A}_\rho$ of the definite set $\mathcal{D}_\rho$

$$\mathcal{A}_\rho := \{A + iB \mid A, B \in \mathcal{D}_\rho\}$$

is a C^*-algebra with the property [24]

$$\mathcal{A}_\rho := \{A \in \mathcal{A} \mid \rho(AB) = \rho(BA) = \rho(A)\rho(B) \text{ for all } B \in \mathcal{A}\}.$$

We require that $\mathcal{A}_\rho$ is a maximal set of observables which at some instant t possess values, that is we require that the definite set $\mathcal{D}_\rho$ is maximal in the sense that

$$\mathcal{D}_\rho \subseteq \mathcal{D}_\varphi \text{ for some state } \varphi \in \mathcal{A}^* \text{ implies } \rho = \varphi.$$

If $\mathcal{A}$ is a C^*-algebra with identity and with no one-dimensional representation, then a state ρ on $\mathcal{A}$ is pure if and only if its definite set $\mathcal{D}_\rho$ is maximal [25].

The ontic interpretation of a dynamical C^*-system presupposes that at every instant $t \in \mathbb{R}$ there is a maximal definite set $\mathcal{D}_t$ of observables. The corresponding complex span $\mathcal{A}_t \subseteq \mathcal{A}$ defines a unique C^*-homomorphism $\rho_t : \mathcal{A}_t \longrightarrow \mathbb{C}$ which we interpret as a valuation map for the observables that are actualized at the instant t. Any observable $A \in \mathcal{A}_t$ possesses at time t the dispersion-free value $\rho_t(A)$. The functional ρ_t has a unique state extension to an extremal, normalized positive linear functional on the C^*-algebra $\mathcal{A}$ [24]. This uniquely given pure state is called the *ontic state* of the C^*-system at the instant t.

That is, ontic states are represented by (and identified with) pure states. It follows that an intrinsic potential property represented by an observable $A \in \mathcal{A}$ is actualized at time t if and only if $\rho_t(A^2) = \{\rho_t(A)\}^2$ where the extremal normalized positive linear functional $\rho_t \in \mathcal{A}^*$ represents the *ontic state* at time t. This delineation fixes the *ontology* of quantum endophysics. Our reference to an *independent reality* makes only sense as a *theoretical construct*. The *intrinsic ontic interpretation* is a *strongly objective* theory in the sense of d'Espagnat [18] since in the first place it makes no reference to observers or probabilities. It may describe reality in itself *but not the phenomena we observe*. The restriction of this ontic interpretation of algebraic quantum mechanics to the classical part of the system[6] corresponds to the generally adopted realistic individual interpretation of the traditional classical physical theories. Hence the adopted immanent ontology is not radically different from the ontology traditionally accepted for classical physical theories.

[6]The classical part of a C^*-system with the C^*-algebra $\mathcal{A}$ is given by the center $\mathcal{Z}(\mathcal{A})$ of $\mathcal{A}$. The C^*-system with the commutative C^*-algebra $\mathcal{Z}(\mathcal{A})$ is a classical quantum system.

Epistemic interpretation of exo–quantum mechanics

A theory which describes observable phenomena cannot keep the human means of data processing out of consideration, but these means are not described by the C^*-algebra of intrinsic observables. The observables which describe the outcomes of measurements are context-dependent, they are represented by *positive operator-valued measures* of the W^*-algebra $\mathcal{M}$ of *contextual observables.* This algebra is not intrinsically given but can be *constructed* from the context-independent C^*-algebra $\mathcal{A}$ by a faithful Hilbert-space representation $\pi(\mathcal{A}) \subseteq \mathcal{B}(\mathcal{H})$ of $\mathcal{A}$ by specifying a new contextual topology by selecting a *folium of contextually preferrend intrinsic states.* The weak closure of the C^*-algebra $\pi(\mathcal{A})$ acting on the Hilbert space $\mathcal{H}$ is W^*-isomorphic to the W^*-algebra $\mathcal{M}$ of contextual observables. In this contextual description, the statistical states are represented by the *normal* positive linear functionals on the W^*-algebra $\mathcal{M}$.

The W^*-algebraic formalism describes the *empirical reality,* it is context-dependent hence only *weakly objective,* in the sense that for a given context there is intersubjective agreement[7]. While the nonoperational individual and ontic interpretation is *fully deterministic* and intrinsically richer than an exophysical statistical description, any of the possible operational exophysical statistical descriptions is necessarily contextual but without exceptions *irreducibly probabilistic.* The primary probabilities of quantum mechanics [26] manifest themselves only in the interaction with *external classical systems.*

Our ability to describe the world cannot go farther than our ability to isolate objects. A realistic operational description of quantum systems is possible if and only if there are no Einstein–Podolsky–Rosen–correlations between the object system and the observing system. Only if we can abstract deliberately from these factually existing Einstein–Podolsky–Rosen–correlations, we can investigate the material world by compartmentalization. A realistic description of an individual quantum system is possible if and only if there are no Einstein–Podolsky–Rosen–correlations between the object system and its environment. Therefore I adopt the following definition of an object [27–32]:

> *An object is defined to be an open quantum system, interacting but not Einstein–Podolsky–Rosen–correlated with the environment.*

It follows that objects are exactly those quantum systems for which at every

[7]The same is true for the quantum-logics approach. The corresponding orthomodular lattice is given by the projection lattice of the contextual W^*-algebra. A representation-independent description (corresponding to the C^*-algebra of intrinsic observables) does not exist in quantum logics.

instant a maximal description in terms of pure states is possible. *An object is something having individuality and potential properties*, so that we can interpret a pure quantum state of an object as an individual state. Here the notion of an 'individual state' refers to a mode of being, describing exophysical characteristics existing independently of any observation, while the notion of a 'pure state' refers to a merely mathematical concept, meaning an extremal positive linear functional on the algebra of observables. Note that the exophysical individual state depends on the breaking of the holistic symmetry of the world by division and abstraction. Over and above, every exophysical description requires a *tensor-product decomposition* but such a decomposititition is not God-given. The usual Hamiltonian tensor-product structure refers to *bare* particles and to *bare* fields whereas the object–environment tensor-product structure refers to contextual *dressed* entities. A contextual quantum object appears as an object *not in spite*, but *because* it interacts with its environment. In particular, classical properties are the result of the interaction of an object with its environment. *Without an appropriate background the concept of a quantum object makes no sense.*

It would be unreasonable to expect that the dynamics of an exosystem is governed by a Hamiltonian or a bidirectionally deterministic time evolution. This dynamics cannot be postulated but has to be derived from the intrinsic endophysical time evolution. In an exophysical description, it is in principle possible to eliminate the environmental variables and to write down the dynamics of an individual object in terms of the object observables alone. In general, this reduced dynamics is given by a *stochastic* and *state-dependent* equation of motion. Both the stochastic behavior and the state-dependence have not to be put in by hand, but they can be *derived* from the fundamental *linear* endophysical dynamics. The chaotic behavior arises from the initial values of the unobserved degrees of freedom of the environment, resulting in a stochastic classical force acting on the object. If the spectral distribution of the autocorrelation of this force is absolutely continuous, then the environment forgets the initial conditions completely so that the stochastic force is usually *completely nondeterministic*. The state-dependence is due to feedback effects from the polarization of the environment by the quantum object. If the dynamics of this individual quantum object can be represented in terms of the irreducible Hilbert-space formalism, then the dynamics of the ontic state can be representend by a trajectory $\Psi \mapsto \Psi_t$ of the state vector whose time evolution is given by a *nonlinear stochastic integro-differential equation* for the state vector Ψ_t. In particularly simple models, one gets a nonlinear stochastic Schrödinger equation in the sense of Itô.

All objects we discuss in empirical science are *contextual objects*, their existence depends both on the environment, and on the abstractions we

are forced to make in every scientific discussion. It is a theorem of algebraic quantum mechanics that an object exists only if its environment is classical[8]. The meaning of the notion 'classical' depends, however, on our abstractions and is therefore context-dependent. That is, in a quantum world there are no intrinsic context-independent objects besides the whole universe of discourse. Contextual objects are abstraction-dependent, but they are not free inventions. They represent *patterns of reality,* yet they are *not* building stones of reality. Elementary or composed "particles" like electrons, atoms or molecules are not primary but rather secondary and derived. Electrons, atoms or molecules do not simply *exist,* they appear only under special conditions–they are *contextual* systems.

In order to go from the universally valid endophysical description to a contextual exophysical description, one has to introduce in addition *regulative principles like* the Baconian rejection of the existence of final processes, our presupposed freedom to create initial conditions, or the feasibility of "detached observers". The chosen observational tools determine a certain context which in algebraic quantum mechanics is characterized by a *new topology in the space of the intrinsic states*[9]. An exophysical description of contextual objects cannot give us complete knowledge of the endophysical independent reality. Contextual objects depend on the contextually selected topology but are independent of a human consciousness, *they are real relative to the chosen context.* An exophysical description is neither absolutely true nor absolutely false, but we may say that it is correct *relative to the chosen way of describing reality.* Yet exophysical descriptions are not unique, they depend on the neglect of some really existing Einstein–Podolsky–Rosen–correlations. Therefore there are always different exophysical descriptions which according to purely endophysical criteria are logically equivalent. No single exophysical description reveals the whole independent reality with its non-Boolean event structure but projects *some aspects* of this reality onto a *Boolean context.* The material reality has many complementary Boolean descriptions, each being valid from its own perspective. There is only one reality, yet there are many legitimate viewpoints, hence many equally legitimate but complementary descriptions of nature.

[8] *Theorem: Let $\mathcal{A}$ and $\mathcal{B}$ be two C^*-algebras and $\mathcal{C} = \mathcal{A} \otimes \mathcal{B}$ their minimal tensor product. Every pure state γ on $\mathcal{C}$ is of the form $\gamma = \alpha \otimes \beta$ for some pure states α of $\mathcal{A}$ and β of $\mathcal{B}$ if and only if either $\mathcal{A}$ or $\mathcal{B}$ is commutative* ([20], theorem 4.14). This theorem implies that a nonclassical open C^*-system is an object if and only if its environment is classical. Clearly, every classical C^*-system is an object.

[9] This new topology is different from the intrinsic C^*-topology and can also be characterized by a *folium* of preferred states which in turn characterize the *normal* states of the W^*-closure of the associated Hilbert-space representation.

Conclusions

Except from the fact that present-day physics has nothing to say about the relation between matter and spirit and is not in the position to avoid the Cartesian split, one of the most important open problems of nonrelativistic quantum theory is the proper description of individual open quantum objects in interaction with their environment. This is mainly a problem of *mathematical physics*, not of philosophy. If we are able relinquishing untenable presuppositions and if we accept the holistic structure of the material reality, the philosophical problems associated with quantum mechanics are not radically different from those of science in general. It is not realism that is refuted by quantum mechanics, but atomism and the idea of the existence of context-independent objects. The context-dependence of every description of reality is inevitable, even in classical physics; it is enforced by Tarski's theorem which implies the necessity of an *exophysical metalanguage*. Due to entanglement effects, individual quantum objects are always abstraction-dependent entities. Contextual objects represent *patterns of reality,* yet they are *not* building stones of an independent reality. According to quantum theory, *a consistent variant of scientific realism cannot postulate an independent existence of building blocks like quarks, electrons, atoms or molecules.* The non-Boolean event structure of quantum reality forces us to give up the classical idea that all potential properties of a quantum object can be actualized at the same instant. The nonseparability and nonlocality of the material world are not compatible with the ontology adopted in classical physics. Due to its holistic nature, quantum reality is more elusive and leads to an amazing variety of complementary descriptions.

References

[1] PAULI, W. (1956). *Die Wissenschaft und das abendländische Denken.* In: Europa - Erbe und Aufgabe. Internationaler Gelehrtenkongress, Mainz 1955. Ed. by M. Göhring. Wiesbaden. Franz Steiner Verlag. Pp. 71–79.

[2] BOHR, N. (1955). *The unity of knowledge.* In: The Unity of Knowledge. Ed. by L. Leary. New York. Doubleday.

[3] NEUMANN, J. VON (1932). *Mathematische Grundlagen der Quantenmechanik.* Berlin. Springer.

[4] DECAMP, W. H. (1989). *The FDA perspective on the development of stereoisomers.* Chirality 1, 2–6.

[5] EDDINGTON, A. S. (1946). *Fundamental Theory.* Cambridge. Cambridge University Press.

[6] AMANN, A. and U. MÜLLER–HEROLD. (1986). *Momentum operators for large systems.* Helv. Phys. Acta 59, 1311–1320.

[7] AMANN, A. (1987). *Broken symmetries and the generation of classical observables in large systems.* Helv. Phys. Acta 60, 384–393.

[8] AMANN, A. (1988). *Chirality as a classical observable in algebraic quantum mechanics.* In: Fractals, Quasicrystals, Chaos, Knots and Algebraic Quantum Mechanics. Ed. by A. Amann, L. Cederbaum and W. Gans. Dordrecht. Kluwer. Pp. 305–325.

[9] PÖTTINGER, J. (1989). *Global quantities in algebraic quantum mechanics of infinite systems: Classical observables or parameters?* J. Math. Phys. 30, 361–368.

[10] AMANN, A. (1991). *Chirality: A superselection rule generated by the molecular environment?* J. Math. Chem. 6, 1–15.

[11] TARSKI, A. (1956). *Logic, Semantics, Metamathematics.* Oxford. Clarendon Press.

[12] TARSKI, A. (1969). *Truth and proof.* Scientific American 220, 63-77.

[13] RÖSSLER, O. E. (1987). *Endophysics.* In: Real Brains, Artificial Minds. Ed. by J. L. Casti and A. Karlqvist. New York. North-Holland. Pp. 25–46.

[14] FINKELSTEIN, D. and S. R. FINKELSTEIN. (1983). *Computational complementarity.* Int. J. Theor. Phys. 22, 753–779.

[15] FINKELSTEIN, D. (1988). *Finite physics.* In: The Universal Turing Machine. A Half-Century Survey. Ed. by R. Herken. Hamburg. Kammerer & Unverzagt. Pp. 349–376.

[16] CURIE, P. (1894). *Sur la symétrie dans les phénomènes physiques.* Journal de Physique 3, 393. Reprinted in: Oeuvres de Pierre Curie. 1908. Paris. Gauthier-Villars. Pp. 118–141.

[17] ESPAGNAT, B. DE (1986). *Physics and reality.* In: Atti del Congresso "Logica e Filosofia della Scienza, oggi". Vol. II. Epistemologia e logica induttiva. Bologna. CLUEB. Pp. 73–78.

[18] ESPAGNAT, B. DE (1989). *Reality and the Physicist. Knowledge, Duration and the Quantum World.* Cambridge. Cambridge University Press.

[19] EINSTEIN, A. (1949). *Reply to criticism.* In: Albert Einstein: Philosopher–Scientist. Ed. by P. A. Schilpp. Evanston, Illinois. Library of Living Philosophers. Pp. 665–688.

[20] TAKESAKI, M. (1979). *Theory of Operator Algebras I.* New York. Springer.

[21] REICHENBACH, H. (1944). *Philosophic Foundations of Quantum Mechanics.* Los Angeles. University of California Press.

[22] MISRA, B. (1967). *When can hidden variables be excluded in quantum mechanics?* Nuovo Cimento A 47, 841–859.

[23] KADISON, R. V. and I. M. SINGER. (1959). *Extensions of pure states.* Amer. J. Math. 81, 383–400.

[24] ANDERSON, J. (1979). *Extensions, restrictions, and representations of states on C^*-algebras.* Trans. Amer. Math. Soc. 249, 303–329.

[25] STØRMER, E. (1968). *A characterization of pure states of C^*-algebras.* Proc. Amer. Math. Soc. 19, 1100–1102.

[26] PAULI, W. (1954). *Wahrscheinlichkeit und Physik.* Dialectica 8, 112–124.

[27] PRIMAS, H. (1980). *Foundations of theoretical chemistry.* In:Quantum Dynamics of Molecules. The New Experimental Challenge to Theorists. Ed. by R. G. Woolley. New York. Plenum Press. Pp. 39– 113.

[28] PRIMAS, H. (1981). *Chemistry, Quantum Mechanics and Reductionism.* Berlin. Springer, second edition 1983.

[29] PRIMAS, H. (1987). *Contextual quantum objects and their ontic interpretation.* In: Symposium on the Foundations of Modern Physics, 1987. The Copenhagen Interpretation 60 Years after the Como Lecture. Ed. by P. Lahti and P. Mittelstaedt. Singapore. World Scientific. Pp. 251–275.

[30] PRIMAS, H. (1990). *Mathematical and philosophical questions in the theory of open and macroscopic quantum systems.* In: Sixty-Two Years of Uncertainty: Historical, Philosophical and Physical Inquiries into the Foundations of Quantum Mechanics. Ed. by A. I. Miller. New York. Plenum. Pp. 233–257.

[31] PRIMAS, H. (1990). *Realistic interpretation of the quantum theory for individual objects.* La Nuova Critica 13-14, 41–72.

[32] PRIMAS, H. (1991). *Necessary and sufficient conditions for an individual description of the measurement process.* In: Symposium on the Foundations of Modern Physics 1990. Quantum Measurement Theory and its Philosophical Implications. Ed. by P. Lahti and P. Mittelstaedt. Singapore. World Scientific. Pp. 332–346.

Logic, Methodology and Philosophy of Science IX
D. Prawitz, B. Skyrms and D. Westerståhl (Editors)

SOME REFLECTIONS ON THE STRUCTURE OF OUR KNOWLEDGE IN PHYSICS

HOWARD STEIN

The University of Chicago

I want to use this occasion to make some remarks of a rather general kind about the character of our knowledge in physics. I do this with some diffidence: I run the risk that what I shall say may seem—may indeed be—largely platitudinous. But there are some points that seem to me important, even if obvious; and also seem, even if obvious, to be not widely recognized, or not held firmly in view, in current philosophical discussion; so I have decided that the risk is worth taking.

That already has something of the air of the introduction to a sermon; sanctimoniousness may be another risk. But let me nevertheless hazard a rough diagnosis of the reason why some things that are (in my view) true, important, and obvious tend to get lost sight of in our discussions. I think "lost sight of" is the right phrase: it is a matter of perspective, of directions of looking and lines of sight. As at an earlier time philosophy was affected by a disease of system-building—the *ésprit de système* against which a revulsion set in toward the end of the last century—so it has (I believe) in our own time been affected by an excess of what might be called the *ésprit de technique*. I see this as having two chief kinds of manifestation. One has to do with details: a tendency both to concentrate on such matters of detail as allow of highly formal systematic treatment (which can lead to the neglect of important matters on which sensible even if vague things can be said),[1] and (on the other hand), in treating matters of the latter sort, to subject them to quasi-technical elaboration beyond what, in the present state of knowledge, they can profitably bear. The second principal manifestation lies in the way we treat the efforts of our forebears and contemporaries: namely, we often discuss their work less in the hope of drawing instructive insight from it than

[1] Wittgenstein's famous aphorism, "Was sich überhaupt sagen läßt, läßt sich klar sagen," although inspiring is unfortunately false; for the maxim he bases on it I would propose a more modest one: not "Wovon man nicht reden kann, darüber muß man schweigen," but "Wovon man nichts beleuchtendes zu sagen findet, darüber schweige man lieber!"

as a source of doctrines to analyze, contrast, elaborate, or destroy—in any case, to serve as material for the further exercise of technique.

Of course I do not think that such is the deliberate practice of philosophers; nor do I intend to devote this talk to the presentation of an indictment. But let me say just a little more about the matter in general; for it seems to me to present certain instructive ironies.

In the first place, what I have described can be characterized rather precisely as a species of scholasticism—which is about as far as may be imagined from what the advocates of a new spirit of philosophy intended to stimulate. In so far as the word "scholasticism," in its application to medieval thought, has a pejorative connotation, it refers to a tendency to develop sterile technicalities—characterized by ingenuity out of relation to fruitfulness; and to a tradition burdened by a large set of standard counterposed doctrines, with stores of arguments and counterarguments. In such a tradition, philosophical discussion becomes something like a series of games of chess, in which moves are largely drawn from a familiar repertoire, with occasional strokes of originality—whose effect is to increase the repertoire of known plays. This was especially unfortunate in the later middle ages, when (in particular in natural philosophy) potentially very fruitful new ideas were introduced—which, however, remained as mere curiosities among the opinions of the commentators on the physics of Aristotle.[2]

On the other hand, what I am speaking of can also be regarded as itself a kind of *ésprit de système*: "local," one might say, and technical, in contrast with the global and "romantic" mentality of the nineteenth-century "systems." Among the unfortunate results of such practice is the frustration of that hope which so signally characterized our predecessors earlier in this century: the hope for a cumulative and progressive philosophy, to the advance of which many workers would contribute in collaboration among contemporaries and development by successors. Of course, in reaction to that hope of our predecessors, it is now vigorously contended in some quarters not only that the hope for philosophy was a delusion, but that science itself lacks the cumulative and progressive character that had been presumed for it. These doctrines, so far as they concern science, seem to me absurd; I shall therefore not say very much on the subject—although I shall say a little, because the absurdity of such a view of science is one of the things I consider important even if obvious. But if it is conceded that science makes progress, it might still be questioned whether philosophy does or can. One of the main theses I want to defend, with examples, is that philosophy indeed *has* made *very*

[2]Cf. on this subject the very instructive account in Clavelin 1968, ch. 2; especially Clavelin's evaluation of the general character and limitations of the natural philosophy of the fourteenth-century schools of Oxford and Paris (pp. 121ff.).

important progress, but that this is seriously obscured by what I have called problems of perspective.

Now, philosophical progress that is not recognized as such by practising philosophers is clearly progress of a precarious sort; so—acknowledging that my diagnosis may be incorrect—I hope it will seem at least forgivable, believing as I do, that I should come to you to proclaim that the sky is not falling.

The first serious platitude I want to present is this: If Wittgenstein's early standard of clarity is impossible to meet; if the hopes of the logical empiricists for a philosophy built up with the rigor and exactness of mathematics upon a basis that is—if not entirely secure epistemologically—at least entirely precise in both structure and content, have failed; and if nonetheless we do not wish to abandon the attempt to achieve such clarity as is possible, or wish to abstain from the use of rigorous techniques where they are fruitful, then there is an obvious rough distinction that we ought never to lose sight of in philosophical work: namely, what I shall just call the distinction between *presystematic* and *systematic* considerations. Accordingly, I emphasize now that in speaking of the "structure" of our knowledge in physics, I am using the crucial words very broadly: I do not presuppose an exact notion of "structure," and in applying the vague presystematic notion to "our knowledge in physics," I am construing the word knowledge in a wide and ambiguous sense. The reflections I am proposing have as their object (a) our knowledge in physics as an *achieved result*: knowledge as *the knowledge we have of* X; (b) our knowledge as susceptible of *justification* or *defense*—that is, as involving a structure of "evidence" for its asserted contents; and (c) knowlege—science—as (to appropriate a word of Isaac Levi's) an *enterprise*: an activity aimed at increasing our knowledge in sense (a), by means appropriate to the constraints of (b). But, again, this is a presystematic description, and I neither promise nor threaten you with even a sketch of an actual *theory* under these three heads.

With regard to the structure of our achieved knowledge in physics, there is a point that struck me with great force many years ago, in the course of my own attempts to learn something of the subject. To present it to you, it will be useful to refer to an early attempt of Carnap's to give a schematic view of the structure of physical knowledge; I quote from his retrospective description in the Schilpp volume:

> In an article on the task of physics [1923][3] I imagined the ideal system of physics as consisting of three volumes: The first was to contain the basic physical laws, represented as a formal axiom

[3]The reference is to the article CARNAP 1923.

> system; the second to contain the phenomenal-physical dictionary, that is to say, the rules of correspondence between observable qualities and physical magnitudes; the third to contain descriptions of the physical state of the universe for two arbitrary time points. From these descriptions, together with the laws contained in the first volume, the state of the world for any other time-point would be deducible,... and from this result, with the help of the rules of correspondence, the qualities could be derived which are observable at any position in space and time. (CARNAP 1968, p. 15.)

This is familiar logical-empiricist doctrine of the earliest vintage, and it foreshadows much of what continued to be Carnap's view of the matter (of course, the implied Laplacian determinism would not have been maintained after the development of quantum mechanics); in particular, the crucial distinction between the "observational" and the "theoretical" could not be more emphatically posed than in this image of the separate volumes. Carnap immediately remarks: "The distinction between the laws represented as formal axioms and the correlations to observables was resumed and further developed many years later in connection with the theoretical language."

The issue now familiarly associated with that distinction is that of the "theory-dependence" of observations. As a subject for philosophical commentary, this issue continues to present virtually limitless opportunities; and I have felt the temptation to expatiate on the matter here to some degree. But I ask myself, how much profit is now to be gained from such discussion? The matter has been very widely treated. One may hope to put a point more trenchantly than has been done before, perhaps even to find a new turn of argument; but hardly, by subtle technical analysis, to effect a real transformation of the subject.

Instead of something subtle, I want to suggest something crude. In Carnap's Platonic myth of the three volumes of physics, consider what the first and second volumes might look like. I submit that there is no difficulty at all in envisaging the first. Carnap says that it is to contain "the basic physical laws, represented as a formal axiom system." I should not wish to insist on the notion of logical formality, which seems to me to have been overemphasized by the logical empiricists; so let me just substitute the phrase, "a mathematical system." It would be inappropriate to demur that we do not possess a mathematical formulation of all "the basic physical laws"—or even a unified mathematical formulation of all the basic physical laws we know—because Carnap is explicitly presenting what I have just called a "myth": an image of "the ideal system of physics." The first volume is conceived simply as having the form of a treatise on theoretical physics. There are

many such treatises in existence, some of them very good indeed; and it is even possible to learn branches of theoretical physics by reading them. This is what I discovered in my student days: I had a strong desire to gain a real understanding of the theory of relativity, both special and general, and after some frustrating attempts came to try Weyl's great work on the subject, *Raum-Zeit-Materie* (WEYL 1923)—from which I had earlier been deterred by a remark I had read of Leon Chwistek's, to the effect that Weyl's book was spoiled by an objectionable philosophy (CHWISTEK 1948, p. 3). The book was a triple revelation for me: it put the physical principles of the special and general theories of relativity—and also, as a preliminary, those of Maxwellian electrodynamics—in what seemed to me an astonishingly clear light; it opened my eyes to a new perspective on mathematics; and, in the process (in view particularly of the fact that the idiosyncrasies of Weyl's philosophy in no way obstructed these clarifications), it altered my conception of what the philosophy of physics could be. At any rate, this is one example among several in my experience of a book that reasonably resembles Carnap's ideal first volume (of course restricted to a more modest scope), and that succeeds not only as a systematic formulation but even as a pedagogical instrument.

When we turn to Carnap's second volume, the situation is drastically different. Carnap says the "phenomenal-physical dictionary" it contains is to make it possible to derive, from the data in the third and the laws in the first, *"the qualities ... which are observable at any position in space and time."* But nothing remotely like this exists, for however restricted a domain of physics. I shall return to the point; but for now I should like to consider a less demanding alternative to that dictionary: granted that it is possible to learn the principles of parts of theoretical physics from books in which those principles are presented in a systematic mathematical framework, is it analogously possible to learn corresponding parts of *experimental* physics? My own experience has been that it is at the least very much harder. My belief is that it is, in practice today (that is, with the help of the existing literature), very nearly impossible: I have never found a single book on experimental physics comparably instructive with those I have found on physical theory. My suspicion is that it may be impossible even in principle. It is hard, but possible, to learn theory by self-study from books; it is surely much harder to learn experimental techniques without a teacher to help one acquire skills; but what I suspect to be impossible is to learn the principles of experiment without *actual* experience with the relevant *instruments.*

That may seem banal; but what strikes me is that it stands in odd contrast with our clichés about the theory-dependence of observation. In a famous passage Duhem said that, in the case of physics, "it is impossible to

leave outside the laboratory door the theory that we wish to test" (DUHEM 1954, p. 182). Technical assistants, however, can be taught to perform experiments, and to report the results of those experiments in usable form, without teaching them the theories those experiments are designed to test. In any case, my point is that, whether or not bringing the theory inside the laboratory door is necessary, it is certainly very far from sufficient.

Now it seems to me that this has a rather interesting consequence, not only for the logical empiricist view of the structure of physical knowledge, but for post-positivist views as well: in a certain sense, in my opinion (here and elsewhere too), the critique of logical empiricism by its opponents has fallen off center. My own view is that in the rough sense Carnap was willing to adopt from the time he abandoned the more primitive versions of the empiricist thesis, there is no great difficulty in defining an "observational" vocabulary: an "observation-language" in Carnap's sense is the language in which we ordinarily conduct the business of daily life, and the only theory it is dependent upon is the theory that there are ordinary objects[4] with such properties as we habitually ascribe to them. There are also systems of concepts of the sort that constitute the framework of fundamental physical theories; so, referring again to my example, I may say that a book like *Raum-Zeit-Materie* demonstrates the existence of theoretical vocabularies distinct from the observational. Thus I argue, on the basis of these crude and banal considerations, that Carnap was right to make and to emphasize this distinction. I also believe that his philosophic career consists to a considerable degree in a series of genuinely instructive attempts to do better justice to the character of the distinction. But I think too that there was a fundamental bar to success along any of the routes Carnap essayed. For he always assumed that "the observation language" is more restricted than, and included in, a *total* language that *includes an observational part and a theoretical part, connected by deductive logical relations.* And this, I think—I do not say by virtue of some basic principle I can identify, but simply, at the present time, *de facto*—is not the case: there is no department of fundamental physics in which it is possible, in the strict sense, to *deduce* observations, or observable facts, from data and theory. So I suggest that the principal difficulty is not that of how to leave the theory outside the laboratory door, but that of how to get the laboratory inside the theory.

Well, how *do* we do it? For of course we do put theory and experiment in relation to one another; otherwise it would be impossible to test theories,

[4]In delivering this address, I interpolated here, with a gesture at the apparatus in question, the words: "including such objects as microphones"—anticipating an objection that might be raised, and indeed was raised during the discussion period, concerning this point; see the Supplementary Note at the end.

and impossible to apply them. It would also, I should add, be impossible to *understand* a theory, as anything but a purely mathematical structure—impossible, that is, to understand a theory *as* a theory of physics—if we had no systematic way to put the theory into connection with observation (or experience). This has been taught us not only by the philosophers we usually call empiricists, but also (for instance) by Kant: "Gedanken ohne Inhalt sind leer, Anschauungen ohne Begriffe sind blind" (KANT 1781/7, p. 51/75). So it might be asked of me—and I did in fact ask of myself—how I succeeded in learning any *physics* from Weyl's book.

The short and simple answer is that Weyl first of all connects his exposition of the new theories he expounds with older physical theories I already knew something of, and secondly describes—I shall say "schematically," and return to comment on this word later—a few experiments that bear critically upon the theories he is developing. But that reply is not very instructive, without some indication of (a) how this is done at all, in view of the difficulties I have claimed lie in the way of drawing logical inferences between theoretical statements and observational ones, and (b) how—or to what extent—it suffices to establish "physical understanding" (Kantian *Inhalt*) for a theory. To enrich the discussion of all this, I want to turn to a much earlier physical theory—not just to an earlier "paradigm," but to what may be called the grand archetype of all that we call physics: namely, the theory presented in the first and third books of Newton's *Principia*. I am going to try to say, in brief compass, what this theory (roughly speaking, in our own terms, the conjunction of Newtonian mechanics and Newtonian theory of gravitation) *is*, as a theory of *a mathematical structure discernible in the world of phenomena, of observations, of experience*—and to do so in a way that adheres to the basic conceptual framework introduced by Newton himself; and also to say something about how both what I have just called the "conceptual framework," and the theory formulated within it, were discovered—or "invented"—by Newton. (Thus I mean to touch upon another aspect of the "structure" of physical knowledge: the question of its *advancement*, or knowledge as an *enterprise*.)

Newton tells us in the preface to the *Principia* that he is proposing in it a certain "method of philosophy" (that is, of natural philosophy: of physics). This method consists in investigating the phenomena of nature—in particular, of motions—with a view to determining what Newton refers to both as "the forces of nature," and "the natural powers"; and it involves the working hypothesis that all natural phenomena result from the action of such forces. Newton says: "[A]ll the difficulty of philosophy seems to consist in this, from the phænomena of motions to investigate the forces of Nature, and then from these forces to demonstrate the other phænomena." He goes on to say that

in the first two books—whose character he describes as "mathematical"—general propositions are developed to facilitate this end, and that in the third book, containing the theory of gravity, he gives *an example* of such investigation. And he adds, in an expanded statement both of the program he is advocating and of the standing he attributes to it in philosophy:

> I wish we could derive the rest of the phænomena of Nature by the same kind of reasoning from mechanical principles. For I am induced by many reasons to suspect that they may all depend upon certain forces by which the particles of bodies, by some causes hitherto unknown, are either mutually impelled towards each other and cohere in regular figures, or are repelled and recede from each other; which forces being unknown, Philosophers have hitherto attempted the search of Nature in vain. But I hope the principles here laid down will afford some light either to that, or some truer, method of Philosophy. (NEWTON 1729, vol. 1, third and fourth pages of the Author's Preface [pages unnumbered].)

I have quoted this passage so often that to do so may seem a mannerism on my part; but it continues to strike me, in its clarity, economy, and what I may call its *philosophical truth of method,* as not only instructive but a shining example.

The passage does however demand some explication. Let me call your attention to one phrase that deserves to be puzzled over. Newton says that he suspects the phenomena of nature all depend upon "certain forces, *by* which the particles of bodies, *by some causes hitherto unknown,*" are urged either towards or away from each other—thus what we call "central forces." But is there not one "by" too many here? Should not *the forces themselves* be called the causes? What sense does it make to speak of a force "*by* which, *by some cause,* bodies are impelled"? Or as an alternative, may it not be appropriate to drop the perplexed notion of cause altogether, expecting the theory itself (including its empirical interpretation, however such interpretation is managed) to give an adequate explication of the *systematic, technical* concept of "force," without any need to cloud positive science by such a metaphysical notion as "cause"?

Hold that question for a while in suspension; I want first to describe how, as I see it, Newton actually proceeded in the development of his theory, and to give an account of the actual system he propounds. (Of course this must be in significant measure speculative; and within the constraints of such a paper as this, necessarily sketchy.)

At the time of the investigation that gave us the *Principia,* Newton had

good reasons for regarding *accelerations* as of critical significance in the interactions of bodies.[5] In fact, nearly two decades earlier Newton had already understood this point well enough to motivate a calculation of the acceleration of the moon, a derivation from Kepler's so-called third law of the implied relation among accelerations (assuming uniform circumferential speeds in circular orbits), and a comparison, on the basis of his result, of the moon's acceleration with that of falling bodies on the earth (cf., e.g., WESTFALL 1980, pp. 151-152). But in 1684 he did something very much more far-reaching.[6] By a purely mathematical, kinematical, analysis he demonstrated, in effect, that the so-called three laws of planetary motion of Kepler[7] are equivalent to the following pair of propositions:

1. Each of the bodies in any one system (that is: planets around the sun; satellites around a planet) has, at each instant of its motion, an acceleration that is a function of *position relative to the central body* alone: namely, having its direction towards that central body, and with a magnitude that varies inversely as the square of the distance from the central body. (In particular, then, in each such system there is a well-defined *field of acceleration,* and the acceleration is independent of any special characteristics of the particular planet or satellite.)

2. In the course of the motion, each body remains within a bounded distance from its center of accelerations.[8]

[5]Notably: Galileo's propositions (by then well-confirmed) about the motion of bodies in free and oblique fall, and of projectiles; the more elaborate application of Galilean principles to constrained motion under the influence of weight in Huygens's great work on the pendulum clock; the successful application of notions derived—once again—from Galileo's theory of fall to the phenomena described as "centrifugal force." (For a brief account of Huygens's investigations as providing significant background for Newtonian mechanics—a background cited by Newton in his scholium to the laws of motion in the *Principia*—see STEIN 1990a, pp. 20–26.)

[6]For a fairly detailed analysis of Newton's argument for universal gravitation see STEIN 1990b.

[7]*Not* so called by Newton: he does refer the "harmonic law" for the primary planets to Kepler; and he records the law of areas for the primary planets, and the harmonic law for the satellites of Jupiter (in the second edition, also for those of Saturn) among the results established by astronomical observations (without explicit reference to Kepler). As for the law of ellipses, Newton does not admit this at all into his catalogue of results secured by observation.

[8]Besides compressing the formulation of this result, I have drawn certain inferences that Newton does not make explicit, but does make use of (and which can be justified on the basis of his theorems).

The second of the two propositions above is needed to exclude the case of open orbits (parabolic or hyperbolic)—to exclude them, that is, not (of course) from occurring in nature, but from the scope of Kepler's laws.

This result is easy to understand, and it is familiar: "Ah, yes!" the sophomore will say, "the law of gravitation—*I* know that; so *that's* how Newton got it!" But of course no: this is not the law of gravitation at all; it is only how Newton *began* to "get it." What he did next (and had already thought of in the 1660's) was to apply his result to the moon, on the assumption that if the position of the moon were made to vary *away from its actual orbit,* it would continue to "explore," as it were, the same inverse-square acceleration-field about the earth that it does in its actual orbital motion; and in particular, he calculated what that acceleration would be at the surface of the earth. The result agreed well with the observed acceleration of falling terrestrial bodies (or rather, with the value of that acceleration derived by Huygens from his careful observation of pendulums). And Newton concluded that the acceleration of falling bodies and that of the moon are effects of "the very same force" (NEWTON 1729, vol. 2, p. 217 [in the proof of Prop. IV, Book III]), or (as he puts it elsewhere) that they are manifestations of one and the same "active Principle" or "general Law of Nature" (NEWTON 1730, p. 401). Since the effect familiar in terrestrial bodies is called that of "weight"—"gravity"—this "active principle," "natural power," or "force of nature" is called by Newton the force of gravity.

Still, that is only a word: what *is* the alleged "principle" or "general law of nature"? Two things are clear enough from what I have already rehearsed: (1) If the line of thought is correct, then the principle in question must assign to any body subject to it, under given relevant circumstances, an acceleration that is independent of that particular body, and depends only upon its geometrical situation (relative to those " relevant circumstances"); for this characteristic, inferred for the *existing* motions from the observed phenomena, has already been assumed in arriving at the identification of the force on the moon with terrestrial weight. (2) By the same token, it is clear that the principle must involve accelerations directed towards certain centers, with magnitudes that vary inversely with the square of the distance from those centers.

None of this is especially subtle (although of course the mathematical analysis that underlies it was pathbreaking for both its methods and its results); and the conclusion that Newton had uncovered a new "principle" of the kind so far characterized was greeted at the time with acclaim and no serious controversy. But for Newton that was not the last or the most crucial step. At this point I want to indulge in some speculation about his state of mind.

As I see it (judging both from the actual sequence of propositions and arguments in the *Principia,* and from the evidence of the circumstances surrounding the development of the work), Newton asked himself two interrelated—

perhaps not clearly distinguished—questions: (a) *What can be the cause* of such an effect as these inverse-square centripetal accelerations, affecting both the planets and all earthly bodies? (b) If there is a principle that certain bodies are affected by inverse-square accelerations directed towards certain other bodies as "centers"; and if among those centers are the sun, the earth, Jupiter, and Saturn, and among the bodies affected are all the planets (towards the sun), all the satellites (towards their planets), and all terrestrial bodies (towards the earth); can one arrive at any conclusion as to *which bodies in general* are centers of such acceleration, and which bodies in general are subject to it—towards which centers?

It is at this point, I should argue, that the new method of philosophy was born. Let me contrast Newton with Huygens. Huygens read the *Principia,* admired it enormously, but thought that just here Newton went wrong. Huygens himself had a ready answer to the questions I have just put. To the first—what can be the cause?—his answer was that "the philosophy of the present day" (which is to say, "modern physics") teaches us that the causes of all natural effects are to be sought in the impinging of matter upon matter (cf., e.g., HUYGENS 1690, pp. 2–3); and so a motion of some kind of ambient medium must be conceived that can give rise to a pressing of bodies towards a center. He had already proposed a theory of weight based on such a hypothesis; and now concluded from Newton's results that one must investigate further just what kind of motion of the medium could produce an inverse-square variation of the force. To the second question, his answer was much simpler: there is, he said, absolutely no evidence that there are any other centers of such acceleration than the ones already identified—except that one will naturally generalize, and say that every star and every planet is a center of the sort. Thus it is precisely about the stars and the planets that one should assume the existence of ambient matter in such a state as to produce this effect.

What Newton did reflected, in contrast, what might be called a respectful skepticism about the demands of "the philosophy of the present day." He had devoted most of the second Book of the *Principia* to an analysis of the behavior of fluid media and of bodies moving through fluid media; and had concluded that there are insuperable obstacles to any attempt to reconcile the observed motions of planets, satellites, and comets, with the existence of any such medium of appreciable density occupying the interplanetary spaces. He did not say that he quite despaired of a "mechanical explanation"—i.e., one in terms of the impinging of bodies upon one another—of the astronomical phenomena; but he did say that one ought not to build positive conclusions in physics on the demand that such explanations be forthcoming. So he reflected upon the situation without any regard

for the mechanical "hypotheses." On the other hand, he allowed himself an extremely bold—one might almost say, a reckless—use of a principle he had extracted from results of work in mechanics by several investigators, including Huygens and himself. This is the principle of the conservation of momentum, discovered by Huygens as a true substitute for the false principle of conservation enunciated by Descartes, but never placed by Huygens in a central theoretical position; a principle to which Newton, emphasizing as he did *acceleration* as a fundamental parameter of natural processes, gave the form (actually somewhat more restrictive) of the third law of motion. In particular, Newton argued thus: If body A is subject to (here using my own term rather than Newton's) an "acceleration-field" directed towards body B, of magnitude dependent only upon position, then what Newton calls the "motive measure" of the force on A is the product of its *mass* by this acceleration—thus the "motive force" (at a given position) is proportional to the mass of the body acted upon. According to the third law of motion, there must be an equal and opposite motive force on something—in the ordinary formulation, which is Newton's own, on whatever *exerts* that force on A. What does? This Newton explicitly, repeatedly, emphatically says he does not claim to know. And yet he takes the step—this is what I have "almost" called reckless—of asserting that there is an equal and opposite force exerted *upon the "central" body B.*

This postulate, together with some simple qualitative considerations, led Newton inescapably, in a very few steps, to the law of universal gravitation: that is, to the conclusion that the answer to question (b) above is that *all* bodies are subject to the "principle" of gravitation and that *all* bodies are centers of gravitational acceleration-fields. (That the "motive measure" of the force is directly proportional to the mass of the body acted upon is equivalent to the proposition that the "accelerative measure" is independent of that body; that the motive force is also proportional to the mass of the gravitational center follows from the third law of motion—so that, as Newton emphasizes, there ceases to be any difference in status between the two bodies concerned: one is dealing with an *interaction,* whose participants enter it *symmetrically.* That the force is inversely proportional to the square of the distance is a conclusion already reached; and the law is formulated in a way that is complete *and perfectly general* [cf. the fuller discussion in STEIN 1990b].)

This is not to say that the gravitational principle is *established* by the argument I have outlined; only that it is *found*—"invented," as the seventeenth century would say—by that argument. More is found as well: namely, the terms of the Newtonian program; it remains to put that into a systematic framework.

But I have not yet given Newton's answer to question (a): what can be the *cause* of such an effect as this—a universal attraction between all particles of matter? Huygens, believing in the necessity of "mechanical" causes, thought it clear that *nothing* could cause such an effect; and, seeing no evidence for universal gravitation, rejected it on those grounds. Newton's answer is rather subtle. He says: "The cause of gravity is what I do not pretend to know."[9] But he is certainly interested in the question; he adds that he wants more time to consider it, and of course he did eventually publish a speculation about it (NEWTON 1730, pp. 350ff. [Book 3, Query 21]). He amplifies his view most significantly near the end of the long final Query in the third Book of the *Opticks,* in his fullest discussion of his conception of a force of nature. He had been attacked for reintroducing, in his theory of gravity, the much-deplored "occult qualities" of the scholastics, by assuming an "occult" cause of the universal attraction. His answer is that he regards his own "active Principles" or forces "not as occult Qualities,... but as general Laws of Nature ...; their Truth appearing to us by Phænomena, though their Causes be not yet discover'd"; and concludes:

> To tell us that every Species of Things is endow'd with an occult specifick Quality by which it acts and produces manifest Effects, is to tell us nothing: But to derive two or three general Principles of Motion from Phænomena, and afterwards to tell us how the Properties and Actions of all corporeal Things follow from those manifest Principles, would be a very great step in Philosophy, though the Causes of those Principles were not yet discover'd: And therefore I scruple not to propose the Principles of Motion above-mention'd, they being of very general Extent, and leave their Causes to be found out. (NEWTON 1730, pp. 401–402.)

There is evidence—which I find convincing—that Newton had in fact abandoned all belief in the traditional "mechanical" causation as ultimate. If I am right, the correct reading of his words here is this: When we have found such principles of motion as Newton's program envisages, and as he has given an example of in the law of gravitation, we have *(ipso facto)* discovered something about what one calls "causation." In any given case, such a principle may have underlying deeper (e.g., "mechanical") causes; and this is a proper subject of inquiry. Then again, it may not; for the ultimate cause (in Newton's view)—"the very first Cause, which certainly is not mechanical"—is the direct legislative action of God (whose laws are self-executing): which we may immediately translate, simply, as "the ultimate constitution of nature." And Newton's words in the preface, referring to his hope to "afford

[9]Letter to Bentley, 17 January 1692(o.s.)/3(n.s.); see TURNBULL 1961, p. 240.

some light either to that, or some truer, method of philosophy," bear two interrelated meanings: first, he hopes that the principles laid down in the *Principia* will facilitate the investigation of nature, either by the means he has suggested or by some modification in which those principles are still of use; second, he hopes that those principles help to explicate the constitution of nature, whether or not they prove adequate to its *ultimate* constitution.

As to the principles themselves, the mathematical structure they involve is reasonably clear (although there are certain important ambiguities, as I shall explain in a moment). Newton found it impossible to codify dynamical theory without presupposing the structures he calls "absolute time" and "absolute space." So we have—in our own terms—the four-dimensional manifold of space-time, given with the product structure: $\mathbf{S} \times \mathbf{T}$, where $\mathbf{S}$ is a three-dimensional Euclidean space and $\mathbf{T}$ a one-dimensional affine space. We also need to posit a further "space," which I shall call $\mathbf{B}$, the set of "bodily points"; this must have the structure of a measure space—the measure, following Newton, we call *mass*. The postulate that all the phenomena of nature depend upon *configurations and motions of bodies* takes the form of the assumption that the entire history of nature is represented by a mapping from the Cartesian product $\mathbf{B} \times \mathbf{T}$ into $\mathbf{S}$: the *kinematical history*. There is a problem about the finer specification of the space $\mathbf{B}$ (for example, to demand that $\mathbf{B}$ have the structure of a differentiable manifold is appropriate to some classical contexts, but highly inappropriate to others). Newton himself considered it "probable," he tells us, that matter consists ultimately of *rigid indivisible particles*; this implies that $\mathbf{B}$ is a disconnected topological space, each of whose connected components has the structure of a compact three-dimensional metrically Euclidean manifold with boundary, and that the kinematical mapping is, for each instant t of time, isometric.

A related subtlety concerns the further requirements to be placed on the kinematical map. Newton would not suppose this map to be everywhere smooth with respect to the time as argument—for he would expect occasional (although, he tells us, very infrequent) *impacts* of the rigid fundamental particles. At any rate, it is immediate that wherever this mapping *is* smooth, it determines (for each bodily point at each such instant of time) the associated velocity and acceleration vectors; and—subject to suitable conditions on the structure of $\mathbf{B}$ and of the kinematical mapping as a function of its body-point as well as its time argument—it is also clear that one will be able to define a motive force (or force-density, or force-*measure*: most generally, a "force-distribution") over such points of $\mathbf{B}$ at such times.

But that concept of motive force, which differs from a purely kinematic concept only in that it involves the mass as a coefficient, is not Newton's central notion of force; the framework so far described does *not* define a

Newtonian mechanical system. (Indeed, the third law of motion has made no appearance yet.) To complete the account requires a new concept, which carries its own—"extra-systematic" if not presystematic—marginal gloss: namely, the concept of what Newton calls a "force of nature." His own way of putting it (only a little paraphrased) is this: the motive measure of the force on a body at a given time is the resultant of a system of what he calls "impressed forces" (the set may be infinite—even continuous, so that the "resultant" becomes an integral); each of these component impressed forces is the *exercise* upon the body in question of a "force of nature" (this is the connotation of his term "impressed force": the "impression" *upon* a body *of* a "natural power"). But the forces of nature are to be known through *general laws* of nature (as Newton says, an example of this is given in the third Book of the *Principia*—the first example of the kind ever discovered, the "invention" and "proof" of which, I am suggesting, is what motivated Newton to elaborate this conceptual framework itself). And these laws—the search for them is the proposed "method of philosophy"—are to take the form of *laws of interaction* between *pairs* of bodies, in which each body enters symmetrically in the sense of the third law.

So the specification of a Newtonian system requires the specification of the structure of the space **B** of bodily points and the specification of a *set* of laws of interaction of the indicated type; the motive force-distribution associated at any instant with the actual motion is to be the vector-sum of all the impressed force-distributions at that instant. Newton hopes that a small set of laws of interaction will suffice for an account of all of nature: "To derive two or three general Principles of Motion from Phænomena, and afterwards to tell us how the Properties and Actions of all corporeal Things follow from those manifest Principles, would be a very great step in Philosophy."

Now, something a little odd has happened in my own argument. We were considering the question of how to get the laboratory (or observatory)—the phenomena—into the theory. My discussion of Newton has largely been a discussion of how Newton got *from* phenomena *to* theory. I have described the theoretical framework—corresponding, as I put it, to Carnap's "first volume"—and have provided an extra-systematic commentary on the development, the motivation, and thus in a certain sense the intended "meaning," of this abstract theoretical structure (which itself is constituted by certain spaces and certain mappings or functions). The dialectic, one might say, has moved from phenomena to pure forms; and this seems, as I have said, to be opposite to the direction we were concerned with.

A closely related point is that we left Newton's theory of gravitation in a peculiar condition. In describing Newton's own path to that theory, I said that he took a very questionable turn, and that the argument by which he

found the law of universal gravitation is not an argument that genuinely *establishes* the law—or establishes that there is any principle at all of *universal* attraction. (I hope it is clear that Huygens's objections were very sane and reasonable.) How, then, does—or did—Newton's theory really get "established"?

Again I shall have nothing startlingly new to tell you about this. Let me first confront the radical objection that the theory did not get established—that no theory ever gets established. You will not expect me to settle this issue; and I do not want to quarrel over mere words: in this case, what "established" should mean. I want to take for granted that we do not now believe Newton's theory of gravitation to be a correct general theory; that we *do* now believe that there is a "universal principle" of gravitation (and, indeed, we place it among the "fundamental forces"); and that we also believe—with enormous confidence—that Newton's theory is a very accurate approximation for a very wide range of applications, in which indeed it is our only usable theory: applications that include planetary astronomy (the recent solar eclipse occurred right on the dot) and such space-travel as has so far been accomplished. (By the way, having studied Newton, I was very much struck at the time of the first moon-landing in 1969 that after nearly three centuries Newton's experiment of bringing, not indeed the moon itself, but at least a piece of the moon, down to the earth's surface and weighing it was to be performed. Of course the whole space program relies crucially upon the law of gravitation in the management of space vehicles. And—despite, let me say, the views either of Popper or of anti-cumulativists—no one then thought that gravitational theory was being put to the test: no one at the time had any doubt at all that the law of gravitation was going to work properly; all anxiety concerned either the adequacy of the engineering or the firmness of the moon's surface.)

Next a point bearing on the question of the theory and the laboratory: it is hardly possible to maintain that the theory of universal gravitation was established by testing it in the laboratory. Cavendish's experiment, for example, can certainly not be regarded as having established the theory. To be sure, it was impressive *confirmation* of Newton's theory when Cavendish was able to demonstrate the existence of an otherwise unsuspected force between two bodies; but that the force in question was gravitational in nature—in origin—in other words, was "caused" by the same principle that is responsible for weight—could not conceivably have been even surmised from the experiment in the absence of the theory (here indeed the theory could not be "left outside the laboratory door"!). (One should add that the experiment also made a capital contribution to the *content* of theory by allowing a determination of the gravitational constant—on the *assumption,*

of course, that the Newtonian theory was correct and that the Cavendish force was gravitational.)

Obviously, then, the really central evidence must be astronomical; and this, I suppose, as I have said, is no great surprise. It would be of some interest to consider just how astronomical evidence—which, after all, consists in observations of rather special, mostly rather large, bodies—could possibly support such an astonishing proposition as Newton's (we forget, I think, just how extravagant a proposition it is, because we have been taught from childhood that it is so—it becomes even a kind of mark of our enlightenment to believe it). But I have to paint here with a broad brush, so I set this question aside, and just consider how any evidence at all gets connected with the abstract mathematical framework I have sketched—once again apologizing for the obviousness of the answer I shall give.

Let me underscore the point that there can be no thought of *deducing observations* within that framework. To do so in the strict sense, one would need to have a physical theory of the actual observer, and to incorporate it into the Newtonian framework. I certainly do not want to say that there is a reason "in principle" why such a thing can *never* be done, for *any* possible (future) physical framework; but everyone knows that Newton could not do it, and that we—in the best versions of our own physics—cannot do it. Even waiving the theory of the observer, it is clear that all astronomical observations are intermediated by *light*; therefore, to deduce anything like observations, one would have to include the theory of light within the framework. Moreover, the light traverses the earth's atmosphere, and is usually received through a telescope; so we need the theory of atmospheric refraction and the theory of the instrument also—we are in the vicinity of the problem of the systematic treatise on experimental physics. In actual fact, the experimental physics is treated separately as a discipline in its own right, that is partly an art: an affair of both knowledge and manipulative and perceptual skill. But the possibility of connecting this art with the theory is closely connected with a certain possibility *within* the mathematical structure that is the theoretical framework: using a word I have introduced earlier, the possibility of representing experiments, and of representing the observer, "schematically." Kant used the word "schematism"—in a way I confess to finding rather obscure—for a process that intermediates between concepts of the highest abstractness (his "pure concepts of the understanding") and sensible contents; my use of it here is vaguely similar (but I hope *not* obscure): where Kant speaks of "schematizing the category to the manifold of intuition," I want to speak (as it were conversely) of "schematizing the observer within the theory"; but the intention is analogous: to secure empirical content—content within experience—for an abstract structure.

Fancy talk, but a simple idea that will be found perfectly familiar. One represents the observer within the spatio-temporal framework by a world-line (or a system of world-lines). Putting in—for the gravitational theory of the solar system—the world-lines of the planets and satellites, as calculated from suitable initial data, one can then determine at each instant all the relevant angles between lines drawn from the observer to the bodies of the system (including, if the theory is properly handled, with the earth represented as an extended body and its rotation treated systematically, lines from the observer to terrestrial landmarks). As a first approximation, such lines are treated as lines of sight. With more sophistication on the observational side, the results are turned over to the experts in observational astronomy, who will take such account as they are able to of atmospheric refraction, of aberration of starlight, and so on. But so far as the fundamental theory is concerned—or rather, so far as *mathematically defined structures and rigorous arguments* are demanded—the "schematic" representation of observers, experiments, and observations, is, I believe, as far as we know how to go.

Let me suggest a few further reflections upon and consequences of all this.

First, in the account I have given, the distinction between what is purely mathematical and what is not has certainly played a central role—closely related to the distinction I have argued we really do need between observational and theoretical "languages." That there is indeed such a distinction, and that it is of fundamental importance, is one of those things that seem to me quite obvious. But it is often denied. It is denied, for instance, by Quine, who (in contrast to Carnap) sees here a "continuum" within which no sharp boundaries can be drawn. This, I submit, is simply wrong. Newtonian mechanics, *in its application to the empirical world,* is a theory that gives very good results in a very wide domain, but that can no longer be defended as correct without restriction both as to domain and as to degree of precision. On the other hand, Newtonian "rational" mechanics, *as a purely mathematical theory,* stands on an unshaken footing and continues to offer a field for useful and deep rigorous investigations. That is the distinctive nature of mathematics, *qua* mathematics: it is, as such, *not* about the given, natural world.

Second, important modifications both of Newtonian space-time theory and of Newtonian dynamics are possible *within classical physics*: It is well known that the product structure of space-time as $\mathbf{S} \times \mathbf{T}$ is demonstrably *inappropriate* to the theory—that it can and should be replaced by the structure of a four-dimensional affine manifold, with an affine "time-projection" and a three-dimensional Euclidean structure on the associated space of vectors with time-projection zero. (Of course one then has to change the description

of the kinematical mapping: it goes from $\mathbf{B} \times \mathbf{T}$ into space-time, commuting with the projection onto $\mathbf{T}$.) Within that structure, velocity vectors are no longer definable; but acceleration vectors are, and therefore also motive forces—the reformulation of the theory is unambiguously determined, and its correctness is *demonstrable* from the *original* formulation (with the help of the principle of Galilean relativity—proved by Newton as a theorem).

As to the other modification, it is even more familiar: A whole series of classical investigators, including Lagrange, Gauss, Hamilton, and Jacobi, found alternative ways of formulating the dynamical law of a Newtonian system. These formulations are not all equivalent; rather, they all generalize a certain common domain, and generalize it in different directions. And the generalizations have very important physical significance; for example, the Maxwellian electromagnetic field is not representable as a Newtonian system, but is representable as a Lagrangian or Hamiltonian one. But of course, Lagrange and Hamilton were consciously building upon and transforming Newton's principles. The result is a transformation of the concept of a "natural power" or "force of nature": such a force is now to be given, not by a law of motive force characterizing action-reaction pairs, but (for instance) by a Hamiltonian function. This surely deserves some recognition as a remarkable fulfillment of Newton's hope: that his principles might "afford some light, either to [his own], or some truer, method of philosophy." The fulfillment becomes all the more remarkable when one considers that, although Newtonian forces have little place in our own most fundamental physics, Hamiltonian and Lagrangian functions—or operators—are at the heart of those theories.

This brings me to another general point. It has been my experience that many philosophers balk at what they may think of—perhaps quite justly—as the rather "Platonic" notion of "general principles—or laws—of motion" as having in some sense a kind of "reality" and even "efficacy": "What can it mean," I have been asked in connection with Newton, to talk—as Newton quite explicitly does—of a 'force of nature' *as* a *law* of nature?" It is surely important to note that that is exactly the way physicists do talk today: when one says that, at the fundamental level, there are "four forces" (or fewer than four, in the light of the unifications that have been made or proposed), that has nothing to do with Newtonian "impressed motive forces," but it has everything to do with laws of nature, forms of interaction, Hamiltonians. It might be rejoined that this is an interesting sociological fact about physicists, but that it cuts no philosophical ice. I said in another context that I do not claim to give reasons of philosophic principle, but to call attention to what seem to me obvious but important *facts* that deserve philosophers' attention; and this is another one. It is not just a question of how physicists talk: it

is a question of what, *de facto,* in the history of physics to date, has tended most to *persist as stable* (and, it appears, reliable) in what we think we know about the world. One of the things that have persisted is the more or less far-reaching and more or less precise (but, in application, always approximate) correctness of theories as applied to domains of phenomena; another is the "forms"—more precisely, certain aspects of the forms—that are characterized by what Carnap called "frameworks."

In fact, if we ask, say, of the physics of the end of the seventeenth century, what of all it told us about the world we can still regard as "true" or as having proved itself "real," the answer is—I use the word yet again—striking: Not Newton's hard particles, not Leibniz's material continuum, not Huygens's ether—indeed, hardly anything to which most philosophers would accord "ontological" status. (In particular, of course, not "space." If one reflects on what quantum field theory has told us about the characteristics of the only thing *in the physical world* that can be regarded as "empty" or "pure" space, its difference from anything earlier centuries conceived is startling indeed.) And of fundamental processes: no impacts of atoms, no pressures of continuous media, no immediate and instantaneous actions at a distance—indeed, no instants at all! And yet, although Huygens's ether has gone, the "form" of the propagation of light to which he contributed a first crude sketch is still discernible in the theory of electromagnetic waves, and through that theory—again transformed—in the quantum theory of electrodynamic and optical processes. One could go on in this vein; I hope I am right in thinking that the point really is obvious, and only needs to have attention called to it; I must come to an end, and there are still some things to say.

First, I want to mention the issue of the "incommensurability" of theories. That is a metaphorical term; in an appropriate interpretation, the doctrine may be true. But in any case, one does *compare* theories—as, of course, the analysis first developed by the Greek geometers allows one to compare *magnitudes* that are (technically) called "incommensurable"; and in so far as forms discernibly persist through the transformation of our theories, such comparisons form a most vital part of science. If we make the assumption that the human race will survive for another millennium, and in circumstances conducive to the advance of knowledge, then I should predict with great confidence—not that quarks and leptons will continue to be regarded as the most basic particles (I don't predict the contrary—I am perfectly agnostic); not that quantum field theory, or general relativity, will retain its fundamental role (on *this* point, I would hope very much for a radical advance)—but that the forms of these theories will be clearly *discernible in,* clearly *related to,* the structures of whatever theories supersede them.

The second point has to do with a special bearing of the crude account I

have given upon the "structure" of our knowledge in the sense of *epistemology*: it is the simple remark that our understanding of our own relation to the world is mediated by our ability to place ourselves, however "schematically," within our conception of the course of nature. And it is a very interesting exercise, within the successive frameworks of Newtonian space-time (*without* absolute space—so that "geometrical" relations hold *only among simultaneous events,* and space is as it were constantly evanescent) and special relativity (in which, by contrast there is no such thing at all as simultaneity), to consider the epistemology of "geometrical knowledge." It is possible not only to see interesting parallels, as well as contrasts, between the two accounts, but to draw rather instructive conclusions about the way in which our "intuitions" of space and time relate to—and presumably result from—our experience of the "real" physical structures. (For a discussion of this point, see STEIN 1991, pp. 155–162.)

One would like to say a similar thing about quantum mechanics—particularly in respect of what it tells us about the structure of *causation,* and our "intuitions" of causation. But this we cannot do. In *this* theory, we just do not know how to "schematize" the observer and the observation. This is a quick way to characterize what I regard as still the basic unsolved philosophical problem of "interpreting" the theory: on a previous occasion, I have expressed the view that "the difficulties [quantum mechanics] presents arise from the fact that *the mode in which this theory 'represents' phenomena* is a radically novel one" (STEIN 1989, p. 59). In other words, here the difficulty of getting the laboratory inside the door of the theory is of a new—and I think still not understood—order.

And on that unresolved dissonance I close.

SUPPLEMENTARY NOTE:

Two questions raised during the discussion at the Congress deserve to be noted.

One of these concerned the point made in the paper concerning the place of "observation" within a theory: it was asked whether, instead of the notion there sketched of the "schematized observer," one could not as well—and in closer accord with traditional (e.g., logical empiricist) terminology—speak of an "idealized" theory of the observer.

To this suggestion I have no serious objection; and I hope it is apparent in the paper itself that I acknowledge a great debt to the logical empiricists, and especially to Carnap, for helping me to clarify my own thoughts about physics. But it has to be understood that the "idealization" involved is an idealized theory *of the observer* in, so to speak, a Pickwickian sense. For instance, in the astronomical example given above, the observer is represented

("ideally" or "schematically") merely by a space-time *locus*. The observational astronomer will infer something about the manipulations to make of the telescope in order to point it so as to receive light from a particular star or planet—but this inference is not one that can be made *within* the theory that incorporates the "idealized observer," because the manipulations in question are not even *describable* in the language of that theory—if they were, a good part of the "idealization" would have been removed ("de-idealized"). Moreover, it should be noted that to be able to infer, even "ideally," that under certain circumstances an observation *will be made,* one would have to include in the ideal theory terms that distinguish conscious from unconscious states of the observer, open from closed eyes, directions of looking, etc. (and noted, in particular, that "ceteris paribus" is not an expression that lends itself to deployment in the context of mathematical argumentation!). Thus unlike, say, the "theory of ideal gases," which does include notions such as temperature, volume, and pressure, central to the study of actual gases, the "theory of idealized observers" would perforce omit those notions that are crucial to the characterization of actual observations. Once this point is well understood, the choice of the word "schematized" or "idealized" is immaterial.

The second question concerned the conjunction of my remark that "technical assistants ... can be taught to perform experiments ... without teaching them the theories those experiments are designed to test," and the closely related claim, expressed just afterwards, that "an 'observation-language' in Carnap's sense is the language in which we ordinarily conduct the business of daily life, [etc.]." It is of course true that technical assistants—and the expert experimenters they assist—need to be masters of a technical discipline, which will include a vocabulary unknown to most of us in "ordinary life." In referring at that point of my talk to the microphone as an example of such an "ordinary object" (with "ordinary properties" that we habitually ascribe to them), I had just this consideration in mind. For what the experimenter needs to be expert in is how to recognize and use the relevant *instruments*; and these—with their properties (including what might be called quirks: their idiosyncrasies and the pitfalls involved in using them, the "other things" that are not always "equal")—*become* familiar (and even "ordinary") in the course of training and use. The microphone is an example of an instrument that has become familiar to most people in the course of the past century or so, although no such thing existed a century and a half ago. But the training and familiarization required for expertness in experimental physics today typically does *not* require a deep study of fundamental physical theories; and, conversely, most theorists today would be lost in a laboratory. (Note that to say this is not to take a stand on the question

of the degree of specialization, in experiment or theory, that is desirable in the education of a physicist. But the state of affairs that *actually obtains* clearly has implications for the structure of the knowledge we *actually have* in physics.)

References

CARNAP, RUDOLF, 1923, *Über die Aufgabe der Physik und die Anwendung des Grundsatzes der Einfachstheit,* Kantstudien 28, pp. 90–107.

CARNAP, RUDOLF, 1963, *Intellectual Autobiography,* in Paul Arthur Schilpp, ed., The Philosophy of Rudolf Carnap (La Salle, Illinois: Open Court).

CHWISTEK, LEON, 1948, *The Limits of Science* (London: Kegan Paul, Trench, Trubner & Co.).

CLAVELIN, MAURICE, 1968, *La philosophie naturelle de Galilée* (Paris: Librairie Armand Colin).

DUHEM, PIERRE, 1954, *The Aim and Structure of Physical Theory,* trans. Philip P. Wiener (Princeton, N. J.: Princeton University Press).

HUYGENS, CHRISTIAAN, 1690, *Treatise on Light,* trans. Silvanus P. Thompson (1912; reprint, Chicago: University of Chicago Press, 1945).

KANT, IMMANUEL, 1781/7 *Kritik der reinen Vernunft,* 1st/2nd eds.

NEWTON, ISAAC, 1729, *The Mathematical Principles of Natural Philosophy,* trans. Andrew Motte, 2 vols. (reprint, London: Dawson's of Pall Mall, 1968).

NEWTON, ISAAC, 1730, *Opticks,* 4th ed. (reprint, New York: Dover, 1952).

STEIN, HOWARD, 1989, *Yes, but ... Some Skeptical Remarks on Realism and Anti--Realism,* Dialectica 43, pp. 47–65.

STEIN, HOWARD, 1990A, *On Locke, 'the Great Huygenius, and the incomparable Mr. Newton',* in Philosophical Perspectives on Newtonian Science, ed. Phillip Bricker and R. I. G. Hughes, (Cambridge, Mass.: MIT Press), pp. 17–47.

STEIN, HOWARD, 1990B, *'From the Phenomena of Motions to the Forces of Nature': Hypothesis or Deduction?* PSA 1990: Proceedings of the 1990 Biennial Meeting of the Philosophy of Science Association, vol. 2, pp. 209-222.

STEIN, HOWARD, 1991, *On Relativity and Openness of the Future,* Philosophy of Science 58, pp. 147–167.

TURNBULL, H. W., 1961, (ED.), *The Correspondence of Isaac Newton,* vol. 3, (Cambridge: Cambridge University Press).

WESTFALL, RICHARD S., 1980, *Never at Rest: A Biography of Isaac Newton* (Cambridge: Cambridge University Press).

WEYL, HERMANN, 1923, *Raum-Zeit-Materie,* 6th ed. (unaltered from the 5th ed., 1923), (Berlin, Heidelberg, New York: Springer-Verlag, 1970).

Logic, Methodology and Philosophy of Science IX
D. Prawitz, B. Skyrms and D. Westerståhl (Editors)

THE LIMITS OF BIOLOGY

GERHARD VOLLMER

Technical University, Braunschweig

1. A question of metabiology

From time to time, we run into discussions of a specific kind and into questions and answers such as the following:

- In a discussion on Nessie: Can you *guarantee* that there is no dinosaur left in Loch Ness? Whereupon the answer might well be: Well, to guarantee the non-existence of an animal transcends the limits of biology.
- In a discussion on Descartes' machine theory of organisms: Do we really *know* that animals feel pain? Does such a claim not go across the limits of biology (or of natural science, of empirical science, of science in general)?
- In discussions on man's place in nature: The evolutionary ladder, the phylogenetic tree, the traditional "scala naturae", or simple complexity considerations, show that man is *superior* to all other living systems (and all the more to inanimate systems). Do we, in making such *evaluative* statements, again trespass the limits of biology?
- As a final example, take the question: Are we *obliged* to preserve on earth as many species as possible? Can such an obligatory claim be justified by biology, or does that go beyond the limits of biology?

In all these cases we seem to run into "the limits of biology", into areas where biologists are no longer competent. What are these limits?

Questions such as these, though posed by biologists, are not genuinely biological questions; at least, they are not answered by way of biological methods, let's say, by outdoor observations or by experiments in a bioscientific laboratory. Questions as to the character of a discipline are rather part of *metascience*, here of philosophy of science. Hence our considerations will be less biological than metabiological.

2. Where the limits don't lie

In trying to specify where the limits of biology do in fact lie, it might be worthwhile first to point out where they do *not* lie.

The limits of biology do not lie where, for some time, they have been supposed to lie: *biology is not imperfect physics.* Philosophy of science has started mainly from physics as the paradigmatic science and was tempted to extend the standards developed there to all sciences. From this perspective, biology could indeed appear as a rather dubious discipline:

- The set of its objects and, therefore, the area of applications is markedly smaller than that of physics and of physical laws.
- Biological laws are much more difficult to find than physical ones.
- Most biological laws seem to allow for exceptions, they are not universally valid even in the field of competence of biology.
- Explanations are less compelling, and many evolutionary facts don't admit of any explanation at all.
- Predictions are difficult, in some cases even completely impossible.
- Therefore biological theories can be confirmed, but hardly refuted.
- According to Popper's criterion of falsifiability — a good empirical theory must be prone to being refuted by experience — biology, and first of all evolutionary biology, would offer nothing but at metaphysical research programme.
- Biological theories are less mathematized and less axiomated than physical ones.

If this characterization were correct and taken seriously, the limits of biology would be determined by the degree to which it meets the standards of physics. Seen from this perspective, biology would appear as a rather questionable science. This perspective, however, is not the only possible one, and, above all, not the only correct one. What could prevent us from turning the table and looking at physics as "lifeless", as "dry", as poor in details, or as awfully abstract? If measured by the numerosity of its object classes, biology is even superior to physics.

By this symmetrization, I don't propose the opposite evaluation, but rather caution against such ratings in general. Only then shall we be able to see and to value the methodological autonomy of biology. And only then will we be able to properly assess the *limits* of biology.

3. Different kinds of limits

A discipline may be limited in several ways: There might be

- theoretical-cognitive limits ("What can we *know*?"),
- limits of curiosity ("What do we *want* to know?"),
- practical-technical limits ("What can we *do*?"),
- ethical-moral limits ("What are we *allowed* to do?").

These limits are not independent of each other. We may distinguish them though not separate them. What we produce, change or prevent, very much depends on our knowledge; and technical progess is, vice versa, a pace-maker for scientific progress. And very often moral limits are recognized and felt only if knowledge and power have reached a certain threshold. This entanglement notwithstanding, we shall try to treat our four questions separately.

We might also ask to what extent the limits of biology are, at the same time, limits of physics, of natural science, of empirical science, of science in general, or of any rational enterprise. It would turn out that most limits of biology apply to, are even characteristic of, all science. But we won't dig too deeply into that problem.

4. Does biology offer certain knowledge?

We might as well extend this question to the more general one whether there is certain knowledge at all. Since we shall deny that question, we need not consider biology separately.

For centuries, people were convinced that certain knowledge existed. Many pathways seemed to lead there: holy scriptures or religous dogmas, divine revelation or Platonic vision, evident axioms or valid inferences, innate ideas or synthetic a priori judgements, experience and reason, observation and experiment, induction or deduction.

At all times, however, there were also sceptics questioning the possibility of certain knowledge. More and more roads to knowledge were found uncertain, subjective, or impassable. Nowadays, the appeal to superhuman authorities appears irrational and dogmatic; intuition and evidence cannot be guaranteed to be intersubjective; and sensory illusions and mass psychoses would depreciate our sensory evidence even if it were intersubjective. Logic and mathematics are structural sciences that owe their certainty — as far as they exhibit such certainty at all — precisely to the fact that they don't even try to describe the world. Success and corroboration don't warrant truth, since occasionally even error may lead to success. Inductive inferences are not necessarily truth-preserving; supposed laws

of nature often prove to be false; and synthetic a priori judgements don't seem to exist.

The arguments for or against the existence, or at least the possibility, of unshakable knowledge cannot be presented here. 2500 years of epistemological critique, however, seem to teach one thing: *certain knowledge about the world doesn't exist.* Whenever we try to find definite proofs, ultimate foundations or final justifications, we find ourselves caught in the notorious Munchausen trilemma, this triple impasse of logical circle, infinite regress and dogmatic break-off. Knowledge in the traditional sense, certain knowledge about the world, ultimate foundations are utopian ideas; all approaches to realize them have failed with sobering regularity.

Biology can't help that. As all science is fallible, preliminary, tentative, or hypothetical, biological knowledge is likewise. From this insight we should not, however, conclude that scientific knowledge, being uncertain, was just speculative and therefore worthless. Between certainty and mere speculation there is a wide spectrum. Philosophy of science tries hard to specify criteria by which theories should be judged and by which rational theory choice is rendered possible. Here necessary and desirable criteria may be distinguished. Necessary features of a good theory in empirical science are non-circularity, consistence, explanatory power, testability and test success; desirable are, in addition, simplicity, applicability and others. Though all these criteria are not sufficient to secure the certainty of scientific hypotheses once dreamed of, they can nevertheless serve to mark out scientific hypotheses as admissable and successful, even as *reliable* or trustworthy.

5. Will biology ever be completed?

Certainty and completeness are different properties. Even if biology does not yield *certain* knowledge, it could still solve all its problems by *preliminary* answers. But even that will never be the case.

Objects of biology are not only plants and animals living now, but also all their phylogenetic forerunners. Therefore, a complete biology should embrace not only descriptions of what there is, but a reconstruction of phylogeny as well. How and why did those highly complex organisms which we find now and which we represent ourselves originate? How did every species, every organ, every tissue, every function, in short every organismic trait come into being? And why? All that would have to be asked and answered by a complete biology.

But there are two million different living species, and even they rep-

resent, according to serious estimations, only one percent of all species which ever existed on earth. To describe and to explain phylogenetically two hundred million species with so many traits and combinations of traits — this is evidently a task that can't ever be performed. A phylogenetic explanation not only requires a description of the evolutionary *path* following which a specific trait originated, not only all initial and all intermediate steps, it must also exhibit the prevailing *selective conditions*, the species- and gene-preserving *functions* of all such traits including again their respective initial and intermediate stages.

Thus, biology will never be complete in this sense. This is true even if physics should come to such a closure. This fundamental incompleteness of biology might be looked at as an advantage or as a disadvantage: as an advantage because, for biologists, the stuff from which questions are made ("der Stoff, aus dem die Fragen sind") will never be exhausted, as a disadvantage because research in biology is a Sisyphean task. Meanwhile at least, it doesn't seem that the science of life could become boring.

6. Does biology provide ultimate explanations?

One of the most important aims of science is to give explanations. Explanations of what? Explanations of all facts which seem to be in need of explanation. Now, what are explanations? Occasionally we are told, to explain something means to reduce it to something familiar. This is not always true. Sometimes — and these are just the great moments of science — scientists frame *new*, so far unheard of, hypotheses by which they then manage to explain either new facts, or facts well-known but hitherto unexplained. Thus, Thomas Hunt Morgan explains many facts of inheritance by using Johannsen's *new* concept of 'gene' and, above that, by framing *new* hypotheses with regard to such genes. And Watson and Crick, by introducing the so far *unknown* or at least unidentified double helix, are able to explain the observations of X-ray diffraction and many more findings. Such explanations are then reductions to something unknown.

Known or not — obviously every explanation not only contains what is explained (the explanandum), but also something by which it explains (the explanans), something to which the explanandum is reduced. The explanatory part — mostly a combination of general laws and special initial and boundary conditions — may then, on its part, become the object of why-questions, hence of deeper explanations. Obviously, there may exist chains, nets, and hierarchies of explanations in which one or more elements serve the purpose of explaining others.

May such a chain, may such a net end somewhere in a natural way? An *infinite* continuation is impossible for practical reasons. And an explanatory *circle* making recourse to facts which were already found to need explanation themselves, would be logically fallacious, a typical vicious circle. An ultimate explanation, then, would be one whose explanans neither needs nor allows for further why-questions. (In philosophy of nature, we could also ask for an *ultimate cause*, for a cause which doesn't have or need any further cause as if, for instance, it could be its own cause (causa sui).)

There is, however, no fact and no factual claim where the why-question would make no sense. Ultimate explanations are therefore impossible, and biology cannot supply them either. It may be that we are not interested in a further explanation; it may be that we don't succeed in finding it although we are interested; and it may be that the explanandum is purely accidental and therefore unexplainable. For whatever reason we have no further explanation — ultimate explanations do not exist.

And yet, biologists talk about "ultimate causes"! How come? The meaning of 'ultimate' is quite different here. *Ultimate* causes in this sense are opposed to *proximate* causes. Proximate causes are, as a rule, physiological *mechanisms*, and proximate explanations make clear how — on the physiological level — a trait is realized or a function is performed. An organismic trait is an *ultimate* cause if it has survival value for the organism (or for its genes), if it enhances its fitness, if it is *functional*.

Whereas physics doesn't care for functions, biology does. Thus, we may say — paradoxically enough — that although there are no ultimate explanations in any science, in biology there are. This is due to the ambiguity of 'ultimate'. It would be preferable to use the word 'teleonomic' and to talk about teleonomic explanations. But since Julian Huxley proposed the ultimate/proximate distinction and since Ernst Mayr made it popular, there is little hope that biologists will change their vocabulary.

7. Are there facts unexplainable by biology?

We did stress that with respect to every fact the question "why?" is perfectly legitimate. From this pervasive legitimacy it does not follow that we always know the answer. Are there facts which are described, but not explained, by biology? Such facts do indeed exist. We may divide them into three groups.

The first group comprises facts which are explained not by biology but by another discipline. Thus, not only search physicists into the origin of stars, but likewise biologists search into the origin of living systems.

However, whereas physicists give a physical answer to their star question, biologists don't get a biological answer to their life question. Genetics and developmental biology, it is true, explain (tentatively) how from individuals new individuals arise, and the theory of evolution explains (tentatively) how from species new species arise. But how the *first* organisms could or did arise, they don't explain. They are unable to do so because they *presuppose* the existence of living systems, of species, of life. Evidently the first living systems could not originate from living systems (because then they would not have been the first ones), but only from non-living systems. And to non-living systems the laws of biology do not, of course, apply yet. Therefore, the origin of life can and will be explained, if at all, only by physics and chemistry. In view of the usual and useful division of labor between biology and physics (on which later), this limitation of biology is easily understood and easily tolerated.

The second group of unexplained facts embraces *chance events* and their consequences. Chance events have no causes and, therefore, no explanations. (The phrase "this can only be explained by chance" must, if permitted at all, be understood *metaphorically*.) It is true that chance events are, as a rule, not completely lawless; they obey statistical laws. Such laws are, however, applicable only to whole classes of events. They cannot explain singular events.

In biology chance events play a constitutive role. The immensely large number of existing species, and even the totality of all living systems having once existed or existing now, is still forbiddingly small compared to the number of all the different organisms which could exist in principle. From the huge spectrum of possible living systems only a minute fraction will be realized even in the farthest future. How are the systems to be realized selected from the domain of the possible ones? We know that this selection occurs under the constitutive influence of several *chance factors*: undirected mutations, fluctuations of population size, random recombinations of genes. Thus biological systems always exhibit accidental aspects which cannot be described, explained or predicted by deterministic or probabilistic laws. Therefore, the limits of repeatability, explainability and predictability are much narrower in biology than in physics. That evolutionary biology could not make testable predictions at all (as, following Popper, some people claim) is, however, not true.

The third group of unexplained facts has been discovered only recently. This is the behavior of *chaotic systems*. A system is called chaotic if *arbitrarily small* alterations of the initial conditions may lead to completely different behavioral results. This is also possible in deterministic systems (deterministic chaos), especially if the system exhibits, as organisms usu-

ally do, feedback and hence nonlinear behavior. Since every measurement is inaccurate to a certain degree, the future of a chaotic system is not always predictable and very often not even explainable afterwards. Just as if nature wanted to compensate for that, chaotic systems open up the *chance* that, despite their fairly chaotic behavior, they might still be described or even understood, at least qualitatively, by deterministic laws. Thus, paradoxically enough, even chaos brings order to biology!

Chaotic behavior could prevail in cell-to-cell contacts, in embryogenesis and, more generally, in morphogenesis, in protein interactions, in the formation of patterns, especially of spirals (sunflower, pine cones, leaf arrangements), in the formation and perturbation of physiological rhythms, in processes in the brain and in the central nervous system, as well as in some illnesses, e.g. in cancer, and finally in whole ecosystems with their characteristic stability problems.

8. Limits of understanding?

The concept of understanding has many facets.

We may, first of all, understand linguistic expressions: words, sentences, theories. We understand a *word* if we know its meaning. (We don't define 'meaning' here.) We understand a *sentence* if we know the words occurring in it and if we know which relation it establishes between them, hence if we know, what it claims, states, commands, forbids, asks, and so on. We understand a *theory* of empirical science, e.g., the theory of heredity, if we understand its main concepts and propositions and if we know which problems it solves and to what degree it solves them better, or worse, than competing theories. It is obvious that for these kinds of understanding there may be limits; however, they do not particularly concern biology and will not be further discussed here.

Apart from linguistic expressions, we also try to understand real systems. For *non-living* systems, 'understanding' is essentially synonymous with 'explanation'. I understand an *object*, e.g., a carbon atom, if I know its special properties, and if I can describe and explain these properties, especially its structure and its behavior. Sometimes, however, we also want to know how a carbon atom comes into being, perhaps even how it can be manufactured. I understand a *process*, e.g., a sun eclipse, if I know how and why it occurs and why it occurs that way and not differently.

In *living* systems, we must add to these properties their *functional* traits. I understand blood circulation if I (cannot only explain it, but if I also) know what it is good for, which function it has, how it secures or enables the organism's survival. In that sense we may also understand plants

and animals. Here again, the limits of understanding coincide with the limits of explanation (functional explanations included). And *complete* understanding — where nothing would be left to ask — is as unattainable as are ultimate explanations.

In the interhuman area, however, we use a still more ambitious concept of understanding. To understand a *human being* obviously means more than to know and to explain his or her traits, his coming into being or her life-serving functions. We know about ourselves that, over and above all that, we do have ideas, memories, intentions, motives, feelings, emotions. There, we have *direct* access at most to our own mental states and processes. We are, nevertheless, ready to ascribe such "mental life" to other humans as well. Therefore, I understand a *human being* only if I also know her inner states, especially her feelings and motives. I understand his *actions* if I know his motives, that is, if I know the wishes and aims that made him act. Sometimes, we even feel that, in order to understand somebody, we should duplicate his or her feelings.

Doubtless this understanding has *limits*. Sometimes, we don't even understand ourselves. It is even more difficult to put oneself, so to speak, into the thoughts and feelings of other people, to have, in a verbal sense, fellow-feeling, com-passion, or sym-pathy. Strictly speaking, we can never know for certain what another person is feeling or thinking, not even whether she feels or thinks at all. At any rate, we cannot prove it. But, as we know, I cannot even prove, to you or to me, that I existed already yesterday. Therefore, from these considerations *no specific limit* follows for interhuman understanding. It can always be increased and improved upon.

9. Do we understand animals?

The motives which induce us to impute feelings and ideas to other *human beings*, all lie in their behavior: in their gestures, in their facial expressions, in their nonverbal utterances, and of course in what they say. In doing that we make the obvious conjecture that if their behavior is similar to ours, similar inner states and processes are at work. This inference by analogy is, as we surely know, not conclusive; it is, nevertheless, one of the scientist's standard tools. In the case of interhuman behavior, it is so natural and subjectively inevitable that Karl Bühler and Konrad Lorenz liked to speak of a *Du-Evidenz* (the evidence of the thou): We cannot help seeing in our human vis-a-vis another person with intentions, thoughts, and feelings.

Quite independently, this inference by analogy is strongly supported by

our knowledge about our biological relatedness and about the similarity of our brains and nervous processes. Since there are — due to age, sex, race or culture — varying degrees of similarity, our understanding of fellow-man and fellow-women reaches varyingly far.

In a weaker form these arguments apply to our relation with *animals*. It is true, they cannot talk to us, they don't communicate with us in our language; but there are even human beings where this is not possible, and, what is more, language is not the only access to others, hence not always necessary. With animals we share the environment, with the higher animals moreover a long evolutionary past. Our sense organs and central nervous systems are phylogenetically related and therefore similar to varying degrees. The longer our common history is, the later the phylogenetic ramification has taken place, the greater are our similarities and, therefore, the chances for sym-pathy, for understanding.

There is no serious doubt, then, that higher animals may suffer and feel pain. In discussions about experiments with or cruelty to animals, about keeping animals in cages or hens in batteries, the problem is not whether animals may suffer; we think we know that, and as biologists we think we can show (though not prove!) and explain it. Therefore we must check how we may diminish or *prevent* such suffering. Here again the biologist, especially the neurobiologist, is qualified: (s)he can judge whether an experiment with animals will sufficiently advance our knowledge, whether a simpler organism would do, whether a living animal is really needed, whether there is a more considerate treatment, whether narcosis, local anesthesia, or nerve cutting might bring relief to the animal. First of all, however, it must be clear whether and how far we are ready to *put up* with animal suffering in view of our other goals and values. This ethical or moral question cannot be answered by the methods of biology alone. Nevertheless, the biologist's knowledge and competence plays a decisive role in such a discussion.

10. Limits of curiosity?

Curiosity and playfulness are vital drives in higher organisms, especially in man. They are essential because they make individual *learning* possible. Thus, environmental conditions and, even more important, environmental changes to which genetic programming could never prepare us, are easily mastered. Curious and playful are, first of all, the youngsters. Man, however, distinguishes himself from all animals by staying curious and open to the world up to his greatest age. Looked at from ethology, man keeps a typical juvenile trait even as an adult. (Therefore Konrad Lorenz,

borrowing from zoology, liked to use the term "neoteny".) 'Homo ludens' (Huizinga) is not the only appropriate characterization of man, but nevertheless quite to the point. Even science owes its existence to human curiosity. And since there will be human beings again and again, who want not only to learn known facts, but to discover new things, human curiosity in this sense is *without limits*.

Biologists, however, are wont to think in cost-benefit relations. Even if our curiosity is unlimited, it may cost more and more to satisfy this curiosity. In fact, scientific progress becomes more and more expensive. Scientific discoveries may be likened to the treasures of the soil: Nearly all raw materials which could *easily* be gained have been used up by now, especially those on the surface of the earth. In order to get more of them, we must dig deeper and deeper. Likewise, in science nearly all *simple* discoveries have been made such that further progress needs more and more education and more and more technical tools. Therefore it could perfectly well happen that the satisfaction of our curiosity would not compensate for the respective costs. In his book "The paradoxes of progress", the molecular biologist Gunther Stent calls this effect the "principle of diminishing return". Economists know that phenomenon as "marginal utility". Even in biology with its inexhaustible wealth of unsolved problems it could happen that we stop fundamental research, not for moral reasons but for cost-benefit considerations. This crucial point is far from being reached, and it is even impossible to say exactly where it is situated. Moreover, it may be shifted by changing practical needs and by the extension of our technical abilities. But it certainly exists.

11. Limits due to useful division of labor

As we have seen, biology has a *richer* spectrum of questions than physics. We could as well express this fact by saying that physics *limits* itself in its questions. That organismic structures support the survival of an individual, of its genes, or of a species, and that they are *useful* in that sense, cannot, of course, escape the physicist. Even so, physics does not use or introduce concepts like function, utility, fitness: they are reserved to biology. The reason is not that physics couldn't say anything about organisms. The physical laws are not restricted to non-living objects. If an organism could violate the law of gravitation or the conservation of energy then these laws would be false; they are claimed to be *universally* valid.

This, then, is the decisive difference between physics and biology: phys-

ics investigates *all* systems, non-living and living ones, and it searches for laws which apply to *all* these systems. Those phenomena, hoewever, which are found only with organisms, and those laws which apply only to them, are traditionally reserved for biology. One limit of biology which is historically conditioned lies in the fact that it does just not care for non-living systems.

This limit, however, is not fixed once and for all. For, which systems are alive, or even better, which systems are *said* to be alive, is itself dependent on new discoveries and useful conventions. When it was found out recently that RNA molecules may replicate biochemists were still free to regard these molecules either as *living* (because they can replicate) or as *non-living* (because they don't evolve to higher systems). Language and intuition cannot anticipate such a decision because they are not "tailored" for such borderline cases.

A similar division of labor as that between physics and biology (or more precisely: between chemistry and biology) obtains between biology and psychology. Again it is impossible to draw a sharp line between these two disciplines: When comparative ethology was still in its beginnings, it was, tellingly enough, called 'animal psychology', operating precisely in the open area between biology and psychology, hence in the former no man's land between the natural sciences and the humanities. It is, however, usual and *suitable* that biology *restricts* itself to scientific methods and, thereby, to such traits which are common to all or many organisms, traits which can be objectified and which can be investigated without introspection (although the latter might be useful even there). Just as physics investigates and applies to living systems, biology also investigates organisms with conciousness (including man), but no mental phenomena as such. Yet again such concepts as conditioned reaction, learning, aggression, or the existence of a discipline like psychobiology, show that a rigorous borderline between biology and psychology does simply not exist.

12. Limits due to wise self-limitation

Obviously, biology as a natural science — and even more general: as an empirical science — excludes certain questions which are asked elsewhere. Questions as to the purpose of the universe, to the goal of being, to the meaning of life, to a creator or ruler of the world, to the roots of validity, or to moral justifications, are not only not answered in biology: they are not even posed there. Inside empirical science, questions are legitimate only if they concern *facts* and if they have at least a *chance* to be answered in the framework of the methods of empirical science.

Again, the borderline is not sharp. In fact, the methods of empirical science, its material and mental tools, its aims and claims, its domains of competence and of application, have drastically changed in the course of time. Isaac Newton (1643–1727), the creator of physics as a science, is still convinced that from time to time God must fix the planetary system in order to preserve it from instability and collapse. The French physicist Maupertuis (1698–1759) interprets the newly discovered extremal principles of mechanics as scientific evidence for the activity of a wisely planning creator and as a physical instantiation of Leibniz' thesis that this world is the best of all possible worlds. And far into the 19th century, the stunning adaptation of organismic structures is looked upon as a visible sign of an ordering hand. Not until Charles Darwin (1809–1882) is this "teleological proof for the existence of God" dismantled, the observed adaptation of organisms now being explained from inside biology, first of all by natural selection.

Thus, whereas the borderlines between biology and the neighboring sciences — physics, chemistry, and psychology — are blurred more and more, the borderlines between biology and metaphysics, biology and theology, biology and ethics, have become even sharper. It was finally recognized that the relations supposed or at least hoped for did just not exist and that the empirical sciences owe their success to this very self-limitation.

Thus the claim appears reasonable that the empirical sciences have been successful by fine-tuning both the admissible *questions* and the *methods* permitted in answering those questions. All this does not mean, of course, that the disciplines characterized here as different and separable, had nothing to do with each other. To the peculiar relation between biology and ethics we shall come back.

13. Limits of feasibility?

There is no doubt that the quest for power is — besides pure curiosity — the main motive for the scientist. Often enough, practical needs, possible applications, technical progress, "social relevance", determine the interests of scientists and, first of all, of their financiers.

Nevertheless, man can obviously not do all he wants to do (quite independently of the question whether he can *desire* what he wants to do). Where are the limits of feasibility, where do they lie in biology?

One important limit is set by the laws of nature. Laws of nature are (or describe) regularities in the behavior of real systems. They tell us what, under specified conditions, will happen. Other kinds of behavior are then, given the same conditions, impossible. Therefore, we may as

well interpret the laws of nature as impossibility statements: the law of energy conservation implies the impossibility of a perpetuum mobile; from the law of entropy increase it follows that heat cannot "of itself" pass over from cold to warm; and according to Nernst's heat theorem (the third law of thermodynamics), it is impossible to reach the absolute zero of temperature. Similarly it is, according to Hardy-Weinberg's law, impossible to eliminate a recessive hereditary disease by removing all pure disease carriers. Since, however, all knowledge is preliminary and fallible, we cannot exclude such possibilities with absolute certainty. Even a law such as the conservation of energy, well-tried, never refuted and intimately interwoven with all empirical science, could in principle turn out to be false. Thus, even claims to the impossible are endowed with the proviso of possible error.

On top of that, many claims to the impossible have turned out to be erroneous in the history of science. Thus it was claimed that man could not live above 3000 meters (Cauchy), that the chemical composition of stars could never be found out (Comte), that aeroplanes should be impossible (Siemens), that rockets could not accelerate in empty space (New York Times), that organic substances could not be synthesized from inorganic ones (vitalism), and so on. All these assertions on supposed impossibilities, on supposed limits of feasibility, were found to be erroneous.

This negative score should warn us. We may confidently declare impossible whatever contradicts the laws of nature; what is, however, possible or impossible inside the framework of natural laws, is quite difficult to determine. Will it be possible to clone human beings? To grow a mammal completely outside a placenta? To synthesize a whole organism from inanimate matter? To decode completely the human genome and to modify it deliberately? To cure hereditary diseases, to eradicate AIDS, to prevent cancer? There are no laws of nature which would exclude in principle such possibilities. Our knowledge is limited, especially our knowledge about the future of our knowledge — and of our abilities.

In the long run, however, the decisive question will not be what we are *able* to do, but what we are *allowed* to do.

14. Biology as a "science of the century"

"Die Jahrhundertwissenschaft" ("The science of the century") is the title of a book by the German historian of science Armin Hermann. As we might expect he presents physics as the most important science of our century. Physics was indeed decisive for the *first* half of our century. In 1900, Max Planck laid the foundations for quantum theory, possibly the

greatest revolution physics has ever seen. The first half of our century ends with the use of nuclear reactors on the one side, of nuclear weapons on the other.

For the *second* half of our century, however, *biology* seems to be the dominant science. In 1952, Watson and Crick find the double helix, and molecular biology has made unforeseen progress since. And again we feel that the second half of our century also ends with rather ambivalent progresses, this time of applied biology.

In 1978, another German author, Jost Herbig, opens a book on genetic engineering with the following words: "Biology has reached the critical stage of a science: it is constructing nature. The age of synthetic biology has now begun." Perhaps it is this what makes a science a science of the century: it constructs nature? Then we could even predict the sciences of the next, of the 21st century: the neurosciences. Will they also construct nature, will they change human beings, will they create brains, will they become synthetic sciences? And will there then ensue another bad awakening? Sciences of the century seem to distinguish themselves by being highly celebrated at the beginning and deeply damned at the end. How come?

The answer is, I suppose, very simple. For thousands of years, man could not do much more than was allowed. In the last centuries, however, the natural sciences developed very fast, even explosively in our century. Along with human knowledge human power increased; whereas what was permitted did not change essentially. Thus, human power by far outgrew what was allowed, and this is a *qualitatively new situation.*

For centuries, it seemed quite unobjectionable for a researcher to satisfy without restraint her thirst for knowledge. The purity of science virtually consisted in ranking truth above all and not caring for applications. Indeed, as long as there were no dangers combined with it, truth rightly could be seen as the upmost good. Warning hints as the biblical tree of knowledge, the magician's apprentice in Goethe's poem, or Mary Shelley's Frankenstein, could be attributed to a far future.

This has now changed. The knowledge of mankind has opened new possibilities which go far beyond the satisfication of urgent needs. We create means and tools that can be used for the weal and woe of mankind. (S)he who nowadays strives just for truth, is looked at as irresponsible. Thus, science meets with *limits* which formerly were known but not felt. What should we do about that?

15. Biology does not supply moral norms

It would certainly be wrong to forbid all research whose results could *possibly* be misused. We can say quite clearly and shortly what then would remain of science: nothing. Even mathematics can be applied, and even the prime numbers, innocent as they seem to be, are of practical and even of military use in modern coding systems.

It would also be misguided to look for values and for norms in the empirical sciences themselves or to try to derive them from scientific findings. Pure norms cannot be gained from pure facts. If you try to do it anyway, you commit the naturalistic fallacy. From the fact alone that a specific behavior has come out from and has been *successful* in evolution, it does not follow, for instance, that it were *good* or *right*. What is *natural* is not automatically *right*.

That descriptive statements alone are not sufficient to yield normative ones, has been thoroughly investigated by logicians and has been shown with sufficient rigor. As we have stressed already, biology, and science in general, owe their success to their self-restriction to the factual and to the fine-tuning of their questions to what is methodically attainable. Being an empirical science, biology is not able to investigate or to yield moral norms; they simply do not lie in its task domain nor in its competence.

Even those norms scientists normally adhere to in their research activity are not sufficient for a general ethical orientation. It is true that the "ethos of science" is exemplary in several respects: it asks you to aim at truth, at objectivity, at precision, it requires symmetrical argumentation, criticizability, internationality, and so on. It is, however, only a *partial* ethos which is not sufficient for the regulation of personal or political relations. That's no wonder: the upmost value of the ethos of science is knowledge; for this it is suitable, and here it is successful. Other values like justice, liberty, or love, are just irrelevant to the ethics of science. Thus moral norms can be gained neither from the results nor from the normative behavior of the natural scientist. Having stressed this, and having identified another limit of biology, we could stop right here. But we want to go one step further.

Man as a social being is absolutely dependent on social norms. Where can, where should he take them from? Should they be supplied or even prescribed to him by others? Should he listen to the priest, to the philosopher, to the lawyer, to the politician? Can someone outside tell the gene technologist what (s)he should or should not do?

This way is sometimes comfortable, but not advisable. The slogan of enlightment is *self-thinking*. It is all right to listen to the arguments of others; but decisions are everyone's own matter. Yet a responsible decision

needs both factual knowledge and moral orientation. Where do they come from, and how do they interact?

16. Facts and norms

From facts alone no norms can follow. Therefore a biologist, searching for practical directives will not get along with biology alone. *Without* factual knowledge, however, things don't work either; it is for this very reason that normative approaches starting from "purely" philosophical positions tend to be far from the mark, being too general, too abstract, too ivory-towered.

What we need are, first of all, one or several *basic norms*. They are, on their part, not justified; ultimate justifications (of norms) are no more feasible than ultimate explanations (of facts). We may hope, however, to meet with unanimous approval for such basic norms. This assent cannot be extorted by way of argument; it can only be stated. From these basic norms more norms are derived *by adding factual knowledge*.

An example might illustrate that point. Suppose we had come to commonly accept the following norm as basic: "We should take care that future generations are not worse off than we are!" (This may be debated; but we must start somewhere.) This norm alone does not prescribe any specific action. Now factual knowledge will inform us that the world population is increasing and that with growing world population the living conditions will deteriorate. (This may again be debated; our issue here is, however, not the *correctness* of factual claims, but rather their *role* in the gaining of moral norms.) Combining now our basic norm with our pertinent factual knowledge, we may derive the norm that we should not multiply further. In combination with additional knowledge about the possibilities of birth control (especially about contraception) more, and more concrete, norms can be derived.

Both parts — basic norms and facts — are indispensable here for the derivation of norms. Therefore the interplay of facts and norms should not be seen *additively*, such as if every term of the sum could already offer something. It should rather be interpreted *multiplicatively*: if one of the two factors is nil, the "product" is also nil — we have nothing then. Only if both elements are combined in an adequate manner, the result can be "positive". Of course, there are more possibilities to combine elements constructively: we may multiply matrices, or cross-breed animals. Multiplication is, however, the simplest model for the cooperation of facts and norms and for their being dependent on each other.

This consideration should make clear what the scientist's genuine contribution to the establishing of norms consists in: (s)he provides the factual knowledge necessary for the derivation of more norms from basic ones. Both this knowledge and these basic norms are *indispensable.* And only our insight into the moral limits of biology enables us to see in its true light the constitutive role of biology even for ethical-moral issues.

Logic, Methodology and Philosophy of Science IX
D. Prawitz, B. Skyrms and D. Westerståhl (Editors)

COGNITIVE SCIENCE AS REVERSE ENGINEERING SEVERAL MEANINGS OF "TOP-DOWN" AND "BOTTOM-UP"

DANIEL C. DENNETT

Center for Cognitive Studies, Tufts University, Medford, MA 02155

The vivid terms, "Top-down" and "Bottom-up" have become popular in several different contexts in cognitive science. My task today is to sort out some different meanings and comment on the relations between them, and their implications for cognitive science.

Models and methodologies

To a first approximation, the terms are used to characterize both research methodologies on the one hand, and models (or features of models) on the other. I shall be primarily concerned with the issues surrounding top-down versus bottom-up methodologies, but we risk confusion with the other meaning if we don't pause first to illustrate it, and thereby isolate it as a topic for another occasion. Let's briefly consider, then, the top-down versus bottom-up polarity in models of a particular cognitive capacity, language comprehension.

When a person perceives (and comprehends) speech, processes occur in the brain which must be partly determined bottom-up, by the input and partly determined top-down, by effects from on high, such as interpretive dispositions in the perceiver due to the perceiver's particular knowledge and interests. (Much the same contrast, which of course is redolent of Kantian themes, is made by the terms "data-driven" and "expectation-driven").

There is no controversy, so far as I know, about the need for this dual source of determination, but only about their relative importance, and when, where, and how the top-down influences are achieved. For instance, speech perception cannot be entirely data-driven because not only are the brains of those who know no Chinese not driven by Chinese speech in

the same ways as the brains of those who are native Chinese speakers, but also, those who know Chinese but are ignorant of, or bored by, chess-talk, have brains that will not respond to Chinese chess-talk in the way the brains of Chinese-speaking chess-mavens are. This is true even at the level of perception: what you hear — and not just whether you notice ambiguities, and are susceptible to garden-path parsings, for instance — is in some measure a function of what sorts of expectations you are equipped to have. Two anecdotes will make the issue vivid.

The philosopher Samuel Alexander, was hard of hearing in his old age, and used an ear trumptet. One day a colleague came up to him in the common room at Manchester University, and attempted to introduce a visiting American philosopher to him. "THIS IS PROFESSOR JONES, FROM AMERICA!" he bellowed into the ear trumpet. "Yes, Yes, Jones, from America" echoed Alexander, smiling. "HE'S A PROFESSOR OF BUSINESS ETHICS!" continued the colleague. "What?" replied Alexander. "BUSINESS ETHICS!" "What? Professor of what?" "PROFESSOR OF BUSINESS ETHICS!" Alexander shook his head and gave up: "Sorry. I can't get it. Sounds just like 'business ethics'!"

Alexander's comprehension machinery was apparently set with too strong a top-down component (though in fact he apparently perceived the stimulus just fine).

An AI speech-understanding system whose development was funded by DARPA (Defense Advanced Research Projects Agency), was being given its debut before the Pentagon brass at Carnegie Mellon University some years ago. To show off the capabilities of the system, it had been attached as the "front end" or "user interface" on a chess-playing program. The general was to play white, and it was explained to him that he should simply tell the computer what move he wanted to make. The general stepped up to the mike and cleared his throat — which the computer immediately interpreted as "Pawn to King-4." Again, too much top-down, not enough bottom-up.

In these contexts, the trade-off between top-down and bottom-up is a design parameter of a model that might, in principle, be tuned to fit the circumstances. You might well want the computer to "hear" "Con to Ping-4" as "pawn to King-4" without even recognizing that it was making an improvement on the input. In these contexts, "top-down" refers to a contribution from "on high" — from the central, topmost information stores — to what is coming "up" from the transducers or sense organs. Enthusiasm for models that have provision for large top-down effects has waxed and waned over the years, from the euphoria of "new look" theories of perception, which emphasized the way perception went "beyond

the information given" in Jerry Bruner's oft-quoted phrase, to the dysphoria of Jerry Fodor's (1983) encapsulated modules, which are deemed to be entirely data-driven, utterly "cognitively impenetrable" to downward effects.

David Marr's (1982) theory of vision is a prime example of a model that stresses the power of purely bottom-up processes, which can, Marr stressed, squeeze a lot more out of the data than earlier theorists had supposed. The issue is complicated by the fact that the way in which Marr's model (and subsequent Marr-inspired models) squeeze so much out of the data is in part a matter of fixed or "innate" biases that amount to presuppositions of the machinery — such as the so-called rigidity assumption that permits disambiguation of shape from motion under certain circumstances. Is the rigidity assumption tacitly embodied in the hardware a top-down contribution? If it were an optional hypothesis tendered for the nonce by the individual perceiver, it would be a paradigmatic top-down influence. But since it is a fixed design feature of the machinery, no actual transmission of "descending" effects occurs; the flow of information is all in one inward or upward direction. Leaving the further discussion of these matters for another occasion, we can use the example of Marr to highlight the difference between the two main senses of "top-down". While Marr, as I have just shown, was a champion of the power of bottom-up models of perception (at least in vision), he was also a main spokesperson for the top-down vision of methodology, in his celebrated three-level cascade of the computational, the algorithmic and the physical level. It is hopeless, Marr argued, to try to build cognitive science models from the bottom-up: by first modeling the action of neurons (or synapses or the molecular chemistry of neurotransmitter production), and then modeling the action of cell assemblies, and then tracts, and then whole systems (the visual cortex, the hippocampal system, the reticular system). You won't be able to see the woods for the trees. First, he insisted, you had to have a clear vision of what the task or function was that the neural machinery was designed to execute. This specification was at what he called, misleadingly, the computational level: it specified "the function" the machinery was supposed to compute and an assay of the inputs available for that computation. With the computational level specification in hand, he claimed, one could then make progress on the next level down, the algorithmic level, by specifying an algorithm (one of the many logically possible algorithms) that actually computed that function. Here the specification is constrained, somewhat, by the molar physical features of the machinery: maximum speed of computation, for instance, would restrict the class of algorithms, and so would macro-architectural features

dictating when and under what conditions various subcomponents could interact. Finally, with the algorithmic level more or less under control, one could address the question of actual implementation at the physical level.

Marr's obiter dicta on methodology gave compact and influential expression to what were already reigning assumptions in Artificial Intelligence. If AI is considered as primarily an engineering discipline, whose goal is to create intelligent robots or thinking machines, then it is quite obvious that standard engineering principles should guide the research activity: first you try to describe, as generally as possible, the capacities or competences you want to design, and then you try to specify, at an abstract level, how you would implement these capacities, and then, with these design parameters tentatively or defeasibly fixed, you proceed to the nitty-gritty of physical realization.

Certainly a great deal of research in AI — probably the bulk of it — is addressed to issues formulated in this top-down way. The sorts of questions addressed concern, for instance, the computation of three-dimensional structure from two-dimensional frames of input, the extraction of syntactic and semantic structure from symbol strings or acoustic signals, the use of meta-planning in the optimization of plans under various constraints, and so forth. The task to be accomplished is assumed (or carefully developed, and contrasted with alternative tasks or objectives) at the outset, and then constraints and problems in the execution of the task are identified and dealt with.

This methodology is a straightforward application of standard ("forward") engineering to the goal of creating artificial intelligences. This is how one designs and builds a clock, a water pump, or a bicycle, and so it is also how one should design and build a robot. The client or customer, if you like, describes the sought for object, and the client is the boss, who sets in motion a top-down process. This top-down design process is not simply a one-way street, however, with hierarchical delegation of unrevisable orders to subordinate teams of designers. It is understood that as subordinates attempt to solve the design problems they have been given, they are likely to find good reasons for recommending revisions in their own tasks, by uncovering heretofore unrecognized opportunities for savings, novel methods of simplifying or uniting subtasks, and the like. One expects the process to gravitate towards better and better designs, with not even the highest level of specification immune to revision. (The client said he wanted a solar-powered elevator, but has been persuaded, eventually, that a wind-powered escalator better fits his needs.)

Marr's top-down principles are an adaptation, then, of standard AI

methodology. Another expression of much the same set of attitudes is my distinction between the intentional stance, the design stance and the physical stance, and my characterization of the methodology of AI as the gradual elimination of the intentional through a cascade of homunculi. One starts with the ideal specification of an agent (a robot, for instance) in terms of what the agent ought to know or believe, and want, what information-gathering powers it should have, and what capacities for (intentional) action. It then becomes an engineering task to design such an intentional system, typically by breaking it up into organized teams of sub-agents, smaller, more stupid homunculi, until finally all the homunculi have been discharged — replaced by machines. A third vision with the same inspiration is Allen Newell's distinction between what he calls the knowledge level and the physical symbol system level. It might seem at first that Newell simply lumps together the algorithmic level and the physical level, the design stance and the physical stance, but in fact he has made the same distinctions, while insisting, wisely, that it is very important for the designer to bear in mind the actual temporal and spatial constraints on architectures when working on the algorithmic level. So far as I can see, there is only a difference in emphasis between Marr, Newell and me on these matters.

What all three of us have had in common are several things:

(1) stress on being able (in principle) to specify the function computed (the knowledge level or intentional level) independently of the other levels.
(2) an optimistic assumption of a specific sort of functionalism: one that presupposes that the concept of the function of a particular cognitive system or subsystem can be specified. (It is the function which is to be optimally implemented.)
(3) A willingness to view psychology or cognitive science as reverse engineering in a rather straightforward way.

Reverse engineering is just what the term implies: the interpretation of an already existing artifact by an analysis of the design considerations that must have governed its creation.

There is a phenomenon analogous to convergent evolution in engineering: entirely independent design teams come up with virtually the same solution to a design problem. This is not surprising, and is even highly predictable, the more constraints there are, the better specified the task is. Ask five different design teams to design a wooden bridge to span a particular gorge and capable of bearing a particular maximum load, and it is to be expected that the independently conceived designs will be very similar: the efficient ways of exploiting the strengths and weaknesses of

wood are well-known and limited.

But when different engineering teams must design the same sort of thing a more usual tactic is to borrow from each other. When Raytheon wants to make an electronic widget to compete with General Electric's widget, they buy several of GE's widget, and proceed to analyze them: that's reverse engineering. They run them, benchmark them, x-ray them, take them apart, and subject every part of them to interpretive analysis: why did GE make these wires so heavy? What are these extra ROM registers for? Is this a double layer of insulation, and if so, why did they bother with it? Notice that the reigning assumption is that all these "why" questions have answers. Everything has a raison d'etre; GE did nothing in vain.

Of course if the wisdom of the reverse engineers includes a healthy helping of self-knowledge, they will recognize that this default assumption of optimality is too strong: sometimes engineers put stupid, pointless things in their designs, sometimes they forget to remove things that no longer have a function, sometimes they overlook retrospectively obvious shortcuts. But still, optimality must be the default asssumption; if the reverse engineers can't assume that there is a good rationale for the features they observe, they can't even begin their analysis.

What Marr, Newell, and I (along with just about everyone in AI) have long assumed is that this method of reverse engineering was the right way to do cognitive science. Whether you consider AI to be forward engineering (just build me a robot, however you want) or reverse engineering (prove, through building, that you have figured out how the human mechanism works), the same principles apply.

And within limits, the results have been not just satisfactory; they have been virtually definitive of cognitive science. That is, what makes a neuroscientist a cognitive neuroscientist, for instance, is the acceptance, to some degree, of this project of reverse engineering. One benefit of this attitude has been the reversal of a relentlessly stodgy and constructive attitude among some neuroscientists, who advocated abstention from all "speculation" that could not be anchored firmly to what is known about the specific activities in specific neural tracts — with the result that they often had scant idea what they were looking for in the way of functional contribution from their assemblies. (A blatant example would be theories of vision that could, with a certain lack of charity, be described as theories of television — as if the task of the visual system were to produce an inner motion picture somewhere in the brain.)

But as Ramachandran (1985) and others (e.g., Hofstadter — see Dennett, 1987) were soon to point out, Marr's top-down vision has its own blind spot: it over-idealizes the design problem, by presupposing first

that one could specify the function of vision (or of some other capacity of the brain), and second, that this function was optimally executed by the machinery.

That is not the way Mother Nature designs systems. In the evolutionary processes of natural selection, goal-specifications are not set in advance — problems are not formulated and then proposed, and no selective forces guarantee optimal "solutions" in any case. If in retrospect we can identify a goal that has been optimally or suboptimally achieved by the evolutionary design process, this is something of a misrepresentation of history. This observation, often expressed by Richard Lewontin in his criticism of adaptationism, must be carefully put if it is to be anything but an attack on a straw man. Marr and others (including all but the silliest adaptationists) know perfectly well that the historical design process of evolution doesn't proceed by an exact analogue of the top-down engineering process, and in their interpretations of design they are not committing that simple fallacy of misimputing history. They have presupposed, however — and this is the target of a more interesting and defensible objection — that in spite of the difference in the design processes, reverse engineering is just as applicable a methodology to systems designed by Nature, as to systems designed by engineers. Their presupposition, in other words, has been that even though the forward processes have been different, the products are of the same sort, so that the reverse process of functional analysis should work as well on both sorts of product.

A cautious version of this assumption would be to note that the judicious application of reverse engineering to artifacts already invokes the appreciation of historical accident, sub-optimal jury-rigs, and the like, so there is no reason why the same techniques, applied to organisms and their subsystems, shouldn't yield a sound understanding of their design. And literally thousands of examples of successful application of the techniques of reverse engineering to biology could be cited. Some would go so far (I am one of them) as to state that what biology is, is the reverse engineering of natural systems. That is what makes it the special science that it is and distinguishes it from the other physical sciences,

But if this is so, we must still take note of several further problems that make the reverse engineering of natural systems substantially more difficult than the reverse engineering of artifacts, unless we supplement it with a significantly different methodology, which might be called bottom-up reverse engineering — or, as its proponents prefer to call it: Artificial Life.

The Artificial Life movement (AL), inaugurated a few years ago with a conference at Los Alamos (Langton, 1989), exhibits the same early enthu-

siasm (and silly overenthusiasm) that accompanied the birth of AI in the early 60's. In my opinion, it promises to deliver even more insight than AI. The definitive difference between AI and AL is, I think, the role of bottom-up thinking in the latter. Let me explain.

A typical AL project explores the large scale and long range effects of the interaction between many small scale elements (perhaps all alike, perhaps populations of different types). One starts with a specification of the little bits, and tries to move towards a description of the behavior of the larger ensembles. Familiar instances that predate the official Artificial Life title are John Horton Conway's game of Life and other cellular automata, and, of course, connectionist models of networks, neural and otherwise. It is important to realize that connectionist models are just one family within the larger order of AL models.

One of the virtues of AL modeling strategies is a simple epistemic virtue: it is relatively easy to get interesting or surprising results. The neuroscientist Valentino Braitenberg, in his elegant little book, *Vehicles: Experiments in Synthetic Psychology* (1984), propounded what he called the law of uphill analysis and downhill synthesis, which states, very simply, that it much easier to deduce the behavioral competence of a system whose internal machinery you have synthesized than to deduce the internal machinery of a black box whose behavioral competence you have observed. But behind this simple epistemological point resides a more fundamental one, first noted, I think, by Langton.

When human engineers design something (forward engineering), they must guard against a notorious problem: unforeseen side effects. When two or more systems, well-designed in isolation, are put into a supersystem, this often produces interactions that were not only not part of the intended design, but positively harmful; the activity of one system inadvertently clobbers the activity of the other. By their very nature unforseeable by those whose gaze is perforce myopically restricted to the subsystem being designed, the only practical way to guard against unforeseen side effects is to design the subsystems to have relatively impenetrable boundaries that coincide with the epistemic boundaries of their creators. In short, you attempt to insulate the subsystems from each other, and insist on an overall design in which each subsystem has a single, well-defined function within the whole. The set of systems having this fundamental abstract architecture is vast and interesting, of course, but — and here is AL's most persuasive theme — it does not include very many of the systems designed by natural selection! The process of evolution is notoriously lacking in all foresight; having no foresight, unforeseen or unforeseeable side effects are nothing to it; it proceeds, unlike human engineers, via

the profligate process of creating vast numbers of relatively uninsulated designs, most of which, of course, are hopelessly flawed because of self-defeating side effects, but a few of which, by dumb luck, are spared that ignominious fate. Moreover, this apparently inefficient design philosophy carries a tremendous bonus that is relatively unavailable to the more efficient, top-down process of human engineers: thanks to its having no bias against unexamined side effects, it can take advantage of the very rare cases where beneficial serendipitous side effects emerge. Sometimes, that is, designs emerge in which systems interact to produce more than was aimed at. In particular (but not exclusively) one gets elements in such systems that have multiple functions.

Elements with multiple functions are not unknown to human engineering, of course, but their relative rarity is signaled by the delight we are apt to feel when we encounter a new one. One of my favorites is to be found in the Diconix portable printer: This optimally tiny printer runs on largish rechargeable batteries, which have to be stored somewhere: inside the platen or roller! On reflection, one can see that such instances of multiple function are epistemically accessible to engineers under various salubrious circumstances, but one can also see that by and large such solutions to design problems must be exceptions against a background of strict isolation of functional elements. In biology, one encounters quite crisp anatomical isolation of functions (the kidney is entirely distinct from the heart, nerves and blood vessels are separate conduits strung through the body), and without this readily discernible isolation, reverse engineering in biology would no doubt be humanly impossible, but one also sees superimposition of functions that apparently goes "all the way down". It is very, very hard to think about entities in which the elements have multiple overlapping roles in superimposed subsystems, and moreover, in which some of the most salient effects observable in the interaction of these elements may not be functions at all, but merely byproducts of the multiple functions being served.

If we think that biological systems — and cognitive systems in particular — are very likely to be composed of such multiple function, multiple effect, elements, we must admit the likelihood that top-down reverse engineering will simply fail to encounter the right designs in its search of design space. Artificial Life, then, promises to improve the epistemic position of researchers by opening up different regions of design space — and these regions include the regions in which successful AI is itself apt to be found!

I will mention one likely instance. A standard feature of models of cognitive systems or thinkers or planners is the separation between a central

"workspace" or "working memory" and a long term memory. Materials are brought to the workspace to be considered, transformed, compared, incorporated into larger elements, etc. This creates what Newell has called the problem of "distal access". How does the central system reach out into the memory and find the right elements at the right time? This is reminiscent of Plato's lovely image of the aviary of knowledge, in which each fact is a bird, and the problem is to get the right bird to come when you call! So powerful is this image that most modelers are unaware of the prospect that there might be alternative images to consider and rule out. But nothing we know in functional neuroanatomy suggests anything like this division into separate workspace and memory. On the contrary, the sort of crude evidence we now have about activity in the cerebral cortex suggests that the very same tissues that are responsible for long term memory, thanks to relatively permanent adjustments of the connections, are also responsible, thanks to relatively fleeting relationships that are set up, for the transient representations that must be involved in perception and "thought". One possibility, of course, is that the two functions are just neatly superimposed in the same space like the batteries in the platen, but another possibility — at least, an epistemic possibility it would be nice to explore — is that this ubiquitous decomposition of function is itself a major mistake, and that the same effects can be achieved by machinery with entirely different joints. This is the sort of issue that can best be explored opportunistically — the same way Mother Nature explores — by bottom-up reverse engineering. To traditional top-down reverse engineering, this question is almost impervious to entry.

There are other issues in cognitive science that appear in a new guise when one considers the difference between top-down and bottom-up approaches to design, but a consideration of them is beyond the scope of this paper.

REFERENCES

1. BRAITENBERG, V., 1984, *Vehicles: Experiments in Synthetic Psychology*, Cambridge, MA: MIT Press/A Bradford Book.
2. DENNETT, D. C., 1987, *"The Logical Geography of Computational Approaches: A View from the East Pole,"*, in M. Harnish and M. Brand, eds., Problems in the Representation of Knowledge, Tucson, AZ: University of Arizona Press.
3. FODOR, J., 1983, *The Modularity of Mind*, Cambridge, MA: MIT Press/A Bradford Book.
4. LANGTON, C., 1989, *Artificial Life*, Redwood City, CA: Addison-Wesley.
5. MARR, D., 1982 *Vision*, San Francisco: Freeman.

6. NEWELL, A., YOST, G., LAIRD, J. E., ROSENBLOOM, P. S., and ALTMANN, 1990 *"Formulating the Problem Space Computational Model,"*, presented at the 25th Anniversary Symposium, School of Computer Science, Carnegie Mellon Univ. 24–26 September ,1990, forthcoming in R. Rashid, ed., ACM Pressbook, Reading, PA: Addison-Wesley.
7. RAMACHANDRAN, V. S., 1985 *Guest Editorial in Perception*, 14, pp.97–103.

Logic, Methodology and Philosophy of Science IX
D. Prawitz, B. Skyrms and D. Westerståhl (Editors)

LOGIC AND THE FLOW OF INFORMATION

JOHAN VAN BENTHEM

Institute for Logic, Language and Computation, University of Amsterdam

1. From propositions to procedures

At the core of standard logic is the notion of a declarative sentence, whose truth conditions in varying situations are the prime target of investigation. Of course, actual linguistic communication involves transient discourse and cognitive change, but this dynamics remains an 'extrinsic' feature of the use that is made of logical propositions. But gradually, a reversal of priorities is taking place in the literature, and many authors have focused instead on the potential for information change inherent in propositions. There are different sources for this trend (which can be traced as an undercurrent far back into this century). In particular, in linguistics, dynamic information flow occurs at various levels. In categorial parsing, categories serve as procedures acting on each other consecutively to produce sentence meanings (Moortgat 1988, van Benthem 1991), at the sentence level, processing of anaphoric dependencies involves shifting variable assignments, quite like the workings of imperative computer programs (Barwise 1987, Groenendijk & Stokhof 1991), and finally, discourse has an obvious dynamic global structure with a sequential game-like character (Hintikka 1973, Hintikka & Kulas 1983). Taken together, these observations suggest that natural languages are more like programming languages, serving various cognitive purposes, than like standard declarative formal languages. This view reflects a more general move in contemporary philosophy, away from static 'knowledge' to dynamic 'cognition', putting cognitive procedures like updating, retraction or revision of information at centre stage (Gärdenfors 1988, Harman 1985) rather than static representational structures. (Of course, such an interest in cognitive change still presupposes some account of standard cognitive content.) In other words, one is moving from 'extrinsic' dynamics to 'intrinsic' dynamics at the core of a logic of information flow.

The purpose of this paper is to put forward a general perspective on these matters, inspired by dynamic logic in computer science (Harel 1984), and

then to identify some salient general logical issues emerging behind many specific systems for linguistic and general cognitive purposes published so far. Notably, our 'procedural turn' leads to a reappraisal of traditional notions like 'logical constant' and 'valid inference', while also raising new issues of 'logical architecture'. We shall consider both a very general procedural logic and various possible specializations, showing that traditional methods of analysis still apply, when given an appropriate new twist. Our presentation follows the main lines of van Benthem 1991, Part VI (to which we refer for many details in what follows), while adding some further refinements and results obtained in the meantime. We shall not propose any specific system for performing the new cognitive tasks, but rather concentrate on foundational issues concerning their design.

Dynamics within classical logics

An immediate entry to procedural thinking takes its departure from any basic text book in standard logic. Consider Tarski's well-known definition for truth of a formula ϕ in a model $M = (D, I)$ under some variable assignment a. Its atomic clause involves a static test whether some fact obtains, but intuitively, the clause for an existential quantifier $\exists x$ involves shifting an assignment value for x until some verifying object has been found. But then, we may also make the latter process explicit, by assigning to each formula a binary relation consisting of those transitions between assignments a which result in its succesful verification. Moreover, eventually, other components of the truth definition admit of dynamization too. For instance, shifting interpretation functions I are involved in ambiguous discourse (van Deemter 1991) or questioning, and eventually, shifting model structures M make sense too in the dynamics of domain change across sentences (Westerståhl 1984). One immediate question arising on this point of view is how to interpret the standard logical constants. Some stipulations seem clear: for instance, most people would agree on letting conjunction stand for sequential 'composition' of transition relations, while disjunctions would amount to some kind of 'choice'. But we shall analyze the options in a more principled way later on.

Another point of departure from standard logic lies in what are the best-known classical information-oriented model structures, namely the possible worlds models for intuitionistic logic proposed by Kripke. Here, worlds stand for information states, and the intuitive picture is that of a cognitive agent traversing such states in the quest for knowledge. Again, intuitively, intuitionistic formulas refer to transitions in this information pattern. E.g., to see that $\neg\phi$ holds, one has to inspect all possible extensions of the current state for absence of ϕ. As before, this dynamics may also be made an explicit part

of the logic, by creating a system of cognitive transitions, such as 'updates' taking us to some minimal extension where a certain proposition has become true. Standard intuitionistic logic is then a forward-looking system, of Dutch mathematicians who never forget and never err, whereas ordinary mortals will display a zigzagging traveling pattern of cognitive advances and retreats, including 'downdates' and 'revisions'. Providing an explicit dynamic system here may even be viewed as taking the original 'constructivist' motivation to its logical conclusion.

Dynamics of inference

Finally, let us take a look at what this dynamic view of logic would mean for the archetypal inferential setting:

$$\frac{P_1 \ldots P_n}{C} \qquad \text{conclusion C follows from premises } P_1 \ldots P_n$$

What is the sense of this when all propositions involved are procedures changing information states? One natural explanation would be as follows. The premises of an argument invite us to transform our initial information state, and then the resulting transition has to be checked to see whether this 'warrants' the conclusion (in some suitable sense). Later on, we shall give a number of ways in which this may be made precise. For the moment, one consequence of this view needs to be pointed out. If the sequence of premises is a complex instruction for achieving some cognitive effect, then its presentation will be crucial. The sequential order of premises matters, the multiplicity of their occurrence matters, and each premise move has to be relevant. And this will bring us into open conflict with even the most basic 'structural rules' of standard logic (allowing us to disregard such aspects in classical reasoning). Think of meeting a date, where one has all the right moves available: flowers, tickets, sweet talking, kisses, and imagine the various ways in which successful seduction might fail by Permutation of actions, Contraction of identical actions, or Monotonic Insertion of arbitrary additional actions. Of course, in certain settings, deviations from classical reasoning will be slight, for instance, when all premise actions correspond to tests, or steady updates. But in general cognition, our information may be more complex, with the information prior to inference also containing retractions ("no, forget about A after all") or qualifications ("unless B, that is"). And then, a more delicate dynamic logic becomes imperative.

2. General dynamics: relational algebra and arrow logic

Relational algebra as procedural logic

Underlying many specific systems of dynamic logic is the usual mathematical concept of a 'state space'

$$(S, \{R_p \mid p \in P\}).$$

States may range here from cognitive constructs such as assignments or sets of possible worlds to real physical situations. Over these, there will be an 'atomic repertoire' of basic actions that can be performed, such as shifting the value in some register, adding or removing a world, kicking some round object. What atomic actions are appropriate depends on the particular choice of states, of course. Moreover, in particular settings, certain broad constraints on possible actions may be imposed from the start. For instance, updates are often assumed to satisfy a principle of 'idempotence': repeating them is unnecessary, that is $\forall xy : Rxy \rightarrow Ryy$. (This makes updating different from physical activities like kicking, or explaining something to one's students.)

On top of the atomic repertoire, which is taken for granted, there is a 'procedural repertoire' of various operations for creating compound actions, which we use for designing our programs or plans. Examples of such procedural operations are sequential composition, choice, iteration, but the literature also knows more exotic proposals. One example is the negation test of Groenendijk & Stokhof 1991, which reads as follows:

$$\neg R = \{(x, x) \mid \text{for no } y, (x, y) \in R\}.$$

Another case are the directed functions of categorial grammar, whose procedural force is as follows:

$$\begin{aligned} A\backslash B &= \{(x, y) \mid \text{for each } z \text{ with } (z, x) \in A,\ (z, y) \in B\} \\ B/A &= \{(x, y) \mid \text{for each } z \text{ with } (y, z) \in A,\ (x, z) \in B\} \end{aligned}$$

Of course, the basic choices made may also influence our freedom here. For instance, if all admissible actions are to be idempotent, then composition need not always be a safe combination, while choice or iteration do preserve idempotence.

What is happening here is a move from a standard Boolean Algebra of propositions to a Relational Algebra. The standard procedural repertoire in relational algebras is as follows:

Boolean operations	$-$ (complement)	$\cap$ (intersection)	$\cup$ (union)
Ordering operations	$\circ$ (composition)	$^{\vee}$ (converse)	

with a distinguished diagonal relation $\triangle$ standing for the sweet achievement of 'dolce far niente'. (At the other extreme, the Boolean structure also provides a universal relation T of 'random activity'.) These operations are definable in a standard predicate logic with variables over states:

$$\begin{array}{lcl} -R & : & \lambda xy\bullet\ \neg Rxy \\ R \cap S & : & \lambda xy\bullet\ Rxy \wedge Sxy \\ R \cup S & : & \lambda xy\bullet\ Rxy \vee Sxy \\ R \circ S & : & \lambda xy\bullet\ \exists z(Rxz \wedge Szy) \\ R^{\vee} & : & \lambda xy\bullet\ Ryx \end{array}$$

The expressive power of this formalism shows in that it can define many other proposed procedural operators. In particular,

$$\begin{array}{lcl} \neg R & : & \triangle \cap -(R \circ T) \\ A\backslash B & : & -(A^{\vee} \circ -B) \\ B/A & : & -(-B \circ A^{\vee}) \end{array}$$

The literature on Relational Algebra contains many results concerning axiomatization of valid identities between such relational expressions, as well as expressive power of various choices of operators (see Németi 1991). Some of these become relevant to our general procedural logic, as will be shown below.

Modal arrow logic and categorial grammar

In the long run, existing Relational Algebra would not be our favourite candidate for analyzing dynamic logic. Transition relations record rather little about the internal structure of processes, and some more delicate form of 'process algebra' (Milner 1980) will probably be needed sooner or later. Moreover, there are a number of mathematical complications in the subject as it exists, having to do with an insistence on set-theoretic relations consisting of ordered pairs. But intuitively, dynamic relations rather seem to consist of 'transitions' or 'arrows' as objects in their own right. Therefore, our preference would be to use a more abstract Arrow Logic (van Benthem 1989, Venema 1991). This may be viewed as a modal logic over 'arrow frames'

$$(W, C, F, I)$$

with a set W of 'arrows', a ternary relation C of 'composition', a binary relation F of 'conversion' and a unary predicate I for 'identical arrows'. The basic truth definition then explains the notion $M, x \models \phi$ (formula ϕ holds for the arrow x), so that formulas will describe sets of arrows, i.e.,

'transition relations' in the new sense. For instance, some key clauses will be as follows:

$$\begin{array}{lll} M,x \models A \cap B & \text{iff} & M,x \models A \text{ and } M,x \models B \\ M,x \models A \circ B & \text{iff} & \text{there exist } y,z \text{ such that } Cx,yz \text{ and} \\ & & M,y \models A \text{ and } M,z \models B \\ M,x \models A^{\vee} & \text{iff} & \text{there exists } y \text{ such that } Fxy \text{ and } M,y \models A \\ M,x \models \Delta & \text{iff} & Ix. \end{array}$$

Arrow Logic is a minimal theory of composition of actions, which may be studied completely by well-known techniques from Modal Logic (cf. also Roorda 1991, Vakarelov 1991). Standard principles of Relational Algebra then express certain constraints on arrow patterns, which can be determined in the usual style through 'frame correspondences' (van Benthem 1985). For instance, an algebraic law like $(A \cup B)^{\vee} = (A^{\vee} \cup B^{\vee})$ is a universally valid principle of 'modal distribution' on arrow frames, but $(A \cap B)^{\vee} = (A^{\vee} \cap B^{\vee})$ expresses the genuine constraint that the conversion relation F be a partial function, whose idempotence would be expressed by the modal axiom $A^{\vee\vee} = A$. For technical convenience (and no more), we shall assume henceforth that there is an idempotent (and hence injective) conversion function f available in arrow frames.

It may be of interest now to see what dynamic content is expressed by some basic categorial laws of natural language. Here is a sample, demonstrating the use of modal correspondence techniques.

PROPOSITION.

$$\begin{array}{lll} A \bullet (A \backslash B) \Longrightarrow B & \textit{expresses that} & \forall xyz : Cx,yz \longrightarrow Cz,f(y)x \\ (B/A) \bullet A \Longrightarrow B & \textit{expresses that} & \forall xyz : Cx,yz \longrightarrow Cy,xf(z). \end{array}$$

Together, these two principles express the basic 'rotations' that can be made in composition (e.g., they imply the familiar law $\forall xyz : Cx,yz \longrightarrow Cf(x),f(z)f(y)$).

Proof. In its arrow transcription, the first categorial principle has the modal form $(A \circ \neg(A^{\vee} \circ \neg B)) \rightarrow B$. Consider any arrow frame satisfying the stated constraint on its composition relation C. Let the antecedent $(A \circ \neg(A^{\vee} \circ \neg B))$ be true at x (under some valuation), and suppose that B fails at x. By the truth definition for $\circ$, there exist arrows y,z with Cx,yz, A true at y and $\neg(A^{\vee} \circ \neg B)$ true at z. But Cx,yz implies $Cz,f(y)x$, and we have $A^{\vee}$ true at $f(y)$ (by the truth definition for $^{\vee}$ and idempotence of f), $\neg B$ true at x. Therefore $A^{\vee} \circ \neg B$ is true at z: which is the required contradiction. Conversely, suppose that our categorial law holds in a frame. Consider any situation Cx,yz. Define the following valuation V on the

relevant proposition letters: $V(A) = \{y\}$, $V(B) = W - \{x\}$. Evidently, B fails at x, and hence so does the modal antecedent $A \circ \neg(A^{\vee} \circ \neg B)$. That is, either A must fail at y (which is impossible by the definition of V) or $\neg(A^{\vee} \circ \neg B)$ fails at z. Then there must be u, v with Cz, uv, $A^{\vee}$ true at u, $\neg B$ true at v. As A is only true at y, injectivity of f implies that $A^{\vee}$ can only be true at $f(y)$, whence $u = f(y)$. As B is only false at x, also $v = x$. But then we have $Cz, f(y)x$, as desired.

The second categorial equivalence may be proved in the same manner. But there is a more general observation to be made. Both categorial principles shown above exhibit a special form: they are so-called 'Sahlqvist formulas' in this binary modal logic. This may be seen by rewriting, e.g., the first to $(A \circ \neg(A^{\vee} \circ \neg B)) \wedge \neg B \rightarrow \bot$, or equivalently to $(A \circ \neg(A^{\vee} \circ B)) \wedge B \rightarrow \bot$, where the antecedent has its positive occurrences of A, B only in 'existential surface positions'. Thus the 'substitution algorithm' of van Benthem 1985 applies, which produces first-order corresponding conditions automatically (cf. also Venema 1991 for this technique). For instance, in this particular case, the formula generates a prefix $\forall xyz : Cx, yz \longrightarrow$, and a substitution '$A = \{y\}, B = \{x\}$' which produces exactly the above frame condition. □

Which frame conditions on C and f are expressed by further categorial laws, such as Geach Composition or Montague Raising? A precise procedural counterpart for the basic 'Lambek Calculus' of categories has been determined in van Benthem 1992A (the above notations suffice). Thus, the above dynamic reading of the catgorial operations translates categorial logics into natural corresponding arrow logics. The precise effect of this translation across the total landscape of Categorial Grammar remains to be investigated.

One could also reverse the perspective here, trying to embed Arrow Logic into a Lambek Calculus with operators $\backslash$, $/$, $\circ$, suitably enriched with Boolean operators $\neg$, $\wedge$, $\vee$ and an 'identity constant' *id*. Its deductive principles are the obvious union of categorial and Boolean laws, together with some suitable axioms concerning *id*. For instance, one possible rendering of relational conversion is as follows:

$$R^{\vee} = \neg(R \backslash \neg id).$$

In this way, procedural principles also acquire categorial content. Here is an illustration:

- The procedural principle $(R \cup S)^{\vee} = (R^{\vee} \cup S^{\vee})$ translates into the equivalence

$$\neg((R \vee S) \backslash \neg id) \longleftrightarrow \neg(R \backslash \neg id) \vee \neg(S \backslash \neg id).$$

The latter exemplifies a derivable law of the Boolean Lambek Calculus, viz.

$$(A \vee B)\backslash C \longleftrightarrow (A\backslash C) \wedge (B\backslash C).$$

- The procedural principle $R^{\vee\vee} = R$ translates into a less straightforward categorial law concerning *id*. For instance, one half would state essentially that $\neg((\text{“}R^{\vee}\text{”} \circ \neg R) \wedge id)$.

Thus, there exists an evident duality between categorial logics and procedural logics, whose further exploration must be foregone here.

Conclusion. Relational Algebra is a useful paradigm for bringing out general options of design for procedural systems of logic. Nevertheless, some more abstract framework like Arrow Logic will be desirable eventually. Either way, procedural logic may be studied using standard semantic tools from Modal Logic.

3. Logical constants as operators of control

Logical constants in standard logic are the key operators forming new propositions out of old ones. In dynamic logic, logical constants will be the key operators of control, combining procedures. Now, much of the recent literature still has a conservative bias, in that the only issue raised is 'what the standard logical constants mean' in a dynamic setting. But in fact, the latter allows for finer distinctions than the standard one, so that there may not be any clear sense to this question. Thus, it has to be analyzed on its own merits. For instance, standard 'conjunction' really collapses various notions: sequential composition, but also various forms of parallel composition. Likewise, standard 'negation' may be either some test as above, or merely an invitation to make any move refraining from some forbidden action ("anything, as long as you leave your father alone"). And also, there will be natural logical operators in the dynamic setting which lack classical counterparts altogether, such as conversion or iteration of procedures.

Logicality as permutation invariance

Nevertheless, there is a general perspective relating the two notions. Intuitively, 'logical' operators do not care about specific individual objects involved in their arguments. This is also true for procedural operators. What makes, say, a complement $-R$ a logical negation is that it works uniformly on all 'arrow patterns' R, in contrast to a negative social operator like 'Dutch' whose action depends on the content of its relational arguments ("Dutch

dining" means making one's guests pay for themselves, "Dutch climbing" is running to the top ahead of one's companions just at the finish). The common mathematical generalization involves *invariance under permutations* π of the underlying set of relevant individuals (here, states):

- Declarative propositions denote unary properties / sets of states, and hence propositional operators satisfy

$$\pi[O(X, Y, \ldots)] = O(\pi[X], \pi[Y], \ldots)$$

- Dynamic procedures denote binary relations / sets of ordered pairs of states, and hence procedural operators satisfy the same schema:

$$\pi[O(R, S, \ldots)] = O(\pi[R], \pi[S], \ldots).$$

A procedural hierarchy

This mathematical condition still leaves a host of possible relational operators. To get a finer view of the options, a more 'linguistic' perspective may be taken, scrutinizing the form of definition for relational operators. For instance, the earlier examples had definitions in a first-order language having variables over states and binary relation letters for procedures. Now, one reasonable measure of complexity is the number of variables essentially employed in such a defining schema, which tells us what is the largest configuration of states involved in determining the action of the operator. For instance, intersection of relations employed only two variables, whereas composition involved three. And the resulting 'finite variable levels' provide an obvious Procedural Hierarchy of complexity against which we can measure proposed procedural operations. (Of course, some infinitary version of the first-order language will be needed to include operators like iteration and its ilk.) Here are some facts about this hierarchy, provable using model-theoretic Ehrenfeucht-Fraissé games:

PROPOSITION.

- *The usual similarity type of Relational Algebra is functionally complete for all relational operators with a three-variable defining schema.*
- *Each n-variable level has a finite functionally complete set of operators.*
- *There is no finite functionally complete set of algebraic operators for the whole procedural hierarchy at once.*

Even finite-variable layers still contain a host of less plausible operators, through contrived definitions. But then, further constraints may be imposed, say of some reasonable computational character. One example is the well-known condition of

Continuity $\qquad O(\ldots, \bigcup_{i\in I} R_i, \ldots) = \bigcup_{i\in I} O(\ldots, R_i, \ldots).$

This forces the operation to determine its values 'locally', by inspecting single transitions (using the fact that $R = \bigcup\{\{(x,y)\} \mid Rxy\}$):

PROPOSITION.

- *Each continuous operation can be written in an existential form (displayed here for the two-argument case $O(R,S)$ only) $\lambda xy\bullet\ \exists zu(Rzu \wedge \exists vw(Svw \wedge$ 'Boolean combination of identities in $\{x,y,z,u,v,w\}$'$))$.*
- *For each fixed arity, there are only finitely many continuous permutation-invariant relational operators.*

Examples of continuous operations are Boolean intersection and union, as well as relational composition and converse. A non-example is Boolean complement.

Continuity in this strong form rules out too much, although it does describe a special 'natural kind' of logical operator. (Belnap 1977 proposes a weaker notion of 'Scott continuity' admitting more candidates.) Therefore, other 'computational' constraints on logicality of procedural operators become of interest too. First, logical constants should not generate 'unfeasible' transitions:

Feasibility Transitions for defined relations must be reachable through some finite sequence of basic actions.

Like Continuity, this rules out complement, while accepting conjunctions, disjunctions or compositions of procedures.

Next, Feasibility can be strengthened by a constraint which illustrates a broadly applicable line of thinking. Consider the important notion of 'simulation' of one process via another, which is crucial to computation. One well-known candidate for this purpose is 'bisimulation' in the usual sense of having a relation C between states in two transition models $(S, \{R_p \mid p \in P\})$, $(S', \{R'_p \mid p \in P\})$ satisfying the following back-and-forth clauses:

- if xCx', xR_py, then there exists some y' with yCy', $x'R'_py'$
- if xCx', $x'R'_py'$, then there exists some y with yCy', xR_py

Logical constants should not 'disturb' such connections:

Simulability Any simulation for basic actions must automatically be one for complex actions defined by logical constants.

Logicality and simulation

With bisimulation in the ordinary sense as a measure of process equivalence, the effect of Simulability will be to rule out essentially all but the 'regular program operations' $\cup$ (Boolean union), $\circ$ (composition), * (infinitary Kleene iteration), together with the following functions of 'domain' $\Diamond$ and 'counter-domain' $\neg\Diamond$:

$$\begin{aligned} \Diamond(R) &= \lambda xy\bullet\ x = y \wedge \exists z Ryz \\ \neg\Diamond(R) &= \lambda xy\bullet\ x = y \wedge \neg\exists z Ryz. \end{aligned}$$

(The latter is the earlier 'test negation'. Note that in fact, $\Diamond R$ is defineable as $\neg\neg R$.) This amounts to the repertoire of Propositional Dynamic Logic (cf. Section 5 below), couched in purely relational terms. A straightforward induction shows that all 'regular modal procedures' defined in this way have the required property vis-à-vis bisimulations. Moreover, here is one kind of converse result (disregarding infinitary matters), adapting an observation from the modal folklore:

PROPOSITION. *Two states x, y in two finite transition models M_1, M_2 (respectively) can be connected by some bisimulation between M_1, M_2 iff they belong to the domains of the same regular modal procedures.*

Proof. The new direction is from right to left. Define a binary relation C between the state domains of M_1, M_2 by setting

$$u\ C\ v \quad \textit{iff} \quad \begin{array}{l} u, v \text{ belong to the domains of the} \\ \text{same regular modal procedures.} \end{array}$$

It suffices to show that C is a bisimulation. So, assume that uCv and consider any R-successor u' of u in M_1, where R belongs to the atomic repertoire. We have to find some R-successor v' of v matching u' in C. Suppose now that each of the finitely many possible R-successors v' of v in M_2 fails to do the job. That is, there is some regular modal procedure π with either $u' \in \text{domain}(\pi)$ and $v' \notin \text{domain}\ (\pi)$ (1), or vice versa (2). But then, consider the following procedure:

$$R \text{ composed with all } \Diamond\pi \text{ of case (1) and all } \neg\Diamond\pi \text{ of case (2).}$$

This is a regular modal procedure with u in its domain, and hence so is v. But this will require the existence of some R-successor of v in M_2 distinct from all v' above: a contradiction. (Unions of procedures have dropped out of the definition here, because of the symmetric form of the preservation condition. Compare also the next Subsection on the pure $\{\neg, \circ\}$ repertoire.) $\square$

This result expresses a kind of 'maximality property' for regular modal operations with respect to Simulability. More sophisticated characterizations

may be found as well, without the restriction to finite models, using the preservation theorems for bisimulation invariance found in general Modal Logic (van Benthem 1985; cf. also Section 6).

Analyzing special repertoires

Whatever the most general notion of procedural 'logicality' may be, there are at least natural subkinds, such as the earlier continuous operators. Conversely, with special sets of operators from the literature, one can try to determine their specific semantic characteristics. An example is the procedural repertoire $\{\neg, \circ\}$ of Groenendijk & Stokhof 1991. This seems more special than the regular modal operations, in that there is no union of procedures. Nevertheless, the difference is a more delicate one. Again, this is seen most clearly in a two-level propositional modal logic, having both propositions and procedures in its language:

- The operations $\neg, \circ$ are both regular, and they suffice to embed the propositional component into the procedural one viá the test mode ?:

$$\begin{aligned} ?(\phi \wedge \psi) &= ?(\phi) \circ ?(\psi) \\ ?(\neg\phi) &= \neg ?(\phi) \\ ?(< \pi > \phi) &= \neg\neg(\pi \circ ?(\phi)) \end{aligned}$$

Thus, at least at the level of propositions or their corresponding tests, this repertoire provides all Boolean operations, including union.

- Adding an explicit operation of union $\cup$ to the $\{\neg, \circ\}$ repertoire results only in addition of outermost unions of $\{\neg, \circ\}$ programs, because of the valid equivalences

$$\begin{aligned} (R \cup S) \circ T &= (R \circ T) \cup (S \circ T) \\ R \circ (S \cup T) &= (R \circ S) \cup (R \circ T) \\ \neg(R \cup S) &= \neg(R) \circ \neg(S). \end{aligned}$$

These two observations establish a virtual equivalence between a standard finitary propositional dynamic logic and a relational algebra based on the above two operations.

Next here is a case of a genuinely different repertoire. In order to exclude unions (i.e., procedural 'choice') more radically, the following semantic characteristic may be used. Consider a 'direct product' of two transition models,

whose domain is the Cartesian product of the state sets, with this stipulation for its atomic repertoire:

$$(x, x')R(y, y') \quad \text{iff} \quad xRy \text{ and } x'Ry'.$$

Arbitrary products may be defined in the same manner. Then, it is easy to show that

- All procedures formed from atomic ones using only the repertoire $\{\Diamond, \circ, \cap\}$ are *invariant for direct products* in the sense of the above equivalence; whereas the latter may fail for $\cup$ and $\neg$.

The reason is that, more generally, all first-order formulas constructed from atoms $\{Rxy, x = y\}$ using $\wedge, \exists$ are invariant for direct products. A further admissible construction here is the universal quantifier $\forall$: but this does not seem to make much sense procedurally, and may be ruled out, e.g., by insisting on the earlier Continuity.

Without proof we state a sample outcome of an earlier model-theoretic analysis here, using binary operations for convenience.

- The logical procedural repertoire satisfying Continuity plus Product Invariance consists of all operations definable by the following schema:

$\lambda xy \bullet \exists zu(Rzu \wedge \exists vw(Svw \wedge$ 'conjunction of identities in $\{x, y, z, u, v, w\}$')).

Finally, a semantic analysis of special atomic repertoires may be of interest too. For instance, in the above-mentioned paper, basic actions π are all 'propositional tests' or 'random assignments', satisfying the identity $\pi \circ \pi = \pi$. Moreover, these are both symmetric relations. (These particular properties are not preserved by the procedural repertoire $\{\neg, \circ\}$, but others are.) Algebras over special kinds of binary relation have also been studied in the recent mathematical literature (cf. Németi 1990).

Conclusion. Logical constants in procedural logic are its basic operators of control, whose structure is much richer than that found in standard logic. The art is now to bring out further intuitions of logicality so as to motivate natural finite bases. For the latter task, ordinary model-theoretic analysis is still a useful tool.

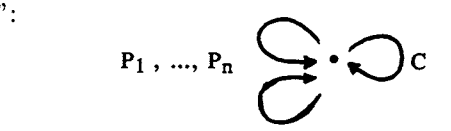
4. Varieties of inference

The standard explication of valid inference demands 'transmission of truth': "in every situation where all premises are true, so is the conclusion". And with classical propositions, this is a most reasonable basic option (although

natural modifications have been proposed since for heuristic purposes in Artificial Intelligence, witness Makinson 1991). If one wants to approximate the standard style in a dynamic setting, then the following seems appropriate. Each procedure may have its 'fixed points', being those states at which it loops (the state already 'satisfies' the goal of the procedure, so to speak). Thus, we can formulate a

classical style: "in all models, each state which is a fixed point for all premises is also a fixed point for the conclusion":

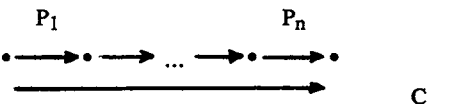
$\mathrm{fix}(P_1) \cap \ldots \cap \mathrm{fix}(P_1) \subseteq C$



But there is also a genuine

dynamic style: "in all models, each transition for the sequential composition of the premises must be admissible for the conclusion":

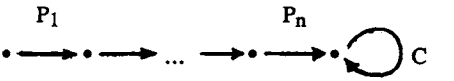
$P_1 \circ \cdots \circ P_n \subseteq C$



And some people prefer a compromise between the two styles:

mixed style: "first process all premises consecutively, then test if the conclusion is satisfied by the resulting state":

$\mathrm{range}(P_1 \circ \cdots \circ P_n) \subseteq \mathrm{fix}(C)$



Thus, there appears to be a genuine variety of dynamic styles of inference, reflecting different intuitions and possibly different applications.

Capturing styles via structural rules

One way of defining a basic 'style of inference' is through its general properties, expressed in the usual 'structural rules'. For instance, the above classical style has all the general properties of standard inference:

	/	$C \Longrightarrow C$	Reflexivity
$X \Longrightarrow D \quad Y, D, Z \Longrightarrow C$	/	$Y, X, Z \Longrightarrow C$	Cut Rule
$X, P_1, P_2, Y \Longrightarrow C$	/	$X, P_2, P_1, Y \Longrightarrow C$	Permutation
$X, P, Y, P, Z \Longrightarrow C$	/	$X, P, Y, Z \Longrightarrow C$	Contraction
$X, P, Y, P, Z \Longrightarrow C$	/	$X, Y, P, Z \Longrightarrow C$	Contraction
$X, Y \Longrightarrow C$	/	$X, P, Y \Longrightarrow C$	Monotonicity

By contrast, the dynamic style satisfies only Reflexivity and Cut. Indeed we have several representation results:

PROPOSITION.

- *{Monotonicity, Contraction, Reflexivity, Cut} completely determine the structural properties of classical inference*
- *{Reflexivity, Cut} completely determine dynamic inference.*

By way of illustration, we prove the second result in more detail. Let us look at models in which the 'propositions' involved in our sequents are interpreted as arbitrary binary relations, while a sequent is 'true' if the above inclusion holds for the composition of its premises in its conclusion. Then there is an obvious notion of semantic consequence $\Delta \models \sigma$ among sequents: truth of all sequents in Δ should imply that of σ.

PROPOSITION. *Reflexivity and Cut completely axiomatize valid consequence among dynamic sequents.*

Proof. Evidently, Reflexivity and Cut are valid on the above semantic interpretation. Conversely, suppose that some sequent σ cannot be derived from a set Δ using these two principles. Then let all finite sequences of basic syntactic items occurring in sequents be our underlying state set, and define the following map * taking basic items to binary relations:

$C^* = \{(X, XY) \mid Y \Longrightarrow C \text{ is derivable from } \Delta \text{ using Reflexivity and Cut}\}$

Then we have that, for sequents $X = X_1, \ldots, X_n$,

$$X \Longrightarrow C \text{ is derivable from } \Delta \quad \text{iff} \quad X_1^* \circ \cdots \circ X_n^* \subseteq C^*.$$

'If'. By Reflexivity, $X_1 \Longrightarrow X_1, \ldots, X_n \Longrightarrow X_n$ are all derivable from Δ. Therefore, the pairs $(<>, X_1), (X_1, X_1X_2), \ldots, (X_1 \ldots X_{n-1}, X_1 \ldots X_{n-1}X_n)$ belong to $X_1^*, \ldots, X_n^*$, respectively. So, $(<>, X)$ is in the composition of the consecutive premise relations, and hence it belongs to C^*. But then, by definition, $X \Longrightarrow C$ is derivable from Δ.
'Only if'. Consider any sequence of transitions according to the successive premises: (X, XY_1) ($Y_1 \Longrightarrow X_1$ derivable), (XY_1, XY_1Y_2) ($Y_2 \Longrightarrow X_2$ derivable), etcetera, up to $Y_n \Longrightarrow X_n$. Then, n successive applications of Cut to the derivable sequent $X \Longrightarrow C$ will derive $Y_1 \ldots Y_n \Longrightarrow C$, and hence $(X, XY_1 \ldots Y_n)$ is in C^*, by the definition of *.

The required counter-example now arises by observing that every sequent in Δ is derivable from it and hence true under the intended relational interpretation, whereas the original nonderivable sequent σ has become false. □

But new religions need not be defined by merely listing which old dogmas they accept or reject. Their point may be precisely that these old dogmas are

too crude as they stand. Inferential styles may in fact modify standard structural rules, reflecting a more delicate handling of premises. For instance, the mixed style has none of the above structural properties (counter-examples are easy to produce), but it does satisfy

Left Monotonicity	$X \Longrightarrow C$		/	$P, X \Longrightarrow C$
Left Cut	$X \Longrightarrow D$	$Y, X, D, Z \Longrightarrow C$	/	$Y, X, Z \Longrightarrow C$

These principles even characterize this style of inference:

PROPOSITION. *{Left Monotonicity, Left Cut} completely determine mixed inference.*

Proof. It suffices to give the recipe for the main representation involved. This time, the following map $^\sharp$ from syntactic items to binary relations will work:

$$C^\sharp = \{(X, X) \mid X \Longrightarrow C \text{ is derivable } \} \cup \{(X, XC) \mid \text{ all sequences } X\}.$$

What may be shown now is the following equivalence: $X \Longrightarrow C$ is derivable iff it is valid under this interpretation in the mixed style.
'If'. The pairs $(<>, X_1), \ldots, (X_1 \ldots X_{n-1}, X_1 \ldots X_{n-1}X_n)$ belong to the successive premise relations. Because of mixed validity then, (X, X) must be in $C^\sharp$, which can only mean that $X \Longrightarrow C$ is derivable.
'Only if'. Here is an example, with $n = 4$. Consider the following sequence of 'mixed' transitions for the premises:

$$(U, UX_1), (UX_1, UX_1X_2), (UX_1X_2, UX_1X_2) \text{ (with } UX_1X_2 \Longrightarrow X_3), \\ (UX_1X_2, UX_1X_2X_4).$$

Then we have $X_1 \ldots X_4 \Longrightarrow C$ (by assumption), $UX_1 \ldots X_4 \Longrightarrow C$ (Left Monotonicity), $UX_1X_2X_4 \Longrightarrow C$ (Left Cut, using $UX_1X_2 \Longrightarrow X_3$). That is, the final pair of objects $(UX_1X_2X_4, UX_1X_2X_4)$ is in $C^\sharp$. □

Switching styles

Having different inferential styles available also raises a new issue. How are these styles going to co-exist? In particular, it is natural to ask whether reasoning according to one style may be systematically reduced to reasoning via another. Here a connection emerges with the earlier topic of logical constants. Often, one inferential style can be 'simulated' inside another, through the addition of suitable logical operators. One example is the above classical style. Let us introduce a relational fixed point operator Φ sending

relations R to their diagonal $\lambda xy\bullet\ (Rxy \wedge y = x)$. Then we have the evident equivalence

$P_1, \ldots, P_n$ imply C classically if and only if
$\Phi(P_1), \ldots, \Phi(P_n)$ imply $\Phi(C)$ dynamically.

In the opposite direction, however, there is no similar formula-wise faithful embedding from the dynamic style into the classical style. The reason is that such an embedding would import the Monotonicity of the classical style into the dynamic one (adding translations of dynamic premises would not disturb the classical translation of a dynamic inference). Still there may be more global kinds of embedding that do the trick, translating whole sequents at once (van Benthem 1992B has a survey of various possibilities).

Another form of interplay between structural rules and logical constants arises as follows. One may wonder whether certain structural behaviour can be licensed, not for all propositions, but for special kinds only (cf. Girard 1987). For instance, in the dynamic style, let O be some operator that is to admit of arbitrary monotonic insertion:

$$X, Y \Longrightarrow C \quad / \quad X, O(P), Y \Longrightarrow C.$$

It is easy to show that this can be the case if and only if $O(P)$ is a 'test' contained in the diagonal relation. Here is a slightly less trivial result:

PROPOSITION. *An operator O allows unlimited contraction if and only if for all P, $O(P)$ is either empty or it contains the diagonal relation.*

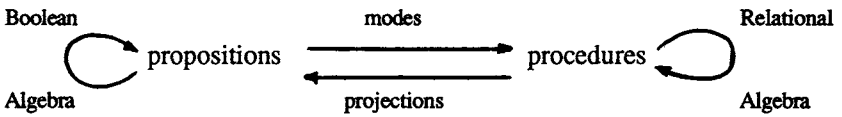
Proof. 'Only if'. If $O(P)$ is empty, then compositions including it are empty, and hence the conclusion of Contraction holds vacuously. If $O(P)$ includes the identity relation, then any relation Y dynamically implies both $Y \circ O(P)$ and $O(P) \circ Y$, whence Contraction holds too.
'If'. Suppose that $O(P)$ allows unlimited contraction under the dynamic interpretation of sequents. If $O(P)$ is not empty, then there exist x, y with $xO(P)y$. Consider any state z. Let R be $\{(y, z)\}$. The sequent $O(P), R, O(P) \Longrightarrow O(P) \circ R \circ O(P)$ is dynamically valid, and hence by Contraction, so is $O(P), R \Longrightarrow O(P) \circ R \circ O(P)$. Hence (x, z) must be in $O(P) \circ R \circ O(P)$, which can only be the case if $zO(P)z$. □

Conclusion. There are different natural styles of dynamic inference, exemplifying various clusters of structural rules, that can be determined via representation theorems. What we need now is some 'abstract proof theory' telling us what clusters are especially natural or useful. Moreover, reductions between inferential styles may be investigated, as a means of understanding the systematic connections between the various options that exist for reasoning.

5. Logical architecture

Combining statics and dynamics

Standard propositions and dynamic propositions both have reasonable motivations. Therefore, there seems little to choose between. Indeed, it is better not to choose at all. Actual inference is a mixture of more dynamic sequential short-term processes and more standard long-term ones, not necessarily working on the same representations. And then, both kinds of system would actually have to be around. In that case, a two-level logical architecture arises:



In this picture, the connections between the two levels have become essential components in their own right. There will be 'modes' taking standard propositions P to procedures with that content, such as 'updating' to make P true, or 'testing' whether P holds already: $\lambda P \bullet \lambda xy \bullet\ Px \wedge y = x$. Running in the opposite direction, there will be 'projections' assigning to each procedure R a standard proposition recording some essential feature of its action, such as the earlier fixed point operator Φ seeing in which states R is already satisfied, or the taking of a set-theoretic range: $\lambda R \bullet \lambda x \bullet\ \exists y Ryx$, seeing where R might still lead.

Thus, we have acquired several new kinds of operators which may be analyzed from a logical point of view, just as much as the preceding ones. This is in fact a quite general phenomenon. If logical architecture becomes important, in systems having several logical calculi at the same time, then there is also an issue of what may be called 'logical management': what is the structure of the possible connections?

Type-theoretic analysis

Despite this diversity, there is also a clear mathematical uniformity. Most of the earlier techniques used in analyzing logical constants make sense for the new categories of operators too, when viewed in a suitable *type-theoretic* perspective. For instance, test or range functions are both 'permutation-invariant' in an obvious extended sense. Moreover, both of them are 'con-

tinuous' in that they commute with arbitrary unions of their arguments. All possibilities of this kind can be classified by previous reasoning:

FACT.

- *The only permutation-invariant continuous modes are those definable by a schema of the form*

$$\lambda P \bullet \lambda xy \bullet \exists z(Pz \wedge \text{'Boolean Condition on } \{=, x, y, z\}\text{'}).$$

'Test' is one example here: $\lambda P \bullet \lambda xy \bullet \exists z(Pz \wedge z = x = y)$. *All other possibilities are simple variations.*

- *The only permutation-invariant continuous projections are those definable by a schema of the form*

$$\lambda R \bullet \lambda x \bullet \exists yz(Ryz \wedge \text{'Boolean Condition on } \{=, x, y, z\}\text{'}).$$

'Fixed Points' is one example here: $\lambda R \bullet \lambda x \bullet \exists yz(Ryz \wedge x = y = z)$, *and so are 'Domain'* $(\lambda R \bullet \lambda x \bullet \exists yz(Ryz \wedge x = y))$ *and 'Range'. The other options are again simple variations.*

Another management question concerning projections is whether there exists some map from procedures to statements preserving all relevant logical structure. In particular, one might expect that composition of procedures will reduce to conjunction of the corresponding statements. But here, a negative result arises (van Benthem 1986):

FACT.

- *There is only one logical Boolean homomorphism from procedures to propositions, namely the diagonal fixed point map* Φ.

  The proof is based on a recipe for 'deflating' (logical) Boolean homomorphisms in a type $((a, t), (b, t))$ *to arbitrary (logical) maps in the type* (b, a), *and then counting the mathematical possibilities there.*

- Φ *does not transform* $\circ$ *into* $\cap$:

$$\Phi(\{(1,2)\} \circ \{(2,1)\}) = \{1\} \neq \emptyset = \Phi(\{(1,2)\}) \cap \Phi(\{(2,1)\}).$$

This issue is related to one raised earlier. The two-level system has inference going on both between propositions and between procedures. Say, propositions have standard inference and procedures the earlier dynamic inference.

Then, we want to see whether one mechanism can be related systematically to the other. One direction here is easy: standard inference may be simulated within the dynamic style using the test mode ?:

$$P_1, \ldots, P_n \models_{\text{class}} C \text{ iff } ?(P_1), \ldots, ?(P_n) \models_{\text{dyn}} ?(C).$$

In the opposite direction, say, the fix point operator 'fix' will not work, for reasons explained earlier: a similar reduction would make dynamic reasoning monotonic.

Static tracing of dynamic procedures

But we can also analyze the situation somewhat differently, using a well-known concept from computer science. Let us trace a procedure through propositions describing successive images of sets of states under its action. Define 'strongest postconditions' as follows:

$$SP(A, R) = R[A].$$

Likewise, there are 'weakest preconditions':

$$WP(R, A) = R^{-1}[A].$$

Then, we can reduce dynamic validity as follows:

PROPOSITION. *$R_1, \ldots, R_n \models_{\text{dyn}} S$ if and only if $SP(A, R_1 \circ \ldots \circ R_n) \models_{\text{class}} SP(A, S)$ for arbitrary sets A.*

Proof. The 'only if' direction follows from the definition of $\models_{\text{dyn}}$ and the monotonicity of SP in its right-hand argument. The 'if' direction follows by considering any pair (x, y) in $R_1 \circ \ldots \circ R_n$, and then applying the second condition to the set $A = \{x\}$. □

Now, it becomes of interest to have a good way of computing weakest preconditions and strongest postconditions. Here are some inductive clauses:

$$\begin{aligned}
SP(A, R \circ S) &= SP(SP(A, R), S) \\
WP(R \circ S, A) &= WP(R, WP(S, A)) \\
SP(A, R \cup S) &= SP(A, R) \vee SP(A, S) \\
WP(R \cup S, A) &= WP(R, A) \vee WP(S, A) \\
SP(A, R^{\vee}) &= WP(R, A) \\
WP(R^{\vee}, A) &= SP(A, R)
\end{aligned}$$

This calculation may be extended to cover the earlier regular modal operations of 'domain' $\Diamond$ and 'counter-domain' $\neg\Diamond$, with clauses such as

$$\begin{aligned} SP(A, \Diamond(R)) &= A \wedge WP(R, T) \\ SP(A, \neg\Diamond(R)) &= A \wedge \neg WP(R, T) \\ WP(\Diamond(R), A) &= A \wedge WP(R, T) \\ WP(\neg\Diamond(R), A) &= A \wedge \neg WP(R, T) \end{aligned}$$

There are no obvious inductive clauses, however, for intersection or complement of programs (let alone, homomorphic behaviour). For instance, we do not have

$$'SP(A, R \cap S) = SP(A, R) \wedge SP(A, S)'.$$

This situation may again be understood by the earlier type-theoretic analysis. SP and WP are transformations from the type of relations to that of functions from statements to statements: i.e., they live in the intensional type

$$((s, (s, t)), ((s, t), (s, t))),$$

having definitions, respectively,

$$\lambda R{\bullet}\lambda A{\bullet}\lambda x{\bullet}\ \exists y(Ay \wedge Ryx) \quad \text{and} \quad \lambda R{\bullet}\lambda A{\bullet}\lambda x{\bullet}\ \exists y(Ay \wedge Rxy).$$

In principle, the type $((s, t), (s, t))$ has more room than (s, t) to acommodate relations (indeed, SP is a bijection between relations and *continuous* maps from sets to sets). Again, we can analyze this larger class of transformations for its most interesting logical inhabitants, and then we find an explanation for the above poverty:

PROPOSITION.

- *The only logical homomorphisms in the type of SP are those defined by a schema of the form $\lambda R{\bullet}\lambda P{\bullet}\lambda x{\bullet}\ R(F(P, x))$, with F a logical map from sets and individuals to pairs of individuals.*

- *The latter are all 'products' of two logical maps from sets P and individuals x to individuals, of which there are essentially just two: 'right projection' to x and 'definite description' $\iota x{\bullet}\ P$ (in case P is a singleton).*

Proof. This may be shown again by the earlier recipe of 'homomorphic deflation', now from type $((s, (s, t)), ((s, t), (s, t))) = ((s{\bullet}s, t), ((s, t){\bullet}s, t))$ to $((s, t){\bullet}s, s{\bullet}s)$, followed by an analysis of invariant candidates. □

Conclusion. It is possible, and desirable, to have logical architectures combining both dynamic and standard inferential styles. This is also a key idea in dynamic logic as found in computer science. What we add here is 'management' as a separate concern: one wants to see to which extent independent reasoning inside the various levels can be related, and for that purpose, the connections between the two levels become independent objects of logical study in their own right. Logical uniformity in exploring this wider domain is guaranteed by taking a suitable type-theoretic perspective.

6. Further informational structure: dynamic modal logic

Cognitive procedures over information patterns

The notions and issues introduced so far are purely procedural, and have nothing to do with information per se. A modest basic step introducing more informational structure consists in endowing state spaces (now thought of as patterns of information states) with the inclusion relation also found in models for intuitionistic or relevant logics:

$$(S, \subseteq, \{R_p \mid p \in P\}).$$

Then, new notions emerge in all components of the earlier architecture, such as

- a new propositional operator

$$\Diamond_{\text{up}}(P) = \lambda x_{\bullet}\, \exists y(x \subseteq y \wedge Py) \qquad \text{'upward modality'}$$

- a new procedural operator

$$\text{forw}(R) = \lambda xy_{\bullet}\, (Rxy \wedge x \subseteq y) \qquad \text{'forward part'}$$

- a new projection

$$\text{fut}(R) = \lambda x_{\bullet}\, \exists y(x \subseteq y \wedge \exists z Ryz) \qquad \text{'future domain'}.$$

In particular, new *modes* may be defined creating dynamic procedures out of standard propositions. Some prominent examples are as follows:

$$\lambda P_{\bullet}\lambda xy_{\bullet}\, x \subseteq y \wedge Py \qquad \text{'loose updating'}$$
$$\lambda P_{\bullet}\lambda xy_{\bullet}\, x \subseteq y \wedge Py \wedge \neg\exists z(x \subseteq z \subset y \wedge Pz) \qquad \text{'strict updating'}$$

Moreover, going through the information pattern in the opposite direction, there are two obvious counterparts for processes of 'cognitive retreat':

$\lambda P \bullet \lambda x y \bullet\ y \subseteq x \wedge \neg Py$ 'loose downdating'
$\lambda P \bullet \lambda x y \bullet\ y \subseteq x \wedge \neg Py \wedge \neg \exists z(y \subset z \subseteq x \wedge \neg Pz)$ 'strict downdating'

Note that downdating is not a converse of updating. The proper duality is this:

$$\text{downdate}(P, \subseteq) = \text{update}(\neg P, \supseteq).$$

This formalism is rich enough to perform all cognitive tasks covered in the well-known system of Gärdenfors 1988. In particular, procedures of 'revision' may be described by combination of updates and downdates. Here is another example. Possibly counterfactual conditional statements $A \rightarrow B$ are often explained via the Ramsey Test: "Assume the antecedent. If this leads to inconsistency, make a minimal adjustment in the background theory so as to restore consistency. Then see if the result implies the consequent". In our models, this procedure becomes: "Downdate strictly with respect to $\neg\Diamond A$. Then update strictly with respect to A. Finally, check if B holds."

Exploring dynamic modal logic

One useful general calculus in this case will be some system of dynamic logic serving as a common generalization of standard modal logic and the earlier relational algebra. That is, the language will have standard propositional and modal operators

$$\{\neg, \wedge, \vee, \Diamond_{\text{up}}, \Diamond_{\text{down}}\}$$

as well as the usual relational repertoire

$$\{-, \cap, \cup, \circ, ^{\vee}, \Delta\}$$

plus suitable modes

$$\{\text{test, loose and strict updates as well as downdates}\}$$

and projections

$$\{\text{fixed point, domain, range}\}.$$

This also contains the usual weakest preconditions $<\pi> P$ via the definition $\text{dom}(\pi \circ ?(P))$.

There are various connections between these operators, witness valid identities like

$$\begin{array}{lcllcl} \subseteq & = & \text{upd}(T) & \Delta & = & ?(T) \\ \text{upd}(P) & = & \subseteq \circ ?(P) & \text{range}(\pi) & = & \text{dom}(\pi^{\vee}) \\ \Diamond_{\text{up}}(P) & = & \text{dom}(\text{upd}(P)) & \text{fix}(\pi) & = & \text{dom}(\Delta \cap \pi) \end{array}$$

But we are not after a least redundant version of this modal system $S4^2$, which may be viewed as a dynamic version of the standard modal logic $S4$. The universal validities of this logic form a general theory of cognitive processes. Notably, one can study various combinations of updating and downdating, noting that:

- update(P) $\circ$ update(P) = update(P)
- and the same holds for strict updates, and for downdates
- an update followed by a downdate need not be idempotent, witness a situation like $\bullet\neg P$ — $\bullet\neg P$ — $\bullet P$

The system also handles various earlier notions of inference. E.g., the dynamic variant $P_1, \ldots, P_n \models_{\text{dyn}} C$ amounts to validity of the implication $<P_1 \circ \ldots \circ P_n> q \longrightarrow <C> q$. And more complex relations between cognitive processes can be formulated too, such as $[\pi_1] < \pi_2 > \phi$ (process π_1 'enables' process π_2 to achieve result ϕ).

There are some obvious technical questions concerning this dynamic version of the standard modal logic $S4$. By general reasoning, its set of universal validities must be recursively enumerable (due to the embedding into first-order logic presented below). De Rijke 1992 presents a complete axiomatization using a slightly stronger formalism including a 'difference operator' D ("truth in at least one different state"). Then, characteristic deductive principles are definitory axioms for the main modes, such as

$$(q \wedge \neg Dq) \rightarrow (< \text{strict-upd}(P) > A \leftrightarrow \Diamond_{\text{up}}(A \wedge P \wedge \Box_{\text{down}}(\Diamond_{\text{down}} q \rightarrow \neg P))).$$

Moreover, there is the issue whether the system is *decidable*. Modal $S4$ has the latter property, full relational algebra over arbitrary relations does not: but what about this intermediate case, which handles only special updating, downdating and test relations?[1]

Without going into deductive detail about dynamic modal logic, one can analyze the characteristic properties of some modes and projections in direct semantic terms (assuming that the operators on propositions and procedures already have their standard interpretation):

PROPOSITION.

- *'Test' is the only permutation-invariant continuous operator satisfying the principles*

$$\begin{array}{lcll} ?(P) & \leq & \Delta & * \\ ?(\neg P) & = & \Delta \cap -?(P) & ** \end{array}$$

[1]Edith Spaan and Maarten de Rijke have settled this question in the negative, by reducing an undecidable tiling problem to $S4^2$–satisfiability.

Proof. By Continuity, it suffices to determine the behaviour of 'test' on singleton arguments $\{x\}$. And since test values are subrelations of the diagonal, by $*$, it suffices to specify the individual states in the image. By permutation invariance, there are only four basic options here: $\{x\}$ itself (1), $\{y \mid y \neq x\}$ (2), the unit set (3) and the empty set (4). Moreover, the choice will be made uniformly for all states x, again by permutation invariance. Now, outcome (4) would result in any test relation being empty: which contradicts $**$. Outcomes (3) and (2) may also be ruled out, by observing that they would allow for two distinct sets $\{x\}$, $\{y\}$ to have overlapping test values, whence some value $?(\neg P)$ would not be disjoint from $?(P)$: another contradiction with $**$. Thus, only the standard interpretation (1) remains. □

A similar kind of argument characterizes a key projection:

PROPOSITION.

- *'Domain' is the only permutation-invariant continuous operator satisfying the principles*

$$\begin{array}{lcll} dom(?(P)) & = & P & \sharp \\ dom(0) & = & 0 & \sharp\sharp \\ dom(R \circ S) & = & dom(R \circ ?(dom(S))) & \sharp\sharp\sharp \end{array}$$

As before, dynamic modal logic allows us to reduce dynamic behaviour of regular procedures to static propositions, via 'weakest preconditions' such as:

$$\begin{array}{lcl} WP(?(P), A) & = & P \wedge A \\ WP(\mathrm{upd}(P), A) & = & \Diamond_{\mathrm{up}}(P \wedge A) \\ WP(\mathrm{downd}(P), A) & = & \Diamond_{\mathrm{down}}(\neg P \wedge A) \end{array}$$

In order to describe the strict variants, a more complex modal language is needed over information patterns, employing two well-known 'temporal' operators UNTIL and SINCE which may be defined by suitable combinations of test, update and domain (de Rijke 1992):

$$\begin{array}{lcl} WP(\text{strict upd}(P), A) & = & \mathrm{UNTIL}(P \wedge A, \neg P) \\ WP(\text{strict downd}(P), A) & = & \mathrm{SINCE}(\neg P \wedge A, P). \end{array}$$

$S4^2$ is only at the bottom end of a ladder of dynamic modal logics. This observation brings us to an even broader formalism over information models.

Dynamic modal logics as fragments of first-order logic

As in standard modal logic, there is a straightforward *translation* from the new dynamic propositions and procedures to unary and binary formulas in

a standard first-order predicate logic over partial orders of states with unary predicates:

$(p)^*$	Px
$(\neg\phi)^*$	$\neg(\phi)^*$
$(\phi \wedge \psi)^*$	$(\phi)^* \wedge (\psi)^*$
$(\Diamond_{\text{up}}\phi)^*$	$\exists y(x \subseteq y \wedge [y/x](\phi)^*)$
$(\Diamond_{\text{down}}\phi)^*$	$\exists y(y \subseteq x \wedge [y/x](\phi)^*)$
$(?P)^\sharp$	$x = y \wedge Py$
$(\text{upd}(P))^\sharp$	$x \subseteq y \wedge Py$
$(\text{strict upd}(P))^\sharp$	$x \subseteq y \wedge Py \wedge \neg\exists z(x \subseteq z \subset y \wedge Pz)$
$(\text{downd}(P))^\sharp$	$y \subseteq x \wedge Py$
$(\text{strict downd}(P))^\sharp$	$y \subseteq x \wedge \neg Py \wedge \neg\exists z(y \subset z \subseteq x \wedge \neg Pz)$
$(\triangle)^\sharp$	$x = y$
$(-\pi)^\sharp$	$\neg(\pi)^\sharp$
$(\pi_1 \cap \pi_2)^\sharp$	$(\pi_1)^\sharp \wedge (\pi_2)^\sharp$
$(\pi_1 \circ \pi_2)^\sharp$	$\exists z([z/y](\pi_1)^\sharp \wedge [z/x](\pi_2)^\sharp)$
$(\pi^\vee)^\sharp$	$[y/x, x/y](\pi)^\sharp$
$(\text{fix}(\pi))^*$	$[x/y](\pi)^\sharp$
$(\text{dom}(\pi))^*$	$\exists y(\pi)^\sharp$

As the first-order theory of partial orders with monadic predicates (an elementary class of models) is recursively enumerable, so is our dynamic logic $S4^2$. And the same holds for any first-order reducible strengthening thereof.

This brings us to a general question of logical design. Modal logics, whether 'static' or 'dynamic', may be viewed as fragments of a full first-order logic over information patterns. And the question is what kinds of fragment are natural for present purposes. Now, several earlier observations may be brought to bear. First, the above translation may be seen to involve essentially only *three* variables over states in any formula. Thus, one view of the matter would be to have a full three-variable fragment, considering all unary and binary first-order formulas $\phi(x), \pi(x, y)$ constructed using only the three variables $\{x, y, z\}$. This establishes a certain 'harmony' between the minimal procedural repertoire found in Relational Algebra and the three-variable {Since, Until} language which has been so prominent in temporal logic. (Conversely, this harmony also amounts to a kind of 'functional completeness' for the dynamic part, which should be strong enough to achieve everything expressible in the static part.) Basically, what we are studying here is the behaviour of cognitive procedures whose action can be described using configurations of no more than three states at any one time. That is, we can specify goal states, while imposing conditions on intermediate states

encountered en route.

This three-variable fragment also has a purely semantic characterization (van Benthem 1991), in terms of the following notion. A 'k-partial isomorphism' between two first-order models is a family of partial isomorphisms of size at most k satisfying the usual Back and Forth properties for addition of new objects on both sides up to length k. Moreover, restrictions of partial isomorphisms in the family are to remain inside it. The relevant result is this

THEOREM. *A first-order formula having its free variables among* $\{x_1, \ldots, x_k\}$ *can be written using these variables only (free or bound) if and only if it is invariant in passing from one model and assignment to another model related to it by some k-partial isomorphism PI and using a PI-matching sequence of objects for the new assignment.*

Specialization to the case $k = 3$ then describes one very natural dynamic modal logic. But there is another, related perspective too. Upon closer inspection, translations of the above {Since, Until} language turn out to involve only part of the full three-variable first-order formalism. This point is even clearer with the basic modal language, which describes a special fragment of the two-variable first-order language over its models, having all quantifiers restricted to relational successors and predecessors, with only unary atoms. There is an independent semantic characterization of the latter fragment too (cf. van Benthem 1985), using an earlier semantic notion:

THEOREM. *A unary first-order formula* $\phi(x)$ *is equivalent to the translation of a basic modal formula if and only if it is invariant for standard bisimulation.*

In the earlier terms, one now restricts attention to 'process simulation' via partial isomorphisms where the next choice in the Back and Forth moves is restricted to successors or predecessors of the previous selection. Moreover, comparisons between matching states concern only atomic propositions.

Inspection of the proof for this preservation theorem and its predecessor shows that there is a recurring pattern here. The crucial step in all cases runs as follows:

> Finite sequences of objects up to some fixed length in two suitably saturated models (e.g., finite ones, as in an earlier argument) are 'connected' if they satisfy the same formulas in the restricted fragment under consideration. Then it is shown how this connection is in fact an appropriate relation of 'bisimulation' between the two models.

This is the precise spot where expressive power of the language and semantic strength of bisimulation meet, suggesting a general 'recipe' for generating preservation results. Here are some illustrations, whose purpose is mainly to give an impression of the general method.

At level $k = 2$, there are basically two options for the above connection. The first has arbitrary Back and Forth moves from single state matches to matched state pairs, and the appropriate formalism needs a strong projection operator to ensure this:

$\lambda x \bullet \exists y \pi(x, y)$ Domain

If one also insists that matched pairs generate two matched individuals, then two modes are needed:

$\lambda xy \bullet \phi(x)$ $\lambda xy \bullet \phi(y)$ Raising

Finally, in order to ensure that matched sequences are truly (partial) isomorphisms, the formalism needs all Boolean operations, as well as relational conversion and identification of arguments. (One might economize on this repertoire, though, by weakening the requirements on 'partial isomorphism'.)

The second main option is to have Back and Forth clauses demanding extension by new individual matchings only, looking at $\subseteq$-successors and $\subseteq$-predecessors. Then essentially, just two weakened versions of the domain projection are needed:

$\lambda x \bullet \exists y(x \subseteq y \wedge \phi(y))$ $\lambda x \bullet \exists y(y \subseteq x \wedge \phi(y))$ Modality

At level $k = 3$, the proper one for the above dynamic logic, similar options emerge. One notion congenial to the intended procedures might be called 'Path Simulation':

> There is a restriction-closed matching between individual states and pairs of states satisfying the following Back and Forth conditions:
>
> - for matched states x, y, selecting a $\subseteq$-successor or $\subseteq$-predecessor z in either model produces an admissible matching xz, yu or vice versa with some $\subseteq$-corresponding state u on the opposite side.
> - for matched pairs xy, zu, selecting a state v in between (along $\subseteq$) leads to a $\subseteq$-corresponding selection w on the opposite side, generating admissible matchings xv, zw and vy, wu (or vice versa).

PROPOSITION. *The complete first-order formalism for invariance under path simulation contains atoms* $\{Px, x \subseteq y\}$, *all Boolean operations, restricted modal existential quantifiers* $\exists y(x \subseteq y \wedge \pi(x, y))$, $\exists y(y \subseteq x \wedge \pi(x, y))$ *(that is,*

domains of the earlier 'forward' and 'backward' parts of π *), as well as a new binary modal quantifier* $\exists z(x \subseteq z \subseteq y \wedge \pi_1(x,z) \wedge \pi_2(z,y))$ *('betweenness').*

Again this outcome may be varied, with weaker versions of simulation capturing restricted 'unary' quantifiers such as $\exists y(x \subseteq y \wedge \phi(y))$ and $\exists z(x \subseteq z \subseteq y \wedge \phi(z))$.

Conclusion. Dynamic modal logics are a joint generalization of standard intuitionistic or other information-oriented logics and relational algebra, encompassing most recent 'cognitive logics' for information processing. There is no single preferred such system, but options for design may be laid out in terms of invariance for 'bisimulation', used as a flexible model-theoretic technique. These logics provide a simple natural 'completion' of constructivist thinking, which can still be studied by standard modal techniques.

7. Towards more realistic systems

All information modelings considered so far have been concerned with transitions in the internal cognitive space of one agent. A general logical architecture for cognition will have to be extended in at least the following ways.

First, more sensitive notions of 'process' are to be introduced to get at finer dynamic phenomena. Examples are the 'failure paths' of Segerberg 1991, the 'full trace models' of Vermeulen 1989 or the 'process algebra' of Milner 1980, Bergstra & Klop 1984. These approaches may be developed in the general logical style advocated here.

Another finer perspective concerns computational complexity. For instance, the above 'modes' are ways of testing or realizing standard propositions that may still be of vastly different complexities. What we want is some understanding of 'minimal cost' for modes with respect to different standard propositions, allowing us to compare them. One relevant viewpoint here is that of 'semantic automata' (van Benthem 1986), which may be viewed as procedural mechanisms checking truth conditions of standard constructions, in particular, various quantifiers. (First-order quantifiers are of finite-state complexity, while computing higher-order ones may involve push-down storage. Moreover, even among first-order quantifiers, e.g., "some" is cheaper than "one".) One could devise similar 'graph automata' operating on the above inclusion patterns of information states. Basic moves are steps along $\subseteq$ or atomic tests $?(p)$, and then, one would want to make comparisons as to complexity of search for various tasks.

Then, the physical world environment is to be brought into the picture if 'real correctness' of cognitive procedures is to be formulated. There are

various proposals to this effect in the recent literature, witness Kamp 1979, 1984 about the triangle 'language—representation—real world' or Barwise 1991 on 'information links' between various abstract and concrete informational systems. The least that should be done to this effect in our case is the introduction of a real world structure in addition to the pattern of information states, with suitable links between these. For instance, one might work with some distinguished 'actual world' in ranges of possible worlds, with some suitable relation of 'instantiation' linking possible worlds to information states. This will enable us to formulate real correctness of cognitive procedures, for instance, by letting the real world be among the possible instantiations of the states along some trajectory in a cognitive state space.

Also, real cognition usually involves the interplay of various agents. This brings in the interplay between different cognitive spaces (and a real world environment). On the modal strategy advocated here, one would need distributed environments as in Halpern & Moses 1989, again with an appropriate general perspective on logical architecture and management. Interestingly, earlier game-theoretical approaches to logic and cognition, like that of Hintikka 1973, had this multi-agent perspective all along.

Finally, cognitive activity is certainly not restricted to the standard business of 'interpretation' and 'inference'. It also involves planning, learning, guessing, querying or searching vis-á-vis patterns of cognitive states in a real world environment. These activities too, with their salient structural properties, will have to be brought eventually within the compass of a genuine logic of information flow.

References

BARWISE, J. (1987), *Noun Phrases, Generalized Quantifiers and Anaphora*, in P. Gärdenfors, ed., Generalized Quantifiers. Logical and Linguistic Approaches, Reidel, Dordrecht, 1-29.

BARWISE, J. (1991), *Information Links*, Department of Philosophy, Indiana University, Bloomington.

BELNAP, N. (1977), *A Useful Four-Valued Logic*, in M. Dunn & G. Epstein, eds., Modern Uses of Multiple-Valued Logics, Reidel, Dordrecht, 8-37.

BENTHEM, J. VAN (1985), *Modal Logic and Classical Logic*, Bibliopolis, Naples / The Humanities Press, Atlantic Heights.

BENTHEM, J. VAN (1986), *Essays in Logical Semantics*, Reidel, Dordrecht, (Studies in Linguistics and Philosophy 29).

BENTHEM, J. VAN (1989), *Modal Logic and Relational Algebra*, Institute for Language, Logic and Information, University of Amsterdam. (To appear in Proceedings Malc´ev Conference, Institute of Mathematics, USSR Academy of Sciences, Novosibirsk.)

BENTHEM, J. VAN (1991), *Language in Action. Categories, Lambdas and Dynamic Logic*, North-Holland, Amsterdam, (Studies in Logic 130).

BENTHEM, J. VAN (1992A), *Dynamic Arrow Logic*, Report LP-92-11, Institute for Logic, Language and Computation, University of Amsterdam (to appear in J. van

Eyck & A. Visser, eds., Logic and Information Flow, Kluwer, Dordrecht).

BENTHEM, J. VAN (1992B), *The Landscape of Deduction*, to appear in K. Dosen & P. Schröder-Heister, eds., Substructural Logics, Oxford University Press.

BERGSTRA, J. & J-W KLOP (1984), *Process Algebra for Synchronous Communication*, Information and Control 60, 109-137.

DEEMTER, K. VAN (1991), *On the Composition of Meaning*, Dissertation, Institute for Language, Logic and Information, University of Amsterdam.

EYCK, J. VAN & F-J DE VRIES (1991), *A Sound and Complete Calcuclus for Update Logic*, Centre for Mathematics and Computer Science, Amsterdam, (to appear in the Journal of Philosophical Logic).

GÄRDENFORS, P. (1988), *Knowledge in Flux. Modelling the Dynamics of Epistemic States*, Bradford Books / MIT Press, Cambridge (Mass.).

GIRARD, J-Y (1987), *Linear Logic*, Theoretical Computer Science 50, 1-102.

GROENENDIJK, J. & M. STOKHOF (1991), *Dynamic Predicate Logic*, Linguistics and Philosophy 14, 39-100.

HALPERN, J. & Y. MOSES (1989), *Knowledge and Common Knowledge in a Distributed Environment*, revised and expanded version, IBM Research Report RJ 4421, (to appear in Journal of the ACM).

HAREL, D. (1984), *Dynamic Logic*, in D. Gabbay & F. Guenthner, eds., Handbook of Philosophical Logic, vol. II, Reidel, Dordrecht, 497-604.

HARMAN, G. (1985), *Change in View: Principles of Reasoning*, The MIT Press / Bradford Books, Cambridge (Mass.).

HINTIKKA, J. (1973), *Logic, Language Games and Information*, Clarendon Press, Oxford.

HINTIKKA, J. & J. KULAS (1983), *The Game of Language*, Reidel, Dordrecht.

KAMP, H. (1979), *Instants, Events and Temporal Discourse*, in R. Bäuerle et al., eds., Semantics from Different Points of View, Springer Verlag, Berlin, 376-417.

KAMP, H. (1984), *A Theory of Truth and Semantic Representation*, in J. Groenendijk et al., eds., Truth, Interpretation and Information, Foris, Dordrecht, 1-41.

MAKINSON, D. (1991), *General Non-Monotonic Logic*, to appear in D. Gabbay et al., eds., Handbook of Logic in Artificial Intelligence and Logic Programming, Oxford University Press.

MILNER, R. (1980), *A Calculus of Communicating Systems*, Springer Verlag, Berlin.

MOORTGAT, M. (1988), *Categorial Investigations: Logical and Linguistic Aspects of the Lambek Calculus*, Foris, Dordrecht.

MOSCHOVAKIS, Y. (1991), *Sense and Reference as Algorithm and Value*, Department of Mathematics, University of California, Los Angeles.

NÉMETI, I. (1990), *Algebraizations of Quantifier Logics. An Introductory Overview*, Institute of Mathematics, Hungarian Academy of Sciences, Budapest.

DE RIJKE, M. (1992), *A System of Dynamic Modal Logic*, Report LP-92-08, Institute for Logic, Language and Computation, University of Amsterdam.

ROORDA, D. (1991), *Resource Logics: Proof-Theoretical Investigations*, Dissertation, Institute for Logic, Language and Computation, University of Amsterdam.

SARASWAT, V., M. RINARD & P. PANANGADEN (1990), *Semantic Foundations of Concurrent Constraint Programming*, Report SSL-90-86, Xerox Parc Research Center, Palo Alto.

SEGERBERG, K. (1991), *Logics of Action*, Abstracts 9th International Congress on Logic, Methodology and Philosophy of Science, Uppsala.

SPOHN, W. (1988), *Ordinal Conditional Functions: A Dynamic Theory of Epistemic States*, in W. L. Harper et al., eds., Causation in Decision, Belief Change and Statistics II, Kluwer, Dordrecht, 105-134.

VAKARELOV, D. (1991), *Arrow Logics*, Abstracts 9th International Congress on Logic, Methodology and Philosophy of Science, Uppsala.

VELTMAN, F. (1991), *Defaults in Update Semantics*, Institute for Logic, Language and Computation, University of Amsterdam (to appear in the Journal of Philosophical Logic).

VENEMA, Y. (1991), *Many-Dimensional Modal Logics*, Dissertation, Institute for Logic, Language and Computation, University of Amsterdam.

VERMEULEN, C. (1989), *A Dynamic Analysis of Reasoning*, Master's Thesis, Philosophical Institute, University of Utrecht.

WESTERSTÅHL, D. (1984), *Determiners and Context Sets*, in J. van Benthem & A. ter Meulen, eds., Generalized Quantifiers in Natural Language, Foris, Dordrecht, 45-71.

Logic, Methodology and Philosophy of Science IX
D. Prawitz, B. Skyrms and D. Westerståhl (Editors)

THE ONTOLOGY OF PHONOLOGY

SYLVAIN BROMBERGER and MORRIS HALLE

Massachusetts Institute of Technology

Introduction

Linguistics, like Gaul, is traditionally divided into three parts, syntax, semantics, and phonology[1], the latter being presumably concerned with the sound aspect of language. The issues we plan to discuss in this paper concern phonology directly and the other two branches indirectly. We take phonological theory to be about the world, about reality, and thus about certain items in the world — certain "particulars," as metaphysicians[2] might put it — whose existence is attested to by the fact that people speak. What is the nature of these "particulars"? In the first part of our talk, we will address that question. Our answer will be that phonology is about concrete mental events and states that occur in real time, in real space, have causes, have effects, are finite in number, in other words are what metaphysicians would call "concrete particulars" closely linked to, but *distinct* from those described by traditional phoneticians. In the second part of our talk we will consider a very different answer, according to which phonology is about types, a certain species of abstract, causally impotent, non-spatio-temporal entities, possibly infinite in number, and distinct from real live utterances. Phonology, like the rest of linguistics, is normally expounded as if it were about types. But does this mean that the discipline is committed to there being such abstract entities as types? We will argue that it isn't.

In the course of our discussion we will use some technical notation but we will keep it to a minimum and will explain it as we go. We will also talk from within a framework that some linguists may reject. That can't be helped. Linguistics is in constant flux and full of controversies. Nothing of real interest in it is conclusively established once and for all and to everybody's satisfaction.

[0]The authors received support from SERC Research Grant no. GRF/42003 during the preparation of this article.

On the nature of tokens

A. The phonological representation of a token

Let us begin by thinking about spoken tokens. And to fix ideas, let us focus on a very specific one, the one that I, the speaker[3], will now produce:

(1) *The merchant sold shelves.*

That token is now history! Only time travel could enable us to ever hear it again. It could, of course be duplicated, but it itself is gone forever. We'll come back to the fact that it could be duplicated, but let us forget about that right now. Let us concentrate on the specific event that happened a few seconds ago. We will refer to it as event (1) since we won't be able to display it again.

Actually many things happened when event (1) occurred. That is why it could be studied by more than one discipline and be analyzed differently by each. So, for instance, noises happened, and event (1) therefore could be investigated under acoustics and given an acoustical analysis. Bodily movements happened, and event (1) could therefore be studied under motor behavior and given an articulatory analysis. Brain and neurological events happened, and (1) could be looked at under neurology and given a neurological analysis. And so on.

However we exhibited event (1) to illustrate a phonological event that is an event that can be examined in the light of phonological theory and given a phonological analysis.

What would such an analysis tell us about (1)?

Well, let us look at how phonologists would represent (1) in the notation of phonology.

They would represent it as follows (the dots represent lines that we omit for present purposes):

(2) (2.a) {[ðə], Art ...} + {[mərtʃənt], Noun ...}+
{Q, Sing ...} + {[sɛl], Vb...} + {Q, Past ...}+
{[ʃɛlf], Noun ...} + {Q, Plur...}
.....

(2.b) {[ðə], Art ...} + {[mərtʃənt], Noun ...}+
{Q, Sing ...} + {[sol], Vb...} + {Q, Past ...}+
{[ʃɛlv], Noun ...} + {Q, Plur...}
.....

(2. c) ðəməˆ tʃntsoldʃɛlvz

In other words, they would represent it as a sequence of lines, a "derivation". Each line would purport to stand for some fact about (1), and the ordering would purport to stand for further facts about it.

What kind of facts?

We are going to go through the derivation step by step to answer that question. But before doing so, we want to describe the general character of that answer.

My production of event (1) was an action. Like other actions, it was therefore brought about by a distinctive kind of mental set — something we will call an "intention". But this term, as we use it, is not to be taken altogether literally. We use it to refer to a familiar kind of purposive mental stance. Think of someone aiming a rifle at a target. That person moved and positioned limbs, head, eyes, etc. in certain ways. But more went on. After all, the movements were not made accidentally, or by way of checking whether the barrel is in line with the butt. The person was set psychologically in a distinct way, i.e. had distinct intentions. More specifically, a person who aims a rifle has certain effects in mind, plans moves in ways calculated to achieve those effects, and, crucially from our point of view, has the intellectual capacity to select those effects and to devise the gestures that achieve them. The uttering of (1), like the aiming of a rifle, also required a distinctive mind set, distinctive intentions on my part, intentions that I could not have formed without certain pre-existing intellectual capacities. Of course, I had many intentions when I produced it: I intended to give you an example, I intended to be understood, I intended to produce a sentence that you have probably never heard before. But only some of my intentions account for the fact that I *pronounced* (1), that (1) was an action of pronouncing something in a language I know, in my idiolect of English. Those intentions are the kinds of facts about (1) that we take (2) to represent.

B. The last line of the derivation

Let us now look at the last line of the derivation (2), that is (2.c).

(2.c) *could* be construed as a phonetic transcription of the utterance (1). Formally it is a string of letters from an alphabet in which each letter traditionally stands for a speech sound[4]. Speech sounds are not unanalyzable entities. They are rather complexes of (phonetic) features. Thus each letter in (2.c) stands for a particular complex of features. In (3) we have given a partial list of the feature composition of some of the component sounds of English.

(3)

p b m f v	t d n s z	k g	
− − − + +	− − − + +	− −	continuant
− − + − −	− − + − −	− −	nasal
− + + − +	− − + − +	− +	voiced
labial	coronal	dorsal	Major articulator

You will readily notice that the three sets of consonants in (3) differ from each other in that each involves action by a different major articulator; i.e. [pbmfv] is produced with active involvement of the lips; [tdnsz], with that of the tongue blade or coronal articulators; and [kg], with that of the tongue body or dorsal articulator. It is the major articulator that stops the air flow from out of the mouth in [−continuant] sounds, but allows it in the sounds that are [+continuant].

In addition to the major articulators the production of consonants involves other articulators as well. In particular, the consonants [m] and [n] are produced with a lowering of the velum, which allows air to flow through the speaker's nasal cavities exciting thereby the resonances of these cavities. In all other consonants, the velum is raised, no air flows through the nasal passages and their characteristic resonances are not excited. This information is reflected by the pluses and minuses in the second line of (3). The third line reflects the behavior of the vocal cords. It is the vocal cords that implement the feature [voiced]. They vibrate in [+voiced] sounds such as [bmvdnzg] and they are stationary in the [−voiced] consonants [pftsk].

We said a moment ago that (2.c) *could* be construed as a phonetic transcription, that is, as a record of articulator movements and positionings. However that is not the way we construe it! We construe (2.c) as standing for a series of intentions that generated those movements. Each letter in (2.c) stands for such an intention, and each of these intentions called for an arrangement of articulators in the expectation of distinctive auditory effects. Each was also the intention to act so as to produce a specific English *speech* sound, and thus required a capacity that I acquired when I acquired English.

Consider, for instance, the [m] in (2.c). It represents an intention (at the time) that called for simultaneously closing my mouth at the lips, lowering my velum, adjusting the stiffness of my vocal folds, and thereby producing a sound "m". That is why the feature notation is appropriate. However it does not represent an intention that merely called for going through all that gymnastics and produce the sound "m". I could have intended that much without intending to produce an English sound, for instance while intending to hum a tune, or imitate a cow, or express mild surprise. The 'm' in (2.c) represents an intention to act so as to produce a specific *English speech* sound, a token of the phoneme /m/. And I

could not have formed that intention, that mind set, had I not acquired English[5].

The other letters in (2.c) stand for similar intentions to utter speech sounds in (1). Let us call them "phonetic intentions[6]." What (2.c) represents is therefore totally unlike what an oscillograph hooked to a microphone might have recorded, and this not only because some information recoverable from oscillograph records such as loudness, rate of speaking, fundamental frequency and other characteristics of the speaker's voice cannot be inferred from (2.c), but crucially because it stands for a different kind of event altogether. (2.c) stands for the occurrence of phonetic intentions. Oscillographs hooked to microphones record the occurrence of noises.

But why not construe (2.c) as standing for the actions that produced the noises rather than for mere intentions? The symbols, after all, were introduced in the discipline for that purpose! We have at least two reasons. The first is conceptual. (2.c), as we shall see in a moment, represents the result of a mental computation. And we don't think that actions can be the results of such computations. Results of computations have content. (2.c) characterizes a content that was executed, but need not have been executed[7]. The second is empirical. When we execute a speech action we take into account and correct for all sorts of momentary impediments and conditions. (2.c) says nothing about such corrections. It only contains linguistic information and takes into account only linguistic knowledge. As a description of the articulator actions it might be false. So we use the traditional symbols, but we don't subscribe to their standard interpretation[8].

C. The first line of the derivation

Let us now turn to (2.a), the first line in the derivation. It represents another series of intentions responsible for event (1), namely, the intentions to use certain words, e.g. the noun 'merchant', the verb 'sell' marked for past tense, etc. in a certain order. This is reflected, for instance, in (2.a) by the clustering of phonetic symbols into larger bracketed word-like units.

Forming the intention to produce (1) clearly required that I know the words I used, and that I retrieve them from memory. So before discussing in more detail how (2.a) relates to (1), let us look at how some linguists represent knowledge of words.

None of us is born with the knowledge of the words of our native language. That children learn words as they develop is obvious and moreover massively documented, and we all know that the process goes on through

life: most of us have only recently acquired such words as *intifada*, *glasnost*, *scud*. The proposition that an essential part of learning a language consists in storing in one's memory (something representable as) a list of words, something we will call "the vocabulary", is therefore one of the most securely founded in all linguistics.

Many words that any speaker of English knows are complex in the sense that they incorporate affixes of various kinds. We illustrate this in (4).

(4) (4.a) shelv-es, child-ren, bough-t, sol-d
(4.b) pre-dis-pose, un-happy, in-secure
(4.c) un-poison-ous-ness, ex-pre-sid-ent, contra-in-dic-ate-d
(4.d) kibbutz-im, hassid-im

In linguistics the term "stem" is used to designate the element to which an affix is added and the term "morpheme" is used as a cover term for both affixes and stems. Like stems, affixes too must be learned and committed to memory (cf. (4.d)).

A speaker's knowledge of morphemes can thus be represented as a list of items containing information about each morpheme stored in memory.

What information?

Obviously information about its meaning, its functional structure, and the thematic roles it assigns. Also about its lexical category, i.e. whether it is a noun, verb, preposition, adjective, conjunction. And certainly information pertaining to how the morpheme is pronounced, that is, phonological information. All this can be thought of as encoded in a "complex symbol", made up of elements that stand for meaning, lexical category, etc. The markers pertaining to how the morpheme is pronounced are of particular interest to us here. We will refer to them as the "identifying index". The vocabulary as a whole can be thus represented as a long list of such complex symbols, each of which contains, among other things, an identifying index[9].

We now turn to the information in identifying indices.

Most morphemes take on the same phonetic form regardless of syntactic and/or morphological contexts. Thus the verb hint shows up in the phonetic form representable as [hInt] whenever uttered. So that string of phonetic symbols is used as its identifying index.

Other morphemes assume different phonetic forms depending on syntactic and/or morphological contexts. For instance, the stems *sell* and *shelf* were pronounced differently in (1) than in the following utterance:

(5) *The merchant sells a shelf.* [sɛl] [ʃ ɛlf]

The identifying index of such stems is also a string of phonetic symbols, namely [sɛl] and [ʃ ɛlf] in the two cases at hand. We will come back to why those particular strings.

Some morphemes, notably the English plural and past tense affixes, not only assume different phonetic forms in different contexts, but these forms can be utterly dissimilar, and sometimes they don't appear phonetically at all!

So note what happens to the plural affix in the following cases:

(6) cat/s, child/ren, kibbutz/im, alumn/i, stigma/ta, geese, moose

and to the past tense morpheme in the following

(7) bake/d, playe/d, dream/t, sol/d, sang, hit

Halle (1990) has dubbed morphemes like the Plural and Past Morphemes which behave in this very irregular fashion "abstract morphemes" and he has used 'Q', a symbol that has no direct phonetic interpretation, as their identifying index[10].

With all this in mind, let us look again at (2.a).

(2.a) is a sequence of complex symbols each made up of an identifying index and other grammatical markers, all copied from the vocabulary.

What facts about event (1) does (2.a) represent?

(2.a) as we said before represents the intention to use certain words, but we can now be more explicit. (2.a) represents the fact that (1) besides being produced by my phonetic intentions in (2.c) was also produced by my intention to use the words retrieved from my vocabulary whose identifying indices (and lexical category) appear in (2.a).

But what are the phonetic symbols doing in (2.a)? Take the initial 'm' in the identifying index of 'merchant'. Does it represent an intention, already present at that stage so to say, to produce a token of the phoneme /m/? Offhand that may seem reasonable. But consider then the 'ɛ' in the identifying index of 'sell'. It can't stand for an intention to produce a token of the phoneme /ɛ/. No such intention was executed in producing (1). Could I have changed my mind between the times I picked the words and pronounced them? That strikes us as a cute but vacuous idea, too literal minded about our use of "intention". As we see it, the role of the phonetic symbols in (2.a) and in (2.c) is very different. In (2.a) they play a computational role. Formulae such as (2.a) have two functions. On the one hand they model an event, represent aspects of that event. On the other hand they are used to compute other formulae in the formalism of our theory. Phonetic symbols appear in (2.a) essentially to simplify computations within the theory. In (2.c) they have that role but they

also represent phonetic intentions. These roles, though connected, are different.

Note that in the vocabulary phonetic symbols could not stand for intentions either. The vocabulary is not a representation of intentions, but of knowledge. But its formulas too enter into computations.

D. The second line of the derivation

Let us now look at (2.b).

(2.b) stands between (2.a) and (2.c). It is like (2.a) except that some of the phonetic symbols in the identifying indices have been changed. Unlike (2.c) it is partitioned, contains syntactic categories labels and occurrences of Q.

What facts about event (1) does (2.b) represent[11]?

It represents a stage between the formation of my intentions to use words, i.e. my intentions represented by (2.a) and the formation of my phonetic intentions, i.e. my intentions represented by (2.c). Unlike (2.a) and (2.c) it does not represent intentions at all, though it does represent a mental set of sorts.

Remember events (1) and (5), the actual utterances? In the earlier one I pronounced the verb one way and in the later one I pronounced the same verb very differently. The facts underlying the difference can be surmised from a vast body of evidence though they also happen to coincide with common beliefs. I know English and that means that I have not only acquired words, but have also acquired rules. In producing these utterances I applied appropriate rules, and this led to different pronunciations of the same verb.

(2.b) stands for a stage in the application of these rules.

We can even tell what stage.

As noted before, the verb *sell* and the noun *shelf* appear in two distinct guises in different utterances. Specifically the verb *sell* undergoes a vowel change in the past tense, and the noun *shelf* undergoes a change of the final consonant, in the plural. In other words, in producing these words we invoke something like the rules in (8).

(8) a. Before [Q,Past] the stem vowel is [o] in the verbs *sell*, *tell*, ...
b. Before [Q, Pl] the stem-final consonant is [+voice] in the nouns *house*, *knife*, *life*, *wife*, *shelf*, *mouth*, ...

(2.b) then represents a stage after the application of (8).

(8) are not the only rules I applied to produce (1). I also applied rules to pronounce the morphemes represented by Q in (2.a). Halle (1990) has argued that the relevant rules are statable roughly as follows:

(9) $Q \rightarrow$ /n/ in env. X — Plural if X is *child*, *ox*, ...
/im/ in env. Y — Plural if Y is *kibbutz*, *hasid*, ...
/i/ in env. Z — Plural if Z is *alumn-*, *radi-*, ...
/ta/ in env. U — Plural if U is *stigma*, *schema*, ...
ϕ in env. V — Plural if V is *mouse*, *moose*, ...
/z/ in env. — Plural

(10) $Q \rightarrow \phi$ in env. X — Past if X is *sing*, *write*, ...
/ta/ in env. Y — Past if Y is *buy*, *dream*, *mean*, ...
/d/ in env. — Past

So (2.b) also represents a stage before the application of those rules!

But what do the phonetic symbols in (2.b) represent? For that matter, what do they represent in the statement of the rules? The answer here is as before. They play a role as symbols in the formal computations of the theory. We conjecture that they also stand for something specific in the production of (1), but if they do, what they stand for is not something clearly understood at this time.

This double role assigned to phonetic symbols, we should point out, has a shortcoming: it slights certain important phenomena. So, for instance, it does not show that between the formation of (2.a)'s referent and (2.c)'s referent a kind of transubstantiation occurred through which mnemonic elements were converted into articulatory ones.

On the nature of types

So far we have concentrated on a single and unique event, the utterance of (1). We have done this because we hold that phonological theory, in so far as it purports to advance knowledge at all, is about such events and about the mental conditions responsible for their occurrence. Those are the sorts of things to which it is ontologically committed, or, as some followers of Quine would put it, those are the kinds of things over which it quantifies.

Our position may strike some as prima facie implausible, as simply conflicting with too many practices of phonologists.

Thus phonologists never mention or try to explain unique events like (1). Their papers and texts mention words, phrases, sentences, phonemes, i.e. types, abstract entities outside time and space, devoid of causal histories and causal consequences. They don't mention utterances, events, or mental states. And though phonologists do sometimes elicit tokens,

they do so only to obtain data about types. That is presumably why they rely on statements which abstract from whatever is peculiar to tokens and clearly fit types. In fact, most phonologists would probably not interpret (2) as about (1), the utterance produced umpteen minutes ago. How could they? How many have ever heard of that utterance! They would implicitly take (2) as about a type possibly attested by something like (1) but attestable as well by other tokens, for instance by the one that I now produce

(11) *The merchant sold shelves.*

And that last token, as Leibniz's indiscernability of identicals tells us, is not only numerically distinct from (1), since it occurred at a different time, but is also numerically distinct from its type, which does not occur in time at all.

Furthermore, phonologists, like all grammarians, strive for theories that neither undergenerate nor overgenerate, theories that predict some items and exclude others. But the relevant items could not be tokens! If they were, any phonological theory, no matter how absurd, could always be trivially confirmed. Imagine, for instance, a theory which predicts that the following is in my language

(12) *The plmotpfu sell yesterday many shelf.*

I would have confirmed that theory simply by having produced the token (12)! What is more, every theory of any interest would be demonstrably false since it would predict an infinite number of tokens for each person, whereas the total number of tokens is bound to be finite. Life is short! Even the most loquacious of us, no matter how long they live, will shut up forever at some point!

That may all be true, nonetheless, we don't believe that there are types! And so phonology can't be about types. We admit (on empirical grounds) internalized grammars, but those exist as mental attributes of concrete, specific human individuals. We admit (on empirical grounds) internalized vocabularies, but those too exist as mental attributes of concrete, specific human individuals; and we admit token events (again on empirical grounds) but those are spatio-temporally located concrete events like (1). But types, we think, belong, with useful fictions like virtual optical images, in the null class.

We can't prove that there are no types. The notion is surely not self-contradictory, or even incoherent. Bromberger (1989) has argued that it is a coherent and even useful one. We just don't see any reason to think that there are any. And we don't accept that phonology provides any evidence for them, or must presuppose their existence.

On the other hand, we do believe that phonology provides overwhelming evidence that tokens cluster into scientifically significant types. That does not imply that there are types besides tokens (not even as sets, or mereological sums of tokens, though such sets and sums may well exist). But it is sufficient to justify most of the practices we have mentioned and to make sense of the demand that theories should neither overgenerate nor undergenerate.

We will now explain that position in more detail.

Instead of producing (1) when I did, I could have produced a very different token. In fact, to fix ideas, here I go:

(13) *Two elephants study in Uppsala.*

And we could now go on and produce a derivation analogous to (2) for this new token. It would be a different derivation from (2). The last line, the first line, the intervening lines, the rules invoked, the items said to represent morphemes retrieved from lexical memory, all would be different. However the two derivations would have one crucial thing in common: they would incorporate answers to the same questions. Different answers, but the same questions.

In other words, event (1) was open to the questions: What morphemes were intended? What was the representation of these morphemes in memory? What articulatory gestures were intended? What rules were invoked in the course of the formation of these intentions? and so on. (2) provides the answers to these questions[12]. The last token (13) was open to the same questions. Its derivation would also provides answers to those questions, but the answers would be different. As we just said, different answers; the same questions. In fact all spoken token events are open to these questions. Some share exactly the same answers. Token event (1) and the token event (11) do. Others don't share the same answers. Token events (1) and (13) don't. Token events that share the same answers are the ones we classify as being of the same type. Those that don't, we classify as being of different types.

That is all there is to talk of types, as far as we are concerned. But that is quite a lot, as we will now show.

Note, for instance, that each token holds specific answers to these questions. We are unmitigated realists about this. We take the fact that each token holds the specific answers it does to each of these questions to constitute truths about the world, not some artifact of our way of looking at things. It is a truth about the world that the answer to "what was the first intended articulation underlying the production of (1)?" is "vibration of the vocal cord (i.e. [+voice]), constriction of the mouth cavity partially

open (i.e. [+continuant]) etc." just as it is a truth about the world that the answer to "How much does Sylvain Bromberger weigh?" is "175 lbs".

Note furthermore that we also take it as a truth about the world — and a very different sort of truth — but not an artifact of our way of looking at things, that (1) has the property of holding such answers at all. We might put this somewhat more technically. It is a truth about the world that event (1) had the determinable property of having intended morphemes. And it is a truth about the world that each spoken token also does. Other events, even events with acoustic properties, don't have that property. Noises made by our coffee pot, or coughs for instance, don't have it. That fact is of the same order as the fact that swinging pendula have periods which standing rocks don't, that positive numbers have square roots, which trees don't, that the manuscript from which we are reading has a certain weight, which the ideas we are expressing don't. Determinable properties, by the way, like period, square root, weight, and so on, are a kind of property presupposed by what-questions such as "What is the period of ...?", "What is the square root of ...?", "What is the weight of ...?" Objects that don't have the property hold no answer to the corresponding question.

(1), (5), (13) , and other tokens are, of course, open to many questions besides the ones answered in derivations such as (2). They hold answers for instance to "At what time was it uttered?", "Where was it uttered?" If we were to use those to compare tokens as to type we would end up with very different typological clusters. But we don't use those. We use questions that define the field of phonology. That may sound more selfrighteous than we intend, so let us put what we have in mind differently: we use the questions that make phonology *as a theoretical field* possible. That there are such questions, by the way, is also an empirical truth about the world!

An analogy that we have used elsewhere, taken from elementary chemistry, may be helpful here. Think for a moment about a sample of water, a real sample that you have "experienced" as they say in California. That sample, like our tokens, is open to a number of questions: "Where was it situated when you experienced it?", "Who owned it?", "What did you do with it?" that *are* of no scientific interest. But it is open to others that are of scientific interest, such as "What is its boiling point?" "What is its freezing point?" "What is its molecular weight?" Other samples of stuff get the same answers to these last questions. They comprise all and only samples of water. Still other samples, though open to the same questions, get different answers. They comprise all and only the samples of other distinct substances, e.g. samples of gold all share one set of answers, samples of mercury share another set of answers, samples of sulfuric acid

share yet another set, and so on. And still other samples of stuff don't hold answers at all to these questions. Pieces of sausage, for instance, or handfuls of mud, or buckets of shoelaces. These don't make up samples of a substance at all! The scientifically interesting questions collect bits of stuff into samples of substances. Not all bits of stuff. Some bits of stuff.

But what makes these questions scientifically interesting?

That should be obvious: the fact that their answers conform to law-like, or at least computable, relationships. There are law-like, or at least computable, relationships, between boiling points, freezing points, and molecular weights. A theory can therefore be constructed on the back of these questions. And a certain attitude can be acquired. For instance, that these samples constitute a "natural" domain, a domain that includes some (water, gold, mercury, etc.) but excludes others (sausage, mud, shoe laces); and that these samples have features that demand similar explanations. Of course these attitudes are warranted only if certain facts obtain, i.e. certain law-like relationships actually hold. For all we once knew, the world might have been otherwise.

A similar story, we believe, applies to utterances. Each utterance is open to a multiplicity of questions. The scientifically interesting ones are those whose answers across tokens stand in law-like, computable relationships to each other. As we noted before, that there are such questions, if there are, is a fact about the world. An interesting fact. We believe there are[13]. If we didn't we would not spend time on linguistics. And we believe that the questions we characterized as defining the domain of phonology are among these questions. If we didn't we would not pursue phonology as we do. All of that then is implicated in our talk about types. And none of it requires us to hold that there are types over and above tokens.

We want to stress one crucial further fact about questions of scientific interest. They are not all given, they are not part of ordinary common sense, they must be smoked out, discovered, and their discovery can be an achievement more revolutionary than the discovery of a new phenomenon. Newton, for instance, discovered many things about the world, but his most important discovery (outside of optics) would not have been possible without the discovery that physical bodies have mass. That was the discovery of a new kind of question (and dispositional property), revolutionary because answers to it turned out to stand in marvelous computable relation to other questions (e.g. "What force is required to accelerate such and such one foot per second per second?"). Aristotle, a very good scientist, did not have the concept of mass, and therefore could not have wondered what the mass of the moon was, could not even know that he did not know what the mass of the moon was, and, of course, could not

have fathomed that the answer to that question was related in systematic ways to the answer to other questions about the moon. Celestial mechanics was beyond even his imagination!

The questions that make phonology possible also had to be discovered. And their discoverers are the heroes of our field: Panini, Rask, Bopp, Saussure, Jakobson, Chomsky. Without them phonology would still consist of pedantically collecting odd curiosities. But the work of discovering the right questions is far from finished!

As we mentioned at the very outset, utterances form a natural domain with other noises, the domain of acoustical theory. To deny this would be like holding that elephants, because they have a sex life, are not, like rocks, physical objects subject to the laws of mechanics! On the other hand to deny that utterances constitute an autonomous domain, the domain of phonology, would be like holding that because elephants are subject to the laws of mechanics, like rocks, they have no sex life! And to deny a-priori that there are systematic relationships between these two domains would be like denying a-priori that there are systematic relationships between the mass of elephants and the character of their sex life. Maybe there are. And some may even be surprising.

We can now tell what we make of the fact that (2), though about (1), contained no information about time, speaker, etc. and could have served for (11) as well. (2) contains only answers to questions of interest from the position of theoretical phonology. It could serve for any token that holds the same answers to those questions. However nothing in this requires that there be types besides tokens. Talk of types then is just a façon de parler[14].

But that still leaves us with the requirement that phonological theory should not overgenerate or undergenerate. How do we construe the prohibition against overgeneration? Realistically: no phonological theory is true of the world that generates derivations (combinations of answers to questions, like (2)) to which no token in our language (i.e. produced as (1) was produced) *can* conform[15]. Phonological theories may not ignore constraints on the production of tokens that are imposed by the internalized rules and vocabulary. There is still no need to assume types.

What of the prohibition against undergeneration? We construe that also realistically: no phonological theory can be true of the world that can't generate derivations to which tokens in our language (i.e. produced as (1) was produced) *could* conform. Phonological theories may not presume more constraints on the production of tokens than are imposed by internalized rules and vocabulary. There is again no need to assume types.

Our construal of these principles is stated with the aid of modalities (ex-

pressed by "could") which vex many semanticists. But that is no reason to admit types. We find admission of types unenlightening as substitutes for these modalities, and at least as vexing.

Two final comments.

Some people may object to our way of looking at phonology on the grounds that it construes phonology as about performance and not about competence[16]. If they mean that we view phonology as about processes in real time responsible for the occurrence of tokens, they are right about our view. But we don't see this as an objection. If they mean that we view phonology as having to take into account contingencies of production over and above those traceable to knowledge of language, then they misconstrue our view. We don't.

Some will object that we have loaded phonology with unwarranted assumptions. Do speakers "really" retrieve morphemes from their memory, invoke rules, go through all these labors when speaking? We think they do. In fact we would like to know more about how they do it. We may be mistaken. Time will tell. But intuition won't. Clearly speakers are not aware of performing such actions. But then we perform many actions like zombies (to borrow a phrase from Ned Block). That is how we learn language, recognize faces, and solve most of our problems.

Some will object that our outlook leaves out entirely that tokens are not only uttered but are also recognized. That is indeed a big hole in our account so far. But it calls for another paper, not abstract types.

Let us then return to the topic of our title the ontology of phonology. What must the "furniture of the world" include if phonological theory, as we conceive it, is to have a chance of becoming true of that world? It is a long list: agents, tokens, phonetic intentions, minds with vocabularies and rules, articulators, and so on, in complicated interrelations. It does not have to include types. And if, perchance, the world does include types, phonology has nothing to say about them[17]. But then, probably no branch of linguistics does.

NOTES

1. We are grateful to Ned Block, Nancy Bromberger, George Boolos, Noam Chomsky, Leonard Clapp, Alec Marantz, Wayne O'Neil, Jean Michel Roy for comments on previous drafts of this paper.

2. See Bromberger and Halle (1989) on why phonology is fundamentally different from syntax and semantics.

3. Though many philosophers of language have views on empirical linguistics, few, if any, have given serious attention to phonology. Recent anthologies and books on the philosophy of language either don't mention phonology at all, or at best perfunctorily restate crude and outdated notions on the subject. This is somewhat surprising since the facts that phonology studies are critical as objects of speech perception or outputs of speech production. But for these facts, there would be no syntax or semantics of natural languages besides sign languages (also neglected by philosophers), and philosophers deliberating about such languages would have to be silent.

   There are a number of explanations for this neglect. To begin with, recent philosophers of language generally belong to an intellectual tradition that admits no essential differences between natural languages and some of their contrived extensions. This was pointed out a long time ago by Strawson (1950), though he had other shortcomings in mind. Philosophic discussions thus generally abstract not only from differences between English, German, Japanese, and other natural languages, but also from differences between these languages and notational systems used in mathematics, logic, physics, chemistry, biology, linguistics, etc. Such notational systems do have a syntax (albeit usually one that has very little in common with the syntax of natural languages), a semantics, and a pragmatics, but happen to have no phonology. Their minimal units are normally ideographs which encode word-like units — rather that phonetic or even orthographic ones — open to many phonologically unrelated pronunciations (if pronounceable at all). Nothing in such notational systems corresponds to the phonologies of natural languages, and nothing about them can thus be captured in an overarching phonological doctrine linked to the overarching semantic and syntactic doctrines studied by philosophers. So it is not surprising that though Frege and his successors include signs in their Sign-Sense-Nominatum triad, they have nothing of interest to say about signs as things uttered and heard. Even philosophers who focus primarily on natural languages belong to that tradition and don't discriminate between aspects peculiar to real languages, i.e. articulated languages whose primary tokens necessarily vanish as soon as produced, and aspects peculiar to conventional notational systems with characteristically enduring tokens.

   Furthermore, philosophers generally seem to believe that there can't be anything of philosophic interest about phonology. This attitude flows naturally from the previous one. The ideographs used by scientists are adopted through open and explicitly published conventions that determine everything true of them qua signs. Since they are also semantically word-like, their subsegments have no autonomous status and raise no philosophic problems. How could there be anything of philosophic interest about the shape of simple numerals, or about the horizontal segment in an inverted A in a quantifier, or in the vertical line of the F for force? It is easy to pass from this outlook to the view that there can't be anything philosophically challenging about the spelling of words, and thence to the view that there can't be anything

philosophically challenging about their pronunciation. And isn't phonology "just" about pronunciation?

Whatever the explanation for philosophers' neglect of phonology, we think that it has a cost. To begin with, we think that it is a mistake to lump all lexical systems together as forming some kind of natural family. It blurs too many differences, and attention to phonology can highlight important ones. Furthermore, we think that no theory about the relation between natural language signs and their referents (or their meaning) can be trustworthy that nonchalantly takes signs for granted. Finally, we think an adequate understanding of the ontology of language — of the objects whose existence constitutes the reality of language — must include an adequate conception of the objects investigated by phonology. More specifically, an adequate conception of language must check our tendency, when we reflect about language, to slip thoughtlessly between talk about individual utterances and talk about types, as if such slips were always innocuous and easily fixed ways of avoiding pedantry. Spoken tokens are transitory events that occur in time and space, that can be perceived, that are shaped by their speaker's occurent intentions, and that are subject to norms fixed in their speaker's mental make-up. Types — if there are types — are abstract entities, neither in time nor in space, devoid of casual histories or causual consequences; hence beyond perception. Types — if there are types — outdistance tokens. Tokens and types — if there are types — are thus utterly different. Conflating them is bound to lead to confusions and incoherence. But giving each its due, and understanding their connection, won't be possible unless we see how the type-token distinction fares in phonology, a topic which we discuss in what follows.

4. We use the first person singular to mention Sylvain Bromberger as producer of tokens displayed during the talk in Uppsala and use the first person plural to mention ourselves, the two authors. We use italic font to indicate displayed spoken tokens. Readers of the paper should thus keep in mind that the lines in italic point to episodes which they can no doubt imagine but which took place in Uppsala when this paper was orally presented.
5. The fact that the utterance consists of sequences of discrete sounds is the insight on which all alphabetic writing systems are based. It may therefore appear to be self-evident. Yet when we last asked our colleagues working on the automatic analysis of speech, we were told that no one has yet found a reliable mechanical procedure that can segment any arbitrary utterance into its constituent sounds.
6. Compare with the availability of click sounds as phonemes to speakers of Bantu, but only as noises to speakers of English.
7. For a related position see Libermann, A. M. and Mattingly, I. G. (1985) (1989) and Bromberger, S. and Halle, M. (1986).
8. Elsewhere we have called this kind of intentional content a "score", to mark the analogy with a musical score, something that can be executed through motions, but need not be executed, and often is not. Inner discourse probably stops at the formation of such "scores".
9. Our interpretation of event (1) commits us to the occurrence of the mental events we call intentions over and above acoustical events and articulatory events. We present some of our reasons in Bromberger, S. and Halle, M. (1986).
10. In talk about identifying indices we must take care to distinguish between what we presume to be "in the mind" of the knower and its representation in the notation

of our theory. We use the term "identifying index" to refer to the representation in the notation of the theory, not what is "in the mind".

11. Other elements in the lexicon may also lack specific phonological content, e.g. PRO, empty complementizers, case, expletives, etc. yet not be represented by a complex symbol that includes any identifying index like Q. Whether or not some morpheme must be represented with an identifying index like Q or no identifying index at all is a contingent matter to be settled on empirical grounds. Halle's proposal is not that we adopt a convention for the sake of giving all complex symbols a common format. It embodies the claim that all unarticulated morphemes are not phonologically equal.
12. Some things we have said so far could be put in Austinian terminology (as in Austin (1975)).
    (2.a) modeled the formation of a *phatic* intention, that is, the intention to produce what Austin called a phatic act, "the uttering of certain vocables or words ... belonging to and as belonging to, a certain vocabulary, conforming to and as conforming to a certain grammar". (We leave out "i.e. noises of certain types" as misleading.) (2.c) modeled a *phonetic* intention, i.e. the intention to produce what Austin called a phonetic act, "the act of merely uttering certain noises" (The "merely" is unfortunate!). (2.b) then models a stage in a mental *process* through which the phatic mental set gets transformed into the phonetic mental set.
13. Strictly speaking, to provide these answers it would have to be supplemented with references to the rules, as it was in the course of our discussion.
14. The law-like computable relationships are those that govern the production of utterances. Not all utterances, but only of utterances produced by invoking rules and a lexicon. We can, of course, produce utterances without such invocation. The crucial fact, not revealed by simple common sense, is that we can produce utterances that do invoke them.
15. We ourselves resort to this façon de parler even in this paper, when, for instance, we speak of morphemes, phonemes, etc.
16. Admittedly this answer requires elaboration. For instance, it seems to avoid commitment to phonological types at the price of commitment to types in the notation of the theory. We think that this appearance can be dispelled, but to do so would require a long discussion of theoretical formalisms. In any case, we are not claiming that there are no abstract entities at all. We are claiming that phonology is not about types.
17. See for instance Geoff Lindsay and John Harris (1990).
18. Chomsky, in a number of writings e.g. Chomsky (1985), distinguishes between two conceptions of language (i.e. of the objects amenable to scientific linguistics): (a) a conception of language as an (infinite) set of expressions (signs) or of expressions paired with meanings, what he calls "E-language" (Actually he subsumes a number of different conceptions of language under E-language, only one of which takes language as a set of types.); (b) a conception of language as a mind/brain state attained under certain contingencies (including exposure to other speakers) and made possible by certain innate brain/mental organizations, what he calls "I-language". He has demonstrated serious shortcomings in all the studies of language offered so far that are based on an E-language concept, and has urged approaches based on the I-language concept. Our paper shares his conviction that I-language is a more appropriate object for scientific study. This is not an accident: we came to our position largely through reflections on his.

CITED REFERENCES

AUSTIN, J. L., 1975, *How To Do Things With Words*, 2d ed., ed. J. O. Urmson and Marina Sbisa. Harvard University Press.

BROMBERGER, SYLVAIN, 1989, *Types and Tokens in Linguistics*, In Alexander George (ed.), Reflections on Chomsky. Basil Blackwell, 58–89.

BROMBERGER, SYLVAIN and HALLE, MORRIS 1986, *On the Relationship between Phonetics and Phonology*, In Joseph S. Perkell and Dennis H. Klatt (eds.), Invariance and Variability in Speech Process. Lawrence Erlbaum Associates, Publishers, 510–520.

BROMBERGER, SYLVAIN and HALLE, MORRIS 1986, *Why Phonology is Different*, Linguistic Inquiry 20 (1), 51–70.

CHOMSKY, NOAM, 1985, *Knowledge of Language*, Praeger.

HALLE, MORRIS, 1990, *An Approach to Morphology*, In Proceedings of North East Linguistic Society. Vol 1, 150–184.

LIBERMAN, A. M. and MATTINGLY, I. G. 1985, *The Motor Theory of Speech Revised*, Cognition 21, 1–36.

LIBERMAN, A. M. and MATTINGLY, I. G., 1989, *Specialization for Speech Perception*, Science 243, 489–494.

LINDSAY, GEOFF and HARRIS, JOHN, 1990, *Phonetic Interpretation in Generative Grammar*, In University College London Working Papers in Linguistics 2, 355–369.

STRAWSON, P. F., 1950, *On Referring*, Mind 235, 320–344.

Logic, Methodology and Philosophy of Science IX
D. Prawitz, B. Skyrms and D. Westerståhl (Editors)

RELATIONAL NOUNS

M.J.CRESSWELL

Department of Philosophy, Victoria University of Wellington, New Zealand

When we give examples of two-place predicates in teaching first-order predicate logic it is possible to follow Arthur Prior (see, e.g., Prior 1956, p. 85) and use transitive verbs, so that

(1) Trevor admires Sally

is an instance of Fxy, with 'admires' replacing F. But it is far more frequent to use examples like

(2) Arabella is the mother of Dan

in which the predicate is made up using a noun and a preposition, most likely the preposition 'of'. The noun 'mother' on its own, occurring in a sentence like

(3) Arabella is a mother

appears to be semantically a one-place predicate such that (3) means that Arabella is the mother of someone. No doubt for the purposes of logic no harm comes from treating 'is the mother of' as a single two-place predicate and 'is a mother' as a distinct though related one-place predicate, despite the danger illustrated by the favourite sophistic conundrum,

(4) That dog is yours
That dog is a father
$\therefore$ That dog is your father

but if we are in the business of providing a logical treatment of the expressions of natural language then we can hardly rest content. The facts

illustrated in (2) and (3) come from English, but as far as I am aware they are common to Indo-European languages, and I do not consider myself to be addressing an issue peculiar to one language. Nevertheless I shall restrict myself to English. If we take seriously the apparent fact that 'of Dan' is a prepositional phrase then we will want the relation between (2) and (3) to be exactly analogous to that between

(5) That is the house at Pooh Corner

and

(6) That is a house.

The difficulty of maintaining this analogy has led most semantical theories to treat (2)/(3) as a quite different phenomenon from (5)/(6). While what I say may not be enough to settle the issue I shall at least be able to provide a formal framework within which the parallel can be maintained, so that we can no longer offer as a reason for the disanalogy that no *semantics* could treat (2) and (3) like (5) and (6). As motivation for this treatment consider some facts noticed in Partee 1989. Consider

(7) The enemy is well-supplied.

If we are in Wellington's army then the enemy is the French. But if we are in Napoleon's army it is the British and their allies. In (7) there is no preposition and yet it seems nevertheless that the context has to supply a point of view to judge who is the enemy. But it is not just context which does this; rather, as with most positions in a sentence which have a value filled by context, the position sometimes plays the role of marking a bound variable. Partee's example is

(8) Every soldier faced an enemy.

(8) could certainly mean that each of Wellington's soldiers faced a (French) enemy soldier, while each of Napoleon's soldiers faced a (British) enemy soldier. There is no single context which supplies the enemy.

The problem with all these cases is that the noun in question appears to have the same syntax whether it occurs as part of an explicit relational statement or not. So the problem is how to accommodate a semantical property which is not mirrored in the syntax. This kind of problem is not new, and it has a long-standing solution by means of the use of semantical

indices. The clearest examples are modal and tense logics. In a temporal language the sentence

(9) It is now four o'clock

does not express a fixed once-and-for-all proposition. If it did it would either be necessary or impossible. It expresses what some have called an open proposition. This is a function from times to (closed) propositions, where these latter may be thought of as sets of worlds. And even these latter have been thought, as in Montague 1974, p. 153, to be functions from worlds to truth values. (And a temporal but non-intensional language could treat propositions as functions from times to truth values.)

What can those who favour an extensional language do? In the early days the fashion was to consider only what Quine 1960b, p. 193f, calls 'eternal sentences' and refuse to give a definite meaning to sentences like (9). But another course, favoured by Taylor 1977, is to treat (9) as an open sentence with a free variable having the form

(10) It is four o'clock at t.

Since (10) contains a free variable it does not have a fixed truth value but only satisfaction conditions. It is true for the assignments to t which give it four o'clock, and false otherwise. The difference between (9) and (10) is not semantic, it is just that what (10) puts into the syntax (9) 'hides' in the semantics. The use of semantical indices is to keep relativity out of the syntax. What I propose to do therefore is see how we might explain the semantics of (2) and (7) using semantical indices.

In part III of Cresswell 1990 I shewed how to express quantification in a propositional language using sequences of individuals as semantical indices and operators of the kind considered in Quine 1960a, Kuhn 1980 and others. This language is as powerful as an intensional predicate language with two-place generalized first-order quantifiers operating on predicates made by λ-abstraction. Such languages can be considerably more powerful than ordinary first-order languages since they may contain two-place 'quantifiers' like *most*, which are known to be inexpressible in ordinary first-order languages.

The syntax of a propositional language $\mathcal{L}$ is extremely simple. The atomic symbols are of two kinds

(i) Simple sentence symbols
(ii) For each n, a set (possibly empty) of n-place sentential functors.

We may impose the requirement that there be only finitely many atomic symbols. The usual restrictions apply to ensure that all the atomic symbols are distinct, and that none is itself a sequence of other symbols. Thus we have what Montague 1974 p. 225 calls a disambiguated language. Some authors like to relax this restriction but I will not. The rules for generating complex symbols are equally simple:

FR1 Every simple sentence symbol is a well-formed formula (wff).

FR2 If δ is an n-place sentential functor and $\alpha_1, \ldots, \alpha_n$ are n wff, not necessarily distinct, then $\delta\alpha_1 \ldots \alpha_n$ is a wff.

Since $\mathcal{L}$ contains no bound variables there is no distinction between closed and open wff. For that reason the word 'sentence', sometimes used for closed wff, could have been used in place of wff. Provided that we know, for each functor δ, how many places it has, and provided the functor precedes all its arguments there is no need for parentheses. It is sometimes convenient, however, to place a functor between its arguments. Thus we have $(\alpha \supset \beta)$ rather than $\supset \alpha\beta$. If this is done parentheses become necessary. For realistic syntax it is perhaps better to think of wff as trees, or to represent them as having a somewhat more complex set-theoretical structure as I did in Cresswell 1973, but for the issues of this paper the wff can be those and only those sequences of the symbols of $\mathcal{L}$ which can be generated by FR1 and FR2.

But although $\mathcal{L}$ has a simple syntax it has a complex semantics. In chapter 13 of Cresswell 1990 I motivated propositional languages by first introducing languages with generalized quantifiers operating on λ-expressions and then proving that the propositional language was equally powerful. Here I will present the semantics directly. I will, however, make a simplification. The principal purpose of Cresswell 1990 is to shew that natural language requires the power of full quantification over possible worlds, and the propositional languages introduced in Part III had as semantical indices not only sequences of individuals but also sequences of worlds and sequences of times. In this paper I shall use only a sequence of individuals, though $\mathcal{L}$ will still be an intensional language and truth will be relative to a world and a time. An interpretation for $\mathcal{L}$ will be a triple $\langle W, D, V\rangle$. Let W be the class of all world-time pairs and D the class of 'things'. I say 'things' rather than 'individuals' only to make clear that D is whatever we quantify over, whether it is metaphysically an individual or a higher-order entity. It is my belief that reference to higher-order entities need not require a higher order syntax. Though again, for the purposes of *this* paper, nothing is lost by taking D to consist solely of individuals, whatever they are. The final member of an interpretation for $\mathcal{L}$ is a function V which connects expressions of $\mathcal{L}$ with their meanings. The idea is

that the meaning of every sentence is a set of triples of the form $\langle w, u, \sigma \rangle$, where $w \in W, u \in D$ and σ is a sequence such that for $1 \leq n$, $\sigma(n) \in D$. (Alternatively we could let $u = \sigma(0)$ and let the triple be replaced by a pair $\langle w, \sigma \rangle$. The role of $\sigma(0)$ is rather special, so I have here signalled it separately rather than use $\sigma(0)$ as I did in Cresswell 1990.) Call the set of all such sets of triples P, for proposition, though in the sense of open proposition introduced in Cresswell 1973 — if $a \in P$ then a is the value of a sentence of $\mathcal{L}$. V must satisfy the following:

(a) For atomic sentence $\alpha, V(\alpha) \in P$

(b) For n-place functor δ, $V(\delta)$ is a function ω such that for $a_1, \ldots, a_n \in P$, $\omega(a_1, \ldots, a_n) \in P$

(c) For complex wff $\delta\alpha_1 \ldots \alpha_n$, $V(\delta\alpha_1 \ldots \alpha_n) = V(\delta)(V(\alpha_1), \ldots, V(\alpha_n))$.

I promised you a complex semantics. But I have delivered simplicity. The complexity comes when we see how this apparently simple semantic framework deals with what in standard treatment requires a complex syntax.

First look at a predicate. Syntactically in $\mathcal{L}$ *admires* would be a wff. $V(admires)$ would therefore be a set of triples of the form $\langle w, u, \sigma \rangle$. In particular it would be that set a such that $\langle w, u, \sigma \rangle \in a$ iff $\sigma(1)$ admires $\sigma(2)$ at w. Now *admires* is semantically a two place predicate, but when we use it in a sentence we are usually interested in dealing with its arguments one at a time. Thus in the standard example

(11) Everyone admires someone

it makes a difference whether we first form the predicate 'admires someone' and then let 'everyone' apply to this, or first form the predicate 'everyone admires' which then 'someone' applies to. Of course 'everyone' must operate on $\sigma(1)$ while 'someone' operates on $\sigma(2)$. In English some of this is indicated by word-order, but other languages use case-endings and at the underlying logical level we need a way of indicating this explicitly. This is where the second term of the $\langle w, u, \sigma \rangle$ triple comes in. It marks the 'evaluation individual', which indicates what is currently being abstracted on. We require a family of abstraction operators. For each n, Abs_n is a one-place sentential functor with the following semantics:

(12) $V(Abs_n)$ is the function ω such that for $a \in P$, and any $\langle w, u, \sigma \rangle$ triple, $\langle w, u, \sigma \rangle \in \omega(a)$ iff $\langle w, u, \sigma(u/n) \rangle \in a$, where $\sigma(u/n)$ is just like σ except that $\sigma(n) = u$.

I hope you will agree that *now* the complexity is forthcoming. The idea is this. When thinking of the predicate *admires* from the point of view of its first argument we are thinking of it as Abs_1*admires*, for that is true of a given u relative to w and σ, iff u admires $\sigma(2)$, because $\langle w, u, \sigma\rangle \in V(Abs_1 admires)$ iff $\langle w, u, \sigma(u/1)\rangle \in V(admires)$, iff u admires $\sigma(2)$ in w. And $\langle w, u, \sigma\rangle \in V(Abs_2 admires)$ iff $\sigma(1)$ admires u in w. We can express the semantics of *everyone* and *someone* as

(13) $V(everyone)$ is the function ω such that for $a \in P$, $\langle w, u, \sigma\rangle \in \omega(a)$ iff for every person $v \in D$, $\langle w, v, \sigma\rangle \in a$.

(13a) $V(someone)$ is the function ω such that for $a \in P$, $\langle w, u, \sigma\rangle \in \omega(a)$ iff there exists a person $v \in D$ such that $\langle w, v, \sigma\rangle \in a$.

It should be clear that nothing in the semantical *framework* for $\mathcal{L}$ prohibits 'quantifiers' of any number of places operating with virtually no constraints. Thus, if 'most' means 'more than half' then

(14) $\langle w, u, \sigma\rangle \in V(most)$ (a, b) iff there are more $v \in D$ such that $\langle w, v, \sigma\rangle \in a$ and $\langle w, v, \sigma\rangle \in b$ than there are $v \in D$ such that $\langle w, v, \sigma\rangle \in a$ and $\langle w, v, \sigma\rangle \notin b$.

What makes us call these 'quantifiers' is simply that they 'abstract' on the second term in the triple. For any $u, v, \in D$ and $a \in P, \langle w, u, \sigma\rangle \in V(everyone)(a)$ iff $\langle w, v, \sigma\rangle \in V(everyone)(a)$. This is analogous to the way a standard quantifier binds its variable. We'll now apply this to (11) showing how to get all four possible meanings:

(15) $\forall x \exists y\, x$ *admires* y

(16) $\exists y \forall x\, x$ *admires* y

(17) $\forall y \exists x\, x$ *admires* y

(18) $\exists x \forall y\, x$ *admires* y

I don't claim that these are equally natural readings of (11) (though I *did* claim on p. 91f of Cresswell 1973 that they are all *possible*) but they are all things we might want to express. In $\mathcal{L}$ they can be expressed as, respectively,

(19) *everyone* Abs_1 *someone* Abs_2 *admires*

(20) *someone* Abs_2 *everyone* Abs_1 *admires*

(21) *everyone* Abs_2 *someone* Abs_1 *admires*

(22) *someone* Abs_1 *everyone* Abs_2 *admires*

If you think of Abs_1 as a nominative case marker and Abs_2 as an accusative case marker you can see that (19)–(22) exhibit a natural structure for inflected languages. I will go through (19).

One of the features of $\mathcal{L}$ is that its wff represent not only those expressions which are complete sentences in the surface language but other expressions like nouns and verbs as well. Indeed, as I argued on p. 226 of Cresswell 1990 the syntactic freedom offered by $\mathcal{L}$ as an underlying logical language frees natural language syntax from heavy *semantic* constraints and enables it to develop autonomously to a much larger extent than truth-conditional semantics has in the past permitted. Where ϕ, as a wff of $\mathcal{L}$, represents a context-independent sentence — and I will assume that (11) depends on times and worlds but not on other contextual features, even though quantifiers are frequently restricted by context — then whether or not a particular $\langle w, u, \sigma\rangle \in V(19)$ should depend only on w and not on u or σ. And the result we want is that $\langle w, u, \sigma\rangle \in V(19)$ iff in world/time w everyone admires someone.

$\langle w, u, \sigma\rangle \in V(19)$ iff for every person $v \in$ D:

$\langle w, v, \sigma\rangle \in V(Abs_1$ *someone* Abs_2 *admires* $)$

iff

$\langle w, v, \sigma(v/1)\rangle \in V($ *someone* Abs_2 *admires*$)$

iff there exists a person $s \in D$ such that

$\langle w, s, \sigma(v/1)\rangle \in V(Abs_2$ *admires*$)$

iff

$\langle w, s, \sigma(v/1, s/2)\rangle \in V($*admires*$)$

iff v admires s in w. And this is the meaning we require. In $\mathcal{L}$ proper names also are one-place functors, so that e.g., $\langle w, u, \sigma\rangle \in V(Arabella)(a)$ iff $\langle w, Arabella, \sigma\rangle \in a$.

All this has been by way of preamble to the problem of dealing with nouns like *mother* and *enemy*. I will begin by looking at an analysis of (8), in the sense in which it means that every soldier faced an enemy of that soldier, in a language with generalized quantifiers taking as arguments

complex predicates obtained by λ-abstraction. In such a language *soldier* will be a one-place predicate and *enemy* a two-place predicate. Then I will shew how to express this meaning in an indexical language $\mathcal{L}$ in which there is no syntactic difference between these two words — both will be sentence symbols — but the relativity appears in the semantics. In a language with λ-abstraction (7) becomes

(23) *every* $(\lambda x$ *soldier* $x)(\lambda x$ *an* $(\lambda z$ *enemy* $zx)(\lambda z$ *faced* $xz))$

In (23) *every* and *an* are both interpreted as two-place quantifiers *every* as universal and *an* as existential. I realize that in the case of *an* this is almost certainly not completely faithful to an adequate semantic treatment. (See the discussion in chapter 10 of Cresswell 1988 of work by Hans Kamp 1983, and Irene Heim, 1983.) I will also not make any reference to the proper analysis of the past tense of *faced* and will pretend that the sentence is simply true or false in any given possible world. The interpretation of (23) is standard. (See for instance Cresswell 1973 pp. 135–139 or chapter 13 of Cresswell 1990.)

In order to see what is going on in the interpretation of (23) think of the phrase

(24) $(\lambda x$ *an* $(\lambda z$ *enemy* $zx)(\lambda z$ *faced* $xz))$

as obtained by the combined operations of abstraction and identification of x and y in the expression

(25) *an* $(\lambda z$ *enemy* $zx)(\lambda z$ *faced* $yz)$

In $\mathcal{L}$ this means using a pair of abstraction operators and (24) becomes

(26) $Abs_1 Abs_2$ *an* $(Abs_1$ *enemy* $)(Abs_2$ *faced* $)$

(23) then becomes

(27) *every* $(Abs_1$ *soldier* $)(Abs_1 Abs_2$ *an* $(Abs_1$ *enemy* $)(Abs_2$ *faced* $))$

I have inserted brackets into (27) to make its structure clear but if you remember that in $\mathcal{L}$ *every* and *an* are two-place functors, Abs_1 and Abs_2 one-place functors, and *soldier* and *faced* sentence symbols, the bracketing is not in fact required. The meanings of Abs_1 and Abs_2 are given in (12), the meaning of *faced* (ignoring its tense) is, mutatis mutandis, just as

for *admires* described above. *every* and *an* would be the appropriate generalizations to the two-place case of *everyone* and *someone*:

(28) $\langle w, u, \sigma \rangle \in V, (every)(a, b)$ iff for every v such that$\langle w, v, \sigma \rangle \in a$, $\langle w, v, \sigma \rangle \in b$

(29) $\langle w, u, \sigma \rangle \in V\,(an)(a, b)$ iff there is at least one v such that $\langle w, v, \sigma \rangle \in a$ and $\langle w, v, \sigma \rangle \in b$.

For *soldier* we have, assuming that it is not a relative noun,

(30) $\langle w, u, \sigma \rangle \in V(soldier)$ iff $\sigma(1)$ is a soldier in w.

Notice that in this case it would have been simpler just to have $\langle w, u, \sigma \rangle \in V(soldier)$ iff u is a soldier in w. Then we need not have had Abs_1 *soldier* in (25) but could have used only *soldier*. This is connected with the fact that in (23) (λx *soldier* x) is equivalent to *soldier*. However, this simplification is restricted to those simple sentence symbols which are semantically one-place predicates. It would thus be unavailable for relational nouns like *enemy*, to say nothing of more complex predicate expressions, whether realized as nouns or as verb phrases. Since there seems to be no *syntactic* marking of relational nouns it is better to assume that Abs_1 would always be used and that the first argument of even non-relational nouns be $\sigma(1)$.

The final thing to do for the evaluation of (27) is to treat *enemy*

(31) $\langle w, u, \sigma \rangle \in V(enemy)$ iff $\sigma(1)$ is an enemy of $\sigma(2)$ in w.

Now to (27):

$\langle w, u, \sigma \rangle \in V(27)$ iff for every $v \in D$ such that $\langle w, v, \sigma \rangle \in V(Abs_1\ soldier)$, $\langle w, v, \sigma \rangle \in V(26)$

And $\langle w, v, \sigma \rangle \in V(Abs_1\ soldier)$ iff $\langle w, v, \sigma[v/1] \rangle \in V(soldier)$, iff v is a soldier in w. So all we have left to shew is that $\langle w, v, \sigma \rangle \in V(26)$ iff in w, v faced someone who is an enemy of v in w.

Now by two applications of $V(Abs_n)$, $\langle w, v, \sigma \rangle \in V(26)$ iff

(32) $\langle w, v, \sigma[v/1, v/2] \rangle \in V(an\ (Abs_1\ enemy)\,(Abs_2\ faced))$

iff, by $V(an)$ there exists $s \in D$ such that

(33) $\langle w, s, \sigma[v/1, v/2]\rangle \in V(Abs_1\ enemy)$

and

(34) $\langle w, s, \sigma[v/1, v/2]\rangle \in V(Abs_2\ faced)$

Now (33) holds iff $\langle w, s, (\sigma[v/1, v/2])[s/1]\rangle \in V(enemy)$. But $(\sigma[v/1, v/2])[s/1] = \sigma[s/1, v/2]$ and so (33) holds iff s is an enemy of v in w. And (34) holds iff $\langle w, s, (\sigma[v/1, v/2])[s/2]\rangle \in V(faced)$. But $(\sigma[v/1, v/2])[s/2] = \sigma[v/1, s/2]$, and so (34) holds iff v faced s in w. So (32) holds iff there exists $s \in D$ such that s is an enemy of v in w and v faced s in w. Which is to say that $\langle w, v, \sigma\rangle \in V(26)$ iff in w, v faced an s which is one of v's enemies, which is just what we require.

What should be said about cases like (6) in which the second argument of *enemy* is supplied by the context? Here too, as argued in chapter 16 of Cresswell 1990, the indexical treatment has a pay-off, for one can simply treat the second argument as a contextual index. The difference between this treatment and more traditional ones is that what it is an index *for* is now supplied by the meaning of the word in question. This meaning is captured by taking (27) and dropping the occurence of Abs_2 which follows Abs_1. For then the second index of *enemy* would not be picked up by any operator and would remain free as a value to be picked up by the context.

The next task is to look at the role of prepositions like *of* in (2). Again it will be convenient to see how this would be expressed in a language in which *mother* is explicitly a two-place predicate and then look at how to treat *of* as a modifier in an indexical language. In chapter 4 of Cresswell 1985 I shewed how to treat spatial prepositions as predicate modifiers and in chapter 14 of Cresswell 1990 how they become sentential functors in a propositional language. In the formalization of (2), assuming that *mother* is just a two-place noun, the role of *of* is rather simple. In the predicate phrase *mother of Dan* we want *Dan* to apply to the second argument of *mother*. This means that *of* is simply Abs_2. For then, assuming that

(35) $\langle w, u, \sigma\rangle \in V$ (*mother*) iff $\sigma(1)$ is the mother of $\sigma(2)$ in w,

(36) $\langle w, u, \sigma\rangle \in V(Abs_2\ mother)$

iff $\sigma(1)$ is the mother of u. Let

(37) $\langle w, u, \sigma\rangle \in V(Dan)(a)$
iff $\langle w, Dan, \sigma\rangle \in a$

and we have

(38) $\langle w, u, \sigma \rangle \in V(\textit{mother of Dan})$

iff $\langle w, \text{Dan}, \sigma(\text{Dan}/2) \rangle \in V(\textit{mother})$

iff $\sigma(1)$ is Dan's mother in w.

What now about (3)? (3) means, I think, that Arabella is someone's mother. To get this meaning we must insert a 'default' existential quantifier. Such quantifiers are common in many areas though there is dispute about whether they should be handled by a special symbol at the level of logical form or by building the existential quantification into the interpretation of the structure. I have a preference for the former since in the case of a noun like *enemy* the default situation is that the context should supply the argument; and in some cases the same noun can be interpreted in both ways. If we write the default quantifier as $\exists_n$ then we have

(39) $\langle w, u, \sigma \rangle \in V(\exists_n)(a)$ iff there exists $v \in D$ such that $\langle w, u, \sigma(v/n) \rangle \in a$

The one-place predicate 'mother' would then be represented as $\exists_2$ *mother* and

(40) $\langle w, u, \sigma \rangle \in V(\exists_2)(V(\textit{mother}))$

iff there exists $v \in D$ such that $\langle w, u, \sigma(v/2) \rangle \in V(\textit{mother})$, iff there exists a v such that $\sigma(1)$ is v's mother in w.

To interpret (3) then, all that is required is to give a semantics for *is* and combine it with the semantics already given for names like *Arabella* and the quantifier a. In Cresswell 1973 I argued that the 'is' in sentences like (3) is identity. In $\mathcal{L}$ this would mean

(41) $\langle w, u, \sigma \rangle \in V(\textit{is})$ iff $\sigma(1) = \sigma(2)$.

To see how (3) works it is perhaps best to express it first in a language with λ-abstraction

(42) $\textit{Arabella}\ (\lambda x(a(\lambda x \exists_2 \textit{mother}\ x)(\lambda y\, x\ \textit{is}\ y)))$

This means that Arabella is an x such that there exists an x who is a mother and is also a y who is (the first) x. In $\mathcal{L}$ this becomes

(43) *Arabella* $Abs_1 a Abs_1 \exists_2$ *mother* Abs_2 *is.*

Notice that the $\exists_n$ quantifiers apply directly to the terms of σ and not to the evaluation individual. This is to prevent the possibility of ever getting a reading in which they could take anything but narrowest possible scope and will prevent a sentence like

Every mother gets tired

ever meaning that there exists someone such that every one of their mothers gets tired. In (3) it might seem that the existential quantification is done by the *an*, in that

(44) Is the mother here?

may be uttered in a situation in which there is a contextually specified child whose mother is sought. However, this seems merely to reflect that *mother* is functional. In a sentence like

(45) Is Alex a registered student?

uttered at the University of Massachusetts it will hardly justify an affirmative answer to be told that Alex in enrolled at the Victoria University of Wellington.

So it seems that we can get the right truth conditions both in the unmodified cases like (3) and in the modified cases like (2) when analysed as (38). However, we still have a problem. For we have moved from a situation with too much syntactic information to one which may seem to have too little. After all if *mother* were a transitive verb Abs_2 would still be necessary to indicate that *Dan* is its object, for exactly the same reason as it is necessary in (19)–(22).

At this point I should say something about what it is that I hope to achieve. One of the important debates in recent years is the extent to which the logical demands of a truth-conditional semantics should constrain the syntax. If this debate is to make progress it would seem desirable to be able to put a lower limit on semantic constraints. In other words one should attempt to set out a language whose *syntactic* demands are as small as possible. Any compositional semantics will of course demand *some* syntax, and if we are to end up with truth conditions, say in the form of a class of possible worlds, it would seem that this process would have to be performed by functional application. In this sense $\mathcal{L}$ does seem to contain the minimal syntax that semantics demands. Given this syntax

the aim would then be to build it up into something more complex but *only* when this can be done using syntactical arguments from natural language. (Such arguments may of course include evidence that particular syntactic analyses depend on particular semantic facts, but these would need to be established case by case.)

I will shew how the literal spatial uses of prepositions as studied in chapter 4 of Cresswell 1985 also emerge as one-place functors in $\mathcal{L}$. Look at the phrase

(46) *house at Pooh Corner*

as it occurs in the sentence

(47) Christopher Robin and Piglet built the house at Pooh Corner.

(There is of course a sense in which *at Pooh Corner* is a sentence modifier saying where the building took place. That is not the sense I have in mind.)

We want $\langle w, u, \sigma(1)\rangle \in V(46)$ iff $\sigma(1)$ is a house and is at Pooh Corner, where we may assume that Pooh Corner is a particular spatial region. The meaning of *at* may not be completely precise. Must the house be completely contained in the region which is Pooh Corner, or need it merely overlap? Suppose the latter. Then *at* will be a one-place sentential modifier with the meanings

(48) $\langle w, u, \sigma\rangle \in V(at)(a)$ iff u is a place which overlaps $\sigma(1)$ in w and $\langle w, u, \sigma\rangle \in a$.

For *Pooh Corner* we have

(49) $\langle w, u, \sigma\rangle \in V(\textit{Pooh Corner})(a)$ iff $\langle w, \text{Pooh Corner}, \sigma\rangle \in a$.

(46) has the structure

(50) *Pooh Corner at house*

$\langle w, u, \sigma\rangle \in V(50)$ iff

(51) $\langle w, \text{Pooh Corner}, \sigma\rangle \in V(\textit{at house})$ iff, by (48),

(52) Pooh Corner is a place which overlaps $\sigma(1)$ in w and $\sigma(1)$ is a house in w.

There is certainly a difference between *at* and *of*, especially when *of* is written as Abs_2 since Abs_2 occurs in modifying sentences realized as verbs, and then there is no symbol for *of* in $\mathcal{L}$. But another way of looking at *of* is to say that it is a different symbol from Abs_2 but happens to have a semantics which turns out to be the same as Abs_2. Thus

(53) $\langle w, u, \sigma \rangle \in V(of)(a)$ iff $\langle w, u, \sigma(u/2) \rangle \in a$.

If we do this we can look at what *of* and *at* have in common. In both cases they form an expression in which the evaluation index u is related in some way or another with one of the terms of σ. For *of* it is identity with $\sigma(2)$. For *at* it is spatial overlap with $\sigma(1)$. The semantic difference is just that while *at* refers to a substantive relation *of* refers to what is often thought of as a logical relation. This seems to me just right. *of* is, if you like, the same kind of thing as *at* except for having a degenerate meaning. Indeed one of the problems in the analysis of prepositions has been how to put together these apparently different roles. The present account, by making it a matter of content rather than form is able to preserve syntactic uniformity, while at the same time giving a plausible explanation of how what is really a difference in content has appeared to be a difference in form. In a sequel I hope to take up the question of how $\mathcal{L}$ might be adapted to take account of natural language syntax. In this paper I have tried to shew you how little syntax one needs to take care of the semantics of relational nouns.

REFERENCES

CRESSWELL, M. J., 1973, *Logics and Language*, London, Methuen.

———————, 1985, *Adverbial Modification*, Dordrecht, Reidel.

———————, 1988, *Semantical Essays*, Dordrecht, Kluwer.

———————, 1990, *Entities and Indices*, Dordrecht,Kluwer.

HEIM, I., 1983, File change semantics and the familiarity theory of definiteness, *Meaning, Use and Interpretation of Language*, (ed. R. Bäuerle *et al.*), Berlin, de Gruyter, pp. 164–189.

KAMP, J. A. W., 1983, A theory of truth and semantic representation., *Formal Methods in the Study of Grammar*, (ed.) J. A. G. Groenendijk *et al.*), Amsterdam, Mathematische Centrum, pp. 277–322.

KUHN, S. T., 1980, Quantifiers as modal operators, *Studia Logica*, Vol 39, pp. 145–158.

MONTAGUE, R. M., 1974, *Formal Philosophy*, New Haven, Yale University Press.

PARTEE, B. H., 1984, Compositionality, *Varieties of Formal Semantics*, (ed. F. Landmann and F. Veltman) Dordrecht, Foris Publications, pp. 281–311.

PRIOR, A. N., 1956, *Formal Logic*, Oxford, Clarendon Press.

QUINE, W. V. O., 1960a, Variables explained away, *Selected Logic Papers*, New York, Random House, 1966, pp. 227–235.

———————, 1960b, *Word and Object*, Cambridge, Mass, MIT Press.

TAYLOR, B., 1977, Tense and continuity, *Linguistics and Philosophy*, Vol 1, pp. 199–220.

Logic, Methodology and Philosophy of Science IX
D. Prawitz, B. Skyrms and D. Westerståhl (Editors)

REDUCING SELF-INTEREST AND IMPROVING THE RELEVANCE OF ECONOMIC RESEARCH

CLIVE W.J. GRANGER

University of California, San Diego, USA

1.

> "Political economy is the science which investigates the laws of production, the distribution, and the exchange of wealth, so far as these laws depend upon the human mind."
>
> Henry Fawcett
> Manual of Political Economy (2nd Ed.),
> MacMillan, London (1865), p. 54.

I will start with a simplistic, naive viewpoint. I will take a *science* to consist of the accumulated knowledge of the researchers, or scientists, working in the area. The *philosophy of science* is thus essentially the philosophy of these scientists, that is their approach to the field, their choice of topics and their strategies for research. I will take the *main objective* of these scientists to be to learn about some specific part of the real world. Thus, physicists and chemists will try to understand the real physical world, biologists the biological world and so forth. Exceptions to this type of objective might be pure mathematicians and some philosophers who may try to understand only abstract objectives. However, the objectives of individual workers may be different from the whole group. I will assume the following objectives:

(a) for the scientific field (the *discipline* or *main objective*): to study and attempt to understand the actual economy (the phrase "real economy" sometimes has a different, specific meaning in the economic literature and so will be avoided)

(b) for an individual researcher (the *individual objective*): to maximise his or her personal utility, reflected through income, self-satisfaction, reputation; and

(c) for a particular piece of research (*research objective*): to influence the beliefs — and thus probably the behavior — of other workers

and, in economics, of economic agents. Thus a new theorem may influence the research approach taken by others, a new econometric technique is used in applied work or a new empirical result may change the *degree of belief* of an economic agent, measured as a probability, about some statement, such as "a change in money supply affects prices." I will propose later that there are *producers* of results and there are *consumers* of these results and that this division has important implications.

I will interpret the main objective as occurring because of the economist's desire to find solutions to important problems in society relating to the actual economy. Thus, economists should perhaps be classified as "pragmatists," together with medical doctors and engineers. Although we want to know the truth about the economy, we will also be happy if a good enough approximation to the truth can be found that will be useful in alleviating a real world problem. "Truth" can be thought of as a final product, a useful approximation to it is an intermediate product.

The three objectives are related but are not necessarily in agreement. A theme of this paper is to consider the implications, causes and cures of this disharmony, which I believe is weakening the discipline of economist's attention towards the main objective. This leads to inefficiencies in the research effort and consequently a lack of respect for economic research by scientists in other fields, by politicians, the media and the public at large. This may be reflected in the shortage of research funds for economics from the National Science Foundation in the U.S., and for the proposed reductions in the number of academic economists in Britain, for example.

It is worth pointing out here a sharp distinction between the physical and the behavioral or decision sciences, the latter including economics, psychology and sociology. In the behavioral sciences research output can influence and change the behavior of the objects being studied — finding the relationships between smoking and health provide an example — but this cannot occur in the physical sciences. In economics examples are optimal ways to select inventory levels or of reducing investment risk by the use of portfolios. Philosophers of science often use analogies from the physical sciences to make points about economics, for example, but these analogies are often not appropriate because of the adaptability of the economic agents.

2. Economics deals with important questions, the suggested answers to which can affect the quality of life of many millions of people. For example:

(a) what will be the short-run and long-run impacts of changes in the

size of the minimum wage? Should such a minimum exist?

(b) what criteria should a rich, developed country use when deciding which developing countries should receive financial aid, and the extent of that aid?

(c) should a large country use restraint of trade to protect its own industries? How effective is such restraint in achieving a political objective, such as persuading a foreign leader to change a policy?

(d) to what extent should taxes be used to flatten the income distribution, by the use of transfer payments?

(e) how should the savings and loan industry in the United States be regulated and re-structured?

(f) should a developing country, with very limited resources, be concentrating on improving the health and educational levels of its population or on attracting new industrial development by providing a better transportation system?

(g) what is the optimum population size of a country or region? Should immigration be encouraged or not? Should selection criteria be applied to decide who can migrate or not?

There are many other examples. In most cases the economist has a potential impact by giving advice and (conditional) forecasts to politicians, government agencies, company executives, and those of banks and other institutions, such as the International Monetary Fund and the World Bank.

There are many difficulties with the interaction between economists and decision makers, some of which are discussed in Friedman (1986), who suggests as a secondary disciplinary objective "to influence policy." This important topic is not discussed here.

Because economics deals with important questions it is important to society that it is a strong discipline. It needs to be encouraged and nurtured but also to be censured if it acts in ways that are contrary to the common good.

3. There are many types of economists. To illustrate this and to indicate the numbers in different categories the membership of the American Economic Association (AEA) was self-classified by major interest into ten broad categories. Table 1 shows the approximate number of members in each group in 1989, the percentage of the total for each category in 1974 and in 1989, and the change in this percentage. Using the standard notation of K = thousand, the total membership in 1974 was 18.7K, in 1988 it was 20.6K, giving a fifteen year growth of 10%, compared with a U.S. population growth of 15% over this period. This comparison is not completely relevant as not all U.S. economists belong to the AEA and

not all members of the AEA live in the U.S. The table shows a decline in the percentage of members in some of the applied subject areas (100, 800, 900), substantial increases in the statistical and applied micro areas (200, 600) and small changes elsewhere.

TABLE 1

AEA MEMBERSHIP BY SUBJECT CLASSIFICATION

Classification		Number of Members 1989	% Total 1989	% Total 1974	Change 89-74
000	General Economics (Theory, History)	3,460	17.2	17.0	0.2
100	Growth, Development (Planning, Fluctuations, Inflation)	2,200	11.0	12.5	−1.5
200	Economic Statistics (Econometrics)	1,875	9.3	7.6	1.7
300	Money, Fiscal Theory	3,050	15.2	14.5	0.7
400	International, Trade	1,970	9.8	9.3	0.5
500	Administration, Finance (Marketing, Accounting)	1,590	7.9	7.2	0.7
600	Industrial Organization, Technology Change	2,105	10.5	8.0	2.5
700	Agricultural Economics, Natural Resources	1,030	5.1	5.4	−0.3
800	Manpower, Labor, Population	1,620	8.0	10.1	−2.1
900	Welfare, Consumer Economics, Urban, Regional Economics	1,200	6.0	8.2	−2.2

Figures from AE Review, Directory of Members, 1974, 1989.
[All figures approximate.]

Two subclassifications are of interest for the arguments of later sections. In 1989, about 43.5% of classification 000 gave "economic theory" as their main interest, which equates to about 1,500 individuals, whereas

69% of classification 200 identified themselves as econometricians, that is about 1,300 individuals. Thus, in this crude and oversimplified analysis, the number of "economic theorists" and of "econometricians" were about equal in the membership of the AEA in 1989, with these together making up about 14% of the full membership.

It seems that philosophers of science often feel the need to categorize groups. In this case "theorists" can be called "rationalists," (reason is the source of knowledge) and econometricians "empiricists" (experience is the source of knowledge). Virtually all applied economists are also "empiricists" as they use data to build, extend and use models. These are, of course, just generalizations, not quite every theorist is just a rationalist, for example. Most economists are closer to the philosophy of Hume, Locke and Berkeley than that of Descartes, Spinoza and Leibniz.

The fact that there are many types of economists and that different basic approaches are used — both non-empirical and empirical, say — would be viewed by an economist as potentially an advantage if it is understood that research returns are higher if a diversification of approaches are used, as has been shown from portfolio theory. However, this diversification often does not occur in a single piece of research, as will be seen. A piece of theory is usually just rationalist, a piece of applied economics attempts to be sequentially rationalist — starting with some theory — and then empirical — using data. The bridge between these components is little discussed, except in Stigum (1990).

Economics is a fairly active field. In 1989 the Journal of Economic Literature (JEL) listed about 10K research papers published in English. Of the papers, about 10% were classified as "economic theory" and 5% as econometrics (plus forecasting). [The classifications used in the AEA survey and in the JEL are not necessarily the same.] The JEL in 1989 lists roughly 3,200 names as authors or coauthors of papers and 1,600 names for books, suggesting that over four thousand economists published in that year.

4.

> "Political Economy consists of two parts – theory and practice; the science and the art."
>
> Mrs. Marcet
> Conversations on Political Economy (7^{th} Ed.)
> Longmans, London, (1839) page 87.

For the purpose of case of argument I will assume that there are basically three types of economic research: theory, econometrics and applied. Thus, it will be assumed that any research paper can be classified into one of

these groups. Any individual researcher may produce papers in various groups, of course. The first two classifications will each be divided into two, giving in:

(a) pure theory (denoted PM theory because of its similarities with pure mathematics), in which variables are given economic labels, and axioms and assumptions are converted into results with no empirical implications by the use of mathematical reasoning.
(b) applicable theory (denoted AM theory), as above but the results do have empirically testable implications and thus may be directly helpful to applied research.
(c) econometric theory (denoted EM theory) which uses mathematical statistics to obtain results about econometric procedures, such as consistency and asymptotic normality of an estimate or the power or relative efficiency of a test.
(d) applicable econometrics (denoted AEM) which suggests applicable procedures, such as estimation techniques for special circumstances, specification searches or tests for causality.
(e) applied economics, which is clearly the largest group and tackles all the real problems involving the actual economy.

The groups may be thought of as having different types of product. The theorists will be said to produce "theorems" as a final product and "theoretical models" as an intermediate product. The theorems will characterize properties of the variables in the models. Theoretical econometricians will also have "theorems" as their final products, which will characterize properties of the "econometric procedures" which are the final product of applicable econometric research. Finally, applied economic research will have "applied models" as an intermediate product and a variety of final products, such as forecasts, policy suggestions, estimates of important parameters such as elasticities and tests of theories. As with all the generalisations used in this paper, this is a vast oversimplification of actuality but it is convenient for what follows. Workers in the (b) and (d) groups will be thought of as providing the new tools to be used by the last group. The PM theorists may well provide useful ways of thinking about questions that are helpful to the AM theorists. The EM theorists will help others appreciate the quality and properties of the econometric procedures. If all of the groups were numerically balanced, were working in unison and in mutually helpful fashions, the main objective of the discipline would be easier to achieve. It is easy to suggest that this is not occurring. However, before discussing these problems it is necessary to mention two other groups of workers. These are the economic, governmental statisticians and non-research economics. This latter group is

probably numerically the largest and consists of economists working in industry, government, institutions such as the World Bank and the IMF and in consulting companies. It is unfair to call their work "non-research", but the vast majority of this work is not published nor publicly available. I think of this work as tackling actual world questions using the products of the applied economic researchers.

There are concerns about the *balance* of published research work in economics. Morgan (1988) gives some numbers on the percentage of papers published in two main-stream economic journals (the American Economic Review (AER) and the Economic Journal (EJ) from Britain) and in similar journals in four other disciplines for period 1982-86 in three categories:

		Economics		Pol.			
		AER	EJ	Sci.	Soc.	Chem.	Phys.
(a)	Mathematical models without data	42	52	18	1	0	12
(b)	Empirical analysis using public data	44	40	51	74	17	41
(c)	Empirical analysis based on experiments and simulations.	6	2	6	3	83	48

[From Morgan (1988), some small categories are not reported.]

Category (a) corresponds to theory and (b) to standard applied work. The balance seems wrong. To use an analogy from section 6, the members of the intelligence corps spend too much time talking to each other and not enough time talking to the army.

5.

> "If anything has become a received idea in recent philosophy of science, it is the thesis that there is no sharp distinction in science between observation and theory; ..."
>
> A. O'Hear
> Introduction to the Philosophy of Science
> Oxford, 1989

The vast majority of data used in economic research is supplied by official statisticians in the form of time series, panels or cross sectional data. There is some discussion with potential users of this data, but what to collect, what to provide and what transformations to use (such as seasonal adjustment) are largely determined by these statisticians. In economics, it is generally true that the data gatherers and the data users

(for research purposes) are separate groups. This may not be true of non research economists. In the physical sciences, where applied researchers generate their own data by the use of experiments, or use casually gathered information, the relevance of the quotation at the top of this section is plausible. Theory is used to decide what experiment to run. In general this is not true in economics and so I believe that most economists would not accept the quotation as being relevant to their own discipline.

This is not the place to discuss the quality of economic data. However, it is worth pointing out that many potentially important economic variables are not gathered or made publicly available. Some of these missing variables are because of the high cost of gathering and manipulating the data, an example is regional consumption figures. Other variables are very difficult to define and measure, such as the "complexity" of an economy, "tastes", technology changes and individual utilities. It has to be accepted that in any piece of applied economic research there will be missing variables, for which no acceptable proxies exist. This will always limit the quality of the applied model.

6.

> "What is lacking (in economics) is an effective means of communication between abstract theory and concrete application."
>
> Barbara Wootton
> Lament for Economics
> 1938, p. 64.

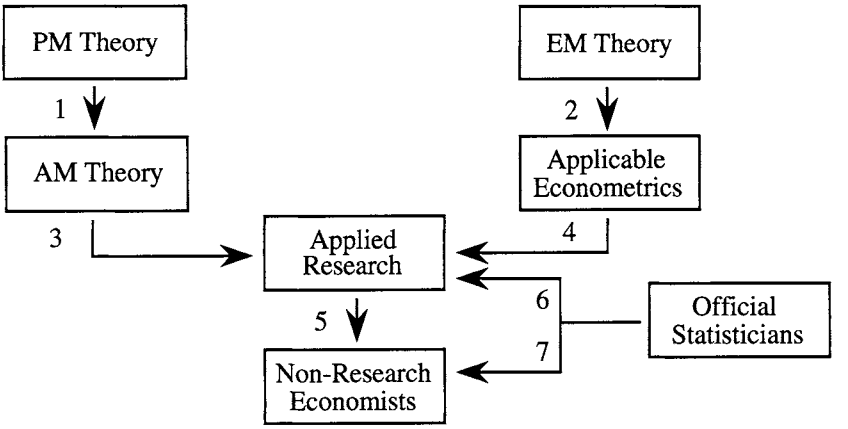


Figure 1. Major Links Between Groups of Economists

The figure shows what I believe to be the major links, or communication channels, between the groups of economics and the official statisticians. It should be noted that some of the links are weak or non existent.

To appreciate the importance of these links, and of the gaps, one can start with the main objective and consider the research strategy employed by an applied economist. Suppose that he or she (henceforth "he" for convenience and with no implications) starts with a well defined actual world problem to be tackled. The traditional research strategy is to find and use a relevant theory to suggest what variables are important, a model specification and simplifying constraints on the model (such as variables to omit or at least expected signs on parameter values).

The data is then gathered and econometric techniques used to estimate parameters and finally the applied worker interprets the results. It can be said bluntly that this research strategy often does not work. The reason is that the econometrician has told the applied worker to evaluate the applied model, and when this is done, this model is usually rejected. The importance of evaluation will be a central theme in this paper. It should be noted that the economic theorists have little to say about the evaluation of applied models as it is usually assumed that the theory model is correct. Econometricians have a different stand, believing that most — or perhaps all — models are incorrect but that improved versions can be achieved by a careful evaluation. This process gives completely different research strategies than the traditional one, some start with a theory model and then evolve away from it as alternatives are considered, whereas other strategies start with pre testing of the data and the pure use of statistical, non theory based models.

This paper will concentrate on two important gaps in the figure;

Gap 1: that between theorists and econometricians, which may lead to advice to applied economists that is difficult to reconcile or may be conflicting, and

Gap 2: that between theorists, but particularly PM theorists, and applied workers.

I would contend that the econometricians are more aware of the problems encountered by applied workers and that they react to these problems. As evidence I would cite the many special estimation procedures that have been developed to deal with particular actual world situations.

An analogy that is graphically useful but cannot be taken too far is to compare the organisation of economic research with the mammalian brain. Suppose that the brain can be considered as having two main parts, the core section which deals with all muscular and nerve activity and which assimilates most of the sensory information reaching the body,

and the outer cerebral cortex, or "little grey cells," which are responsible for thinking and the analysis of the more abstract information reaching the brain. The cerebral cortex is separated into two distinct sections, linked by a nerve bundle called the corpus callosum, and each section has distinct duties. In the analogy, the brain core corresponds to applied economic research, and the two sections of the cortex to theorists and econometricians. In a real brain, if the corpus callosum is cut, the body receives confused instructions, its actions becomes disjointed and inconsistent. One can decide to pick a certain coat from a closet but cannot work out how to move an arm to get it, or an arm can go to the closet but no decision made about which coat to select. In many ways the discipline of economics behaves as though the corpus callosum is inoperative. The theorists and the econometricians do not interact sufficiently in their research to give clear signals to the applied economists about what model specifications and what techniques to use. The existence of this weak link or Gap 1 is particularly puzzling as there exists an organisation, the Econometric Society, with 3,500 members worldwide which includes most of the leading theorists and econometricians.

The second gap is between the theorists, and to a lesser extent also the econometricians, with applied economists. The above quote by Baroness Wootton suggests that this is not a new problem. A different analogy may be helpful. Consider the applied economists as the officers and soldiers of an army engaged in a battle. The theorists make up the intelligence corp. This corp is situated away from the battle, can take an overview of the whole situation, interpret incoming information and pass on relevant digested information and advice to the fighting part of the army. Thus, the members of the corp will spend part of their time talking to each other about the situation and new information and part of the time communicating with the army. Obvious difficulties arise if these two activities get out of balance. If most corp members do nothing but talk to each other, and even appear to forget that the army exists, then the army may not do well in the battle. The equivalence in economic research is when most theory papers are written in a technical, difficult to appreciate style, make no attempt to suggest empirical implications and do not try to convince applied economists that the theorems are relevant or useful. Thus criticism applies to most theory papers but not all, the work on European unemployment by Jaques Dreze and on the theory of demand by Werner Hilderbrand are counterexamples. A similar criticism can be applied to some papers in theoretical econometrics, such as asymptotic theory with no indication of the kind of sample size required to make this theory a useful approximation, or complicated and

uncheckable conditions for a property to hold, such as for the invertibility of a bilinear time series model. The next three sections discuss why these two gaps exist, given the present research strategies of the various groups.

7. A dominant feature of any actual economy is its complexity. It consists of many economic units individuals and families making many decisions about consumption, investments, work patterns, savings and so forth. The decisions are effectively made independently, probably in some maximising mode, but the available choices of one agent will depend on the previous choices of many other agents, leading to externalities and shortages of some goods, such as available space on a commuter road. Interacting with, and limiting the behavior of, these decision makers are many other agents, in corporations, institutions such as universities, and local and federal governments. In a large economy, such as the United States, there may be nearly a hundred million families and two million corporations. Decisions are made frequently affecting production, employment, taxes, prices, wages, savings, factory location and many other variables that are generally accepted to lie within the purview of Economics. For example, the Statistical Abstract of the United States, 1989 takes over 800 pages and well over a thousand tables just to *summarise* the current and recent U.S. economy. It is difficult to measure the complexity of an economy, although it is probably generally accepted that a big economy is more complex than a small one and a current economy is more complex than those in the past. One aspect of complexity is the amount of interaction, or communication, between agents and this is certainly increasing.

A further analogy can be made with a mammalian brain. Both a (large) economy and a (small) brain contain several hundred million units, the agents or the cells, that both act independently yet are linked. There are several different types of units, which have memories and act dynamically.

A further aspect of complexity, which occurs in a discipline such as economics but not in a physical science, is the dynamics of the object being studied. The working of an economy is affected and constrained by its institutions, such as bonus or regulatory agencies, and these evolve through time for a variety of reasons. As a consequence, and also through learning and increases in the efficiency of the use of human capital, the behaviour of economic agents also continually changes what kind of products are bought, investments made or careers sought depend not only on prices, interest rates or wages, but also on accumulated knowledge which can affect personal utility functions. It follows that economic theories also evolve, not only by changing existing theories or through the empirical

evaluation process, but also because new theory is required to investigate new situations. It follows that the philosopher of social science concerned with economics is criticizing a moving target. There is little impact of criticism of economic theories, or even methodology, of thirty, twenty or even ten years ago, although much writing in the philosophy of science cites work of such antiquity or even earlier.

Because of the immense complexity of an actual economy it is generally accepted that it is necessary to adopt a *simplifying strategy* in order to conduct sensible research on it. This strategy usually consists of the adoption of a set of simplifying assumptions, although these are not always stated explicitly. An example of such a strategy is to study just a part of the macro economy, such as considering the relationship between wages, prices, productivity and employment rate. The simplifying assumption is that such a subsystem can be usefully analysed, whilst ignoring all the other thousands of variables, without any important distortions of the results. Sufficient conditions for this to be true, in terms of the concepts of exogeneity, super exogeneity and non-causality are given in Engle, Hendry and Richards (1983) although these conditions can only be tested in a limited fashion. The use of simplifying assumptions typically imply that the research workers aim is to produce a model or theorem that, at best, only approximates the actual economy. The quality of the approximation cannot be judged just from theoretical considerations or by simulation, but may be judged from a suitable empirical analysis, using data not used in the model. An exact fit should never be expected.

A possible model for the brain is to assume that all N brain cells are identical, to work out what reaction that cell would have to a particular input, multiplying this reaction by N and then to claim that this will be the whole brain's reaction. As far as I know, no such model for the brain exists or would be taken seriously. Cells are not identical, they interact, share common inputs, and information is passed from one cell to another, possibly after some analysis. Any model that ignores all of this activity is likely to be completely inadequate.

However, such a model not only exists in economics — called the representative agent theory — but it forms the basis of many of the dominant theories. Here, economic agents are assumed to be identical and to make their decisions as though they exist in isolation — individual Robinson Crusoe economies. This type of theory is an example of one based on a particular simplifying assumption which produces an approximation to reality. It may be able to provide worthwhile predictions about the effect of an increase of income tax on consumption, say, but can make no prediction about the tax change on the income distribution. Representative

agent theory is a frequently used simplification assumption because it is a way of overcoming the important question of how micro theory can be aggregated into macro theory. Other aspects of the aggregation problem are discussed in section 10.

One simplification that is necessary for applied economists and many econometricians is going from the actual economy to the observed economy. Only some parts of the economy are observable and measured or estimated. The quantity of available data is enormous but there are still many variables which economists would like to see but are not available, such as some regional series and a frequent, complete input-output table.

8. A consequence of the use of a simplifying strategy with a true complex system is that every model (and most theories) can *at best* be good approximations to the truth. It follows that all models are ultimately falsifiable given enough data and computer effort.

A great deal of effort is made to falsify theories by econometricians and applied economists. A good example is the efficient market hypothesis which says that consistent positive profits cannot be made by investment on the stock market or any such speculative market. The theory is easily understood and is obviously sensible — if it were clearly not true, the stock market would be a money pump and everyone would be rich. Until the late 1970's the accumulated evidence from empirical evaluations was in favour of the hypothesis but more recent work involving more data, stronger computers and new statistical techniques have largely been against the hypothesis. Of course, in a stochastic world no theory can be said to be true or false, but the weight of evidence can be in favour or not. Some economic theories have been shown to be not falsifiable in that they are so flexible that some form of the theory is consistent with virtually any set of data. Rothschild (1990) says that to the question "what restrictions does economic theory (the assumption that rational agents maximize) place on asset prices?" he gives the answer "almost none", which is similar to the answer Sonnenschein gave when he asked the same question about (excess) demand functions.

As Redman (1990) has stated, economists generally do not try to falsify their own theories (or empirical techniques) but because of the strong competition for publications/promotions/career improvement it is very common for other economists to criticize and attack successful theories and techniques of others. I would claim that it is this continual discussion and evaluation of both new and old results that give the discipline of economics whatever strengths it has.

9.

"In the year 1854 the (British) Chancellor of the Exchequer confidently affirmed that the income tax would be gradually reduced, and would be entirely abolished in the year 1860"

Fawcett
op. cit. p. 532.

"Now everybody will get to work"

Alfred P. Sloan, President
General Motors, Oct. 27, 1929.

"There is no cause for worry. The high tide of prosperity will continue."

Andrew W. Mellon
Secretary of the (U.S.) Treasury
Sept., 1928.

Perfect foresight is a common assumption in economic theory, it being assumed that all economic agents can forecast some parts of the economy without any error into the distant future. It is an example of an approach to simplification that is controversial — the use of "clearly incorrect assumptions" (henceforth CIA). Of course no assumption is "clearly incorrect" to all economists! I will use the term to include any assumption that appears to be untrue to most scientists who have given at least a casual glance at economic data. Representative agent theory is clearly based on CIA's.

A closely related class are the "possibly incorrect assumptions" (PIA's) which are also usually made for convenience or as a simplification. Examples are the assumption of normality by econometricians, the use of a Cobb-Douglas production function in an applied study, the use of rational expectations or an assumption of linearity in a theory or applied model. PIA's can be correct and in many cases are testable by using economic data.

These types of assumptions can be discussed within the contexts of a research strategy and the evaluation of theories or models. For PM theory any assumptions are acceptable and need not be justified. AM theory can use CIA's to explore a new theory but needs to discuss the robustness of the theory to changes in the assumptions to make the results potentially useful for applied work. In fact, AM theory should be judged as being incomplete until this robustness question is addressed. If a small perturbation of the initial assumptions leads to a large change in the implications of the theory, such as a bifurcation, then this theory is unlikely to be useful in practical research. An example of the relaxation of assumptions is to consider first consumer choice theory under the CIA of certainty and then to consider uncertainty, as discussed in Stigum (1990), chapters 10 and 19.

10. One of the main reasons for the gap, or weak link, between theorists and econometricians are that they are likely to use different simplifying assumptions and also to use assumptions that the other group view as CIA's. Examples of assumptions often used by theorists that most econometricians view as clearly incorrect are:

(a) the economy is deterministic, rather than being inherently stochastic;
(b) there are no unobservable variables, so that first order conditions are exactly true, whereas an econometrician would expect an error term to be present;
(c) there is perfect foresight, whereas most of us find forecasting to be difficult and for there to be substantial forecast errors;
(d) if a variable is stochastic then it is stationary, whereas data analysts find many variables to be trending, have "unit-root" long run components and important seasonals; an example is the theory of inventories;
(e) the economy is static, rather than being inherently dynamic.

Assumptions often used by econometricians which may be viewed as CIA's by theorists include:

(f) linearity or very simplistic forms of nonlinearity;
(g) systems are small rather than large; and
(h) no use is made of an underlying optimizing behavior by economic agents when considering model specification.

Of these differences (a) is critical. I believe it is true to say that virtually every applied economist and certainly all econometricians believe that the economy is stochastic rather than deterministic. Further, they believe that economic variables cannot be decomposed into a deterministic part, to which all theory relates, plus an unknowable additive "measurement error." The stochastics come from unexpected shocks or innovations (weather, new technology), reaction to previous shocks and from unobserved, and probably unobservable, variables. There is clear difficulty in communication between a (theorist) who believes — or appears to believe — in a deterministic economy and an applied worker who sees the economy to be modelled as inherently stochastic.

Another important aspect of this gap is that the two groups sometimes appear to be interested in quite different aspects of the economy. A gross generalisation has the theorists being most interested in "equilibrium" and the econometricians in "dynamics", which can be viewed as disequilibrium. The typical theory text, is either micro or macro economies, will contain a great deal of discussion about equilibrium in it's many forms,

whereas most well known econometrics texts make absolutely no mention of the concept, at least according to their indices. The three-volume *Handbook of Econometrics* (edited by Z. Griliches and M. Intriligator, North-Holland, 1984) has no mention of equilibrium in any of the indices. This certainly suggests a major difference of emphasis between the two groups and that econometricians have not developed techniques to test some theories proposed by the theorists, although the use of cointegration goes some way to reducing this problem. It is difficult to define equilibrium in a way that is useful to an applied economist, partly because it is needed for a stochastic system.

A further example of different approaches used by theorists and econometricians is the questions that come from the process of aggregation. Econometricians have concentrated on the effects of aggregating from a group of subeconomies, such as states, to the full economy. Conditions are known where no or little loss of information occurs in some cases. There questions are about going from one set of observable variables, at the state level, to the aggregate observable variables. The theorists are more concerned with aggregation from the micro variables, which are usually unobservable, to macro variables.

11.

> "It is not very constructive to dismiss macroeconomics because it *requires* implausible aggregation assumptions " (emphasis added).
>
> A. Deaton
> New Palgrave Dictionary of Economics
> vol. 1, p. 597

The macroeconomy is the aggregate of all the microeconomies of the individual economic agents and institutions. Some macro variables, such as total consumption, are simple sums of the consumption of the many individual agents. However, relationships do not necessarily aggregate so easily, for example, a non linear relationship between micro variables may become linear, to a close degree of approximation, between the corresponding macro variables. Once more, the theorists and the econometricians have quite different approaches to aggregation. The theorists ask questions about how aggregation can occur so as to preserve at the macro levels a relationship that has been derived at the micro level, leading to the quotation at the top of this section. Of course, the macroeconomy exists and has properties regardless of the theorist's assumptions. The use of micro theory to suggest macro relations is one reason why an understanding of aggregation is so important. Econometricians have been inclined to take a more pragmatic — but stochastic — viewpoint of aggregation. It

can be shown that a strong $(R^2 = 0.99)$ micro relationship can aggregate into a weak $(R^2 = 0.01)$ macro relationship but that a weak micro model $(R^2 = 0.01)$ can lead to a well fitting macro model $(R^2 = 0.99)$. Details can be found in Granger (1987). What matters are the existence, observability and location in the model of common factors, which are factors that enter the micro models of every agent. If these factors are in the residuals of the micro relationships, they can have a profound effect on the model for aggregated variables. This problem is often assumed away by theorists, either taking theories to be deterministic or by taking residuals to be independent of each other, as in representative agent models. The econometricians are equally at fault for not trying to investigate the types of relationships between micro residuals that occur in practice, by looking more at panel data for instance. It follows that the usefulness in practice of using aggregated micro models is still very unclear, leaving the applied economists with difficulties in research strategy.

12. There seems to be one inevitable component of the gap between pure theorists on one hand and applicable econometricians and applied economists on the other, which arises from a basic difference in viewpoint and interest. It is helpful to distinguish between properties of models (strictly of variables generated by a model when used as a data generating process (DGP)) and properties of actual variables. Theorists are inherently interested only in properties of models whereas applied economists are — or should be — interested mostly in properties of variables. Properties of variables include stability, stationarity, ergodicity, forecastability (Granger) or causality using universal information sets, long memory or having near unit roots, capacities (see Nancy Cartwright (1989)), integrated and cointegration. Properties of models include consistency or efficiency of estimates, weak exogeneity, invertibility and encompassing. None of the properties of models have an existence without a model whereas properties of variables can be evaluated directly from data without use of a specific model. If a variable has a particular property, a model can also have this property, in which case the model is "data coherent" with respect to this particular property. Thus, if we realize by looking at the data that dividend payments are never negative, a model that only produced positive dividends would be data coherent. It is true that some models are only "potentially data coherent" in that if they are properly identified they may have parameter estimates that produce the required property. A model that is data coherent with respect to some sets of PV's but not others is of some but limited value. For example, a macro model built to explain a stylized fact (assumed to be an actual PV) may be data

coherent to that fact but with little else. The construction of such models can be thought of as a learning exercise for macro theorist rather than as a serious actual world modelling exercise. If successful, its value is clear and the method can be used in later models but it would be dangerous, to try to forecast the economy or to discuss policy implications if there are other important PV's with which the model is not data coherent. It will be an important part of any modelling exercise for the researcher to declare what PV's are considered important and which not, as part of the simplification process.

13.

> "When the hypothesis nor the implications of a theory can be confronted with the real world, that theory is devoid of any scientific interest."
>
> Maurice Allais
> extract from his 1988 Nobel Lecture
> "My Conception of Economic Science"
> Methodus, 2, (1990), 5–7.

The main disagreement between theorists and econometricians is the importance of empirical investigation of the implications of a theory. Many theory papers will not even indicate the relationship of the theory to the actual economy, yet the authors are willing to state "policy implications" of the theory.

There is a simple division of how a result can be evaluated; one can discuss the research strategy used to obtain the result and then evaluate its quality. The research strategy can be discussed on purely intellectual grounds such as how sensible are the assumptions, how robust is the result to changes in the assumptions and how good is the quality of the mathematics used in getting the result. I agree with Allais that, at least for producers in categories 2, 4 and 5, the quality of the result — theorem, procedure or model — can only finally be evaluated by facing it with the actual economy. I will also argue that the intellectual discussions about how a result is obtained are of very little value when evaluating it's quality. A theory or a model should be evaluated directly, not from it's origins. The quality of the mind of a child is not evaluated by considering the intelligence of it's parents — a genius can be born to parents of below average intellect. Similarly a great wine can be made from medium quality grapes and a masterpiece of art can be painted on poor quality canvas. Thus CIAs may be used to produce an important theorem and bad econometrics may produce an excellent model. It is probably a bad research strategy to start with poor ideas and technique, the result obtained will probably be of poor quality but it is not necessarily so.

There are many ways to evaluate and some of them can be illustrated

by considering how to reach an opinion about the quality of a bottle of wine. Some approaches to deciding the wine is not drinkable are:

(a) purely intellectual — "the wine is made by untraditional methods, using poor grapes and it comes from Texas";
(b) arbitrary — "the bottle is ugly";
(c) visual comparison — "it has a funny colour"; and
(d) empirical test — "I've had a taste and its not good."

Some academics are inclined to give a great deal of weight to approach (a) even though society has largely rejected it in most countries — one cannot base an employment decision on race, gender or age of an applicant, for example. Science surely prefers (d), which in economics corresponds to the empirical testing of a theory.

If a theorist provides a theorem *and* a discussion of the empirical implications, the work can be allocated to the AM theory group. Economists, believing in the efficiency of worker specialization, would not expect the theorist to conduct the empirical investigation him or herself. It is the job of the applied economist to do the empirical testing, but this is less likely to occur if the AM theorist does not communicate well and does not provide a clear statement about which is testable about the theory. If very special data is required, such as estimates of the utility functions of each participant in a survey, the testing may well not occur. If the theory can be stated as a specific hypothesis, the tools of mathematical statistics and procedural econometrics can be applied, using the hypothesis as the null H_0, provided also that suitable data is available. There is no need to discuss this part of evaluation further, although the appropriation reaction when H_0 is rejected need further consideration. If the theory cannot be stated as a specific hypothesis, because of impreciseness, such as an unclear definition of "income" say, or because important economic variables are ignored, such as if the theory only applies to closed economies say, then the applied economist has more difficulty evaluating the theory. It is necessary to add further simplifying assumptions to make a version of the theory testable. This should occur less often if the AM theorist keeps in mind the empirical testing question when constructing the theory. It follows that AM theorists should have some familiarity with data availability and with the main limitations of this data.

This attitude will make the AM theorists' task harder but more relevant. A procedure for progressing from sound theory to quality empirical evaluation is discussed in Stigum (1990), although he finds that it takes over a thousand pages to explain and develop the procedure. This perhaps give an indication of the magnitude of the task facing a responsible AM theorist. Various aspects of the empirical testing question are considered

further in sections 14 and 15.

The situation discussed in this section is hardly new, as the following quotation from Keynes (1936) states:

> But although the doctrine itself has remained unquestioned by orthodox economists up to a late date, its signal failure for purposes of scientific prediction has greatly impaired, in the course of time, the prestige of its practitioners. For professional economists, after Malthus, were apparently unmoved by the lack of correspondence between the results of their theory and the facts of observation; a discrepancy which the ordinary man has not failed to observe, with the result of his growing unwillingness to accord to economists that measure of respect which he gives to other groups of scientists whose theoretical results are confirmed by observation when they are applied to the facts.

14.

> "The science of political economy may be divided into two great branches. The theoretic and the practical... . The practical branch is far more arduous."
>
> Nassau W. Senior, 1827
> Selected writings on Economics
> Reprints of Economic Classics
> A. Kelley publisher, New York 1966

I believe that it is generally agreed amongst theorists that really good applied work is much more difficult than good theory. When students run into problems whilst analysing data, the econometric theorist can give a self-satisfied smile and say "that is why I stick to theory." A theorist can select his or her assumptions so that the development of the theory can proceed, but actual economic data is often less obliging and often disobeys these assumptions, non normality being an example.

Not only is good applied work difficult to produce but it seems to be particularly difficult to get the work accepted for publication. It is always easy to find topics that have been neglected or assumptions that are dubious but not tested, or simplification strategies that a reviewer finds unacceptable. This is particularly true if the evaluator is a theorist or an econometrician, who demand use of up to date theoretical results or econometric procedures. As such evaluators are likely to be unfamiliar with the practical problems of applied research or are unsympathetic to the actual difficulties encountered unrealistic demands may be put on the applied workers to get a paper accepted for publication. This may decrease if the evaluation was conducted largely by other applied economists plus consumers of applied research and if evaluation using empirical tests was given more weight.

A difficulty is that a poor quality applied paper, with incorrect theory or out-of-date econometric procedures cannot be allowed to be published, otherwise other applied workers would believe that work of such quality

was acceptable. Just as roses benefit from hard pruning, so does applied work benefit from harsh but fair criticism. Of course, the benefits of severe, constructive criticism will also apply when applied to theoretical work. The discipline's main objective is not reached if only applied work is put through an intensive evaluation process.

In recent years econometricians have paid some attention to the question of how a model can be evaluated (see the collection of papers in Granger (1990) and references given there). There are many in-sample specification tests but to avoid problems of "data-mining" most econometricians prefer out-of-sample or forecasting tests, although when there is insufficient data cross-validation is a computer intensive compromise. The "standard practice" is to specify a model, possibly after a specification search, to estimate it's parameters by minimising a cost function and then to apply various evaluation tests. As these procedures are well documented in the econometric literature there is no need to discuss them here, other than to say that the methods used are still evolving and there is by no means complete agreement amongst econometricians about the best strategies to use.

The method of evaluation may depend on the objectives of the model, so that a model devised just for forecasting should not be evaluated in the same way as a policy model, the difference being evaluation of unconditional versus conditional forecasts.

Some less conventional estimation and evaluation techniques are also employed. An estimated model may be rejected because it does not agree with economic intuition — an estimated coefficient may have the "wrong" sign, for example, or an estimated coefficient may be liked because it has a similar value to estimates from other, possibly less sophisticated, studies.

A new aspect of estimation has arisen with a class of models, called real business cycle models, which use an optimizing representative agent to suggest macro models. The models are usually too complicated to perform a full estimation procedure, and so values of important parameters, including means, variances and autocorrelations, are taken from a variety of separate, but relevant, sources. To simulate the output of the model, if an input series is thought to be white noise, then the mean and variance of an observed white noise is used, together with a normality assumption, to produce the required series. Similarly, autocorrelations are taken from actual series and plugged into the simulation. The process is called "calibration" and assures that the output of the model resembles the actual economy in certain, selected features. It is true of virtually all of the calibration process, including the recently introduced use of spectral analysis to see if the model really does produce a "business cycle", is that it com-

pletely ignores the sequence in which the actual data is generated, so that the information in the "arrow of time" is not employed. If the economic data were given to the calibration in a reversed order it would not affect the values used. The temporal order of data, which is the basic building block of virtually all discussions of causality, is generally considered to be the most important information that is available to a modeller. As the real business cycle theories are still evolving, the preset ones are certainly misspecified. What it means to calibrate a mis specified model is unclear.

15. Much of the discussion about evaluation can be used to consider questions of causality in economics. A statement like "a change in money supply results in a change in prices" has causality implications. One can attempt to evaluate the statement intellectually, by discussing the quality and correctness of the theory used to derive the relationship, and if there is no acceptable theory the correctness of the statement can be rejected. As argued above if the statement is a proposed property of the actual economy the only sound evaluation is by an empirical study. For this, an operational definition of causality is required. If instantaneous relationships are excluded, as I have argued elsewhere that they should, then a forecastability test of causality is viable and practical (see Granger (1988)). Such a test will not necessary be the only one available and may well not capture all parts of the complex issue of causality but it does have enough desirable features to make it worth using.

Thus, I am again separating discussion of the process that leads to the causality statement from the evaluation of the correctness of the statement. Clearly it is not possible to use statistical procedures on unique or rare causal events.

16.

> "Among persons interested in economic analysis, there are tool-makers and tool-users"
>
> A.C. Pigou
> Sydney Ball Lecture, 1929.

It has been suggested above that an objective of economic research is to affect the beliefs, and hence the behaviors, of others. Thus each piece of research will have both a producer and a consumer. The producer will need to market the product, so that information is provided to the potential consumer so that an evaluation can be attempted. The natural questions that follow from this viewpoint are who should evaluate research and how should it be done?

As to who should evaluate, the obvious workers are peers, those working

on similar topics, and potential consumers, those who will be using the work. Using the five categories of economists proposed in section four, the main consumers for the work of each group are:

Producer	Main Consumers
1. PM theorists	Other PM theorists Possibly AM theorists
2. AM theorists	Theorists, Applied Economists
3. Econometric theorists	Applicable Econometricians
4. Applicable Econometricians	Applied Economists
5. Applied Economists	Other Applied, Practicing Economists in government, industry, finance etc.

At present most evaluations of research papers for journal publication, new books or for career promotions by outside reviewers are performed by peers. This suggests that the quality of work by producers in groups 2 to 5 would be improved if consumers were also used as evaluators. [I have no comments on how the work of group 1 producers should be evaluated.] As an analogy, if Chevrolet had its cars evaluated only by Ford and vice-versa, consumers would be worse off than if consumers were involved in the evaluations. Reliance just on peer review ignores the main objective of the discipline.

The acknowledgement of the existence of consumers is particularly relevant for models designed to have policy implications. A policy maker interested in using the recommendations of such a model will surely want to be convinced that the model is relevant for the actual world. Econometricians would also surely claim that this can only occur if the model has been evaluated using data from the actual economy. Precisely how this should be done is still unclear, but some helpful techniques are available.

17.

"The first principle of Economics is that every agent is actuated only by self-interest"

F.Y. Edgeworth
Mathematical Psychics (1881). p.16.

If agents act largely in their own self-interest it should be of no surprise that economic researchers do the same. Their choice of research topics and the research strategy used will maximise a personal utility function, incorporating income, location, citations and so forth, and this behavior

probably does not maximise a social utility function or progress towards the main objective of the discipline. If the incentives of the system do not penalise self-interests, because of the use of peer-review rather than consumer-review, then nothing will change. The situation is similar to the mis-use of a free good, such as the pollution of air and water. Individuals may feel that it is too costly personally to attempt difficult and harshly evaluated applied work than less risky theory and mainstream econometrics. Society may want better applied economic research but the market for this research is not working correctly, largely because society has insufficient input into the decision processes, such as awards of grants, promotions and on publications. The economics solution to the mis-use of free goods is to have a public body, such as a local government, to impose a cost for pollution. Of course this solution does not work if the government consists just of polluters. The equivalence here is for editors, publishers and employers to give encouragement to work aimed at the main objective. It is not being argued that there should be no PM theory for example, but that the *balance* should change.

These ideas are, of course, by no means new, see for example the discussion on "satisficing" in Giere (1988).

To improve matters, there has to be re-evaluation of their objectives by individual workers, of the incentives given by the NSF and other grant givers and by employers, the greater use of consumer evaluators, a concerted effort by editors of important journals to make the evaluation process fairer to applied work and to suggest that theory and econometric papers pay specific attention to how the work can be made useful for applications.

18. In this paper I have identified four major problems with the discipline of economics:

- P1. The gap, or lack of communication, between theorists and econometricians, and particularly the lack of appreciation by theorists of the importance of empirical evaluation of their results.
- P2. The gap between much of economic theory and application.
- P3. The lack of appreciation of applied economic research.
- P4. The lack of attention paid to relevant research on the actual economy by academic economists. This problem is largely due to the application of self interest by researchers in their choice of research topics.

Some partial solutions are:

- S1. Require each research paper to precisely state its objective and how

this fits in with the main objective of the discipline. This will allow appropriate evaluation criteria to be used.

S2. Require each research paper to end by stating the empirical implications of the results, if any. In some occasions this will require theorists to be aware of data availability, limitations and its major properties.

S3. Evaluation of research papers, books and promotion files should be by consumers of the research as well as by peer producers.

S4. The discipline, through major academic or popular journals, should publish regular, say annual, accounts of what has been achieved in terms of solving major problems of actual economics.

S5. Research and fund giving organizations, such as the U.S. National Science Foundation or national governments and central banks, or international organizations such as the World Bank or IMF should occasionally issue challenges to the Economic Profession to solve — or have a substantial impact — on important, actual economy problems. An example would be to reduce unemployment in some disadvantaged section of society. Funds will need to be made available to support the research.

What is most required is wider discussions of the problems outlined above. The proposed solutions could alleviate some of the problems, but what is most needed is a change of attitude by many academic economists, which can perhaps only be achieved by a change in incentives.

REFERENCES

CARTWRIGHT, N. (1989). *Nature's Capacities and Their Measurement*, Oxford University Press.

ENGLE, R. F., D. F. HENDRY and J. F. RICHARD (1983). "Exogeneity", *Econometrica*, 51, 277–304.

FRIEDMAN, M. (1986). "Economists and economic policy". *Econ. Inquiry* 24, 1–10.

GIERE, R. N. (1988). *Explaining Science.* University of Chicago Press.

GRANGER, C. W. J. (1987). "Implications of aggregation with common factors", *Econometric Theory* 3, 208-22.

GRANGER, C. W. J. (1988). "Causality testing in a decision science", Pages 3–22, *Causation, Chance and Credence.* ed. B. Skyrms and W. L. Harper, Kluwer Publishers, Dorvrecht.

GRANGER, C. W. J. (1990). *Modelling Economic Series: Readings in Econometric Methodology*, Oxford University Press.

KEYNES, J. M. (1936). *The General Theory of Employment, Interest and Money*, reprinted by MacMillan for Royal Economic Society, 1971.

MORGAN, T. (1988). "Theory versus empiricism in academic economics: update and comparisons,", *J. of Economic Perspectives*, 2, 159–164.

REDMAN, D. (1991). *Economics and the Philosophy of Science*, Oxford University Press.

ROTHSCHILD, M. (1990). "Economic theory teaches us that economic theory teaches us nothing. The case of asset prices." UCSD working paper.

STIGUM, B. (1990). *Toward a Formal Science of Economics*, MIT Press.

Logic, Methodology and Philosophy of Science IX
D. Prawitz, B. Skyrms and D. Westerståhl (Editors)

A THEORY OF INFERRED CAUSATION*

JUDEA PEARL
THOMAS S. VERMA

Cognitive Systems Laboratory, Computer Science Department
University of California, Los Angeles, CA 90024
judea@cs.ucla.edu and verma@cs.ucla.edu

1. Introduction

The study of causation is central to the understanding of human reasoning. Inferences involving changing environments require causal theories which make formal distinctions between beliefs based on passive observations and those reflecting intervening actions [Geffner, 1989, Goldszmidt and Pearl, 1992, Lifchitz, 1987, Pearl, 1988a, Shoham, 1988]. In applications such as diagnosis [Patil et al., 1982, Reiter, 1987], qualitative physics [Bobrow, 1985], and plan recognition [Kautz, 1987, Wilensky, 1983], a central task is that of finding a satisfactory *explanation* to a given set of observations, and the meaning of explanation is intimately related to the notion of causation.

Most AI works have given the term "cause" a procedural semantics, attempting to match the way people use it in reasoning tasks, but were not concerned with the experience that prompts people to believe that "a causes b", as opposed to, say, "b causes a" or "c causes both a and b." The question of choosing an appropriate causal ordering received some attention in qualitative physics, where certain interactions attain directionality despite the instantaneous and symmetrical nature of the underlying equations, as in "the current causes the voltage to drop across the resistor" [Forbus and Gentner, 1986]. In some systems causal ordering is defined as the ordering at which subsets of variables can be solved independently of others [Iwasaki and Simon, 1986], in other systems it follows the way a disturbance is propagated from one variable to others [de Kleer and Brown, 1986].

*This paper is a modified version of one presented at the Second International Conference conference on the Principles of Knowledge Representation and Reasoning, Cambridge, Massachusetts, April 1991.

Yet these choices are made as a matter of convenience, to fit the structure of a given theory, and do not reflect features of the empirical environment which compelled the formation of the theory.

An empirical semantics for causation is important for several reasons. First, an intelligent system attempting to build a workable model of its environment cannot rely exclusively on preprogrammed causal knowledge, but must be able to translate direct observations to cause-and-effect relationships. Second, by tracing empirical origins we stand to obtain an independent gauge for deciding which of the many logics proposed for causal reasoning is sound and/or complete, and which provides a proper account of causal utterances such as "a explains b", "a suggests b", "a tends to cause b", and "a actually caused b", etc.

While the notion of causation is often associated with those of necessity and functional dependence, causal expressions often tolerate exceptions, primarily due to missing variables and coarse description. We say, for example, "reckless driving causes accidents" or "you will fail this course because of your laziness". Suppes [Suppes, 1970] has argued convincingly that most causal utterances in ordinary conversation reflect probabilistic, not categorical relations[1]. Thus, probability theory should provide a natural language for capturing causation [Reichenbach, 1956, Good, 1983]. This is especially true when we attempt to infer causation from (noisy) observations – probability calculus remains an unchallenged formalism when it comes to translating statistical data into a system of revisable beliefs.

However, given that statistical analysis is driven by covariation, not causation, and assuming that most human knowledge derives from statistical observations, we must still identify the clues that prompt people to perceive causal relationships in the data, and we must find a computational model that emulates this perception.

Temporal precedence is normally assumed essential for defining causation, and it is undoubtedly one of the most important clues that people use to distinguish causal from other types of associations. Accordingly, most theories of causation invoke an explicit requirement that a cause precedes its effect in time [Good, 1983, Reichenbach, 1956, Shoham, 1988, Suppes, 1970]. Yet temporal information alone cannot distinguish genuine causation from spurious associations caused by unknown factors. In fact the statistical and philosophical literature has adamantly warned analysts that, unless one knows in advance all causally relevant factors, or unless one can carefully manipulate some variables, no genuine causal inferences are possible [Cartwright, 1989, Cliff, 1983, Eells and Sober, 1983, Fisher, 1953,

[1] See [Dechter and Pearl, 1991] for a treatment of causation in the context of categorical data.

Gärdenfors, 1988, Holland, 1986, Skyrms, 1980][2]. Neither condition is realizable in normal learning environments, and the question remains how causal knowledge is ever acquired from experience.

This paper introduces a minimal-model semantics of causation which provides a plausible account for how causal models could be inferred from observations. Using this semantics we show that genuine causal influences can in many cases be distinguished from spurious covariations and, moreover, the direction of causal influences can often be determined without resorting to chronological information. (Although, when available, chronological information can significantly simplify the modeling task.) Such semantics should be applicable, therefore, to the organization of concurrent events or events whose chronological precedence cannot be determined with precision, (e.g. "old age explains disabilities") in the spirit of Glymour [Glymour et al., 1987] and Simon [Simon, 1954].

This paper is organized as follows. In Section 2 we define the notions of causal models and causal theories, and describe the task of causal modeling as an identification game scientists play against Nature. In Section 3 we introduce the minimal-model semantics of causation and exemplify its operability and plausibility on a simple example. Section 4 identifies conditions under which effective algorithms exist that uncover the structure of casual influences as defined above. One such algorithm (called IC) is introduced in Section 5, and is shown to be sound for the class of stable distributions, even when some variables are not observable[3]. Section 6 extracts from the IC-algorithm the essential conditions under which causal influences are identified and proposes these as independent definitions of genuine influences and spurious associations, with and without temporal information. Section 7 provides an intuitive justification for the definitions proposed in Section 6, showing that our theory conforms to the common understanding of causation as a stipulation of stable behavior under external interventions. The definitions are shown to be in line with accepted standards of controlled experimentation, save for requiring the identification of "virtual" experimental conditions within the data itself. In Section 8 we invoke the "virtual control" metaphor to elucidate how causal relationships can still be ascertained in the absence of temporal information. We then offer an explanation for the puzzling, yet universal agreement between the temporal and the statistical aspects of causation.

---

[2]Some of the popular quotes are: "No causation without manipulation" [Holland, 1986], "No causes in, no causes out" [Cartwright, 1989], "No computer program can take account of variables that are not in the analysis" [Cliff, 1983].

[3]Proofs can be found in [Verma, 1992].

2. The causal modeling framework

We view the task of causal modeling as an identification game which scientists play against Nature. Nature possesses stable causal mechanisms which, on a microscopic level are deterministic functional relationships between variables, some of which are unobservable. These mechanisms are organized in the form of an acyclic schema which the scientist attempts to identify.

DEFINITION 1 *A* **causal model** *of a set of variables U is a directed acyclic graph* (dag), *in which each node corresponds to a distinct element of U.*

The nodes of the dag correspond to the variables under analysis, while the links denote direct causal influences among the variables. The causal model serves as a blue print for forming a "causal theory" – a precise specification of how each variable is influenced by its parents in the dag. Here we assume that Nature is at liberty to impose arbitrary functional relationships between each effect and its causes and then to perturb these relationships by introducing arbitrary (yet mutually independent) disturbances. These disturbances reflect "hidden" or unmeasurable conditions and exceptions which Nature chooses to govern by some undisclosed probability function.

DEFINITION 2 *A* **causal theory** *is a pair $T = <D, \Theta_D>$ consisting of a causal model D and a set of parameters Θ_D compatible with D. Θ_D assigns a function $x_i = f_i[\mathbf{pa}(x_i), \epsilon_i]$ and a probability measure g_i, to each $x_i \in U$, where $\mathbf{pa}(x_i)$ are the parents of x_i in D and each ϵ_i is a random disturbance distributed according to g_i, independently of the other ϵ's and of any preceding variable $x_j : 0 < j < i$. (The variables are assumed ordered such that all arcs point from lower to higher indices.)*

This requirement of independence renders each disturbance "local" to one parents-child family; disturbances that influence several families simultaneously will be treated explicitly as "latent" variables (see Definition 3).

Once a causal theory T is formed, it defines a joint probability distribution $P(T)$ over the variables in the system, and this distribution reflects some features of the causal model (e.g., each variable must be independent of its grandparents, given the values of its parents). Nature then permits the scientist to inspect a select subset $O \subseteq U$ of "observed" variables, and to ask questions about $P_{[O]}$, the probability distribution over the observables, but hides the underlying causal theory as well as the structure of the causal model. We investigate the feasibility of recovering the topology of the dag, D, from features of the probability distribution.[4]

[4]This formulation invokes several idealizations of the actual task of scientific discovery.

3. Model preferences (Occam's razor)

In principle, U being unknown, there is an unbounded number of models that would fit a given distribution, each invoking a different set of "hidden" variables and each connecting the observed variables through different causal relationships. Therefore with no restriction on the type of models considered, the scientist is unable to make any meaningful assertions about the structure underlying the phenomena. Likewise, assuming $U = O$ but lacking temporal information, he/she can never rule out the possibility that the underlying model is a complete (acyclic) graph; a structure that, with the right choice of parameters can *mimic* (see Definition 4) the behavior of any other model, regardless of the variable ordering. However, following the standard method of scientific induction, it is reasonable to rule out any model for which we find a simpler, *less expressive* model, equally consistent with the data (see Definition 6). Models that survive this selection are called "minimal models" and with this notion, we can construct our definition of *inferred causation*:

"A variable X is said to have a causal influence on a variable Y if a strictly directed path from X to Y exists in every minimal model consistent with the data"

DEFINITION 3 *Given a set of observable variables $O \subseteq U$, a* **latent structure** *is a pair $L = <D, O>$ where D is a causal model over U.*

DEFINITION 4 *One latent structure $L = <D, O>$ is* **preferred** *to another $L' = <D', O>$ (written $L \preceq L'$) iff D' can* **mimic** *D over O, i.e. for every Θ_D there exists a $\Theta'_{D'}$ s.t. $P_{[O]}(<D', \Theta'_{D'}>) = P_{[O]}(<D, \Theta_D>)$.*

Two latent structures are **equivalent**, *written $L' \equiv L$, iff $L \preceq L'$ and $L \succeq L'$.*

Note that the preference for simplicity imposed by Definition 4 is gauged by the expressive power of a model, not by its syntactic description. For example, one latent structure $L1$ may invoke many more parameters than $L2$ and still be preferred, if $L2$ is capable of accommodating a richer set of probability distributions over the observables. One reason scientists prefer simpler models is that such models are more constrained, thus more falsifiable; they

It assumes, for example, that the scientist obtains the distribution directly, rather than events sampled from the distribution. This assumption is justified when a large sample is available, sufficient to reveal all the dependencies embedded in the distribution. Additionally, we assume that the observed variables actually appear in the original causal theory and are not some aggregate thereof. Aggregation might result in feedback loops which we do not discuss in this paper. Our theory also takes variables as the primitive entities in the language, not events which permits us to include "enabling" and "preventing" relationships as part of the mechanism.

provide the scientist with less opportunities to overfit the data hindsightedly and, therefore attain greater credibility [Pearl, 1978, Popper, 1959].

We also note that the set of dependencies induced by a causal model provides a measure of its expressive power, i.e., its power of mimicing other models. Indeed, $L1$ cannot be preferred to $L2$ if there is even one observable dependency that is induced by $L1$ and not by $L2$. Thus, tests for preference and equivalence can often be reduced to tests of induced dependencies which, in turn, can be determined directly from the topology of the dags, without ever concerning ourselves with the set of parameters. (For example, see Theorem 1 below and [Frydenberg, 1989, Pearl et al., 1989, Verma and Pearl, 1990]).

DEFINITION 5 *A latent structure L is* **minimal** *with respect to a class $\mathcal{L}$ of latent structures iff for every $L' \in \mathcal{L}$, $L \equiv L'$ whenever $L' \preceq L$.*

DEFINITION 6 *$L = <D, O>$ is* **consistent** *with a distribution $\hat{P}$ over O if D can accommodate some theory that generates $\hat{P}$, i.e. there exists a Θ_D s.t. $P_{[O]}(<D, \Theta_D>) = \hat{P}$*

Clearly, a necessary (and often sufficient) condition for L to be consistent with $\hat{P}$, is that the structure of L can account for all the dependencies embodied in $\hat{P}$.

DEFINITION 7 (INFERRED CAUSATION) *Given $\hat{P}$, a variable C has a* **causal influence** *on E iff there exists a directed path $C \rightarrow^* E$ in every minimal latent structure consistent with $\hat{P}$.*

We view this definition as normative, because it is based on one of the least disputed norms of scientific investigation: Occam's razor in its semantical casting. However, as with any scientific inquiry, we make no claims that this definition is guaranteed to always identify stable physical mechanisms in nature; it identifies the only mechanisms we can plausibly infer from non-experimental data.

As an example of a causal relation that is identified by the definition above, imagine that observations taken over four variables $\{a, b, c, d\}$ reveal two vanishing dependencies: "a is independent of b" and "d is independent of $\{a, b\}$ given c". Assume further that the data reveals *no other* independence, except those that logically follow from these two. This dependence pattern would be typical for example, of the following variables: $a =$ *having cold*, $b =$ *having hay-fever*, $c =$ *having to sneeze*, $d =$ *having to wipe ones nose*. It is not hard to see that any model which explains the dependence between c and d by an arrow from d to c, or by a hidden common cause

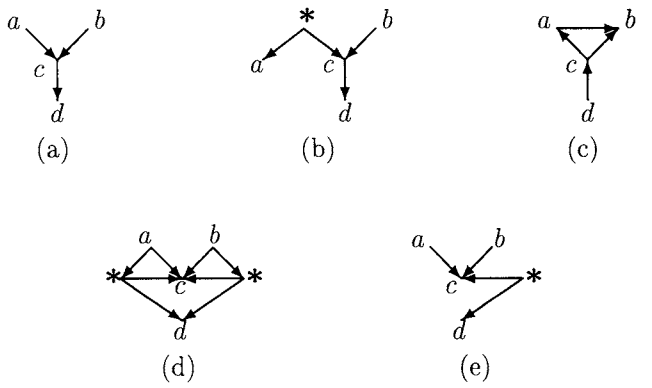


Figure 1: Causal models illustrating the soundness of $c \rightarrow d$. The node $(*)$ represents a hidden variable.

$(*)$ between the two, cannot be minimal, because any such model would be able to out-mimic the minimal model shown in Figure 1(a) (or the one in Figure 1(b)) which reflects all observed independencies. For example, the model of Figure 1(c), unlike that of Figure 1(a), accommodates distributions with arbitrary relations between a and b. Similarly, Figure 1(d) is not minimal as it fails to impose the conditional independence between d and $\{a, b\}$ given c. In contrast, Figure 1(e) is not consistent with the data since it imposes a marginal independence between $\{a, b\}$ and d, which was not observed. (For theory and method of identifying conditional independencies in causal graphs see [Pearl, 1988b] and [Pearl et al., 1989])

4. Proof theory and stable distributions

It turns out that while the minimality principle is sufficient for forming a normative and operational theory of causation, it does not guarantee that the search through the vast space of minimal models would be computationally practical. If Nature truly conspires to conceal the structure of the underlying model she could still annotate that model with a distribution that matches many minimal models, having totally disparate structures. To facilitate an effective proof theory, we rule out such eventualities, and impose a restriction on the distribution called "stability" (or "dag-isomorphism" in [Pearl, 1988b]). It conveys the assumption that all vanishing dependencies

are structural, not formed by incidental equalities of numerical parameters[5].

DEFINITION 8 *Let I(P) denote the set of all conditional independence relationships embodied in P. A causal theory* $T = <D, \Theta_D>$ *generates a* **stable** *distribution iff it contains no extraneous independences, i.e.* $I(P(<D, \Theta_D>)) \subseteq I(P(<D, \Theta'_D>))$ *for any set of parameters* Θ'_D.

With the added assumption of stability, every distribution has a unique causal model (up to equivalence), as long as there are no hidden variables. This uniqueness follows from the fact that the structural constraints that an underlying dag imposes upon the probability distribution are equivalent to a finite set of conditional independence relationships asserting that, given its parents, each variable is conditionally independent of all its non-descendents [Pearl et al., 1989]. Therefore two causal models are equivalent (i.e. they can mimic each other) if and only if they relay the same dependency information. The following theorem, which is founded upon the dependency information, states necessary and sufficient conditions for equivalence of causal models which contain no hidden variables.

THEOREM 1 [VERMA AND PEARL, 1990] *When* $U = O$, *two causal models are equivalent iff their dags have the same links and same set of uncoupled head-to-head nodes*[6].

The search for the minimal model then boils down to recovering the structure of the underlying dag from queries about the dependencies portrayed in that dag. This search is exponential in general, but simplifies significantly when the underlying structure is sparse (see [Spirtes and Glymour, 1991, Verma and Pearl, 1990] for such algorithms).

Unfortunately, the constraints that a latent structure imposes upon the distribution cannot be completely characterized by any set of dependency statements. However, the maximal set of sound constraints can be identified [Verma and Pearl, 1990] and it is this set that permits us to recover sound fragments of latent structures.

---

[5]It is possible to show that, if the parameters are chosen at random from any reasonable distribution, then any unstable distribution has measure zero [Spirtes et al., 1989]. Stability precludes deterministic constraints. Less restrictive assumptions are treated in [Geiger et al., 1990].

[6]i.e. converging arrows emanating from non-adjacent nodes, such as $a \rightarrow c \leftarrow b$ in Figure 1(a).

5. Recovering latent structures

When Nature decides to "hide" some variables, the observed distribution $\hat{P}$ need no longer be stable relative to the observable set O, i.e. $\hat{P}$ may result from many equivalent minimal latent structures, each containing any number of hidden variables. Fortunately, rather then having to search through this unbounded space of latent structures, it turns out that for every latent structure L, there is a dependency-equivalent latent structure called the projection of L on O in which every unobserved node is a root node with exactly two observed children:

DEFINITION 9 *A latent structure* $L_{[O]} = <D_{[O]}, O>$ *is a* **projection** *of another latent structure* L *iff*

1. *Every unobservable variable of* $D_{[O]}$ *is a parentless common cause of exactly two non-adjacent observable variables.*

2. *For every stable distribution* P *generated by* L*, there exists a stable distribution* P' *generated by* $L_{[O]}$ *such that* $I(P_{[O]}) = I(P'_{[O]})$.

THEOREM 2 [VERMA, 1992] *Any latent structure has at least one projection.*

It is convenient to represent projections by bi-directional graphs with only the observed variables as vertices (i.e., leaving the hidden variables implicit). Each bi-directed link in such a graph represents a common hidden cause of the variables corresponding to the link's end points.

Theorem 2 renders our definition of inferred causation (Definition 7) operational; we will show (Theorem 3) that if a certain link exists in a distinguished projection of any minimal model of $\hat{P}$, it must indicate the existence of a causal path in every minimal model of $\hat{P}$. Thus the search reduces to finding a projection of any minimal model of $\hat{P}$ and identifying the appropriate links. Remarkably, these links can be identified by a simple procedure, the IC-algorithm, that is not more complex than that which recovers the unique minimal model in the case of fully observable structures.

IC-Algorithm (Inductive Causation)

Input: $\hat{P}$ a sampled distribution.
Output: **core**$(\hat{P})$ a marked hybrid acyclic graph.

1. For each pair of variables a and b, search for a set S_{ab} such that (a, S_{ab}, b) is in $I(\hat{P})$, namely a and b are independent in $\hat{P}$, conditioned on S_{ab}.
   If there is no such S_{ab}, place an undirected link between the variables, $a - b$.

2. For each pair of non-adjacent variables a and b with a common neighbor c, check if $c \in S_{ab}$.
   If it is, then continue.
   If it is not, then add arrowheads pointing at c, (i.e. $a \rightarrow c \leftarrow b$).

3. Form $\mathbf{core}(\hat{P})$ by recursively adding arrowheads according to the following two rules:[7]
   If there is a directed path from a to b and, in addition there is an edge between a and b, then add an arrowhead to that edge pointing toward b.
   If a and b are not adjacent but there exists a node c that is adjacent to both a and b such that $\overrightarrow{ac}$ and $c - b$, then direct $c \rightarrow b$.

4. If a and b are not adjacent but $\overrightarrow{ac}$ and $c \rightarrow b$, then mark the link $c \rightarrow b$.

The result of this procedure is a substructure called $\mathbf{core}(\hat{P})$ in which every marked uni-directed arrow $X \rightarrow Y$ stands for the statement: "X has a causal influence on Y (in all minimal latent structures consistent with the data)". We call these relationships "genuine" causal influences (e.g. $c \rightarrow d$ in previous Figure 1a).

DEFINITION 10 *For any latent structure L, $\mathbf{core}(L)$ is defined as the hybrid graph*[8] *satisfying (1) two nodes are adjacent in $\mathbf{core}(L)$ iff they are adjacent or they have a common unobserved cause in every projection of L, and (2) a link between a and b has an arrowhead pointing at b iff $a \rightarrow b$ or a and b have a common unobserved cause in every projection of L.*

THEOREM 3 *(soundness) For any latent structure $L = <D, O>$ and an associated theory $T = <D, \Theta_D>$ if $P(T)$ is stable then every arrowhead identified by IC is also in $\mathbf{core}(L)$.*

COROLLARY 1 *If every link of the directed path $C \rightarrow^* E$ is marked in $\mathbf{core}(\hat{P})$ then C has a causal influence on E according to $\hat{P}$.*

6. Probabilistic definitions for causal relations

The IC-algorithm takes a distribution $\hat{P}$ and outputs a dag, some of its links are marked uni-directional (denoting genuine causation), some are unmarked uni-directional (denoting potential causation), some are bi-directional (denoting spurious association) and some are undirected (denoting relationships that remain undetermined). The conditions which give rise to these

[7] $\overrightarrow{ab}$ denotes either $a \rightarrow b$ or $a \leftrightarrow b$, and $a - b$ denotes an undirected edge.
[8] In a hybrid graph links may be undirected, uni-directed or bi-directed.

labelings constitute operational definitions for the various kinds of causal relationships. In this section we present explicit definitions of potential and genuine causation, as they emerge from Theorem 3 and the IC-algorithm. Note that in all these definitions, the criterion for causation between two variables, X and Y, will require that a third variable Z exhibit a specific pattern of interactions with X and Y. This is not surprising, since the very essence of causal claims is to stipulate the behavior of X and Y under the influence of a third variable, one that corresponds to an external control of X. Therefore, our definitions are in line with the paradigm of "no causation without manipulation" [Holland, 1986]). The difference is only that the variable Z, acting as a virtual control of X, must be identified within the data itself. The IC-algorithm provides a systematic way of searching for variables Z that qualify as virtual controls.

Detailed discussions of these definitions in terms of virtual control are given in Sections 7 and 8.

DEFINITION 11 (POTENTIAL CAUSE) *A variable X has a* **potential causal influence** *on another variable Y (***inferable** *from $\hat{P}$), if*

1. *X and Y are dependent in every context.*
2. *There exists a variable Z and a context S such that*
   (i) X and Z are independent given S
   (ii) Z and Y are dependent given S

By "context" we mean a set of variables tied to specific values. Note that this definition precludes a variable X from being a potential cause of itself or of any other variable which functionally determines X.

DEFINITION 12 (GENUINE CAUSE) *A variable X has a* **genuine causal influence** *on another variable Y if there exists a variable Z such that:*

1. *X and Y are dependent in any context and there exists a context S satisfying:*
   (i) Z is a potential cause of X
   (ii) Z and Y are dependent given S.
   (iii) Z and Y are independent given $S \cup X$,

   or,
2. *X and Y are in the transitive closure of rule 1.*

DEFINITION 13 (SPURIOUS ASSOCIATION) *Two variables X and Y are* **spuriously associated** *if they are dependent in some context S and there exists two other variables Z_1 and Z_2 such that:*

1. *Z_1 and X are dependent given S*
2. *Z_2 and Y are dependent given S*
3. *Z_1 and Y are independent given S*
4. *Z_2 and X are independent given S*

Succinctly, using the predicates I and $\neg I$ to denote independence and dependence respectively, the conditions above can be written:

1. $\neg I(Z_1, X|S)$
2. $\neg I(Z_2, Y|S)$
3. $I(Z_1, Y|S)$
4. $I(Z_2, X|S)$

Definition 11 was formulated in [Pearl, 1990] as a relation between events (rather than variables) with the added condition $P(Y|X) > P(Y)$ in the spirit of [Good, 1983, Reichenbach, 1956, Suppes, 1970]. Condition 1(i) in Definition 12 may be established either by statistical methods (per Definition 11) or by other sources of information e.g., experimental studies or temporal succession (i.e. that Z precedes X in time).

When temporal information is available, as it is assumed in most theories of causality ([Granger, 1988, Spohn, 1983, Suppes, 1970]), then Definitions 12 and 13 simplify considerably because every variable preceding and adjacent to X now qualifies as a "potential cause" of X. Moreover, adjacency (i.e. condition 1 of Definition 11) is not required as long as the context S is confined to be earlier than S. These considerations lead to simpler conditions distinguishing genuine from spurious causes as shown next.

DEFINITION 14 (GENUINE CAUSATION WITH TEMPORAL INFORMATION) *A variable X has a causal influence on Y if there is a third variable Z and a context S, both occurring before X such that:*

1. $\neg I(Z, Y|S)$
2. $I(Z, Y|S \cup X)$

DEFINITION 15 (SPURIOUS ASSOCIATION WITH TEMPORAL INFORMATION) *Two variables X and Y are* **spuriously associated** *if they are dependent in some context S, X precedes Y and there exists a variable Z satisfying:*

1. $I(Z, Y|S)$
2. $\neg I(Z, X|S)$

7. Causal intuition and virtual experiments

This section explains how the formulation introduced above conforms to common intuition about causation and, in particular, how asymmetric probabilistic dependencies can be transformed into judgements about asymmetric causal influences. We shall first uncover the intuition behind Definition 14, assuming the availability of temporal information, then (in Section 8) generalize to non temporal data, per Definition 12.

The common intuition about causation is captured by the heuristic definition [Rubin, 1989]: "X is a cause for Y if an external agent interfering only with X can affect Y" .

Thus, causal claims are much bolder than those made by probability statements; not only do they summarize relationships that hold in the distribution underlying the data, but they also predict relationships that should hold when the distribution undergoes changes, such as those inferable from external intervention. The claim "X causes Y" asserts the existence of a *stable* dependence between X and Y, one that cannot be attributed to some prior cause common to both, and one that should be preserved when an exogenous control is applied to X.

This intuition requires the formalization of three notions:

1. That the intervening agent be "external" (or "exogenous")
2. That the agent can "affect" Y
3. That the agent interferes "only" with X

If we label the behavior of the intervening agent by a variable Z, then these notions can be given the following probabilistic explications:

1. **Externality of Z**: Variations in Z must be independent of any factors W which precede X, i.e.,

$$I(Z, W) \quad \forall \quad W : t_W < t_X \tag{1}$$

2. **Control**: For Z to effect changes in Y (via X) we require that Z and Y be dependent, written:
$$\neg I(Z, Y) \tag{2}$$

3. **Locality**: To ensure that Z interferes "only" with X, i.e., that its entire effect on Y is mediated by X, we use the conditional independence assertion:
$$I(Z, Y|X) \tag{3}$$
to read "Z is independent of Y, given X".

Note that (1) and (2) imply (by the axioms of conditional independence [Pearl, 1988b]) that X and Y are dependent, namely, $\neg I(X, Y)$.

Conditions (1) through (3) constitute the traditional premises behind controlled statistical experiments, with Z representing the decision to administer condition $X = x$ to a given unit (or a given subject), and (1) reflecting the requirement that units selected for the experiment be assigned at random to the various experimental conditions. They guarantee that any dependency observed between X and Y cannot be explained away by holding fixed some factor W preceding X (as in Figure 3), hence it must be attributed to genuine causation (as in Figure 2). The sufficiency of these premises is clearly not a theorem of probability theory, as it relies on temporal relationships among the variables. However, it can be derived from probability theory together with Reichenbach's principle [Reichenbach, 1956], stating that every dependence $\neg I(X, Y)$ requires a causal explanation, namely either one of the variables causes the other, or there must be a variable W preceding X and Y such that $I(X, Y|W)$ (see Figure 2). Indeed, if there is no back path from Z to Y through W (Eq. (1)) and no direct path from Z to Y avoiding X (Eq. (3)) then there must be a causal path from X to Y that is responsible for the dependence in Eq. (2)[9].

In non-experimental situations it is not practical to detach X completely from its natural surrounding and to subject it to the exclusive control of an exogenous (and randomized) variable Z. Instead, one could view some of X's natural causes as "virtual controls" and, provided certain conditions are met, use the latter to reveal non-spurious causal relationship between X and Y. In so doing we compromise, of course, condition (1), because we can no longer guarantee that those natural causes of X are not themselves affected by other causes which, in turn, might influence Y (see Figure 3). However, it turns out that for stable distributions, conditions (2) and (3) are sufficient to guarantee that the association between X and Y is non-spurious, thus justifying Definition 14 for genuine causation.

[9]Cartwright [Cartwright, 1989] offers a sufficiency proof in the context of linear models.

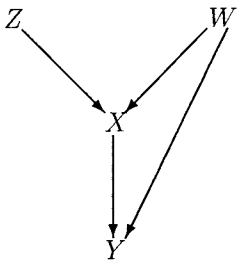


Figure 2

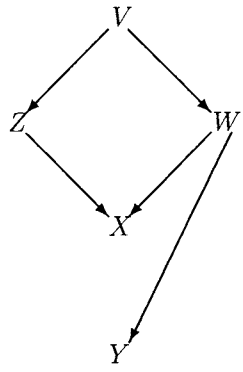


Figure 3

The intuition goes as follows (see Figure 3): If the dependency between Z and Y (and similarly, between X and Y) is spurious, namely, X and Y are merely manifestations of some common cause W, there is no reason then for X to screen-off Y from Z, and condition (2) should be violated. In case condition (2) is accidentally satisfied by some strange combination of parameters, it is bound to be "unstable", as it will be perturbed with any slight change of expiremental conditions.

Conditions (2) and (3) are identical to those in Definition 14, save for the context S which is common to both. The inclusion of the fixed context S is legitimized by noting that if $P(X, Y, Z)$ is a marginal of a stable distribution, then so is the conditional distribution $P(X, Y, Z|S = s)$, as long as S corresponds to variables which precede X.

Definition 14 constitutes an alternative way of recovering causal structures, more flexible than the IC-algorithm; we search the data for three variables Z, X, Y (in this temporal order) that satisfy the two conditions in some context $S = s$, and when such a triple is found, X is proclaimed to have a genuine causal influence on Y. Clearly, permitting an arbitrary context S increases the number of genuine causal influences that can be identified in any given data; marginal independencies and even 1-place conditional independencies are rare phenomenon.

Note that failing to satisfy the test for genuine causation does not mean that such relationship is necessarily absent between the quantities under study. Rather, it means that the data available cannot substantiate the claim of genuine causation. To further test such claims one may need to either conduct experimental studies, or consult a richer data set where virtual

control variables are found.

In testing this modeling scheme on real life data, we have examined the observations reported in Sewal Wright's seminal paper "Corn and Hog Correlations" [Wright, 1925]. As expected, corn-price (X) can clearly be identified as a cause of hog-price (Y), not the other way around. The reason lies in the existence of the variable corn-crop (Z) that, by satisfying the conditions of Definition 14 (with $S = \emptyset$), acts as a virtual control of X (see Figure 2). To test for the possibility of reciprocal causation, one can try to find a virtual controller for Y, for example, the amount of hog-breeding (Z'). However, it turns out that Z' is not screened off from X by Y (possibly because corn prices exert direct influence over farmer's decision to breed more hogs), hence, failing condition 3, Y disqualifies as a genuine cause of X. Such distinctions are important to policy makers in deciding, for example, which commodity, corn or hog, should be subsidized or taxed.

8. Non-temporal causation and statistical time

When temporal information is unavailable the condition that Z precede X (Definition 14) cannot be tested directly and must be replaced by an equivalent condition, based on dependence information. As it turns out, the only reason we had to require that Z precede X is to rule out the possibility that Z is a causal consequence of X; if it were a consequence of X then the dependency between Z and Y could easily be explained away by a common cause W of X and Y (see Figure 2).

The information that permits us to conclude that one variable is not a causal consequence of another comes in the form of an "intransitive triplet", such as the variables a, b and c in Figure 1(a) satisfying: $I(a, b)$, $\neg I(a, c)$ and $\neg I(b, c)$. The argument goes as follows: If we create conditions (fixing S_{ab}) where two variables, a and b, are each correlated with a third variable c but are independent of each other, then the third variable cannot act as a cause of a or b, (recall that in stable distributions, common causes induce dependence among their effects); it must be either their common effect, $a \rightarrow c \leftarrow b$, or be associated with a and b via common causes, forming a pattern such as $a \leftrightarrow c \leftrightarrow b$. This is indeed the eventuality that permits our algorithm to begin orienting edges in the graph (step 2), and assign arrowheads pointing at c. It is also this intransitive pattern which is used to ensure that X is not a consequence of Y (in Definition 11) and that Z is not a consequence of X (in Definition 12). In definition 14 we have two intransitive triplets, (Z_1, X, Y) and (X, Y, Z_2), thus ruling out direct causal influence between X and Y, implying spurious associations as the only explanation for their dependence.

This interpretation of the intransitive triple is in line with the "virtual control" view of causation. For example, one of the reasons people insist that the rain causes the grass to become wet, and not the other way around, is that they can find other means of getting the grass wet, totally independent of the rain. Transferred to our chain $a - c - b$, we can preclude c from being a cause of a if we find another means (b) of potentially controlling c without affecting a [Pearl, 1988a, p. 396].

Determining the direction of causal influences from nontemporal data raises some interesting philosophical questions about the nature of time and causal explanations. For example, can the orientation assigned to the arrow $X \rightarrow Y$ in Definition 14 ever clash with temporal information (say by a subsequent discovery that Y precedes X)? Alternatively, since the rationale behind Definition 14 is based on strong intuitions about how causal influences should behave (statistically), it is apparent that such clashes, if they occur, are rather rare. The question arises then, why? Why should orientations determined solely by statistical dependencies have anything to do with the flow of time?

In human discourse, causal explanations indeed carry two connotations, temporal and statistical. The temporal aspect is represented by the convention that a cause should precede its effect. The statistical aspect expects causal explanations (once accounted for) to screen off their effects, i.e., render their effects conditionally independent[10]. More generally, causal explanations are expected to obey many of the rules that govern paths in a directed acyclic graphs (e.g., the intransitive triplet criterion for potential causation, Section 7). This leads to the observation that, if agreement is to hold between the temporal and statistical aspects of causation, natural statistical phenomena must exhibit some basic temporal bias. Indeed, we often encounter phenomenon where knowledge of a present state renders the variables of the future state conditionally independent (e.g., multi-variables economic time series as in Eq. (4) below). We rarely find the converse phenomenon, where knowledge of the present state would render the components of the past state conditionally independent. The question arises

[10]This principle, known as Reichenbach's "conjunctive fork" or "common-cause" criterion [Reichenbach, 1956, Suppes and Zaniotti, 1981] has been criticized by Salmon [Salmon, 1984], who showed that some events would qualify as causal explanations though they fail to meet Reichenbach's criterion. Salmon admits, however, that when a conjunctive forks does occur, the screening off variable is expected to be the cause of the other two, not the effect [Salmon, 1984, p. 167]. He notes that it is difficult to find physically meaningful examples where a response variable renders its two causes conditionally independent (although this would not violate any axiom of probability theory). This asymmetry is further evidence that humans tend to reject causal theories that yield unstable distributions.

whether there is any compelling reason for this temporal bias.

A convenient way to articulate this bias is through the notion of "Statistical Time".

DEFINITION 16 (STATISTICAL TIME) *Given an empirical distribution P, a statistical time of P is any ordering of the variables that agrees with at least one minimal causal model consistent with P.*

We see, for example, that a scalar Markov-chain process has many statistical times; one coinciding with the physical time, one opposite to it and the others correspond to any time ordering of the variables away from some chosen variable. On the other hand a process governed by two coupled Markov chains,

$$\begin{aligned} X_t &= \alpha X_{t-1} + \beta Y_{t-1} + \xi_t \\ Y_t &= \gamma X_{t-1} + \delta Y_{t-1} + \xi'_t, \end{aligned} \tag{4}$$

has only one statistical time – the one coinciding with the physical time[11]. Indeed, running the IC-algorithm on samples taken from such a process, while suppressing all temporal information, quickly identifies the components of X_{t-1} and Y_{t-1} as genuine causes of X_t and Y_t. This can be seen from Definition 11, where X_{t-2} qualifies as a potential cause of X_{t-1} using $Z = Y_{t-2}$ and $S = \{X_{t-3}, Y_{t-3}\}$, and Definition 12, where X_{t-1} qualifies as a genuine cause of X_t using $Z = X_{t-2}$ and $S = \{Y_{t-1}\}$ of X_t.

The temporal bias postulated earlier can be expressed as follows:

CONJECTURE 1 (TEMPORAL BIAS) *In most natural phenomenon, the physical time coincides with at least one statistical time.*

Reichenbach [Reichenbach, 1956] attributed the asymmetry associated with his conjunctive fork to the second law of thermodynamics. We are not sure at this point whether the second law can provide a full account of the temporal bias as defined above, since the influence of the external noise ξ_t and ξ'_t renders the process in (4) nonconservative[12]. What is clear, however, is that the temporal bias is *language dependent.* For example, expressing Eq.(4) in a different coordinate system (say, using a unitary transformation $(X'_t, Y'_t) = U(X_t, Y_t)$), it is possible to make the statistical time (in the (X', Y') representation) run contrary to the physical time. This suggests that the apparent agreement between the physical and statistical times is a byproduct of human choice of linguistic primitives and, moreover, that the choice is compelled by a survival pressure to facilitate predictions at the expense of diagnosis and planning.

[11] ξ_t and ξ'_t are assumed to be two independent, white noise time series. Also $\alpha \neq \delta$ and $\gamma \neq \beta$.

[12] We are grateful to Seth Lloyd for this observation.

9. Conclusions

The theory presented in this paper should dispel the belief that statistical analysis can never distinguish genuine causation from spurious covariation. This belief, shaped and nurtured by generations of statisticians [Fisher, 1953, Keynes, 1939, Ling, 1983, Niles, 1922] has been a major hindrance in the way of developing a satisfactory, non-circular account of causation. In the words of Gärdenfors [Gärdenfors, 1988, page 193]:

> In order to distinguish genuine from spurious causes, we must already know the causally relevant background factors. ... Further, the extra amount of information is substantial: In order to determine whether C is a cause of E, *all* causally relevant background factors must be available. It seems clear that we often have determinate beliefs about causal relations between events, even if we do not know exactly which factors are causally relevant to the events in question[13].

This paper shows that such extra information is often unnecessary: Under the assumptions of model-minimality (and/or stability), there are patterns of dependencies that should be sufficient to uncover genuine causal relationships. These relationships cannot be attributed to hidden causes lest we violate one of the basic maxims of scientific methodology: the semantical version of Occam's razor. Adherence to this maxim may explain why humans reach consensus regarding the directionality and nonspuriousness of causal relationships, in the face of opposing alternatives, perfectly consistent with experience. Echoing Cartwright [Cartwright, 1989] we summarize our claim with the slogan "No Causes In, Some Causes Out".

From a methodological viewpoint, our theory should settle some of the ongoing disputes regarding the validity of path-analytic approaches to causal modeling in the social sciences [Freedman, 1987, Ling, 1983]. It shows that the basic philosophy governing path-analytic methods is legitimate, faithfully adhering to the traditional norms of scientific investigation. At the same time our results also explicate the assumptions upon which these methods are based, and the conditions that must be fulfilled before claims made by these methods can be accepted. Specifically, our analysis makes it clear that causal modeling must begin with *vanishing* (*conditional*) *dependencies* (i.e. missing links in their graphical representations). Models that embody no vanishing dependencies contain no virtual control variables, hence, the causal component of their claims cannot be substantiated by observational

[13]See also Cartwright [Cartwright, 1989] for a similar position, and for a survey of the literature.

studies. With such models, the data can be used only for estimating the parameters of the causal links once we are absolutely sure of the causal structure, but the structure itself, and especially the directionality of the links, cannot be inferred from the data. Unfortunately, such models are often employed in the social and behavioral sciences e.g. [Kenny, 1979].

On the practical side, we have shown that the assumption of model minimality, together with that of "stability" (no accidental independencies) lead to an effective algorithm of structuring candidate causal models capable of generating the data, transparent as well as latent. Simulation studies conducted at our laboratory show that networks containing tens of variables require less than 5000 samples to have their structure recovered by the algorithm. For example, 1000 samples taken from the process shown in Eq. (5), each containing ten successive X,Y pairs, were sufficient for recovering its double-chain structure (and the correct direction of time). The greater the noise, the quicker the recovery (up to a point).

Another result of practical importance is the following: Given a proposed causal theory of some phenomenon, our algorithm can identify in linear time those causal relationships that could potentially be substantiated by observational studies, and those whose directionality and non-spuriousness can only be determined by controlled, manipulative experiments.

It should also be interesting to explore how the new criteria for causation could benefit current research in machine learning. In some sense, our method resembles a search through a space of hypotheses [Mitchell, 1982] where each hypothesis stands for a causal theory. Unfortunately, this is where the resemblance ends. The prevailing paradigm in the machine learning literature has been to define each hypothesis (or theory, or concept) as a subset of observable instances; once we observe the entire extension of this subset, the hypothesis is defined unambiguously. This is not the case in causal modeling. Even if the training sample exhausts the hypothesis subset (in our case, this corresponds to observing P precisely), we are still left with a vast number of equivalent causal theories, each stipulating a drastically different set of causal claims. Fitness to data, therefore, is an insufficient criterion for validating causal theories. Whereas in traditional learning tasks we attempt to generalize from one set of instances to another, the causal modeling task is to generalize from behavior under one set of conditions to behavior under another set. Causal models should therefore be chosen by a criterion that challenges their stability against changing conditions, and these show up in the data in the form of virtual control variables. Thus, the dependence patterns identified by definition 11 through 14 constitute islands of stability as well as virtual validation tests for causal models. It would be interesting to examine whether these criteria, when incorporated into ex-

isting machine learning programs would improve the stability of theories discovered by such programs.

Acknowledgement

We are grateful to Clark Glymour for posing the problem of equivalence in latent structures. Some of the problems treated in this paper were independently explored by Glymour, Spirtes and Scheines [Spirtes et al., 1989, Spirtes and Glymour, 1991], and we thank them for calling our attention to an oversight in an earlier formulation of the IC-algorithm. Discussions and correspondence with P. Bentler, D. Geiger, C. Granger, M. Hanssens, J. de Leeuw, S. Lloyd, R. Otte, A. Paz, B. Skyrms and P. Suppes are greatly appreciated. This work was supported in part by NSF grant #IRI-9200918, AFOSR grant #900136, and MICRO grants #91-123/4.

References

[Bobrow, 1985] BOBROW, D. (1985). *Qualitative Reasoning about Physical Systems.* MIT Press, Cambridge, MA.

[Cartwright, 1989] CARTWRIGHT, N. (1989). *Nature Capacities and Their Measurements.* Clarendon Press, Oxford.

[Cliff, 1983] CLIFF, N. (1983). *Some cautions concerning the application of causal modeling methods.* Multivariate behavioral research, 18:115 – 126.

[de Kleer and Brown, 1986] DE KLEER, J. and BROWN, J. S. (1986). *Theories of causal ordering.* Artificial Intelligence, 29(1):33 – 62.

[Dechter and Pearl, 1991] DECHTER, R. and PEARL, J. (1991). *Directional constraint networks: A relational framework for causal modeling.* In Proceedings, 12th International Joint Conference on Artificial Intellingence (IJCAI - 91), Sydney, Australia, August, 1991, 1164-1170.

[Eells and Sober, 1983] EELLS, E. and SOBER, E. (1983). *Probabilistic causality.* Philosophy of Science, 50:35 – 57.

[Fisher, 1953] FISHER, R. A. (1953). *Design of Experiments.* Oliver and Boyd, London.

[Forbus and Gentner, 1986] FORBUS, K. D. and GENTNER, D. (1986). *Causal reasoning about quantities.* Proceedings Cognitive Science Society, pages 196 – 207.

[Freedman, 1987] FREEDMAN, D. (1987). *As others see us: A case study in path analysis (with discussion).* Journal of Educational Statistics, 12:101 – 223.

[Frydenberg, 1989] FRYDENBERG, M. (1989). *The chain graph markov property.* Technical Report 186, Department of Theoretical Statistics, University of Aarhus, Denmark.

[Gärdenfors, 1988] GÄRDENFORS, P. (1988). *Causation and the dynamics of belief.* In Harper, W. and Skyrms, B., editors, Causation in Decision, Belief Change and Statistics II, pages 85 – 104. Kluwer Academic Publishers.

[Geffner, 1989] GEFFNER, H. (1989). *Default Reasoning: Causal and Conditional Theories.* PhD thesis, UCLA Computer Science Department, Los Angeles, CA. Also, MIT Press.

[Geiger et al., 1990] GEIGER, D., PAZ, A., and PEARL, J. (1990). *Learning causal trees from dependence information.* In Proceedings, AAAI-90, pages 770 – 776, Boston, MA.

[Glymour et al., 1987] GLYMOUR, C., SCHEINES, R., SPIRTES, P., and KELLY, K. (1987). *Discovering Causal Structure.* Academic Press, New York.

[Goldszmidt and Pearl, 1992] GOLDSZMIDT, M. and PEARL, J. (1992). *Rank-based systems: A simple approach to belief revision, belief update, and reasoning about evidence and actions.* In Proceedings of the Third International Conference on Knowledge Representation and Reasoning, Cambridge, MA, October 1992.

[Good, 1983] GOOD, I. J. (1983). *A causal calculus.* British Journal for Philosophy of Science, 11 and 12 and 13:305 – 328 and 43 – 51 and 88. reprinted as Ch. 21 in Good Thinking University of Minnesota Press, Minneapolis, MN.

[Granger, 1988] GRANGER, C. W. J. (1988). *Causality testing in a decision science.* In Harper, W. and Skyrms, B., editors, Causation in Decision, Belief Change and Statistics I, pages 1 – 20. Kluwer Academic Publishers.

[Holland, 1986] HOLLAND, P. (1986). *Statistics and causal inference.* Journal of the American Statistical Association, 81:945 – 960.

[Iwasaki and Simon, 1986] IWASAKI, Y. and SIMON, H. A. (1986). *Causality in device behavior.* Artificial Intelligence, 29(1):3 – 32.

[Kautz, 1987] KAUTZ, H. (1987). *A formal Theory of Plan Recognition.* PhD thesis, University of Rochester, Rochester, N.Y.

[Kenny, 1979] KENNY, D. A. (1979). *Correlation and Causality.* Wiley, New York.

[Keynes, 1939] KEYNES, J. M. (1939). *Professor Tinbergen's method.* Economic Journal, 49:560.

[Lifschitz, 1987] LIFSCHITZ, V. (1987). *Formal theories of action.* In Workshop of the Frame Problem in AI, pages 35 – 57, Kansas.

[Ling, 1983] LING, R. (1983). *Review of Correlation and Causation* by D. Kenny. Journal of the American Statistical Association, pages 489 – 491.

[Mitchell, 1982] MITCHELL, T. M. (1982). *Generalization as search.* Artificial Intelligence, 18:203 – 226.

[Niles, 1922] NILES, H. E. (1922). *Correlation, causation, and Wright theory of "path coefficients".* Genetics, 7:258 – 273.

[Patil et al., 1982] PATIL, R. S., SZOLOVITZ, P., and SCHWARTZ, W. B. (1982). *Causal understanding of patient illness in patient diagnosis.* In Proceedings of AAAI-82, pages 345 – 348.

[Pearl, 1978] PEARL, J. (1978). *On the connection between the complexity and credibility of inferred models.* International Journal of General Systems, 4:255 – 264.

[Pearl, 1988a] PEARL, J. (1988A). *Embracing causality in formal reasoning.* Artificial Intelligence, 35(2):259 – 71.

[Pearl, 1988b] PEARL, J. (1988B). *Probabilistic Reasoning in Intelligent Systems.* Morgan-Kaufman, San Mateo, CA.

[Pearl, 1990] PEARL, J. (1990). *Probabilistic and qualitative abduction.* In Proceedings of AAAI Spring Symposium on Abduction, pages 155 – 158, Stanford.

[Pearl et al., 1989] PEARL, J., GEIGER, D., and VERMA, T. S. (1989). *The logic of influence diagrams.* In Oliver, R. M. and Smith, J. Q., editors, Influence Diagrams, Belief Networks and Decision Analysis, pages 67 – 87. John Wiley and Sons, Ltd., Sussex, England.

[Popper, 1959] POPPER, K. R. (1959). *The Logic of Scientific Discovery.* Basic Books, New York.

[Reichenbach, 1956] REICHENBACH, H. (1956). *The Direction of Time.* University of California Press, Berkeley.

[Reiter, 1987] REITER, R. (1987). *A theory of diagnosis from first principles.* Artificial Intelligence, 32(1):57 – 95.

[Rubin, 1989] RUBIN, H. (1989). *Discussion of "The Logic of Influence Diagrams" by Pearl at al.* In Oliver, R. M. and Smith, J. Q., editors, Influence Diagrams, Belief Networks and Decision Analysis, pages 83 – 85. John Wiley and Sons, Ltd., Sussex, England.

[Salmon, 1984] SALMON, W. (1984). *Scientific explanation and the causal structure of the world.* Princeton University Press., Princeton.

[Shoham, 1988] SHOHAM, Y. (1988). *Reasoning About Change.* MIT Press, Boston, MA.

[Simon, 1954] SIMON, H. (1954). *Spurious correlations: A causal interpretation.* Journal American Statistical Association, 49:469 – 492.

[Skyrms, 1980] SKYRMS, B. (1980). *Causal Necessity.* Yale University Press, New Haven, CT.

[Spirtes and Glymour, 1991] SPIRTES, P. and GLYMOUR, C. (1991). *An algorithm for fast recovery of sparse causal graphs.* Social Science Computer Review, 9:92-111.

[Spirtes et al., 1989] SPIRTES, P., GLYMOUR, C., and SCHEINES, R. (1989). *Causality from probability.* Technical Report CMU-LCL-89-4, Department of Philosophy Carnegie-Mellon University.

[Spohn, 1983] SPOHN, W. (1983). *Deterministic and probabilistic reasons and causes.* Erkenntnis, 19:371 – 396.

[Suppes, 1970] SUPPES, P. (1970). *A Probabilistic Theory of Causation.* North Holland, Amsterdam.

[Suppes and Zaniotti, 1981] SUPPES, P. and ZANIOTTI, M. (1981). *When are probabilistic explanations possible?* Synthese, 48:191 – 199.

[Verma, 1992] VERMA, T. S. (1992). *Causal Modeling: A graph-theoretic approach.* PhD dissertation, UCLA Computer Science Department, Los Angeles, CA (In preparation).

[Verma and Pearl, 1990] VERMA, T. S. and PEARL, J. (1990). *Equivalence and synthesis of causal models.* In Proceedings 6th Conference on Uncertainty in AI, pages 220 – 227.

[Wilensky, 1983] WILENSKY, R. (1983). *Planning and understanding.* Addison Wesley.

[Wright, 1925] WRIGHT, S. (1925). *Corn and hog correlations.* Technical Report 1300, U.S. Department of Agriculture.

Logic, Methodology and Philosophy of Science IX
D. Prawitz, B. Skyrms and D. Westerståhl (Editors)

BUILDING CAUSAL GRAPHS FROM STATISTICAL DATA IN THE PRESENCE OF LATENT VARIABLES

PETER SPIRTES*

Department of Philosophy, Carnegie-Mellon University

1. Introduction

The problem of inferring causal relations from statistical data in the absence of experiments arises repeatedly in many scientific disciplines, including sociology, economics, epidemiology, and psychology. In addition, the building of expert systems could be expedited if background knowledge elicited from experts could be supplemented with automated techniques using relevant statistics.

Recently, efficient algorithms for determining causal relationships between random variables (in the form of Bayesian networks) from appropriate statistical data when there are no unmeasured or "latent" variables have been discovered. (See Spirtes, Glymour and Scheines 1990, Spirtes and Glymour 1991, Verma and Pearl 1990, and Pearl and Verma 1991.) Inferring causal relations when unmeasured variables are also acting is a much more difficult problem. In many cases it is impossible to infer the structure among the latent variables from statistical relations among the measured variables. But the presence of latent variables can also make it difficult to infer the causal relations among the measured variables themselves. When only two variables, A and B, have been measured, and there is a correlation between the two, this does not suffice to establish whether A causes B, B causes A, or there is a third unmeasured variable causing both A and B. Nevertheless, when other variables are measured, more knowledge about the causal relations between A and B is possible. We will prove in Theorem 2 that there are some circumstances in which it is possible to establish that A causes B,

*I thank C. Glymour and R. Scheines for valuable help with the work described here. Research for this paper was supported by the Naval Personnel Research and Development Center and the Office of Naval Research under contract number N00114-89-J-1964.

rather than that B causes A, or that a third unmeasured variable causes both A and B; and we will prove in Theorem 3 that there are other circumstances in which the possibility that A causes B can be eliminated. The proofs are given in Spirtes, Glymour and Scheines (forthcoming).

2. Directed acyclic graphs

Causal processes among a set of random variables $\mathbf{V}$ are represented by a directed acyclic graph over $\mathbf{V}$, where there is an edge from A to B if and only if A is an immediate cause of B relative to $\mathbf{V}$ (i.e. there is a mechanism by which A causes B that is not blocked by holding fixed any of the other variables in $\mathbf{V}$.) If there is a directed path from A to B in the causal graph, we will say that A is a (possibly indirect) cause of B. (In what follows, we will capitalize random variables, and boldface any sets of variables. We will use the terms "vertices in a graph" and "variables in a graph" interchangably.)

A directed acyclic graph G over a set of random variables $\mathbf{V}$ can also be used to represent the set of probability distributions over $\mathbf{V}$ that satisfy the following two conditions:

Markov Condition: Let $\mathbf{Parents}(X)$ be the set of parents of X in G (i.e. the set of Z such that $Z -> X$ is in G) and $\mathbf{Descendants}(X)$ be the set of descendants of X in a graph G (i.e. the set of Z such that there is a directed path from X to Z in G.) A directed acyclic graph G and a probability distribution P on the vertices $\mathbf{V}$ of G satisfy the Markov condition if and only if for every X in $\mathbf{V}$, X and $\mathbf{V}\backslash(\{X\} \cup \mathbf{Descendants}(X))$ are independent conditional on $\mathbf{Parents}(X)$.

Faithfulness Condition: If G is a directed acyclic graph and P is a distribution over the set of vertices $\mathbf{V}$ in G, then P is faithful to G if and only $< G, P >$ satisfy the Markov condition and every conditional independence relation true in P is entailed by the Markov condition for G.

If a distribution is placed over the exogenous variables (variables of zero indegree) in the causal graph of a causal process, which in turn affect the values of other random variables, the result is a joint distribution over all of the random variables. In that case, we will say that the causal process generated the joint distribution. We assume that the distribution generated by a causal process satisfies the Markov and Faithfulness conditions for the causal graph of that process; we will call this the *Causal Faithfulness Assumption.* In Pearl's terminology (Pearl 1988) the causal graph is a *Bayesian network* of any distribution that it generates.

In a directed graph G, we will write $X -> Y$ if there is an edge from X to Y in G, and we will say that X is *parent* of Y. X and Y are *adjacent in a*

directed graph G if and only if either $X -> Y$ or $Y -> X$ in G. If X and Y are adjacent in G, we will also say that X *is a neighbor of* Y and Y *is a neighbor of* X. In a directed acyclic graph G, an *undirected path* U *from* X *to* Y is a sequence of vertices starting with X and ending with Y such that for every pair of vertices A and B that are adjacent to each other in the sequence, A and B are adjacent in G, and no vertex occurs more than once in U. In a directed acyclic graph G, a *directed path* P *from* X *to* Y is a sequence of vertices starting with X and ending with Y such that for every pair of variables A and B that are adjacent to each other in the sequence in that order, the edge $A -> B$ occurs in G, and no vertex occurs more than once in P. X and Y are *adjacent on path* P (as distinct from adjacent in the graph) if and only if X and Y are adjacent in the sequence P. An *edge between* X *and* Y *occurs in a path* P (directed or undirected) if and only if X and Y are adjacent in P. If an undirected path U contains an edge between X and Y, and an edge between Y and Z, the two edges *collide* at Y if and only if $X -> Y$ and $Z -> Y$ in G. On an undirected path U, Y is a *collider* if and only if there exist edges $X -> Y$ and $Z -> Y$ in U; Y is an *unshielded collider* on U if and only if in addition Z and X are not adjacent in G. X is an *ancestor* of Y and Y is a *descendant* of X if and only if there is a directed path from X to Y. (We count the sequence consisting of a single vertex $< X >$ as a directed path from X to X, so X is its own ancestor and descendant, although it is not its own parent or child.) X, Y, and Z form *triangle* $X - Y - Z$ in G if and only if X is adjacent to Y, Y is adjacent to Z, and Z is adjacent to X in G. A *trek* between X and Y is either a directed path from X to Y, a directed path from Y to X, or a pair of directed paths from some third variable Z to X and Y respectively that intersect only at Z.

Verma and Pearl (see Pearl 1988) have shown how to calculate the conditional independence relations that are entailed by distributions satisfying the Markov condition for a graph G using the d-separability relation. In graph G, a path U between X and Y *d-connects variables* X *and* Y *given a set of vertices* $\mathbf{S}$ not containing X or Y if and only if (i) every collider on U has a descendent in $\mathbf{S}$ and (ii) no other vertex on U is in $\mathbf{S}$. Vertices X *and* Y *are d-separated given a set* $\mathbf{S}$ not containing X and Y if and only if no path d-connects X and Y given $\mathbf{S}$. Disjoint sets of vertices $\mathbf{X}$ and $\mathbf{Y}$ are d-separated given $\mathbf{S}$ in G if and only if every member of $\mathbf{X}$ is d-separated from every member of $\mathbf{Y}$ given $\mathbf{S}$ in G. If distribution P satisfies the Markov and Faithfulness Conditions, then for disjoint sets of vertices $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$, $\mathbf{X}$ is independent of $\mathbf{Y}$ conditional on $\mathbf{S}$ if and only if $\mathbf{X}$ is d-separated from $\mathbf{Y}$ given $\mathbf{S}$ in G (Pearl 1988).

We say that $\mathbf{V}$ is *causally sufficient* if and only if every cause of any two

members of $\mathbf{V}$ is itself in $\mathbf{V}$. If the distribution P is generated by some causal process, then given the Causal Faithfulness Assumption, P is faithful to some directed acyclic graph; however, if a set of measured variables $\mathbf{V}$ is not causally sufficient, the marginal distribution of P over $\mathbf{V}$ may not be faithful to any directed acyclic graph. Our strategy for making inferences about causal relationships when latent variables may be present is to find properties held in common by all directed acyclic graphs that have faithful distributions for which P could be the marginal.

3. Spurious causal dependencies

In a directed acyclic graph G over a set of variables $\mathbf{V}$, if A and B are adjacent in G, then A and B are not d-separated by any subset of $\mathbf{V}\backslash\{A, B\}$. Hence under the assumption of causal sufficiency, either A is a direct cause of B or B is a direct cause of A relative to $\mathbf{V}$ if and only if A and B are independent conditional on no subset of $\mathbf{V}$. (Recently more efficient and reliable algorithms for determining causal structure from statistical data when there are no latent variables have been devised. See Spirtes, Glymour and Scheines 1990, Spirtes and Glymour 1991.) At first glance, it appears that this technique can be generalized to the case where $\mathbf{V}$ is not causally sufficient by inferring from the dependence of A and B conditional on every subset of $\mathbf{V}\backslash\{A, B\}$ that either A is a direct cause of B relative to $\mathbf{V}$, or B is a direct cause of A relative to $\mathbf{V}$, or there is some latent variable L that is a common cause of both A and B. Unfortunately, this is not the case, as the following example shows.

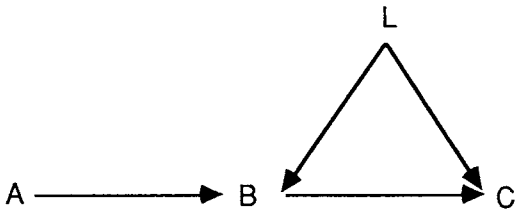


Figure 1: Graph G

Let $\mathbf{V} = \{A, B, C, L\}$ and $\mathbf{O} = \{A, B, C\}$. $\mathbf{O}$ is not causally sufficient because L is a cause of both B and C which are in $\mathbf{O}$, but L itself is not in $\mathbf{O}$. A and C are not d-separated by any subset of $\mathbf{O}\backslash\{A, B\}$, so in any marginal of a distribution faithful to G, A and C are not independent conditional on

any subset of $\mathbf{O}\backslash\{A, B\}$. Nevertheless, A is not a direct cause of C relative to $\mathbf{O}$, C is not a direct cause of A relative to $\mathbf{O}$, and there is no latent common cause of A and C. If the algorithms for the causally sufficient case where applied to $\mathbf{O}$, they would find, erroneously that A and C are adjacent. Our problem is to find a more reliable procedure.

4. Inducing paths

Given a directed acyclic graph G over a set of variables $\mathbf{V}$, and $\mathbf{O}$ a subset of $\mathbf{V}$, Verma and Pearl (1990) have characterized the conditions under which two variables in $\mathbf{O}$ are not d-separated by any subset of $\mathbf{O}\backslash\{A, B\}$. In a directed acyclic graph G over a set of variables $\mathbf{V}$, an undirected path U between A and B is an *inducing path over a subset* $\mathbf{O}$ of $\mathbf{V}$ if and only if every member of $\mathbf{O}$ on U is a collider on U, and every collider on U is an ancestor of either A or B. (We will sometimes refer to members of $\mathbf{O}$ as *observed* variables.)

THEOREM 1 *In a directed acyclic graph G over $\mathbf{V}$, where $\mathbf{O}$ is a subset of $\mathbf{V}$, A and B are not d-separated by any subset of $\mathbf{O}\backslash\{A, B\}$ if and only if there is an inducing path over the subset $\mathbf{O}$ between A and B.*

In Figure 1, the inducing path between A and C over $\mathbf{O} = \{A, B, C\}$ is $< A, B, L, C >$.

5. Inducing path graphs

The inducing paths relative to $\mathbf{O}$ in a graph G over $\mathbf{V}$ can be represented in the following structure described (but not named) in Pearl and Verma (1990). In an *inducing path graph G' for directed acyclic graph G over a subset of variables* $\mathbf{O}$ there is an edge between variables A and B with an arrowhead at A if and only if A and B are in $\mathbf{O}$, and there is an inducing path in G between A and B relative to $\mathbf{O}$ that is into A (i.e. there is an edge in the path with an arrowhead into A.) Note that in an inducing path graph, there are two kinds of edges: $A -> B$ entails that every inducing path over $\mathbf{O}$ between A and B is out of A and into B, and $A <-> B$ entails that there is an inducing path over $\mathbf{O}$ that is into A and into B. This latter kind of edge can only occur when there is a latent common cause of A and B.

We can extend the concept of d-separability to inducing path graphs without modification, if we interpret directed paths in inducing path graphs as paths containing only edges with one arrowhead, and undirected paths as

containing edges with either single or double arrowheads. If G is a directed acyclic graph, G' is the inducing path graph for G over $\mathbf{O}$, and $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{S}$ are disjoint sets of variables included in $\mathbf{O}$, then $\mathbf{X}$ and $\mathbf{Y}$ are d-separated by $\mathbf{S}$ in G' if and only they are d-separated by $\mathbf{S}$ in G.

However, there is one very important difference between d-separability relations in an inducing path graph and in a directed acyclic graph due to the existence of double-headed arrows in the former. In a directed acyclic graph over $\mathbf{O}$, if A and B are d-separated by any subset of $\mathbf{O}\backslash\{A, B\}$ then A and B are d-separated either by $\mathbf{Parents}(A)$ or $\mathbf{Parents}(B)$. This is not true in inducing path graphs. However, we have shown the following.

Let $NA(A, B)$ (mnemonic for non-ancestor) be A if A is not an ancestor of B, and otherwise let it be B. (In an acyclic graph either A is not an ancestor of B or B is not an ancestor of A, so the vertex that is $NA(A, B)$ is not an ancestor of the other vertex.)

If G' is an inducing path graph, and $A \neq B$, V is a member of $\mathbf{D-SEP}(A, B)$ if and only if $A \neq V$ and there is an undirected path U between $NA(A, B)$ and V such that every vertex on U is an ancestor of A or B, and (except for the end points) is a collider on V.

LEMMA 1 *In an inducing path graph G', if A and B are not adjacent then A and B are d-separated by* $\mathbf{D\text{-}SEP}(A, B)$.

The importance of this fact is that we can determine whether A and B are adjacent in an inducing path graph without determining whether A and B are dependent conditional on *all* subsets of $\mathbf{O}$.

If $\mathbf{O}$ is not a causally sufficient set of variables, then although we can infer the existence of an inducing path between A and B if A and B are dependent conditional on every subset of $\mathbf{O}\backslash\{A, B\}$, we cannot infer that either A is a direct cause of B relative to $\mathbf{O}$, B is a direct cause of A relative to $\mathbf{O}$, or there is a latent common cause of A and B. Nevertheless, the existence of an inducing path between A and B relative to $\mathbf{O}$ does contain information about the causal relationships between A and B, as the following lemma shows.

LEMMA 2 *If G is a directed acyclic graph over $\mathbf{V}$ that contains an inducing path relative to $\mathbf{O}$ (included in $\mathbf{V}$) between A and B that is out of A, then there is a directed path from A to B in G.*

It follows from lemma 2 that if $\mathbf{O}$ is a subset of $\mathbf{V}$ and we can determine that there is an inducing path between A and B relative to $\mathbf{O}$ that is out of A, then we can infer that A is a (possibly indirect) cause of B. Hence, if we can infer properties of the inducing path graph over $\mathbf{O}$ from the distribution over

O, we can draw inferences about the causal relationships between variables, regardless of what variables we have failed to measure. In the next section we describe algorithms for inferring properties of the inducing path graph over **O** from the distribution over **O**.

6. Partially oriented inducing path graphs

There are four kinds of edges in a *partially oriented inducing path graph*: $A -> B$, $A \circ\!\!-\!\!> B$, $A \circ\!\!-\!\!\circ B$, and $A <-> B$. We use "$*$" as a metasymbol to represent any of the three kinds of ends (nothing, ">", or "$\circ$") that an edge in a partially oriented inducing path graph can have; the "$*$" symbol itself does not appear in a partially oriented inducing path graph. (We also use "$*$" as a metasymbol to represent the two kinds of ends (nothing or ">") that can occur in an inducing path graph.)

A partially oriented inducing path graph π for directed acyclic graph G with inducing path graph G' over **O** is intended to represent the adjacencies in G', and the orientation of the edges in G' that are common to all inducing path graphs with the same d-connection relations as G'. Let $E(G')$ be the set of inducing path graphs over the same vertices with the same d-connections as G'. It is easy to see that every inducing path graph in $E(G')$ shares the same set of adjacencies.

π is a a partially oriented inducing path graph of directed acyclic graph G with inducing path graph G' over **O** if and only if

1. π and G' have the same vertices, and
2. π and G' have the same adjacencies, and
3. if $A \circ\!\!-\!\!> B$ in π, then $A -> B$ or $A <-> B$ in every inducing path graph in $E(G')$, and
4. if $A -> B$ in π, then $A -> B$ in every inducing path graph in $E(G')$, and
5. if $A *\!\!-\!\!\underline{*B*}\!\!-\!\!* C$ in π, then the edges between A and B, and B and C do not collide at B in any inducing path graph in $E(G')$;
6. if $A <-> B$ in π, then $A <-> B$ in every inducing path graph in $E(G')$.

Note that an edge $A *\!\!-\!\!\circ B$ places no constraints upon the edge between A and B being into or out of B in any subset of $E(G')$.

The adjacencies in an inducing path graph π for G can be constructed by making A and B adjacent in π if and only if A and B are d-connected given every subset of $\mathbf{O}\backslash\{A, B\}$. (If a distribution P is faithful to G, then this amounts to making A and B adjacent if and only if A and B are dependent in P conditional on every subset of $\mathbf{O}\backslash\{A, B\}$.) Once the adjacencies have been determined, it is trivial to construct a partially oriented inducing path graph π for G. Simply orient each edge $A *\!-\!* B$ as $A \circ\!-\!\circ B$. Of course this particular partially oriented inducing path graph π for G is very uninformative about what features of the orientation of G' are common to all inducing path graphs in $E(G')$.

In a maximally oriented partially oriented inducing path graph π for G, an edge $A *\!-\!\circ B$ would appear only if the edge between A and B were into B in some members of $E(G')$, and out of B in other members of $E(G')$. Such a maximally oriented partially oriented inducing path graph π for G could be oriented by the simple algorithm of constructing every possible inducing path graph with the same adjacencies as G', throwing out the ones that do not have the same d-connection relations as G', and keeping track of which orientation features are common to all members of $E(G')$. Of course, this is completely computationally infeasible.

Our goal is to state an algorithm that constructs a partially oriented inducing path graph for a directed acyclic graph G that contains as many orientations as possible, while remaining computationally feasible. The algorithm we propose is divided into two main parts. First, the adjacencies in the partially oriented inducing path graph are determined. Then the edges are oriented in so far as possible. In order to state the algorithm, several more definitions are needed.

In a partially oriented inducing path graph π:

1. B is a *definite non-collider* on undirected path U if and only if B is an endpoint of U or there exist vertices A and C on U such that either $A <\!- B *\!-\!* C$, $A *\!-\!* B -\!> C$, or $A *\!-\!* \underline{*B*} -\!* C$ on U.

2. A is a *parent* of B if and only if $A -\!> B$ in π.

3. B is a *collider* along path $< A, B, C >$ if and only if $A *\!-\!> B <\!-\!* C$ in π.

4. An edge between B and A is *into* A if and only if $A <\!-\!* B$ in π.

5. An edge between B and A is *out of* A if and only if $A -\!> B$ in π.

6. *E is a definite discriminating vertex for C with respect to triangle $A - B - C$ using path P and vertex B*, if and only either the edge

between A and C is into A and the edge between B and A is out of A, or the edge between A and C is out of A and the edge between B and A is into A, and E is a closest vertex to A such that

a. E is not adjacent to B, and

b. P is an undirected path from E to A not containing B or C, and every vertex between E and A is a collider or a definite non-collider, and

c. for every vertex V on P, if V' is adjacent to V on P and between V and A on P, then $V *\!\!-\!\!> V'$ in G, and

d. every vertex V on P between E and A is adjacent to B in G, and

e. except for the endpoints of P, if V is a collider on P then $V -\!\!> B$ in G, and if V is a definite non-collider on P, then $V <\!\!-* B$ in G.

Figure 2 illustrates the concept of a definite discriminating vertex.

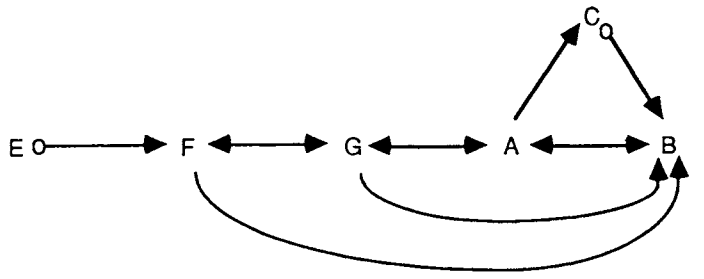


Figure 2: E is a definite discriminating vertex for C with respect to triangle $A - B - C$ using vertex B and path $< E, F, G, A >$.

Causal Inference Algorithm[1]

If G is a directed acyclic graph over $\mathbf{V}'$, and $\mathbf{V}$ is a subset of $\mathbf{V}'$, the input to the algorithm is the set of d-separation relations involving just members of $\mathbf{V}$ that is true in G. Let $\mathbf{A}_Q(A, B)$ denote the set of vertices adjacent to A or to B in graph Q, except for A and B themselves. (Since the algorithm is continually updating Q, $\mathbf{A}_Q(A, B)$ is constantly changing as the algorithm progresses.)

[1] As I explain in more detail in Section 8, the Causal Inference Algorithm uses some ideas from the Inductive Causation algorithm described in Pearl and Verma (1990).

A) Form the complete undirected graph Q on the vertex set $\mathbf{V}$.

B) If A and B are d-separated by any subset $\mathbf{S}$ of $\mathbf{V}$, remove the edge between A and B, and record $\mathbf{S}$ in $\mathbf{D}(A, B)$.

C) Let F be the graph resulting from step B. Orient each edge as $\circ\!-\!\circ$. For each triple of vertices A, B, C such that the pair A, B and the pair B, C are each adjacent in F but the pair A, C are not adjacent in F, orient $A *\!-\!* B *\!-\!* C$ as $A *\!-\!> B <\!-\!* C$ if and only if B is not in $\mathbf{D}(A, C)$, and orient $A *\!-\!* B *\!-\!* C$ as $A *\!-\!\underline{*B*} -\!* C$ if and only if B is in $\mathbf{D}(A, C)$.

D) repeat

> If there is an edge $A *\!-\!> B$, and an edge $B *\!-\!* C$, A and C are not adjacent, and there is no arrowhead into B, then orient $B *\!-\!* C$ as $B -\!> C$,
>
> else if there is a directed path from A to B, and an edge $A *\!-\!* B$, orient $A *\!-\!* B$ as $A *\!-\!> B$,
>
> else if V is a definite discriminating vertex for M using R in triangle $P - M - R$ then
>
> > if M is in $\mathbf{D}(V, R)$ then mark M as a non-collider on subpath $P *\!-\!\underline{*M*} -\!* R$
> > else orient $P *\!-\!* M *\!-\!* R$ as $P *\!-\!> M <\!-\!* R$.
>
> else if $P *\!-\!\underline{> M*} -\!* R$ then orient as $P *\!-\!> M -\!> R$.

until no more edges can be oriented.

Unfortunately, the Causal Inference Algorithm as stated is not practical for large numbers of variables because of the way the adjacencies are constructed. While it is theoretically correct to remove an edge between A and B from the complete graph if and only if A and B are d-separated by some subset of $\mathbf{O}\backslash\{A, B\}$, this is impractical for two reasons. First, there are too many subsets of $\mathbf{O}\backslash\{A, B\}$ on which to test the conditional independence of A and B. Second, for discrete distributions, unless the sample sizes are enormous there are no reliable tests of independence of two variables conditional on a large set of other variables.

Remember, however, that in an inducing path graph if A and B are d-separated by any subset of $\mathbf{O}$, then they are d-separated by $\mathbf{D} - \mathbf{SEP}(A, B)$. Unfortunately, until we have actually constructed the inducing path graph we do not know which variables are in $\mathbf{D} - \mathbf{SEP}(A, B)$. Nevertheless, as the partially oriented inducing path graph is constructed, we can determine

that some variables are definitely not in $\mathbf{D}$ -- $\mathbf{SEP}(A, B)$. This reduces the number and size of the subsets of $\mathbf{O}$ that have to be checked in order to determine whether A and B are adjacent in the inducing path graph.

We will determine which edges to remove from the complete graph in three stages. First, we will remove the edge between A and B if they are independent conditional on subsets of neighbors of A and B. This will eliminate many, but perhaps not all of the edges that are not in the inducing path graph. Second, we will orient edges by determining whether they collide or not. Third, using the partially oriented inducing path graph π that we have constructed thus far, we will form two sets of vertices **Possible-D-SEP**(A, B, π), and **Possible-D-SEP**(B, A, π) one of which includes every vertex that could possibly be in **D-SEP**(A, B). (We need two such sets because we cannot determine from the partially oriented inducing path graph constructed thus far whether A is a descendant of B or B is a descendant of A.) Finally, we will remove the edge between A and B if A and B are independent conditional on any subset of either **Possible-D-SEP**(A, B, π) or **Possible-D-SEP**(B, A, π). Once we have obtained the correct set of adjacencies, we will unorient all of the edges, and then proceed to re-orient them. For a given partially constructed partially oriented inducing path graph π, **Possible-D-SEP**(A, B, π) is defined as follows.

If $A \neq B$ in a partially oriented inducing path graph π, V is in **Possible-D-SEP**(A, B, π) if and only if $V \neq A$ and there is an undirected path U between A and V in π such that for every subpath $< X, Y, Z >$ of U either Y is a collider on U, or Y is not a definite non-collider on U and X, Y and Z form a triangle in π.

Fast Causal Inference Algorithm

If G is a directed acyclic graph over $\mathbf{V}'$, and $\mathbf{V}$ is a subset of $\mathbf{V}'$, the input to the algorithm is the set of d-separation relations involving just members of $\mathbf{V}$ that is true of G. Let $\mathbf{A}_Q(A, B)$ denote the set of vertices adjacent to A or to B in graph Q, except for A and B themselves. (Since the algorithm is continually updating Q, $\mathbf{A}_Q(A, B)$ is constantly changing as the algorithm progresses.)

A.) Form the complete undirected graph Q on the vertex set $\mathbf{V}$.

B.) n = 0.
repeat
repeat
select a pair of variables X and Y that are adjacent in Q such that $\mathbf{A}_Q(X, Y)$ has cardinality greater than or equal to n, and a subset $\mathbf{S}(X, Y)$ of $\mathbf{A}_Q(X, Y)$ of cardinality n, and if X and Y are d-separated by some subset of $\mathbf{S}(X, Y)$ delete the edge between X and Y from Q, and record the subset in $\mathbf{D}(X, Y)$
until all variable pairs X and Y such that $\mathbf{A}_Q(X, Y)$ has cardinality greater than or equal to n and all subsets $\mathbf{S}(X, Y)$ of $\mathbf{A}_Q(X, Y)$ of cardinality n are exhausted.
n = n + 1.
until for each pair of adjacent vertices X, Y, $\mathbf{A}_Q(X, Y)$ is of cardinality less than n.

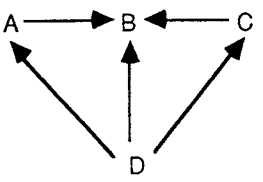
C.) Let F' be the graph resulting from step B. Orient each edge as $\circ\!-\!\circ$. For each triple of vertices A, B, C such that the pair A, B and the pair B, C are each adjacent in F' but the pair A, C are not adjacent in F', orient $A *\!-\!* B *\!-\!* C$ as $A *\!-\!> B <\!-\!* C$ if and only if B is not in $\mathbf{D}(A, C)$, and orient $A *\!-\!* B *\!-\!* C$ as $A *\!-\!* \underline{B*} -\!* C$ if and only if B is in $\mathbf{D}(A, C)$.

D.) For each pair of variables A and B connected by an edge in F', if A and B are d-separated by any subset of **Possible-D-SEP**$(A, B, F')\backslash\{A, B\}$ or any subset of **Possible-D-SEP**$(B, A, F')\backslash\{A, B\}$ remove the edge between A and B.

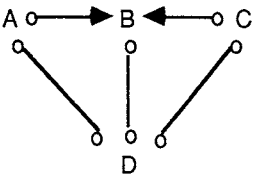
The algorithm then orients an edge between any pair of variables X and Y as $X \circ\!-\!\circ Y$, and proceeds to re-orient the edges in the same way as steps C and D of the Causal Inference Algorithm. The correctness of the algorithm is proved in Spirtes, Glymour and Scheines (forthcoming).

The example in Figure 3 shows that the algorithm is not complete, i.e. there are edges that are not oriented by the algorithm, whose orientation is common to all of the inducing path graphs with the same d-connection relations as the inducing path graph G.

In Figure 3, the edge between D and B is not oriented by the Fast Causal Induction Algorithm, even though there is an arrowhead at B in every inducing path graph with the same d-connection relations as G. We could of course simply add another orienting rule to handle this case.



Inducing Path Graph G



Partially Oriented Inducing Path

Figure 3:

7. Preservation theorems

Using the partially oriented inducing path graph output by the Fast Causal Inference Algorithm, and the inferences about graphs that can be drawn from inducing path graphs, we have the following two results. If the d-connections true of a directed acyclic graph G and involving just the variables in $\mathbf{O}$ are the input to the Causal Inference Algorithm, we will refer to the partially oriented inducing path graph over $\mathbf{O}$ output by the Causal Inference Algorithm as the CI partially oriented inducing path graph of G over $\mathbf{O}$.

THEOREM 2 *If π is a partially oriented inducing path graph of directed acyclic graph G over $\mathbf{O}$, and there is a directed path U from A to B in π, then there is a directed path from A to B in G.*

A *semi-directed path from A to B* in partially oriented inducing path graph π is an undirected path from A to B in which no edge contains an arrowhead pointing towards A (i.e. if X and Y are adjacent on the path, and X is between A and Y on the path, then there is no arrowhead at the X end of the edge between X and Y and there is no arrowhead at A.)

THEOREM 3 *If π is the CI partially oriented inducing path graph of a directed acyclic graph G over $\mathbf{O}$, and there is no semi-directed path from A to B in π, then there is no directed path from A to B in G.*

As an example of the application of the Fast Causal Inference Algorithm, suppose that the causal structure depicted in Figure 4 is the true causal structure among a set of variables related to breathing dysfunction, and that all of the variables except those in boxes, (Environmental Pollution and Genotype) are measured. (I am not proposing this graph as a model of breathing dysfunction; we constructed it merely to illustrate the application of Theorems 2 and 3.) The partially oriented inducing graph over the

measured variables constructed by the Fast Causal Inference Algorithm is depicted in Figure 5.

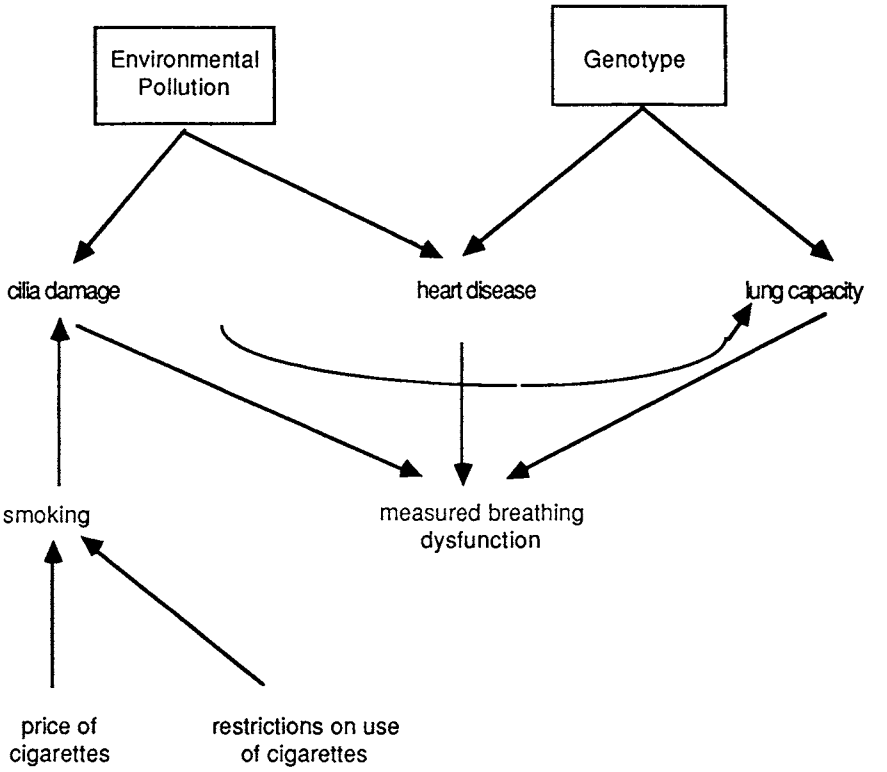


Figure 4: Causal Graph of Breathing Dysfunction

By applying Theorem 2, we infer that smoking does cause breathing dysfunction. By applying Theorem 3, we infer that smoking does not cause heart disease.

Note that in order to infer that smoking causes breathing dysfunction, it is necessary to measure two causes of smoking (whose collision at smoking orients the edge from smoking to cilia damage.) In general, this suggests that in the design of studies intended to determine if there is a causal path from variable A to variable B, it is useful to measure not only variables that might mediate the connection between A and B, but also to measure possible causes of A.

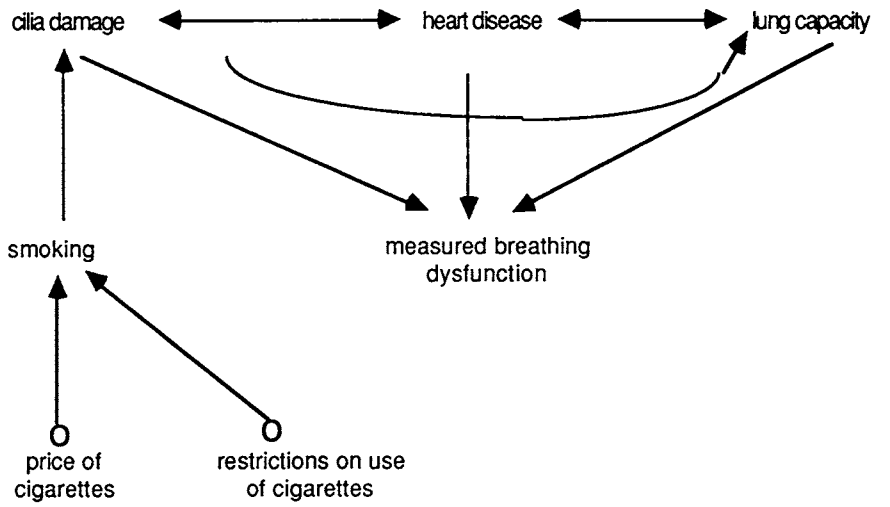


Figure 5: Partially Oriented Inducing Graph of Breathing Dysfunction Over Measured Variables

8. Historical note

In a series of papers (Pearl and Verma 1990, Pearl and Verma 1991, Verma and Pearl 1990, and Verma and Pearl 1991) Verma and Pearl describe an "Inductive Causation" algorithm that outputs a structure that they call a *pattern* (or sometimes a *completed hybrid graph*) of a directed acyclic graph G over a set of variables $\mathbf{O}$. Their algorithm differs from the Causal Inference Algorithm in two main respects. First, early versions of the algorithm did not distinguish between $A -> B$ and $A \circ\!\!-\!\!> B$; this distinction was introduced (in a different notation) in Spirtes and Glymour (1990a). Second it does not use definite discriminating vertices to orient any edges. (And unlike the Fast Causal Inference Algorithm it cannot be applied to large numbers of variables because it requires testing the independence of some pairs of variables conditional on every subset of $\mathbf{O}\backslash\{A, B\}$.) The most complete description of their theory appears in Pearl and Verma (1990). The key ideas of an inducing path, an inducing path graph, and the proof of (what we call) Theorem 1 all appear in this paper. Unfortunately, the two main claims that they make about patterns in this paper are both false.

In order to state their claims we need the following definitions. A pattern over $\mathbf{O}$ contains three kinds of edges: directed edges (e.g. $A -> B$), undirected edges (e.g. $A - B$), and bi-directed edges (e.g. $A <-> B$.) Directed

paths and descendants are defined in a pattern the same way they are defined in acyclic directed graphs; however, an undirected path in a pattern can contain bidirected edges and undirected edges as well as directed edges. Edges between A and B, and B and C, collide at B on an undirected path in a pattern if both edges have arrowheads at B. A and B are *h-separated* by $\mathbf{S}$ in a pattern π if and only if there is no undirected path between A and B in which every collider has a descendant in $\mathbf{S}$, and no non-collider is in $\mathbf{S}$.

Verma and Pearl claimed first (lemma A.2 in their paper) that if π is the pattern of a directed acyclic graph G over $\mathbf{O}$, and A and B are in $\mathbf{O}$, for all $\mathbf{S}$ included in $\mathbf{O}$, A and B are h-separated by $\mathbf{S}$ in π if and only if A and B are d-separated by $\mathbf{S}$ in G. Their second claim (Theorem 2 in their paper) was that any two directed acyclic graphs with the same pattern over $\mathbf{O}$ were "equivalent", i.e. they entailed the same d-separation relations involving just variables in $\mathbf{O}$. The following example shows that both of these claims are false.

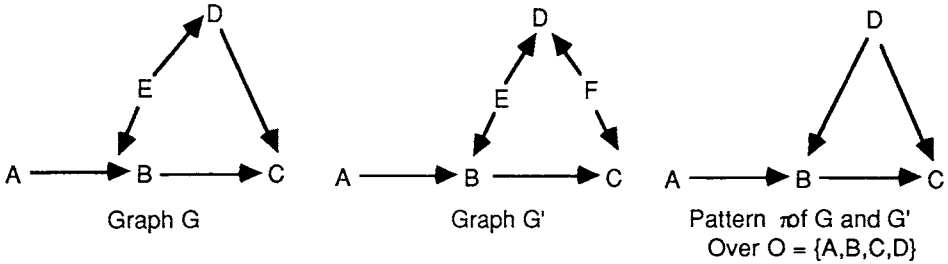


Figure 6:

According to the Verma-Pearl algorithm both G and G' in Figure 6 have the pattern π over $\mathbf{O} = \{A, B, C, D\}$ depicted in Figure 6. (The edge between C and D in the pattern is oriented from C to D in order to avoid a cycle involving B, C, and D.) However in G A and C are d-separated by $\{B, D\}$ but not by $\{B\}$, whereas in G' A and C are d-separated by $\{B\}$ but not by $\{B, D\}$. Hence G and G' have different d-separation relations among variables in $\mathbf{O}$ even though they have the same pattern. Moreover, A and C are h-separated by $\{B, D\}$ in the pattern of G', even though they are not d-separated by $\{B, D\}$ in G'.

Even though the patterns over $\mathbf{O}$ generated by the Verma-Pearl algorithm for G and G' are identical, the partially oriented inducing path graphs over $\mathbf{O}$ generated by the Causal Inference Algorithm for G and G' are different. This is because in both cases A is a definite discriminating vertex, and hence the edge between C and D is oriented differently in the CI partially oriented

inducing path graph of G and the CI partially oriented inducing path graph of G'. The output of the Causal Inference Algorithm for G and G' over $\mathbf{O}$ is depicted in Figure 7.

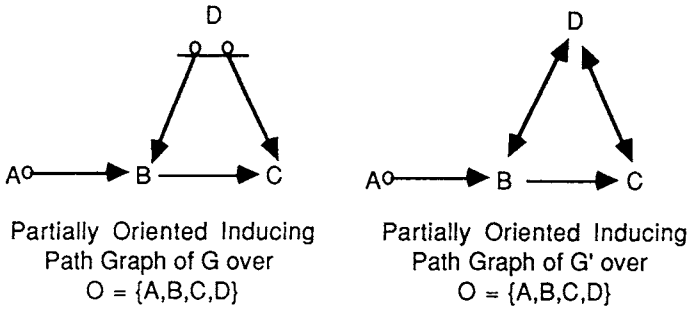


Figure 7:

While the proofs of Verma and Pearl's two main claims about patterns contained fallacies, we have used several of their proof techniques in our proofs.

References

J. PEARL, *Probabilistic Reasoning in Intelligent Systems*, Morgan Kaufmann, 1988.

J. PEARL and T. VERMA, *A Formal Theory of Inductive Causation*, Technical Report, R-155, Department of Computer Science, University of California at Los Angeles, October, 1990.

J. PEARL and T. VERMA, *A Theory of Inferred Causation*, in Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference, ed. by J. Allen, R. Fikes, and E. Sandewall, Morgan Kaufmann, San Mateo CA, 1991.

P. SPIRTES, C. GLYMOUR and R. SCHEINES, *Causality from Probability*, Proceedings of the Conference on Advanced Computing for the Social Sciences, Williamsburg, Va. 1990.

P. SPIRTES, C. GLYMOUR and R. SCHEINES, *Causality from Probability*, in G. McKee, ed. Evolving Knowledge in Natural and Artificial Intelligence, Pitman, 1990.

P. SPIRTES and C. GLYMOUR, *Causal Structure among Measured Variables Preserved with Unmeasured Variables*, Laboratory for Computational Linguistics Technical Report No. CMU-LCL-90-5, August, 1990a.

P. SPIRTES and C. GLYMOUR, *An Algorithm for Fast Recovery of Sparse Causal Graphs*, Social Science Computer Review, 9, 1991.

P. SPIRTES, C. GLYMOUR and R. SCHEINES, *Causality, Prediction and Search*, forthcoming from Springer-Verlag.

T. VERMA and J. PEARL, *On Equivalence of Causal Models*, Technical Report, R-150, Department of Computer Science, University of California at Los Angeles, April 1990.

T. VERMA and J. PEARL, *Equivalence and Synthesis of Causal Models*, Technical Report, R-150, Department of Computer Science, University of California at Los Angeles, March, 1991.

Logic, Methodology and Philosophy of Science IX
D. Prawitz, B. Skyrms and D. Westerståhl (Editors)

COHERENT INFERENCE AND PREDICTION IN STATISTICS

WILLIAM D. SUDDERTH*

School of Statistics, University of Minnesota, Minneapolis, MN 55455, U.S.A.

1. Introduction

The arguments of de Finetti and Ramsey, which justify the use of the probability calculus for certain numerical measures of degrees of belief, are, I think, at least as familiar to an audience of philosophers as to one of statisticians. Questions of "temporal coherence", whether Bayes' formula is only correct instantaneously or can be used to modify opinions through time, have, I gather, been widely debated by philosophers (cf. Skyrms (1990)). What may be new to this audience and what I will describe is the application of simple coherence arguments to standard problems in statistics.

After a brief review of some of de Finetti's ideas, I will introduce some standard terminology and then discuss two notions of coherence for statistical inference and predictions.

2. de Finetti's theory of coherence

Let $\mathcal{E}$ be a collection of events. For my purposes, $\mathcal{E}$ can be regarded as a collection of subsets of the set Ω of possible outcomes of some experiment like a die toss. (However, de Finetti (1974) has warned us about the uncritical acceptance of this identification of events and sets.) Suppose that, to each event A in $\mathcal{E}$, a bookie assigns a price $P(A)$ to a ticket worth one dollar if the actual outcome ω of the experiment is in A and worth nothing if ω is not in A. Thus the net payoff to a gambler who purchases such a ticket will be

$$A(\omega) - P(A) \tag{2.1}$$

*Research sponsored by National Science Foundation Grant DMS-8911548.

where $A(\omega)$ is 1 or 0 accordingly as ω belongs to A or not. To encourage "fair" prices, we require that the bookie be willing to buy as well as sell tickets so that a gambler can also have $-[A(\omega) - P(A)]$ as payoff. We also require that the bookie be willing to buy or sell arbitrary quantities of a given ticket. The resulting payoffs are of the form

$$a[A(\omega) - P(A)] \tag{2.2}$$

where a is a real number corresponding to the quantity of tickets on A purchased by the gambler. The total payoff to a gambler who buys a_i tickets on A_i for $i = 1, \ldots, n$ will be

$$\varphi(\omega) = \sum_{i=1}^{n} a_i[A_i(\omega) - P(A_i)]. \tag{2.3}$$

Call the bookie or the price function P *coherent* if there is no φ of the form above which is positive at every ω.

The basic result of de Finetti is that P is coherent if and only if P is consistent with a finitely additive probability measure. If $\mathcal{E}$ is an algebra of sets, there are elementary and entertaining arguments from de Finetti (1937) to prove that coherence is equivalent to the usual axioms:

(a) $P(\Omega) = 1$

(b) $0 \leq P(\Omega) \leq 1$

(c) $P(A \cup B) = P(A) + P(B)$ if $A \cap B = \emptyset$

If $\mathcal{E}$ is not an algebra, a coherent P can always be extended to an algebra so as to be a finitely additive probability.

Several authors (see, for example, Heath and Sudderth (1972) or Skyrms (1984)) have remarked that if a gambler is allowed to make countably many bets, then P must be countably additive to avoid a sure loss. However, de Finetti (1972, p. 91) considered such arguments to be circular because they rely on the usual conventions about infinite sums which are tantamount to an assumption of countable additivity.

Notice that the collection of payoff functions φ as in (2.3) form a linear space since any real multiple of such a payoff function is itself one and the sum of two such payoff functions is one also. Furthermore, if P is coherent, then the expectation (i.e. the integral) of every payoff φ is zero because, for φ as in (2.3),

$$E\varphi = \sum_{i=1}^{n} a_i[EA_i - P(A_i)] = \sum_{i=1}^{n} a_i \cdot 0 = 0.$$

This suggests a useful generalization.

Let Φ be a linear space of bounded, real-valued functions defined on Ω. Think of Φ as the collection of possible payoff functions from a bookie to a gambler and call the bookie *coherent* if there is no φ in Φ which has a positive infimum. (A φ as in (2.3) can have only finitely many values. If such a φ is everywhere positive, then it is bounded away from zero.) The result in this more general framework is that coherence is equivalent to the existence of a finitely additive probability measure P defined on all subsets of Ω such that $E\varphi(= \int \varphi dP)$ is zero for every φ in Φ . (This equivalence is an easy consequence of Lemma 1 in Heath and Sudderth (1978).)

The de Finetti theory treats conditional probability by the device of called–off bets. Suppose A and B are events and B is not empty. Let $P(A \mid B)$ be the bookie's price for a ticket worth one dollar if ω is in A but with the convention that the transaction is called off if ω is not in B. The net payoff for this called–off bet is

$$\psi(\omega) = B(\omega)[A(\omega) - P(A \mid B)]. \tag{2.4}$$

As above we require that multiples and finite sums of payoff functions be payoff functions. If the bookie is coherent, then there is a P such that

$$E\psi = 0.$$

But

$$\begin{aligned} E\psi &= \int \{B(\omega)A(\omega) - B(\omega)P(A \mid B)\} dP(\omega) \\ &= P(A \cap B) - P(B)P(A \mid B) \end{aligned}$$

and the usual multiplication rule

$$P(A \cap B) = P(B)P(A \mid B)$$

follows. The rule can also be proved using simple betting arguments as in de Finetti (1937).

3. Statistical models, inferences, and predictions

A basic problem of statistics is to infer something about a parameter or state of nature θ after observing the value of a random variable x whose distribution p_θ depends on θ. The family $\{p_\theta\}$ of possible distributions for x is called a *statistical model.* Another basic problem is to predict the value of a second variable y after observing x.

For a simple example, think of θ as a real-valued physical constant and suppose x is a measurement of θ subject to a random error ε so that

$$x = \theta + \varepsilon.$$

If ε has a standard normal distribution, then the distribution p_θ for x is normal with mean θ and standard deviation 1, or $N(\theta, 1)$ for short, and has probability density function

$$f(x \mid \theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2}$$

Now suppose you observe a particular value of x. What can you say about θ or about the value of a subsequent measurement y?

A statistician, who takes the majority view of probabilities as being limiting values of relative frequencies, might construct *confidence intervals* for θ. Whatever the value of θ may be, the probability under p_θ that x lies between $\theta - 1.96$ and $\theta + 1.96$ is .95. The statistician solves the inequalities

$$\theta - 1.96 < x < \theta + 1.96$$

for θ getting

$$x - 1.96 < \theta < x + 1.96.$$

The interval from $x - 1.96$ to $x + 1.96$ is called a 95% *confidence interval* for θ, the word "confidence" being used partly to avoid making probability statements about θ. However, these confidence numbers are consistent with a distribution q_x for θ which is $N(x, 1)$.

The great statistician R.A. Fisher (1956) would do similar algebra to arrive at the same q_x which he called the "fiducial distribution" of θ.

The Bayesians interpret probability as a measure of degree of belief and are willing to treat θ as a random variable with some distribution π prior to observing x. The inference about θ is made by calculating the conditional distribution q_x of θ given x. This conditional distribution is called the *posterior*. If, as in the example, each p_θ has a probability density function $f(x \mid \theta)$ and if π also has a density $g(\theta)$, then q_x has a density $h(\theta \mid x)$ given by Bayes' formula

$$g(\theta \mid x) = \frac{f(x \mid \theta) g(\theta)}{\int f(x \mid \varphi) g(\varphi) d\varphi}$$

or, more simply,

$$g(\theta \mid x) \propto f(x \mid \theta) g(\theta).$$

In general, a posterior q_x for a prior π cannot necessarily be calculated using Bayes' formula. It is required, as a conditional distribution, to satisfy

the formula

$$\iint \varphi(\theta, x) p_\theta(dx)\pi(d\theta) = \iint \varphi(\theta, x) q_x(d\theta) m(dx)$$

for bounded functions φ, where m is the marginal distribution of x defined by

$$\int \psi(x) m(dx) = \iint \psi(x) p_\theta(dx)\pi(d\theta)$$

for bounded ψ. (By the way, in order to avoid inappropriate technicalities, I have not explicitly assumed the functions φ and ψ to be measurable and I will likewise omit mention of such assumptions below. More technical treatments can be found among the references.)

Not all Bayesians are willing to require that prior distributions satisfy all of the conventional axioms for probability. A few, like de Finetti, prefer not to assume countable additivity and many others use *improper* priors which assign infinite total mass to the space of θ–values rather than mass 1 (cf. Jeffreys (1961) and Hartigan (1983)). If, in our example, π is taken to be improper with density $g(\theta)$ identically equal to 1, then the *formal posterior* obtained from Bayes' formula has density

$$g(\theta \mid x) = f(x \mid \theta)$$

which agrees with the confidence and fiducial distributions. This posterior cannot be obtained from a proper, countably additive π, but it does correspond to the posterior for a finitely additive probability π on the real line which is invariant under translations (Heath and Sudderth (1978)).

For the rest of this paper, a statistical *inference* $\{q_x\}$ will be, as in the example above, a family of probability distributions for θ indexed by the values of x. Thus the notions of a model $\{p_\theta\}$ and an inference $\{q_x\}$ are mathematically symmetric in θ and x. However, x is observable by assumption and θ is typically not observable.

Consider now the problem of predicting the value of a second variable y after observing x assuming that p_θ is the probability distribution for the pair (x, y). A *prediction* is defined here to be a family $\{r_x\}$ of probability distributions for y indexed by the values of x. (These were called "predictive inferences" in Lane and Sudderth (1984).) The prediction of a frequentist statistician might be calculated using "prediction intervals" for y which are analogous to confidence intervals. R.A. Fisher generated a "fiducial distribution" for y. A Bayesian, equipped with a prior distribution π for θ, predicts y from x by calculating r_x, the conditional distribution of y given x. This conditional distribution is called the *predictive distribution*. If all the variables have densities, then r_x has a density $h(y \mid x)$ given by

$$h(y \mid x) = \int f(y \mid x, \theta) h(\theta \mid x) d\theta$$

where $f(y \mid x, \theta)$ is the conditional density for y given x and θ and $h(\theta \mid x)$ is the posterior density for θ as before.

For a simple example, suppose again that θ is a constant and also that

$$\begin{aligned} x_1 &= \theta + \varepsilon_1 \\ &\vdots \\ x_n &= \theta + \varepsilon_n \\ y &= \theta + \varepsilon_{n+1} \end{aligned}$$

are variables corresponding to $n+1$ measurements for which the errors $\varepsilon_1, \ldots, \varepsilon_{n+1}$ are independent, standard normal variables. You observe $x = (x_1, \ldots, x_n)$ and must predict y. Now y is $N(\theta, 1)$. So it is tempting to estimate θ by $\overline{x}$, the average of $x_1, \ldots, x_n$, and then take our predictive distribution r_x to be $N(\overline{x}, 1)$. However, $y - \overline{x}$ is $N(0, 1 + n^{-1})$ which results in prediction intervals for y consistent with a $N(\overline{x}, 1 + n^{-1})$ and the same prediction can be obtained using Fisher's fiducial argument or by calculating the predictive distribution corresponding to the improper prior density $g(\theta)$ which is identically equal to 1.

It seems more in keeping with de Finetti's original work to concentrate on predictions of observables rather than inferences about unobservables. In practise, inference and prediction are closely linked and models, which use unobservables, seem indispensable for predictions (cf. Geisser (1991)).

4. The statistician as bookie I

Suppose a statistician has an inference $\{q_x\}$ for the model $\{p_\theta\}$. Regard the model as given, but think of the statistician as a bookie and the inference as a price function on subsets of Θ, the set of possible values for θ. A gambler can choose a subset A^x of Θ and an amount $b(x)$ (positive or negative) to bet on the event that θ belongs to A^x. The payoff from this bet is

$$b(x)[A^x(\theta) - q_x(A^x)]. \tag{4.1}$$

The function b is restricted to be bounded. Also, the transaction between the gambler and the bookie is thought of as a contract conditional on the value of x. When x is observed, the gambler pays $b(x)q_x(A^x)$ dollars for $b(x)$ tickets on A^x and receives $b(x)$ dollars if θ turns out to be in A^x. The gambler is allowed to make a finite number of these bets and have total payoff of the form

$$\varphi(\theta, x) = \sum_{i=1}^{n} b_i(x)[A_i^x(\theta) - q_x(A_i^x)]. \tag{4.2}$$

The gambler's expected payoff under the model is

$$E(\theta) = \int \varphi(\theta, x) p_\theta(dx). \tag{4.3}$$

The inference $\{q_x\}$ is called *coherent* I if there is no φ as in (4.2) whose expected payoff $E(\theta)$ has a positive infimum; i.e. there is no "expected sure win" for the gambler.

This definition of coherence is from Heath and Sudderth (1978) but is based on the ideas of Freedman and Purves (1969). It is in the spirit of statistical decision theory where decision functions are evaluated in terms of their expected loss under the model (Wald (1950)). The next section gives a "sure loss" criterion closer in spirit to that of de Finetti. Here is a characterization of coherent I inferences from Heath and Sudderth (1978).

THEOREM 4.1 *An inference is coherent I if and only if it is the posterior distribution for some finitely additive prior π for θ.*

The idea of the proof is quite simple. Take Φ to be the collection of all functions $E(\theta)$ as in (4.3) which correspond to expected payoffs. The definition of coherence I is that the linear space Φ contains no function with a positive infimum and this means there is a finitely additive probability π on θ which gives every function in Φ expectation zero. This is the π for which $\{q_x\}$ is a posterior.

To get a better understanding of what it means for an inference to be coherent I, return to the simple measurement model in which p_θ is $N(\theta, 1)$. Consider four possible inferences: q_x^1 is $N(x/2, 1/2)$ (i.e. normal with mean $x/2$ and standard deviation $1/\sqrt{2}$), q_x^2 is $N(x, 1)$, q_x^3 is $N(x, 1/2)$, q_x^4 is $N(x+1, 1)$. To show an inference is coherent, it suffices to find a prior for which it is the posterior. If we take π_1 to be $N(0, 1)$, a calculation based on Bayes' formula shows q_x^1 to be the posterior and therefore coherent. The inference q_x^2 is the one discussed in section 3 where it was mentioned that it is the posterior of a finitely additive prior and thus coherent. The last two inferences are not coherent and the easiest way to see this is to find an expected sure win by exploiting inconsistencies between the model and the inferences. For instance, one can use normal probability tables to see that

$$\begin{aligned} p_\theta\{x : |x - \theta| > 1\} &= .32 \\ q_x^3\{\theta : |x - \theta| > 1\} &= .16. \end{aligned} \tag{4.4}$$

Now take $A^x = \{\theta : |x - \theta| < 1\}$ and $b(x) = 1$. So, for

$$\begin{aligned} \varphi(\theta, x) &= A^x(\theta) - q_x^3(A^x) \\ &= A^x(\theta) - .16, \end{aligned}$$

we have

$$E(\theta) = .32 - .16 = .16,$$

an expected sure win. Similarly, the inconsistency

$$p_\theta\{x : x > \theta\} = .5$$

and

$$q_x^4\{\theta : x > \theta\} = .16$$

leads to an expected sure win for the gambler.

The last example is due to M. Stone (1976) who gave it and a number of other examples to show the danger of using improper priors. The incoherent inference q_x^4 is, in fact, the formal posterior of an improper prior with density $g(\theta) = e^\theta$. Recall that the coherent inference q_x^2 is also the formal posterior of an improper prior. Even for our simple example, it is not easy to characterize those improper priors which lead to coherent inferences (cf. Heath and Sudderth (1989)).

Suppose now that $\{r_x\}$ is the statistician's prediction of the variable y from x based on the model $\{p_\theta\}$, where p_θ is the distribution for the pair (x, y) when the parameter has value θ. Regard the prediction as a price function on subsets of Y, the set of possible values for y. This time a gambler can choose a subset A^x of Y and an amount $b(x)$ to bet on the event that y belongs to A^x. The payoff is exactly as in (4.1) with θ replaced there by y and q_x by r_x. The function b must be bounded and a finite number of such bets are permitted. So the gambler can have any payoff function

$$\varphi(x, y) = \sum_{i=1}^{n} b_i(x)[A_i^x(y) - r_x(A_i^x)]. \tag{4.5}$$

The expected payoff under the model is

$$E(\theta) = \int \varphi(x, y) p_\theta(d(x, y)). \tag{4.6}$$

The prediction $\{r_x\}$ is *coherent* I if none of these expected payoffs has a positive infimum.

Here is a characterization from Lane and Sudderth (1984).

THEOREM 4.2 *The prediction $\{r_x\}$ is coherent I if and only if it is the predictive distribution of y given x corresponding to some finitely additive prior π for θ.*

In our prediction example from the previous section, $x = (x_1, \ldots, x_n)$ and y correspond to $n+1$ independent $N(\theta, 1)$ variables. The prediction in which r_x^1 is $N(\overline{x}, 1 + n^{-1})$ is coherent because it can be shown to be the predictive distribution for a finitely additive π which is translation invariant on the line. The plug-in prediction in which r_x^2 is $N(\overline{x}, 1)$ is not coherent because

$$p_\theta\{(x, y) : |\overline{x} - y| > \sqrt{1 + n^{-1}}\} = .32$$

and

$$r_x\{y : |\overline{x} - y| > \sqrt{1 + n^{-1}}\} < .32.$$

The gambler can exploit this inconsistency with a bet on

$$A^x = \{y : |\overline{x} - y| > \sqrt{1 + n^{-1}}\}$$

to get an expected sure win.

5. The statistician as bookie II

In the previous section, the model $\{p_\theta\}$ was assumed to be given. This is in keeping with statistical practise, at least in textbooks, and even Bayesians sometimes regard the model as being "objective" in some sense (Savage (1962, p.16)). In real problems, the model, like the inference, is usually supplied by the statistician and represents the opinions of the statistician.

So assume now that our statistician is responsible for the model $\{p_\theta\}$ as well as the inference $\{q_x\}$. The gambler can make bets on θ as before to get a payoff $\varphi(\theta, x)$ as in (4.2) but can also make bets on x to get a payoff

$$\psi(\theta, x) = \sum_{j=1}^{m} d_j(\theta)[C_j^\theta(x) - p_\theta(C_j^\theta)] \tag{5.1}$$

where the C_j^θ are subsets of the set X of possible values for x and the d_j are bounded, real-valued functions on Θ. Say that $\{p_\theta\}$ and $\{q_x\}$ are *coherent* II if there is no payoff $\varphi(\theta, x) + \psi(\theta, x)$ with a positive infimum; i.e. no "sure win" for the gambler. It's obviously harder for the gambler to get a sure win than an expected sure win. On the other hand, more bets are available. It turns out that the characterization of coherence is the same.

THEOREM 5.1 *A model $\{p_\theta\}$ and inference $\{q_x\}$ are coherent II if and only if $\{q_x\}$ is the posterior for some finitely additive prior π under the model $\{p_\theta\}$.*

Recall how, for our simple measurement model, the inconsistency reflected in (4.4) was used to construct an expected sure win thereby showing $\{q_x^3\}$ to be incoherent I. Now that the gambler can bet on x as well as θ, we can get a sure win by setting

$$A^x = \{\theta : |x - \theta| > 1\}, C^\theta = \{x : |x - \theta| > 1\},$$

$$\varphi(\theta, x) = A^x(\theta) - q_x^3(A^x) = A^x(\theta) - .16,$$

$$\psi(\theta, x) = -1[C^\theta(x) - p_\theta(A^x)] = -C^\theta(x) + .32.$$

Now $A^x(\theta) = C^\theta(x)$ and so

$$\varphi(\theta, x) + \psi(\theta, x) = .32 - .16 = .16,$$

a sure win.

The situation is completely analogous for predictions. The statistician's model $\{p_\theta\}$ now gives distributions for (x, y) and so payoff functions for bets on (x, y) are of the form

$$\psi(\theta, x, y) = \sum_{j=1}^{m} d_j(\theta)[C_j^\theta(x, y) - p_\theta(C_j^\theta)]$$

and the payoffs $\varphi(x, y)$ for bets on y based on the prediction $\{r_x\}$ are just as before of the form in (4.5). The model $\{p_\theta\}$ and prediction $\{r_x\}$ are called *coherent* II if there is no payoff $\varphi(x, y) + \psi(\theta, x, y)$ with a positive infimum. Again the characterization is the same.

THEOREM 5.2 *A model $\{p_\theta\}$ and prediction $\{r_x\}$ are coherent II if and only if $\{r_x\}$ is the predictive distribution of y given x corresponding to some finitely additive prior π under the model $\{p_\theta\}$.*

The proofs of the theorems in this section rely again on the fact that the collection Φ of possible payoffs is a linear space.

6. Are standard statistical procedures coherent?

As we have seen, coherent inferences and predictions have a simple characterization as being conditional distributions based on priors and these priors may be only finitely additive. Nevertheless it is not always easy to determine whether a particular procedure is coherent.

One important class of problems consists of those in which the parameter space Θ is a group and the model $\{p_\theta\}$ is a *translation family* in the sense

that, for some random variable z with values in Θ, p_θ is the distribution of $x = \theta z$. Thus $\theta = xz^{-1}$ and, for a *pivotal* inference, we take q_x to be the distribution of xz^{-1}. Our simple measurement model has this structure with Θ equal to the additive group of real numbers, z standard normal, and $x = \theta + z$. The pivotal inference here has q_x equal to the distribution of $x - z$ which is $N(x, 1)$. Whether the pivotal inference is coherent depends on the properties of the underlying group Θ. If Θ is locally compact and amenable, which means there is a finitely additive, group invariant, probability defined on Θ, then the pivotal inference is coherent and is, in fact, a posterior for a group invariant probability (Heath and Sudderth (1978)). If the group is not amenable, the pivotal inference need not be coherent. (See M. Stone (1976) for some nice examples.) As it turns out, many of the standard inferences of classical statistics (e.g. in analysis of variance) correspond to pivotals for amenable groups and are coherent. However, nonamenable groups occur in multivariate analysis and it seems likely that some of the standard inferences are incoherent.

As was pointed out in section 4, incoherent inferences can also occur as posteriors of improper priors and improper priors are often used by Bayesians in applications.

7. Related notions

Recently, Regazzini (1987) and Berti, Regazzini, and Rigo (1991) have formulated another notion of coherence for statistical inferences which is based on a theory of finitely additive conditional probability. It seems to be somewhat easier to be coherent under their definition of coherence. For example, all of the inferences $\{q_x^i\}, i = 1, \ldots, 4$, considered in section 4 for the simple measurement model are coherent in their sense. A different way of evaluating inferences has been developed by Eaton (1982, 1986), who treats inferences as decision rules and raises the question as to which inferences are admissible for an appropriate class of loss functions. It seems to be somewhat harder to be admissible in Eaton's sense than to be coherent in the sense of this paper.

The theory of coherence, as I have explained it, is based on an economic metaphor involving fictitious transactions. It is interesting that ideas quite similar to coherence are studied in the theory of finance where an "almost sure win" is called an "arbitrage opportunity" or a "free lunch". The book by Duffie (1988) provides a useful introduction.

References

BERTI, P., REGAZZINI, E. and RIGO, P. (1991). *Coherent statistical inference and Bayes theorem.* Annals of Statistics, 19, 366-381.

DE FINETTI, B. (1937). *La prevision: ses lois logiques, ses sources subjectives.* Annales de l'Institut Henri Poincaré, 7, 1-68. English translation in: Studies in Subjective Probability, eds. H.E. Kyberg Jr. and H.E. Smokler, Wylie, New York, 1964.

DE FINETTI, B. (1972). *Probability, Induction, and Statistics.* Wylie, New York.

DE FINETTI, B. (1974). *Theory of Probability.* Wylie, New York.

DUFFIE, D. (1988). *Security Markets: Stochastic Models.* Academic Press, San Diego.

EATON, M.L. (1982). *A method for evaluating improper prior distributions.* Statistical Decision Theory and Related Topics III, vol. 1, 329-352. Academic Press, San Diego.

EATON, M.L. (1986). *Admissibility in fair Bayes decision problems, I. General theory.* University of Minnesota School of Statistics Technical Report 482.

FISHER, R.A. (1956). *Statistical Methods and Scientific Inferences.* Oliver and Boyd, Edinburgh.

FREEDMAN, D.A. and PURVES, R.A. (1969). *Bayes method for bookies.* Annals of Mathematical Statistics, 40, 1177-1186.

GEISSER, S. (1991). *Predictive Approaches in Statistics,* to appear.

HARTIGAN, J.A. (1983). *Bayes Theory.* Springer-Verlag, New York.

HEATH, D. and SUDDERTH, W. (1972). *On a theorem of de Finetti, oddsmaking, and game theory.* Annals of Mathematical Statistics , 43, 2072-2077.

HEATH, D. and SUDDERTH, W. (1978). *On finitely additive priors, coherence, and extended admissibility.* Annals of Statistics, 2, 333-345.

HEATH, D. and SUDDERTH, W. (1989). *Coherent inference from improper priors and from finitely additive priors.* Annals of Statistics, 2, 907-919.

JEFFREYS, H. (1961). *Theory of Probability,* 3rd edition. Oxford Press, London.

LANE, D. and SUDDERTH, W. (1984). *Coherent predictive inference.* Sankhya Series A, 46, 166-185.

RAMSEY, F.P. (1926). *Truth and probability,* reprinted in Studies in Subjective Probability. eds. H.E. Kyberg Jr. and H.E. Smokler. Wylie, New York, 1964.

REGAZZINI, E. (1987). *De Finetti's coherence and statistical inference.* Annals of Statistics, 15, 845-864.

SAVAGE, L.J. ET AL. (1962). *The Foundations of Statistical Inference.* Wiley, New York.

SKYRMS, B. (1984). *Pragmatics and Empiricism.* Yale Press, New Haven.

SKYRMS, B. (1990). *The Dynamics of Rational Deliberation.* Harvard Press, Cambridge.

STONE, M. (1976). *Strong inconsistency from uniform priors.* Journal of the American Statistical Association, 2, 241-250.

WALD, A. (1950). *Statistical Decision Functions.* Wiley, New York.

Logic, Methodology and Philosophy of Science IX
D. Prawitz, B. Skyrms and D. Westerståhl (Editors)

CARNAP'S VOLUNTARISM[1]

RICHARD JEFFREY

Department of Philosophy, Princeton University, Princeton, N. J., 08544, USA

A certain voluntarism was a constant in Carnap's attitude from well before its expression in *Der logische Aufbau der Welt* (1928),[2] through its later expression in "Empiricism, semantics and ontology" (1950),[3] to his death in 1970. Carnap's voluntarism was a humanistic version of Descartes's explanation of the truths of arithmetic as holding because God willed them: not just "Let there be light," but "Let 1+1=2" and all the rest. Carnap substituted humanity for God in this scheme; that's one way to put it, a way Carnap wouldn't have liked much, but close to the mark, I think, and usefully suggestive. Item: Descartes was stonewalling, using God's *fiat* to block further inquiry. It is not for us to inquire why He chose 2 instead of 3. But for our own *fiat* the question is not what it was, but what it will be: choice of means to our chosen ends. This *fiat* cannot be a whim, for this choice will be made through the public deliberations of a constitutional convention, surveying the alternatives and comparing their merits. In *that* sense the choice is conventional.

Philosophically, Carnap was a social democrat; his ideals were those of the enlightenment. His persistent, central idea was: "It's high time we took charge of our own mental lives" — time to engineer our own conceptual scheme (language, theories) as best we can to serve our own purposes; time to take it back from tradition, time to dismiss Descartes's God as a distracting myth, time to accept the fact that there's nobody out there but us, to choose our purposes and concepts to serve those purposes, if indeed we are to choose those things and not simply suffer them. That's a bigger "if" than Carnap would readily acknowledge. A good part of his dispute with Quine centered on it.[4] Philosophically as well as politically Quine generally spoke as a conservative, Carnap as a socialist. For Carnap, deliberate choice of the syntax and semantics of our language was more than a possibility – it was a duty we owe ourselves as a corollary of freedom.

In his last 25 years, Carnap counted specification of c-functions among the semantical rules for languages. Choice of a language was a framework question, a practical choice that could be wise or foolish, and lucky or unlucky, but not true or false. (It will be true that we have chosen a particular framework, but that doesn't make it a true choice. Truth and falsity just don't apply to choices.) If deliberation eventuates in adoption of a framework whose semantical rules specify the confirmation function c^* then we have made it true by fiat, by convention, by reasoned choice, that "Pa" confirms "Pb" to degree 2/3.[5] It is not the individual scientist who chooses the c-function; that is a social choice, a convention specifying the framework within which the scientist works. Thus it is by social fiat that $c(h, e) = p$ for particular sentences h, e, and particular numbers p. The contribution of the individual experimental scientist might be to determine that the sentence e is true in fact, but *that* determination is by observation, not fiat.

Today I'd put Carnap's voluntarism to broader judgmental use. Judgmental probabilities are not generally in the mind or brain, awaiting elicitation when needed. Of course the hypothesis poses no difficulty for Carnap, for whom $c(h, e)$ values come with the framework, and for whom elicitation of your rational judgmental probability for h is is a matter of identifying an e that represents everything you're sure of relevant to h. And of course it does pose a difficulty for those subjectivists who think that probabilities are "in the mind" in some simple sense. But I think the important question is not whether we have probabilities in mind, but whether we can fabricate useful probabilistic proxies for whatever it is we have in mind.[6] The question is whether it's feasible and desirable for us to train ourselves to choose probabilities or odds or Bayes factors, etc., as occasions demand, for use in our practical and theoretical deliberations. The question is not whether we are natural Bayesians but whether we can *do* Bayesianism, and, if so, whether we should. For Carnap, this last is the practical question: whether, after due consideration, we will.

Recall the broad outlines of Carnap's philosophical development.

It was through his engagement with the program of logical analysis floated by Russell in 1915 that Carnap began to affect the shape of 20th century philosophy. The program aimed at bringing to philosophy a certain method or attitude, resembling that of the sciences in its focus on progress in solving problems rather than on defense of doctrines. Progress brought with it the doctrinal flux that soon saw the phenomenalism of Carnap's *Logische Aufbau der Welt*[7] yield to the physicalism of his *Logische Syntax der Sprache*[8] and saw his early deductivism give way to the probabilism of his last 25 years, during which his work in semantics

fostered a flowering of modal logic.[9] He welcomed real progress and its attendant doctrinal flux from whatever source, others no less than himself. The celebrated "Death of Logical Positivism"[10] refers to particular doctrines (*e.g.* phenomenological reductionism) and methods (*e.g.*, syntax) that Carnap and his friends abandoned for reasons rooted in the program itself. Broadcasting out of Europe those of its participants and associates whom it did not kill, Nazi power propagated the movement, which grew and changed rapidly in response to hard challenges. If what fired the *Wiener Kreis* has either died or grown out of recognition, that's largely due to the standard of clarity and care set by works like the *Aufbau*, *Logische Syntax der Sprache*, *Meaning and Necessity*,[11] "Empiricism, semantics, and ontology,"[12] and *Logical Foundations of Probability*.[13]

Carnap's earliest philosophical work, in the years just after the first world war, treated space from a Kantian perspective modified by Einstein's recently enunciated general theory of relativity and by his own characteristic analysis of the disputes over the nature of space as stemming from conflation of different senses of the term (formal, intuitive, physical). This was his Jena doctoral dissertation, restricting the Kant's synthetic apriorism to the local topology of intiuitive space (with the rest softened to conventionalism), and viewing physical space as purely empirical.[14]

In Jena he had attended Frege's lectures on logic, but it was after the war, in 1920, that he read Frege's *Grundgestze der Arithmetik* [15] and was fired by Whitehead and Russell's *Principia Mathematica*.[16] His dissertation had characterized formal space in terms of the logic of relations. Einstein's reduction of the concepts of space and time to concretely conceived local, momentary operations of measurement played a strong rôle with him then as with Russell and with Whitehead, suggesting a general empirical constructivist program in philosophy. In his intellectual autobiography Carnap recalls his excitement in 1921 upon reading Russell's *Our Knowledge of the External World as a Field for Scientific Method in Philosophy* [17], where "I found formulated clearly and explicitly a view of the aim and method of philosophy which I had implicitly held for some time".[18] According to Russell "This method, of which the first complete example is to be found in the writings of Frege, has gradually, in the course of actual research, increasingly forced itself upon me as something perfectly definite, capable of embodiment in maxims, and adequate, in all branches of philosophy, to yield whatever objective scientific knowledge it is possible to obtain."[19] The actual research to which he refers is the construction or reduction or logical reconstruction or analysis in *Principia Mathematica* of the numbers and operations of mathematics in terms of purely logical concepts.

The *Aufbau* was Carnap's serious start on the program of completing that analysis by extending it to the empirical domain. In effect, perhaps, it was a start on the heralded, unforthcoming fourth volume of *Principia*. Russell had described *Scientific Method in Philosophy* [20] as "an attempt to show, by means of examples, the nature, capacity, and limitations of the logico-analytic method in philosophy." He continued: "The central problem by which I have sought to illustrate method is the problem of the relation between the crude data of sense and the space, time, and matter of mathematical physics. I have been made aware of this problem by my friend and collaborator Dr Whitehead ... I owe to him ... the whole conception of the world of physics as a *construction* rather than an *inference*. What is said on these topics here is, in fact, a rough preliminary account of the more precise results which he is giving in the fourth volume of our *Principia Mathematica*." [21]

Volume 3 of *Principia* had laid the foundations for the conceptual apparatus of physics in a theory of measurement (part VI, "Quantity"). As with the characterization of cardinal numbers (*e.g.*, of 2 as a property posessed by the empirical property of being an author of *Principia Mathematica*, or as a set containing the set [Whitehead, Russell]), so the rationals are characterized in empirically applicable terms, as when 2/5 is defined as a relation between relations.[22] That was appropriate since physical magnitudes were analyzed as relations: "We consider each kind of quantity as what may be called a "vector-family," *i.e.*, a class of one-one relations all having the same converse domain, and all having their domain contained in their converse domain. In such a case as spatial distances, the applicability of this view is obvious; in such a case as masses, the view becomes applicable by considering, *e.g.*, one gramme as + one gramme, *i.e.*, as the relation of a mass m to a mass m' when m exceeds m' by one gramme. What is commonly called simply one gramme will then be the mass which has the relation + one gramme to the zero of mass." [23] The fourth volume was to have treated geometry. Russell makes it sound as though the treatment would spell out something like the fourth chapter of *Scientific Method in Philosophy*, "The World of Physics and the World of Sense." Considering Whitehead's views in 1914 of physics, geometry, and sense, and his exclusive responsibility for volume 4, perhaps what he intended for it was an abstract theory that could be specialized to a construction of geometrical concepts out of perceptual ones somewhat as the theory in volume 3 could be specialized to a *physikalische Begriffsbildung*.

However that may be, the *Aufbau* was Carnap's start on the project of enlarging the *Principia* construction to include geometry, empirical science, and everyday knowledge. It was to be a collaborative effort, to

which Nelson Goodman was the earliest and most productive recruit.[24] "The individual no longer undertakes to erect in one bold stroke an entire system of philosophy. Rather, each works at his special place within the one unified science. ... If we allot to the individual in philosophical work as in the special sciences only a partial task, then we can look with more confidence into the future: in slow careful construction insight after insight will be won. Each collaborator contributes only what he can endorse and justify before the whole body of his co-workers. Thus stone will be carefully added to stone and a safe building will be erected at which each following generation can continue to work."[25] Carnap was serious about the importance of rationalization, division of labor, but the suggestion of a centralized allotment of special tasks to individuals is a stylistic artifact; the individual chooses his own special place as Goodman did.

Carnap's 1928 preface was a period piece, like Bridgman's operationalistic call to arms the year before: "We should now make it our business to understand so thoroughly the character of our permanent mental relations to nature that another change in our attitude, such as that due to Einstein, shall be forever impossible. It was perhaps excusable that a revolution in mental attitude should occur once, because after all physics is a young science, and physicists have been very busy, but it would certainly be a reproach if such a revolution should ever prove necessary again."[26]

For Carnap as for Russell, scientific method was a datum, something one had the hang of as a scientifically trained person; its rational reconstruction was not high on the agenda as something to be accomplished before the rational reconstruction of our substantive knowledge could begin. Russell offers general remarks about the education of scientific investigators, and maxims claiming for the scientist such characteristics as honesty and open-mindedness, but neither he nor Carnap then essayed an explicit account of scientific method. Both saw the body of work from Frege to *Principia Mathematica* as a historically given basis and paradigm for scientific work in philosophy analogous to the body of work from Galileo to Einstein in physics. But while these histories included revolutionary change, as in the transition from classical to relativistic physics, and cataclysmic collapse of the sort that Russell's paradox produced in Frege's *Grundgesetze* construction, Carnap's 1928 statement envisaged only slow, irreversible progress. He would soon see his mistake, with Gödel's [27] incompletability proof for the *Principia* project, and Tarski's [28] deflationary rehabilitation of the concept of truth.

In *Logische Syntax der Sprache* Carnap dropped the phenomenalistic reductionism of the *Aufbau* in favor of a version of the physicalism that Neurath had been urging. The project remained that of logical analysis:

"That part of the work of philosophers which may be held to be scientific in its nature ... consists of logical analysis. The aim of logical syntax is to provide a system of concepts, a language, by the help of which the results of logical analysis will be exactly formulable. *Philosophy is to be replaced by the logic of science* — that is to say, by the logical analysis of the concepts and sentences of the sciences, for *the logic of science is nothing other than the logical syntax of the language of science.*"[29]

This was the work Quine read as it issued from Ina Carnap's typewriter in 1932 during his "rebirth in central Europe".[30]

Carnap's last big philosophical project finally addressed the nature of scientific method, in an attempt at rational reconstruction of our way of building scientific and everyday knowledge on experience. This took to a pure extreme the idea, subscribed to by many Bayesians, that it is the difference between your experience and mine that separates your probability function from mine — an idea often characterized by the vaguer saying that there are no unconditional probability judgments, but all are relative to background experience. Characteristically, Carnap took this idea quite seriously, and formulated it simply and clearly: the current probability that any individual i attributes to any hypothesis h in her or his language ought to be $c(h, e_i)$, where c is a "logical" probability function, the same for all rational agents. This constant c is the logical ingredient in logical empiricism, while the variable e_i is the empirical ingredient. Put baldly in Carnap's way the idea is dismissed by Bayesians to whom garbled expressions of it sound like obvious truths. Here as in the *Aufbau*, Carnap was conscious and careful in making moves that others whistle over hastily, eyes averted from the void below.

Like the *Aufbau*, this was the start of a long-range project for which Carnap hoped to recruit collaborators. That was the point of the *Studies in Inductive Logic and Probability* [31] series, to expedite cooperation among scattered collaborators through quick publication of new work on the project. He was optimistic and (Ayer's term) serene, first to last. He died with his logical boots on, at work on the project.

What's wrong with logical empiricism? Quine offered a double answer in "Two dogmas of empiricism": belief in the analytic-synthetic distinction, and reductionism, "the belief that each meaningful statement is equivalent to some logical construction upon terms which refer to immediate experience."[32] For these, Quine substituted the belief that our overall system of (yes/no) judgments is an economizing response to the totality of irritations of our sensitive surfaces. These irritations are not captured by protocol sentences recording sensations, but have their effects here and there in the totality of our assents to and dissents from sentences of various

sorts. The account he gives of the economy of thought involves a circle of terms noticed by Charles Chihara:[33] projectibility, similarity, simplicity. Quine sees that as a virtuous circle, unlike the vicious circle so carefully traced in "Two dogmas". But with Chihara, I find the view clearer from a Bayesian, probabilistic perspective — if not quite the one Carnap was mapping in his last 25 years.[34]

Surface irritations aren't enough. We are active animals, processing our inputs in view of what we take ourselves to be doing as we receive them. We are primates who don't swing from trees but do get sensory inputs on the fly, as an integral part of activities like walking, driving cars, conversing, catching balls, writing, etc. There is no hint of that in the common philosophical examples — observing hard brown tables, sensing yellow patches, "here now headache," and the rest. Mach's illustration, the view from his left eye as he lies on a couch, sums it up.[35] Here's where Anscombe and Hampshire on knowledge without observation are right on the money.[36] It's not just surface irritations that provide our sense of what we're doing and trying. That's the big truth in pragmatism as I see it. Rejecting reductionism won't get us out of the way of our own feet until we also reject the view from behind Mach's moustache.

It strikes me that in adopting Russell's program, Carnap adapted it quite characteristically. Just as he advocated artificial languages, consciously constructed by us to serve our purposes, so his philosophical method was synthesis, not analysis — a fact better understood by detractors dismissing him as a mere engineer than by some of his friends. He saw meanings as human artifacts, but had no reverence for traditional modes of conceptual production and their attendant mythology, for the lore of our fathers. He thought it practical, and essential for progress, to select and abide by linguistic rules that fit our purposes. Carnap was an activist, not only in relation to language but in his insistence on human agency as a prime epistemological perspective. This pragmatism or epistemological activism grew during the work on inductive logic to which he devoted much of his last quarter century, in the course of which he came to see rational deliberation as the primary context for his notion of inductive probability.[37]

By continuing in the contentious scientific spirit that Carnap and Russell urged upon us we can get further than they did; that was the idea all along. What we embrace is not a body of philosophical doctrine but a *de facto* method that still wants definition, "explication." Carnap made a good start on that. Where would he have ended? There is no fact of the matter. The next steps are for us.

With Carnap, I reject the Cartesian view of judgments as acts of flat

belief and disbelief. Seeing those as acts that might be undertaken wisely or rashly, Descartes enunciated a method for avoiding false belief, a discipline of the will "to include nothing more in my judgments than what presented itself to my mind with such clarity and distinctness that I would have no occasion to put it in doubt".[39] He called such acts of the will "affirmations," *i.e.*, acts of accepting sentences or propositions as true. What do "belief", "acceptance", and "affirmation" mean in this context? I don't know. I'm inclined to doubt that anyone else does, either, and to explain the general unconcern about this lack of understanding by familiarity of the acceptance metaphor masquerading as intelligibility, perhaps as follows: "Since it's clear enough what's meant by accepting other things — gifts, advice, apologies — and it's clear enough what's meant by sentences' being true, isn't it clear what's meant by accepting sentences as true? Doesn't Quine make "holding" sentences true the very pivot of his epistemology? And isn't affirmation just a matter of saying 'Yes'?"

Probabilism, be it Carnap's or de Finetti's, replaces the two Cartesian options of affirmation and denial by a continuum of judgmental probabilities in the interval from 0 to 1, endpoints included, or — what comes to the same thing — a continuum of judgmental odds in the interval from 0 to ∞, endpoints included. Zero and one are probabilities no less than 1/2 and 99/100 are. Probability 1 corresponds to infinite odds, 1:0. That's a reason for thinking in terms of odds: to remember how momentous it may be to assign probability 1 to a hypothesis. It means you'd stake your all on its truth, if it's the sort of hypothesis you can stake things on. To assign 100% probability to success of an undertaking is to think it advantageous to stake your life upon it in exchange for any petty benefit. We forget that when we imagine that we'd assign probability 1 to whatever we'd simply state as true.[40]

What is involved in attributing particular judgmental probabilities to sentences? With Carnap in his last decade and with de Finetti, I'd answer in terms of a theory of preference seen as a relation between sentences or propositions: preference for truth of one sentence ("Cameroon wins") to truth of another ("Britain wins").[41] This theory is subjectivistic in addressing only the effects of such probability judgments, without saying how those judgments ought to be arrived at. The theory doesn't prejudge attempts like Carnap's to supply norms for forming such judgments; and indeed Carnap accepted this subjectivistic theory as an account of how judgmental probabilities are to be applied, once formed.

Broadly speaking, probabilism or "Bayesianism" sees making up the mind as a matter of either adopting an assignment of judgmental probabilities or adopting certain features of such an assignment, e.g., the feature

of assigning higher conditional probability to 5 year survival on a diagnosis of ductal cell carcinoma than on a diagnosis of islet cell carcinoma. Some insist on restricting the term "Bayesian" narrowly to those who see conditioning (or "conditionalization") as the only rational way to change the mind; I don't. (See *The Logic of Decision*, Chapter 11.) Rationalistic Bayesianism — hereafter, "rationalism" — is a subspecies of the narrow Bayesianism just noted, according to which there exists a (*logical, a priori*) probability distribution that would define the state of mind of a perfect intelligence, innocent of all experience. Notable subscribers: Bayes, Laplace, W. E. Johnson, J. M. Keynes, Carnap, John Harsanyi.

Rationalism and empiricism are two sides of the same Bayesian coin. One side is a purely rational, "logical" element, a prior probability assignment **M** characterizing the state of mind of a newborn Laplacean intelligence. Carnap spent his last 25 years trying to specify **M**. The other side is a purely empirical element, a comprehensive report D of all experience to date. Together, these determine the experienced Laplacean intelligence's judgmental probabilities, obtained by conditioning the "ignorance prior" **M** by the *Prototokollsatz* D. Thus $\mathbf{M}(H|D)$ is the correct probabilistic judgment about H for anyone whose experiential data base is D.

Radical probabilism makes no attempt to analyze judgment into a purely rational component and a purely empirical component, without residue. It rejects the empiricist myth of the sensuously given data proposition D as well as the rationalist myth of the ignorance prior **M**; it rejects the picture of judgment as a coin with empirical obverse and rational reverse. Let's see why.

On the empirical side, reports of conscious experience are too thin an abstract of our sensory inputs to serve adequately as the first term of the equation

(1) $$\textit{experience} + \textit{reason} = \textit{judgment}$$

COUNTEREXAMPLE: *Blindsight*.[42] In humans, monkeys, etc., some 90% of optic nerve fibres project to the striate cortex at the very back of the brain via the dorsal lateral geniculate nucleus in the mid-brain. But "while the geniculo-striate pathway constitutes the major portion of the optic nerve ... there are at least 6 other branches that end up in the midbrain and sub-cortical regions ..., and one of these contains about 100 000 fibres, by no means a trivial pathway — it is larger than the whole of the auditory nerve ... Therefore, if striate cortex is removed or its direct input blockaded, one should expect that some visual capacity should remain because all of those non-geniculo-striate pathways are left

intact. The paradox is that in man this is usually not so: destruction of occipital cortex ... characteristically causes blindness in that part of the visual field corresponding to the precise projection map of the retina on to the striate cortex ... Admittedly, some primitive visual reflexes can be detected ... but typically the patient himself does not appear to discriminate visual events or to have any awareness of them." The non-geniculo-striate 10% of optic nerve fibres seem to provide visual capacities of which such patients are unaware — capacities which they dismiss as mere guesswork, and which the rest of us need never distinguish as a special category. Thus, although a patient ("D. B.") whose right occipital lobe had been surgically removed "could not see one's outstretched hand, he seemed to be able to reach for it accurately. We put movable markers on the wall to the left of his fixation, and again he seemed to be able to point to them, although he said he did not actually see them. Similarly, when a stick was held up in his blind field either in a horizontal or vertical position, and he was asked to guess which of these two orientations it assumed, he seemed to have no difficulty at all, even though again he said he could not actually see the stick." This sort of thing looks like bad news for the New Way of Ideas, empiricism based on sense data: D. B. has factual sensitivity, a basis for probabilistic judgment, with no corresponding phenomenal sensitivity.

If sense data won't do for the first term of formula (1), perhaps the the triggering of sensory receptors will. Quine seems to think so: "By the stimulation undergone by a subject on a given occasion I just mean the temporally ordered set of all those of his exteroceptors that are triggered on that occasion."[43] The experiential data base D might then correspond to an ordered set of such ordered sets, whence the Carnapian judgmental probability $\mathbf{M}(H|D)$ might be calculable. But no; not even "in principle". The trouble is that the temporal record of exteroception makes no perceptual sense without a correlated record of interoception; thus, interpretation of a record of activity in the optic nerve requires a correlated record of relative orientations of eye, head, trunk, etc. When Quine's bit of neurophysiology is put in context, his exteroceptive data base looks no more adequate for its purpose than did the sense data base it replaced.

COUNTEREXAMPLE: *Proprioception and visual perception.* [44] Exteroceptive nervous activity is interpreted in the light of concurrent interoceptive activity. Thus, optic nerve input is interpreted in the light of concurrent activity in oculomotor brain-stem neurons sending axons directly to the eye muscles, whose activity is coordinated by interactions of nuclei commanding vertical, oblique, and lateral determinants of gaze. The vestibular neurons in the brain stem relay information about head position from

inner ear to oculomotor neurons. ... Head and eye position are related, in turn, to spinal control of posture by the reticular formation. "Without the constant and precise operation of these three systems, we could neither walk and see, nor sit still and read. ... Together with the cerebellum, the integrated activity of these brain-stem systems is responsible for giving sighted animals complex control of their acts." Quite apart from the question of awareness, it seems that the neurological analog of sense data must go beyond irritations of sensory surfaces. In the Cartesian mode it must treat the observer's body as a part of the "external" world providing the mind with inputs to be coordinated with exteroceptive inputs by innate nerological circuitry that is fine-tuned mostly *in utero* and in the earliest years of extrauterine life.

From Carnap to Quine, it is ordinary thing-languages to which physicalists have looked for observation sentences, whose imputed truth values (or probability values) are to be propagated through the confirmational net by conditioning (or generalized conditioning). Quine gestures toward temporally ordered sets of triggered exteroceptors as an empirical substrate for the real epistemological action, Cartesian affirmations of ordinary observation sentences. But the proffered substrate, once mentioned, plays no further rôle in Quine's epistemology. It is anyway incapable of providing an empirical footing for his holdings true until enriched by a coordinated efferent substrate. The full-blown afferent-efferent substrate would provide a footing ("neurological solipsism") upon which holdings true and holdings probable to various degrees could supervene, but it would play no rôle, either. Bag it.

So much for the empirical side of the epistemological coin. On the other side, radical probabilism abandons Carnap's search for the fountain of rationality in a perfect ignorance prior, at the same time abandoning the idea that conditioning, or generalized conditioning, is the canonical way to change your mind. Instead, radical probabilism offers a dynamic or diachronic point of view, from which the distinction between making up your mind and changing it becomes tenuous. The Carnapian motion picture is a sequence of instantaneous frames, your successive complete probability assignments to all sentences of your language, beginning with $\mathbf{M}$ and changing every time a new conjunct is added to your data base: $\mathbf{M}(-), \mathbf{M}(-|D_1), \mathbf{M}(-|D_1\&D_2)$, and so on up to your present assignment, $\mathbf{M}(-|D_1\&D_2\&...\&D_t)$. The radical probabilist picture is less detailed in each frame, and smoother or more structural across frames in the time dimension.

Thus, making up your mind probabilistically involves making up your mind about how you will change your mind. It's not that you must map

that out in fine detail, any more than you must map out your instantaneous probabilities for all sentences of your language, frame by frame. But since it no longer goes without saying that you will change your mind by conditioning or generalized conditioning (probability kinematics) any more than it goes without saying that your changes of mind will be quite spontaneous or unconsidered, these are questions about which you may make up your mind about changing your mind in specific cases. You may decide to change your mind by generalized conditioning on some set of data propositions. According to the laws of probability logic ("the probability calculus") such a decision comes to the same thing as deciding to keep your conditional probabilities on the data propositions constant when your unconditional probabilities for them change. In case your probability for one of the data propositions changes to 1, this reduces to ordinary conditioningon the data proposition you've become sure of.

It needs to be emphasized that becoming sure of a sentence's truth doesn't guarantee that your new conditional probabilities based on on it will be the same as they were before you became sure of it. That's why Carnap required that you condition only on sentences that you regard not only as true but as recording the whole of the relevant truth that you know about. For this to imply constancy of conditional probabilities there must be available to you an infinitely nuanced assortment of data propositions to condition upon. It strikes me as a fantasy, an epistemologist's pipe-dream, to imagine that such nuanced propositions are generally accessible to us. There need be no sentence you can formulate, that fits the description "the whole of the relevant truth that you know about."[45] But the diachronic perspective of radical probabilism reveals a different dimension of nuance that you can actually use in such cases to identify a set of data propositions relative to which you expect your conditional probabilities to be unchanged by an impending observation that you think will have the effect of changing your probabilities for some of the data propositions. That will be a case where updating by probability kinematics is appropriate.

Constancy of conditional probabilities opens other options for registering and communicating the effect of experience, e.g., the option of registering the ratios ("Bayes' factors") $f(A, B)$ of odds between A and B afterward and beforehand:

$$f(A,B) = \frac{Q(A) : Q(B)}{P(A) : P(B)} \tag{2}$$

What's conveyed by the Bayes' factor is just the effect of experience, final odds with prior odds factored out. Others who accept your response

to your experience, whether or not they share your prior opinion, can multiply their own prior odds between A and B by your Bayes' factor to get their posterior odds, taking account of your experience.[46]

EXAMPLE: *Expert opinion.* Jane Doe is a histopathologist who hopes to settle on one of the following diagnoses on the basis of microscopic examination of a section of tissue surgically removed from a pancreatic tumor. (To simplify matters, suppose she is sure that exactly one of the three diagnoses is correct.)

$$A = \text{Islet cell carcinoma}$$
$$B = \text{Ductal cell carcinoma}$$
$$C = \text{Benign tumor}$$

In the event, the experience does not drive her probability for any diagnosis to 1, but does change her probabilities for the three candidates from the following values (P) prior to the experience, to new values (Q):

	A	B	C
P	1/2	1/4	1/4
Q	1/3	1/6	1/2

Henry Roe, a clinician, accepts the pathologist's findings, i.e., he adopts, as his own, her Bayes' factors between each diagnosis and some fixed hypothesis, say, C:[47]

$$f(A,C) = 1/3, \qquad f(B,C) = 1/3, \qquad f(C,C) = 1$$

It is to be expected that, *given a definite diagnosis*, his conditional probabilities for the prognoses "live" (for five years) and "die" (within 5 years) are stable, unaffected by the pathologist's report. For definiteness, suppose those stable probabilities are as follows, where lower case "p" and "q" are used for the clinician's prior and posterior probabilities, to distinguish them from the pathologist's.

$$q(\text{live}|D) = p(\text{live}|D) = .4, .6, .9 \quad \text{when} \quad D = A, B, C$$
$$q(\text{die}|D) = p(\text{die}|D) = .6, .4, .1 \quad \text{when} \quad D = A, B, C$$

Given his prior probabilities $p(D)$ for the diagnoses and his adopted Bayes' factors, these conditional probabilities determine his new odds on 5 year survival.[48] It works out as follows.

$$(3) \qquad \frac{q(\text{live})}{q(\text{die})}$$
$$= \frac{p(\text{live}|A)p(A)f(A,C) + p(\text{live}|B)p(B)f(B,C) + p(\text{live}|C)p(C)}{p(\text{die}|A)p(A)f(A,C) + p(\text{die}|B)p(B)f(B,C) + p(\text{die}|C)p(C)}$$

If the clinician's prior distribution of probabilities over the three diagnoses was flat, $p(D) = 1/3$ for each diagnosis, then the imagined numbers given above raise his new odds on 5 year survival from $p(\text{live}) : p(\text{die}) = 19 : 11$ to $q(\text{live}) : q(\text{die}) = 37 : 13$, so that his probability for 5 year survival rises from 63% to 74%.

Prima facie, the task of eliciting Bayes' factors looks more difficult than eliciting odds, for Bayes' factors are ratios of odds.[49] For the same reason it may seem that the pathologist's Bayes' factor, (posterior odds): (prior odds), cannot be elicited if (as it may well be) she has no definite prior odds. But if her Bayes' factors would be stable over a large range of prior odds, so as to be acceptable by colleagues with various prior odds, her Bayes' factors are as easily elicited as her posterior odds if she can and will adopt definite odds prior to her observation, e.g., in the light of real or imagined statistics. With even priors, $P(A)/P(C) = P(B)/P(C) = 1$, her Bayes' factors would simply be her posterior odds; but if it is only uneven priors that are cogent for her, the extra arithmetic presents no real difficulty.

The example illustrates two contrasts between the radical probabilism advocated here and the phenomenalism I have been deprecating. The less important contrast concerns the distinction between probability and certainty as basic attitudes toward Protokollsätze. The more important contrast concerns the status of those attitudes toward Protokollsätze (or toward what they report) as foundations for all of our knowledge. Here, C. I. Lewis wears the Cartesian black hat better than Carnap: "Subtract, in what we say that we see, or hear, or otherwise learn from direct experience, *all that could conceivably be mistaken*; the remainder is the given content of the experience inducing this belief. If there were no such hard kernel in experience — *e.g.*, what we see when we think we see a deer but there is no deer — then the word 'experience' would have nothing to refer to."[50]

This is the sort of empiricism dismissed above, in which the term "experience" is understood not in its ordinary sense, as the sort of thing that makes you an experienced doctor, sailor, lover, traveller, carpenter, teacher, or whatever, but in a new sense, the sensuously given, in which experience is bare phenomenology or bare irritation of sensitive surfaces. It presupposes a unitary faculty of reason, the same for all subject matter, which, added to the sensuously given, equals good judgment. The formula itself goes back much further than Descartes, *e.g.*, to Galen: "When I take as my standard the opinion held by the most skillful and wisest physicians and the best philosophers of the past, I say: The art of healing was originally invented and discovered by the logos [reason] in conjunction with

experience. And to-day also it can only be practiced excellently and done well by one who employs both of these methods."[51]

But in this formula reason is theory, and experience is gained by purges, surgery, etc., the sort of thing Hippocrates had called dire in his famous "experiment perilous" aphorism. For experience in Galen's formula, C. I. Lewis substitutes the given. Galen's formula is

$$\textit{experience} + \textit{reason} = \textit{medical expertise}$$

There's a similar formula for other kinds of knowledge and technique, with "reason" and "experience" referring to other things than they do in the case of medicine.[52] But Lewis's formula is general:

$$\textit{the given} + \textit{reason} = \textit{good judgment}$$

Here "reason" needs to be understood as something like a successful outcome of the project to which Carnap devoted his last 25 years, of designing a satisfactory general inductive logic. For that I have no hope, for reasons given above under the headings of "blindsight" and "perception and proprioception".

Carnap himself was undogmatic; with high hopes for his program, he offered general, inconclusive arguments as an invitation to join in testing the idea. In fairness to him it should be noted that I haven't tried to present the case for his program here; I've used it, or a simplistic cartoon of it, as a foil for a different program ("radical probabilism") that rejects the analytical basis that I've attributed to Carnap's program, the analysis of good judgment into an a priori probability function representing reason and a propositional data base representing experience.[53]

Radical probabilism doesn't insist that probabilities be based on certainties; it can be probabilities all the way down, to the roots. Modes of judgment (in particular, probabilizing) and their attendant standards of rationality are cultural artifacts, bodies of practice modified by discovery or invention of broad features seen as grounds to stand on. It is ourselves or our fellows to whom we justify particular judgments. Radical probabilism is often faulted as uncritical, *e.g.*, as not requiring the pathologist to justify the Bayes' factors she finds cogent; "Anything goes." But probabilizing — adoption of personally cogent odds, Bayes' factors, and the like, concerning some range of matters, *e.g.*, tumors — is a subject-matter-dependent *techne*, an art of judgment for which honest diligence is not enough. In practice, justification — what makes the histopathologist's personally cogent Bayes' factors cogent for her colleagues as well — is a mish-mash including the sort of certification attested by her framed

diploma and her reputation among relevant *cognoscenti*. (Relevant: the cogencies of a certified, reputable faith healer are not transferrable to me.)[54] Personal cogency may express itself in commitment to specific action (say, excision) that would be thought irresponsible if the probabilistic judgment (odds on the tumor's being benign) were much different. With probability judgment as with judgment of truth and falsity, quality varies with subject-matter; handicappers and meteorologists are mostly useless as diagnosticians.[55] Judgments are *capta*, outputs of human transducers like our histopathologist, in whose central nervous system perceptive and proprioceptive neuronic inputs somehow yield probabilistic judgments about stained cells under her microscope. Although she is far from knowing how that works, she can know *that* it works, pretty well, and know how she learned to work that way — whatever that way may prove to be.

As I see it, radical probabilism delivers the philosophical goods that logical empiricists reached for over the years in various ways. Carnap got close, I think, with his idea of a "logical" c-function encoding meaning relations, but I'd radicalize that probabilism twice, cashing out the idea of meaning in terms of skilled use of observations to modify our probabilistic judgments, and cashing *that* out in terms of Bayes' factors. Carnap's idea of an "ignorance" prior cumulatively modified by growth of one's sentential data base is replaced by a pragmatical view of priors as carriers of current judgment, and of rational updating in the light of experience as a congeries of skills like those of the histopathologist. Described conscious experiences are especially welcome data, for by Bayes' theorem, when the new odds come from the old by conditioning on a data proposition, the diachronic Bayes' factor reduces to the synchronic "likelihood ratio" at the right:[56]

$$\underset{\text{factor}}{\text{Bayes'}}\ \frac{Q(A):Q(B)}{P(A):P(B)} = \frac{P(A|\text{data})/P(B|\text{data})}{P(A)/P(B)} = \frac{P(\text{data}|A)}{P(\text{data}|B)}\ \underset{\text{ratio}}{\text{Likelihood}}$$

Among the virtues of describable experience are utility for teaching or for routinizing skills of probabilizing ("If it's purple, the odds are 7:3 on A against B"), and for thrashing out differences of opinion in the matter. But conscious experience eluding adequate description has some of those virtues. Example: histopathological instruction of medical students using a microscope with two eyepieces and an arrow of light with which the instructor indicates complex features of particular cells. The discussion of blindsight and proprioception was not meant to deny that but to call attention to the considerable rôle of unconscious inputs and inputs resisting description — a rôle we can expect to be far greater than has been noted, precisely because of that unconsciousness and resistance.

In logical positivism (= logical empiricism) the move from verification to probability as a way of cashing out "meaning" goes back to the 1930's, to Reichenbach's *Wahrscheinlichkeitslehre* (Leyden, 1935) and *Experience and Prediction* (Chicago, 1938), and to Carnap's "Testability and meaning" (*Philosophy of Science* 1936, 1937).[57] My own probabilism stems from a fascinated struggle with those sources, begun in Chicago with Carnap in the late 1940's and refocused in Princeton with Hempel in the mid-1950's. I see its departures not so much as a rejection but as a further step in the development of logical empiricism, i.e., the movement, not its particular verificationist stage ca. 1929.[58]

NOTES

1 This paper draws on my *Probability and the Art of Judgment* (Cambridge University Press, 1992) and "After Carnap," Erkenntnis 35 (1991) 255-262.

2 Translated by Rolf George, *The Logical Structure of the World*: University of California Press, 1967.

3 Reprinted in the second edition of Carnap's *Meaning and Necessity*: University of Chicago Press, 1956.

4 This was also evident in their linguistic hobbies: Carnap was a dedicated learner of artificial, would-be international languages, Quine of natural, national languages.

5 Like the truth that "Pa & Pb" logically implies "Pa", that's a truth about the adopted framework, not to be expressed in it but in a metalanguage.

6 The fabricators are not generally the untimate users, but innovators like Bruno de Finetti, who showed us how to use exchangeability. Intelligent users mostly choose such probabilistic constructs ready-to-wear, from catalogs.

7 Felix Meiner: Vienna, 1928, Hamburg, 1961. Transl. Rolf George, *The Logical Structure of the World*, Berkeley and Los Angeles: University of California Press, 1967.

8 Julius Springer, Vienna, 1934. Translation by Amethe Smeaton, *The Logical Syntax of Language*, London: Kegan Paul, Trench, Trubner & Co., 1937.

9 In "Modalities and quantification" (Journal of Symbolic Logic 11 [1946] 33–64) Carnap announced soundness proofs for the simplest (S5) system of propositional modal logic and for quantified S5 relative to a semantics in which models are represented by the state descriptions of his *Introduction to Semantics* (Harvard University Press, Cambridge, Mass, 1942). He left the completeness question open, as did Stig Kanger in *Provability in Logic* (Stockholm, 1957), the first publication proposing retalional model theories. In J. Symbolic Logic 24 (1959) Saul Kripke published soundness and completeness proofs for quantified S5 (pp. 1–15) and annonced results for other systems (pp. 323–324) whose proofs appear in Z. math. Logik und Grundlagen der Math. 9 (1963) 67–96, where footnote 2 gives some early history — as do Hintikka and Kripke in Acta Philosophica Fennica 16 (1963) 65–94.

10 For which many claim reponsibility, starting with Karl Popper. See *The Philosophy of Karl Popper*, La Salle, Illinois: Open Court, 1974, vol. 1, pp. 69–71.
11 Chicago and London: University of Chicago Press, 1947, 1956.
12 Revue International de Philosophie 4 (1950) 20–40.
13 University of Chicago Press, Chicago, 1950, 1962.
14 *Der Raum, ein Beitrag zur Wissenschaftslehre*, Kant-Studien, No. 56, Berlin, 1922.
15 2 vols., Jena, 1893, 1903.
16 3 vols., Cambridge, England, 1910–1913; 2nd ed., 1925–1927.
17 Open Court, Chicago and London, 1915.
18 *The Philosophy of Rudolf Carnap*, P. A. Schilpp (ed.), Open Court, La Salle, Illinois, 1963, p. 13.
19 Russell, Op. cit., p. v.
20 The title of *Our Knowledge of the External world as a Field for Scientific Method in Philosophy* as abbreviated on the spine and cover of the first edition.
21 *Scientific Method in Philosophy*, pp. v–vi. Russell's emphases.
22 An example outside physics: the relation 2/5 holds between the grandparent relation and the great-great-great grandparent relation. See Willard van Orman Quine, "Whitehead and modern logic," *The Philosophy of Alfred North Whitehead*, P. A. Schilpp (ed.), Northwestern University Press, 1941, p. 161.
23 Principia Mathematica, vol. 3, p. 233.
24 See *The Structure of Appearance*, Cambridge, Mass., 1951: Harvard University Press. Goodman reports (p. vii) beginning work on the project within a year or so of the Aufbau's publication.
25 The Logical Structure of the Word, pp. xvi–xvii.
26 Percy Bridgman, *The Logic of Modern Physics*, New York: Macmillan, 1927.
27 "Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I," Monatshefte für Mathematic und Physik, 1931.
28 Summaries in Polish and in German were published in 1931 and 1932. The full paper appeared in German translation much later: *"Der Wahrheitsbegriff in den formalisierten Sprachen,"* Studia Philosophica 1 (1936) 261–405. Carnap would offer this as prime evidence of the need for an international language. (He had in mind something like Esperanto or Ido, not English.)
29 The Logical Syntax of Language, p. xiii, Carnap's emphases.
30 See *The Philosophy of W. V. Quine* (ed. L. E. Hahn and P. A. Schilpp), Open Court, La Salle, Illinois, 1986, p. 12.
31 Berkeley, Los Angeles and London: University of California Press, vol. 1 (Rudolf Carnap and Richard Jeffrey, eds.) 1971, vol. 2 (Richard Jeffrey, ed.) 1980.
32 First paragraph of "Two dogmas of empiricism," W. V. Quine, From a Logical Point of View, Cambridge, Mass: Harvard University Press, 1953.
33 "Quine and the confirmational paradoxes," Midwest Studies in Philosophy, vol. 6, ed. P. French, T. Uehling, Jr. and H. Wettstein, U. of Minnesota Press, Minneapolis, 1981, pp. 425–452.
34 I collaborated with Carnap as a fellow-traveller — toward a destination closer to Bruno de Finetti's: see "Reading Probabilismo," Erkenntnis 31 (1989) 225-37.
35 See Figure 1 in Ernst Mach, *Beitrage zur Analyse der Empfindungen*, Jena, 1886; translation by C. M. Williams, *Contributions to the Analysis of Sensa-*

tions, La Salle, Illinois: Open Court, 1896.

36 E.g., see sec. 11.9 of G. E. M. Anscombe, *Intention* (Oxford: Basil Blackwell, 1957; 2nd ed., Ithaca, N. Y.: Cornell, 1963).

37 See "Inductive logic and rational decisions," Studies in Inductive Logic and Probability, vol. 1 (fn. 25 above).

38 Aufbau, preface.

39 *Discourse on the method...*, part 2.

40 As probabilities p range over the unit interval [0,1], the corresponding odds $p/(1-p)$ range over the extended non-negative reals $[0, 1]$, enhancing resolution high in the scale. Thus, probabilities 99%, 99.9%, 99.99% correspond to odds 99, 999, 9999. But at the low end, where odds $p/(1-p)$ are practically equal to probabilities p, there is no such increase in resolution. Logarithms of odds, increasingly positive above $p = .5$ and symmetrically negative below, yield the same resolution at both ends. (Odds of 99, 999, 9999 become log odds of approximately 2, 3, 4, and at the low end, probabilities of .01, .001 .0001 become log odds of approximately $-2, -3, -4$.)

41 See Richard Jeffrey, *The Logic of Decision*, McGraw-Hill, 1965; University of Chicago Press, 1983, 1990.

42 *Blindsight, a Case Study and Implications*, by L. Weiskrantz, Clarendon Press, Oxford, 1986, pp. 3–6, 24, and 168–9. See also Patricia Smith Churchland, *Neurophilosophy*, MIT Press, Cambridge, Mass., 1986, pp. 224–228.

43 W. V. Quine, *The Pursuit of Truth*, Harvard University Press, Cambridge, Mass., 1990, p. 2.

44 J. Allen Hobson, *The Dreaming Brain*, Basic Books, New York, 1988, pp. 110–112.

45 Given an old probability distribution, P, and a new one, Q, it is an open question whether, among the sentences D in your language for which your conditional probabilities are the same relative to Q as they are relative to P, there are any for which your new probability $Q(D)$ is 1. If so, and only then, Q can be viewed as coming from P by conditioning.

46 Such uses of Bayes' factors were promoted by I. J. Good in chapter 6 of his book *Probability and the Weighing of Evidence*, Charles Griffin, London, 1950. See also *Alan Turing: the Enigma*, by Andrew Hodges, Simon and Schuster, New York, 1983, pp. 196–7. Good promotes *logarithms* of odds ("plausibilities") and of Bayes' factors ("weights of evidence") as intelligence amplifiers which played a rôle in cracking the German "enigma" code during the second world war.

47 Proposed by Schwartz, W. B., Wolfe, H. J., and Pauker, S. G., "Pathology and probabilities: a new approach to interpreting and reporting biopsies," The New England Journal of Medicine 305 (1981) 917–923.

48 See Richard Jeffrey and Michael Hendrickson, "Probabilizing pathology", Proceedings of the Aristotelian Society, v. 89, part 3, 1988/89: p. 217, odds kinematics.

49 In general, elicitation is a process of *drawing forth*. Here, authenticity does not require the elicited Bayes' factors to have been present in the pathologist's mind before the process began; the process may well be one in which she is induced to form a judgment, making up her mind probabilistically.

50 *An Analysis of Knowledge and Valuation*, Open Court, La Salle, Illinois, 1946, pp. 182–3. (Lewis's emphasis.)

51 The first sentences of "On medical experience", translated by Richard Walzer, in Three Treatises on the Nature of Science, Hackett, Indianapolis, 1985, p. 49.

52 Like "experience," "reason" has a different sense (comprehending theory) in Galen's formula from what it has in C. I. Lewis's; see pp. xx–xxxi of Michael Frede's introduction to the Galen book (note 10).

53 For Carnap's program, see his essays in *Studies in Inductive Logic and Probability*, University of California Press: Berkeley, Los Angeles and London, volume 1, Rudolf Carnap and Richard Jeffrey (eds.), 1971, volume 2, Richard Jeffrey (ed.), 1980.

54 In meterology, radically probabilistic methods of assessing and improving the quality of probabilistic judgment have been in use since the 1950's, but such techniques remain largely unknown in medicine and other areas. For an account of such methods, see Morris H. DeGroot and Stephen E. Fienberg, "Assessing probability assessors: calibration and refinement." In *Statistical Decision Theory and Related Topics*, vol. 3. New York: Academic Press, 1982.

55 The practical framework of Bayesian decision analysis is the native ground of such probabilizing. See, e.g., *Clinical Decision Analysis* by Milton C. Weinstein, Harvey V. Fineberg, et al.: W. B. Saunders Co., Philadelphia, London, Toronto, Mexico City, Rio de Janeiro, Sydney, and Tokyo, 1980.

56 The *relevance quotient* $Q(A)/P(A)$ plays the same rôle in updating probabilities that the Bayes' factor plays in updating odds. Where Q comes from P by conditioning on a data proposition, $Q(A)/P(A) = P(A|\text{data})/P(A) = P(\text{data}|A)/P(\text{data})$.

57 But here Carnap's notion of confirmation is not yet definitely probabilistic.

58 I thank Richard Creath, Burton Dreben, Frank Döring, Michael Friedman, Carl Hempel, Saul Kripke, and Michaelis Michael for setting me straight at various points.

Logic, Methodology and Philosophy of Science IX
D. Prawitz, B. Skyrms and D. Westerståhl (Editors)

THE LIMITS OF VINDICATION

HILARY PUTNAM

Department of Philosophy, Harvard University

Reichenbach's sense of the *need* for a justification of induction is poignantly expressed in the closing chapter of *Experience and Prediction*[1]. Reichenbach chides Hume, "He [Hume] is not alarmed by his discovery; *he does not realize that if there is no escape from the dilemma pointed out by him, science might as well not be continued*—there is no need for a system of predictions if it is nothing but *a ridiculous self delusion.* There are modern positivists who do not realize this either. They talk about the formation of scientific theories, but they do not see that, if there is no justification for the inductive inference, *the working procedure of science sinks to the level of a game and can no longer be justified by the applicability of its results for the purpose of actions.* .....If, however, we should not be able to find an answer to Hume's objections within the frame of logistic formalism, we ought to frankly admit that the antimetaphysical version of philosophy led to the renunciation of any justification of the predictive methods of science—led to a *definitive failure of scientific philosophy.*" (*op. cit.*, 346, emphasis added.) Indeed, Reichenbach suggests that not only "scientific philosophy" but life itself would be impossible if induction were not justifiable: "If we sit at the wheel of a car and want to turn the car to the right, why do we turn the wheel to the right? ...if we should not regard the inductive prescription and consider the effect of a turn of the wheel as entirely unknown to us, we might as well turn it to the left as well. I do not say this to suggest such an attempt; the effects of sceptical philosophy applied in motor traffic would be rather unpleasant. But I should say a philosopher who is to put aside his principles any time he steers a motor car is a bad philosopher."

At the same time, Reichenbach insisted that Hume had successfully shown that:

"1. We have no logical demonstration for the validity of inductive inference.

2. There is no demonstration a posteriori for the inductive inference; any

such demonstration would presuppose the very principles which it is to demonstrate."

"These two pillars of Hume's criticism of the principles of induction have stood unshaken for two centuries," Reichenbach tells us[2], "and I think they will stand as long as there is a scientific philosophy."

If the inductive inference cannot be "demonstrated" to lead to successful prediction either deductively or a posteriori, and yet it can still be "justified", that "justification" must do something other than demonstrate that we will obtain successful predictions in the long run by relying on induction; what Reichenbach claimed to have shown was that *induction will lead to the goal of successful prediction in the long run if any method will do so.* Thus, Reichenbach contended, we can concede that Hume did prove 1., and 2., above—that is, Hume did show that there is no possibility of a deductive or an a posteriori proof that the goal of successful prediction can be reached by induction (or by any other method, for that matter); but this important discovery of Hume's does not preclude the possibility that there is a proof of the *conditional proposition* that induction will reach the goal *if* the goal is attainable at all. This way of reconstruing the notion of a "justification of induction" is so radical a departure from what had been sought by Hume's critics (or so the positivists thought) that Herbert Feigl thought one should give it a new name. A proof that induction will lead to success would be a *justification* of induction, in Feigl's terminology; a proof of the conditional proposition that it will lead to success *if success is attainable* would be a *vindication*[3]. And, like Reichenbach, Feigl argued that while induction cannot be *justified* (in the sense Feigl gave to that term[4], nevertheless (he and Reichenbach had shown that) it can be *vindicated.* Moreover, both Reichenbach and Feigl held that *vindication is enough*; In *The Rise of Scientific Philosophy*[5], Reichenbach employs a vivid analogy, "The man who makes inductive inferences may be compared to a fisherman who casts a net into an unknown part of the ocean—he does not know whether he will catch fish, but he knows that if he wants to catch fish he has to cast his net. Every inductive prediction is like casting a net into the ocean of the happenings of nature; we do not know whether we shall have a good catch, but we try, at least, and try by the help of the best means available." What makes induction "the best means available", according to Reichenbach, is that it is the only[6] method concerning which we can *prove* that it will lead to successful prediction if any method will. Since the possibility of the long run success of induction is the necessary condition for the possibility of the long run success of any method at all, we are rationally justified in using induction if we want successful prediction. "We should at least actualize the necessary conditions of success if the sufficient conditions are not within our reach."[7]

Given the magnificent grandeur of what Reichenbach claimed to have achieved—nothing less than to have provided the rational warrant for empirical knowledge which he believed us to so desperately need—it is surprising that there is relatively little in-depth discussion[8] of Reichenbach's vindication argument in the literature, and most of that fails to address what seem to me to be the central issues. In the present contribution, I want to emphasize those issues, as I see them, and to assess the successes and the failures of Reichenbach's vindicatory strategy.

Reichenbach's arguments

Reichenbach imagines an immortal inquirer who is engaged in sampling from an infinite population. For the sake of an example, let us imagine that balls are being drawn successively from an infinite urn, and that each ball is either red or black. The immortal scientist finds that 262 of the first 1000 balls are black (let us say), that 2,489 of the first 10,000 are black, that 25,021 of the first 100,000 are black, ... (Perhaps the limit of the relative frequency of black balls in the infinite series is actually .25.) Reichenbach's Rule of Induction tells the scientist to *keep positing that the limit of the relative frequency of the attribute he or she is interested in is approximately equal (to within any given preselected level of accuracy) to the frequency in the sample of cases so far examined.* (Thus, in our example, if the preselected level of accuracy is "±.01", the scientist should "posit" that the limit of the relative frequency is .262 ± .01 when the first 1000 have been examined, that it is .2489 ± .01 when the first 10,000 have been examined, that it is .25021 ± .01 when the first 100,000 have been examined, ...)

Now, if the relative frequency of the attribute the scientist is interested in approaches a limit at all, then, no matter what the degree of accuracy the scientist is using (±.01, in our example), the relative frequency with which the attribute occurs among the N members of the series so far observed will differ from that limit by less than the degree of accuracy (by less than ±.01) after N reaches a certain size (called "the point of convergence"). In other words, the scientist, using Reichenbach's "Rule of Induction" may make a finite number of mistakes "at the beginning", but *once the point of convergence has been passed all of the posits that scientist makes will be correct.* This is simply a consequence of the definition of the concept "limit" as applied to a series of relative frequencies. (In our example, if the limit of the relative frequency of black balls in the infinite sequence is actually .25, then the first "posit" was incorrect, and the subsequent ones were correct.)

But why does Reichenbach believe that all prediction problems can be reduced to estimates of limits of relative frequencies? Well, consider a case

that might seem recalcitrant to such treatment. Imagine that the balls do not come out of the urn with a random distribution of colors at all. Instead, imagine that we see a black ball come out of the urn, then we see nine red balls, and then we see successive black balls until ninety percent of the observed balls have been black, after which we see successive red balls until ninety percent of the observed balls have been red, then successive black balls until ninety percent of the observed balls have been black ... and so on. This series is such that it happens infinitely often that the relative frequency of black balls among the ones so far observed is .90 and infinitely often that the relative frequency of black balls among the ones so far observed is only .10. Clearly, in this series there is no "limit of the relative frequency" with respect to the attribute "black". But once we "tumble" to what is going on we will have no difficulty in predicting the color of any individual ball that may be examined in the future. Does this not show that successful prediction does not require the existence of limits of the relative frequency?

Reichenbach's strategy with cases like this[9] is to define a new attribute. Thus, suppose we have formulated a law which we believe to be the law obeyed by our sequence (in the case of the example, the law would be that the sequence begins with a black ball followed by a "run" of red balls; and when there is a run of red balls the run always continues until a point is reached at which at least ninety percent of the balls which have been drawn are red, and as soon as this happens a run of black balls always begins; and when there is a run of black balls, the run always continues until a point is reached at which at least ninety percent of the balls which have been examined are black, and as soon as this happens a run of red balls always begins.) If the color of a ball is what it would be if the sequence followed our law, we say the ball has the attribute **C**, and otherwise that it has the attribute **not-C**. Then to know the *reliability* of our law, what we need to know is the relative frequency in the long run (i.e., the limit of the relative frequency) of **C**. And only if such a reliability estimate can be made, Reichenbach argues, does our "law" give us any means of control over what happens. Thus all successful prediction depends on the existence of limits of the relative frequency, although not necessarily on the existence of limits of the relative frequency of the attribute we are ultimately interested in.

One obvious problem is that Reichenbach's vindication deals with success in *the infinite long run.* The relevance of such a vindication to the projects of we all-too mortal members of a species which will not last forever in a universe which may not last forever is highly problematic! Although this is a problem Reichenbach himself discusses, it is not the problem I shall say much about today.[10] Today I will accept the idealization that Reichenbach works with, the idealization to an immortal inquirer who wants to discover limits

of the relative frequency (or methods for making successful predictions), and is willing to accept an arbitrarily long finite initial sequence of failures (or wrong posits).

Given this idealization, the position seems to be this. Reichenbach has shown that if successful prediction (in the indefinitely long run) is possible at all, then some limits of the relative frequency must exist and be capable of being estimated by us. (This is the argument we just reviewed.) And he has observed that it is a tautology that if there are limits of the relative frequency, then by using the "Rule of Induction" we shall eventually make correct posits about them (and no further incorrect ones, once the point of convergence has been reached). Does this not constitute precisely the vindication of induction that we were promised?

The problem of the consistency of Reichenbach's "Rule of Induction"

The problem with Reichenbach's vindicatory argument is so stunningly simple that it is surprising how little it has been discussed. The problem was, in effect, first pointed out by Nelson Goodman, although he did not refer to the issue of vindicating induction but instead pointed out that there was a prior issue, the issue of *saying what induction is.* Straightforward traditional formulations of the rule of induction – including Reichenbach's—*lead to inconsistent predictions.*

To see the bearing of this observation on Reichenbach's argument, consider the analogous problem of justifying a system of deductive logic. One cannot justify such a system merely by proving that all valid formulas in the appropriate vocabulary are theorems of the system. As it stands, such a result is not sufficient. (It is also not necessary, but that is not relevant here.[11]) It is not sufficient because this kind of "completeness" is trivially possessed by all *inconsistent* systems. Before we are impressed by a proof that " if S is valid, then S is a theorem of the system" we need to make sure (by a proof, if possible, and by 'mathematical experience' otherwise) that the system in question is consistent.

The same thing is true in the case of *induction.* Reichenbach's celebrated conditional, "if successful prediction can be made by any method, it can be made by using the Rule of Induction" *needs to be supplemented by a proof that the rule of induction is consistent.* But, unfortunately, Reichenbach's rule is not consistent.

Perhaps Goodman's discovery has not been taken seriously enough by philosophers who have discused Reichenbach's vindication of induction (it has, of course, been taken very seriously in other connections) because Good-

man's predicate *grue*[12] is so "peculiar". But it is difficult to know how such "peculiar " predicates are to be ruled out, or what would be left of Reichenbach's argument if they were. The predicate **C** we used above is just as "peculiar" as "grue" (it takes a calculation to determine whether a given item in a series has to be red or black to have the attribute **C**); but *Reichenbach's demonstration that all successful prediction depends on the existence of limits of the relative frequency turns precisely on the fact that the rule of induction accepts predicates like* **C**. Certainly Reichenbach did not seem to appreciate the seriousness of the problem; in the English edition of his *Theory of Probability*[13], he dismisses the problem with the strange remark that "with respect to consistency, inductive logic differs intrinsically from deductive logic; it is consistent not *de facto* but *de faciendo*, that is, not in its actual status, but in a form to be made."

Indeed, the problem is much more severe than the example given by Goodman would suggest. For the fact is that when we have observerved some finite initial part of a sequence, the data we have obtained can always be reconciled with more than one possible universal hypothesis (this is a special case of the familiar "underdetermination of theory by evidence"). For example, if we have observed that the successive runs of red balls in a certain sequence have the lengths of the prime numbers up to 103, then using Reichenbach's Rule of Induction we might project the hypothesis[14] "the successive runs of red balls will always have the lengths of the successive prime numbers," but we might also project incompatible hypotheses (for example, the hypothesis that the sequence will repeat from the beginning once there has been a run of 103 successive red balls). In this respect, the inconsistency of Reichenbach's Rule of Induction is quite different from the inconsistency of a system of axioms. When a system of axioms is inconsistent, say a system of set theory, it sometimes takes a great deal of work to derive a contradiction from the axioms. And one might (it has been suggested, at any rate) seek to "localize" the contradiction by *modifying the propositional calculus* so that it will no longer be the case that from a single contradiction every formula of the system can be derived (I am thinking of the possibility of employing relevant logic instead of classical truth functional logic). But in the case of Reichenbach's Rule of Induction, whenever the Rule is employed to project a universal hypothesis, it can also be employed to project a directly inconsistent universal hypothesis; there is no need to derive one esoteric contradiction (perhaps involving "grue") and then employ the principle that from a truth-functional contradiction every assertion follows.

I confess that many times over the years I have wondered about what Reichenbach meant by saying that "[inductive logic] is consistent not *de facto* but *de faciendo*, that is, not in its actual status, but in a form to be

made." I think that what he *must* have meant is this: to keep my inductions consistent, what I must do is *choose* ; if I project the hypothesis that all the prime numbers will appear in order (as the lengths of "runs" of red balls), I must not simultaneously project any hypothesis which is inconsistent with that hypothesis *until* the projected hypothesis has been falsified. I may, of course, choose the "wrong" attribute to project initially; I may, to use Goodman's example, initially project "grue" rather than "green"; but the hypothesis that "all emeralds are grue" will be falsified as soon as green emeralds are examined after time t, and then I will (sooner or later) project the more useful attribute "green". (Or so Reichenbach might have thought.) Similarly, in terms of the example I used a moment ago, I may initially project the hypothesis that the series will repeat after a run of 103 red balls has occurred, but I will give up that hypothesis when I get a run of 107 red balls, and then I will (sooner or later) project the hypothesis that the lengths of the successive runs of red balls are the successive primes.

If this is what Reichenbach meant, however, it raises an interesting and very deep question: *What guarantee is there that the right hypothesis (or the corresponding attribute* **C***) will ever be projected*, if there is a right hypothesis?

The example of the clairvoyant

One indication of Reichenbach's thinking here is the discussion of non-inductive methods of prediction in *Experience and Prediction.* Reichenbach imagines an objector who argues that "there might be a clairvoyant who knows every event of a series individually, who could foretell precisely what would happen from event to event" (*op. cit.* 358). And he replies that in such a case we could use induction to discover how reliable the clairvoyant was; in his classes on Inductive Logic at U.C.L.A. this was one of Reichenbach's favorite ways of illustrating his claim that induction must succeed if any method does.

This example, as it stands, does not speak to the problem of *consistency*, but it is easy enough to modify it so that it does. In my "Degree of Confirmation and Inductive Logic,"[15] I introduced a "Method M" in which "effective hypotheses" receive *priority rankings* based upon the time at which they are proposed. (I pointed out[16] that more sophisticated priority rankings are possible.) An effective hypothesis is one which says of each individual in a series whether it has some property **P**; and it does so *effectively* in the sense that it is possible to determine recursively whether the hypothesis implies that the individual in the nth place of the series is **P** or **not-P** (for n=1,2,3...). Thus an effective hypothesis is one that we can actually compare with the

data; we can calculate what the character of the members of the series is supposed to be, and see whether the predictions of the hypothesis agree with the facts as more and more individuals are observed. The following rules defined the method M[17]:

1. Let $P_{t,M}$ be the set of hypotheses considered at time t with respect to a property **M** of the given series. I.e. $P_{t,M}$ is a finite set of effective hypotheses each of which specifies for each member of the series whether or not it is **M**.

2. Let $h_{t,M}$ be the effective hypothesis *accepted* at time t (if any). I.e., we suppose that, at any given time, various incompatible hypotheses have been actually suggested with respect to a given **M**, and have not yet been ruled out (we require that these be *consistent* with the data and with all other accepted hypotheses with respect to other series and other predicates). In addition, one hypothesis may have been accepted at some time prior to t, and may not yet have been abandoned. This hypothesis is called "the accepted hypothesis at the time t".

3. (Rule I:) At certain times t_1, t_2, t_3 ... initiate an *inductive test* with respect to **M**. This proceeds as follows. The hypotheses in $P_{t,M}$ at this time t_i are called the *alternatives*. Calculate the character (**M** or **not-M**) of the next individual on the basis of each alternative. See which alternatives succeed in predicting this. Rule out the ones that fail. Continue until (a) all the alternatives but one have failed; or (b) all the alternatives have failed (one or the other must eventually happen). In case (a) accept the alternative that does not fail. In case (b) reject all the alternatives.

4. (Rule II:) hypotheses suggested in the course of the inductive test are taken as alternatives (unless they have become inconsistent with the data) in the *next* test. I.e., if h is proposed in the course of the test begun at t_3, then h belongs to $P_{t_4,M}$ and not to $P_{t_3,M}$.

5. (Rule III:) If $h_{t,M}$ is accepted at the conclusion of any inductive test, then $h_{t,M}$ continues to be accepted as long as it remains consistent with the data. (In particular, while an inductive test is going on the previously accepted hypothesis continues to be accepted, as long as it is consistent with the data.)

It is built into the method M that one cannot accept mutually inconsistent hypotheses (cf. 2. above). Yet, simple as the method M is, it is easy to show that:

(Completeness property of M:) *If h is an effective hypothesis and h is true, then, using method M, one will eventually accept h if h is ever proposed.*

Thus, modifying Reichenbach's argument, we have been able to show that there is a method which will converge to the right hypothesis if there *is* a right hypothesis (in a certain class), and which is not plagued with consistency problems. Extending this kind of argument to other (and larger) classes of hypotheses has become an important branch of inductive logic in recent years.[18] In this sense, Reichenbach's "vindicationist" strategy *has* born fruit. But has induction really been vindicated?

The method M^*

Although the completeness property of M is certainly an attractive one, it falls short of the vindication Reichenbach hoped for for several reasons. (1) Whether we will get successful prediction using M depends on what effective hypotheses are actually proposed. To show that we can achieve successful prediction if successful prediction is possible at all, we would need to make sure that the immortal inquirer sooner or later considers *every* effective hypothesis. (2) As it stands, M does not accept hypotheses like "the clairvoyant's predictions about the series are always right". Yet, although this hypothesis is not "effective" in the sense that we can *calculate* from it what the character of an arbitrary future member of the series must be, we do have a *procedure* for telling what it will be if this h is correct, namely to ask the clairvoyant. To show that we can achieve successful prediction if successful prediction is possible at all, we would need to make sure that the immortal inquirer also considers (sooner or later) every hypothesis which leads to a procedure for telling what the character of a future member of the series will be, even if the procedure is not a calculation. (3) The assumption that humans cannot compute *nonrecursive* functions, which we have made throughout, would have to be justified without relying on induction. I shall defer discussion of (2) and (3) to the next section; in the present section we shall see how difficulty (1) can be met.

The problem is the following: if we could arrange all testable hypotheses in a single list, and then proceed to test hypotheses in the order in which they occurred in our list (in the way illustrated by the method M), then we could ensure that if there is a correct testable hypothesis, sooner or later it would get accepted. However, the list must be a list of *different*, indeed of *incompatible* hypotheses. To see why, suppose that two of the hypotheses in $P_{t,M}$ are in fact equivalent (or at least compatible) although we mistakenly

believe that they are incompatible. Then, if they happen to both be true, the inductive test prescribed by Rule I will never terminate! I "ducked" this problem when I formulated the Method M by restricting the sets $P_{t,M}$ to sets of *incompatible* hypotheses; but, if our scientific language is rich enough to contain number theory, there will not be any effective method for telling whether an arbitrary pair of effective hypotheses *is* incompatible or not! This is an "undecidable problem" in recursion theory. Moreover, if an effective hypothesis has to make a prediction about *every* member of the sequence, there is no effective method for telling whether a hypothesis *is* "effective" or not; however, we can handle this problem by allowing all hypotheses of the form "(n)(the nth ball in the sequence is red if and only if f(n)=1)", where f is a *partial* recursive function to count as effective; for the partial recursive[19] functions (as opposed to the general recursive functions) can be effectively enumerated.

However, Reichenbach's puzzling remark about inductive logic being "consistent not *de facto* but *de faciendo*, that is, not in its actual status, but in a form to be made" suggests that he was not thinking of proceeding in accordance with a *fixed* priority ordering of the hypotheses at all; he seems to have assumed his immortal inductive inquirer could simply rank order the hypotheses as he went along. But then, there will be no guarantee that the immortal inquirer will consider *every* hypothesis that could lead to successful prediction; and this is what Reichenbach's vindicatory argument assumes.

As I thought about this difficulty, it occurred to me that a present-day Reichenbachian, apprised of the surprising relevance of Gödel's and Church's work on undecidable problems to inductive logic (as Reichenbach himself, of course, was not[20]) might plausibly propose modifying the method M by simply letting the sets $P_{t,M}$ be larger and larger subsets of the set of all hypotheses of the form "(n)(the nth ball in the sequence is red if and only if f(n)=1)", where f is a computable (partial recursive) function (thus giving up the requirement that the sets $P_{t,M}$ consist of *incompatible* hypotheses), and replace Rule I by the following rule:

(Rule I*:) At certain times t_1, t_2, t_3 ... initiate an *inductive test* with respect to **M**. This proceeds as follows. The hypotheses in $P_{t,M}$ at this time t_i are called the *alternatives*. Start calculating the character (**M** or **not-M**) of future individuals (not necessarily the *next* individual[21]) on the basis of each alternative. See which alternatives succeed in these predictions. Rule out the ones that fail. Continue until (a) the alternatives that have still not failed are *not known to be incompatible*; or (b) all the alternatives have failed (one or the other must eventually happen). In case (a) accept *all* the alternatives that have not failed. In case (b) reject all the alternatives.

Call the modified method "Method M^*". Then if the immortal inquirer

uses Method M^* instead of Method M he or she may, indeed, accept two or more mutually inconsistent hypotheses: but there is a useful strategy the immortal inquirer can follow to take care of this eventuality. This strategy, which is familiar to recursion theorists, is (1) to assign a priority ranking *without ties* to all hypotheses (so that the hypotheses within a single $P_{t,M}$ will be assigned different priorities, instead of being treated as a single "clump" as they were in the method M); (2) when the conjunction of two or more accepted hypotheses is discovered to be inconsistent, the one with highest priority continues to be accepted (until and unless it becomes inconsistent with the data), and other hypotheses in the set of inconsistent hypotheses are rejected, starting with the hypothesis of *lowest* priority, until consistency is restored. For example, if h_1, h_2, h_3 are all accepted (where h_1 has a higher priority than h_2, which in turn has a higher priority than h_3) and their conjunction has been discovered to be inconsistent, although the conjunction of the first two is not known to be inconsistent, then we would reject only h_3. (3) If one or more of the hypotheses that survive this process is rejected at a later time, then hypotheses which were rejected because of their inconsistency with hypotheses which have now been falsified are to be reconsidered (i.e., incorporated into the next set $P_{t,M}$) unless they themselves have been falsified by the data in the meanwhile. If we add these rules to M^*, our method has the attractive property:

> (Completeness property of M^*:) *If h is an effective hypothesis and h is true, then, using method M^* one will eventually accept h.*

Moreover, even though it can happen that at one time one does accept hypotheses which are inconsistent using M^*, this is rectified as soon as it is discovered. In this sense, M^* realizes Reichenbach's idea that "inductive logic ... is consistent not *de facto* but *de faciendo*, that is, not in its actual status, but in a form to be made".

The limits of "vindication"

At the beginning of the last section, I pointed out three reasons why the completeness property of M falls short of the justification (or "vindication") Reichenbach hoped for: (1) whether we get successful prediction using M depends on what effective hypotheses are actually proposed. To show that we can achieve successful prediction if successful prediction is possible at all, we need to make sure that the immortal inquirer sooner or later considers *every* effective hypothesis. (2) M does not accept hypotheses like "the clairvoyant's predictions about the series are always right". Yet, although

this hypothesis is not "effective" in the sense that we can *calculate* from it what the character of an arbitrary future member of the series must be, we do have a *procedure* for telling what it will be if this h is correct, namely to ask the clairvoyant. To show that we can achieve successful prediction if successful prediction is possible at all, we need to make sure that the immortal inquirer also considers every hypothesis which yields a procedure for telling what the character of a future member of the series will be, even if the procedure is not a calculation. (3) The assumption that humans cannot compute *nonrecursive* functions, which we have made throughout, has not been justified. Difficulty (1) has now been dealt with. What can we say about (2) and (3)?

Difficulty (3) can be subsumed under (2) in the following way: if it turns out (contrary to what we now believe) that there is a method by which humans (or idealized immortal counterparts of human inquirers) can compute nonrecursive functions, then those procedures, whatever they may be, can be incorporated in a hypothesis of the kind envisaged under difficulty (2), a hypothesis to the effect that a procedure P will always tell us what the character of a future member of a series will be ("tell us" by a method other than algorithmic calculation). Thus we can confine attention to (2).

The general form of hypotheses like the one about the clairvoyant is: "If you do X and get result Y, then the nth member of the sequence will be M" (where X must depend on n). The problem is that is impossible to know how such hypotheses can be "enumerated" when we do not know in advance what predicates they may contain. Indeed, even if we restrict attention to so-called "observation predicates", how can we know what the limits to human powers of observation are? The answer might seem simple, and perhaps it *is* simple, *if we are allowed to use empirical knowledge, knowledge obtained by induction*, but Reichenbach agrees with Hume that "there is no demonstration a posteriori for the inductive inference; any such demonstration would presuppose the very principles which it is to demonstrate". Clearly a "vindicatory" argument any of whose premisses requires a " demonstration a posteriori" would by unacceptable for the same reason.

Again, I believe that Reichenbach's attitude would be that *completeness* of our inductive method, like consistency, is "not *de facto* but *de faciendo*". If the immortal observer learns at some point that there are more predicates needed for prediction, or even more observation predicates, then he suspected, he can simply enlarge his language and adjust his inductive method accordingly. Formally, this means that M^* will be enlarged by allowing the sets $P_{t,M}$ to contain hypotheses of the form " If you do X and get result Y, then the nth member of the sequence will be M" if such hypotheses are proposed, in addition to containing larger and larger subsets of the set of all

effective hypotheses. (It also means that the *language* to which the method M^* is applied is not fixed once and for all.) And again one can obtain a completeness property which says that:

> *If h is a hypothesis of the form specified* ["If you do X and get result Y, then the nth member of the sequence will be M"] *and h is true, then, using method M, one will eventually accept h if h is ever proposed.*

While still keeping the completeness property that

> *If h is an effective hypothesis and h is true, then, using method M^* one will eventually accept h.*

If we confine attention to hypotheses which have the form of universal laws (i.e., if we put aside for the moment the more general question of *confirming statistical laws* which interested Reichenbach), does this not give us everything Reichenbach claimed?

There is no question but that these completeness properties of certain inductive methods are of great theoretical interest. As I have already mentioned, a whole little "industry" has developed to study them. And this is certainly a success of Reichenbach's vindicatory strategy, even if the details are much more complicated that Reichenbach anticipated.

It is not, however, a complete success. Reichenbach's argument about the clairvoyant already has a feature that should now be conspicuous: the immortal inductive inquirer only considers the clairvoyant hypothesis *if he or she encounters a clairvoyant.* That it, Reichenbach does not really attempt to describe a method by which the immortal inquirer will (in the long run) arrive at successful prediction if successful prediction is possible (which was his announced goal); what he really describes is a method by which the immortal inquirer will arrive at successful prediction *if anyone else does* (and the immortal inquirer learns about it). But is this not enough?

Well, it would be enough *if the immortal inquirer could be sure that what methods other people use are independent of the method the immortal inquirer chooses.* But there is no reason to believe that this will be the case.

The problem is this. Reichenbach, in effect, says to the immortal inquirer, "You have nothing to lose (in the long run) by using the Rule of Induction. Using it will only increase your chances of successful prediction; for if either you yourself or anyone else thinks of a way of getting successful prediction, then you will get successful prediction too." In effect, Reichenbach claims that induction is the *dominant strategy.* But if the policies followed by the

other inquirers are influenced by the policy-choice of our immortal inducer, then the immortal inducer may conceivably *fail* to achieve successful prediction in the long run (and everyone else may fail as well) *even though someone — perhaps our immortal inquirer —* would *have achieved successful prediction if our immortal inquirer had* not *used induction*! The reason is that, as we have seen, the immortal inquirer cannot be sure of testing *all hypotheses that could conceivably lead to successful prediction* (or at least, the immortal inquirer cannot do this without relying on empirical laws confirmed by induction.) The hypotheses that the immortal inquirer tests depend on what other people do. But then, it could be the case that if the immortal inquirer had not used induction someone else would have thought of a method which is not even describable in the language of the immortal inquirer as it presently stands, and whose reliability will, therefore, never get tested in the inductive way that Reichenbach describes, because the immortal inquirer's choice of induction brought it about that this method was not even thought of. Indeed, if the immortal inquirer had not been a believer in induction, he himself might have thought of such a method, or been persuaded by someone else to use it. In sum, there is no proof that induction is the dominant strategy, when the policies used by other inquirers are not independent of the policy-choice of the immortal inquirer.

In plain empirical fact, belief in the superiority of any method, belief that any method is identical with "science", does certainly effect the methods and the hypotheses that occur to other people. Thus there is no reason at all to belief in the independence assumption which Reichenbach's argument (in a hidden way) requires. If we did not believe in induction, it is certainly logically possible that other methods would be tried—and logically possible that they would succeed — that will never get tried (and hence never get tested) in the actual world. In short, there is a logically possible world in which (immortal) people use induction and fail to make successful predictions, although those same people *would have made* successful predictions if they had not used induction (using a method which was never tried in that world—never tried *because* induction was persisted in instead). Thus, there is no deductive proof that induction will succeed (even in the long run) provided successful prediction is possible at all, and no deductive proof that using induction is the dominant strategy.

Finally, a word about the problem to which my own Ph.D. dissertation was devoted, the problem of extending Reichenbach's "justification of induction" (i.e., the vindicatory argument) to "the finite case". All of the positive results we have sketched here depend on the availability of unlimited future time; thus it seems clear that even the partial results in the direction Reichenbach wanted that we have been able to obtain in the form

of completeness properties for various inductive methods have no analogues in the finite case. (My own attempt, in the dissertation, at a justification for the finite case was invalidated by the discovery of the inconsistency of the Rule of Induction.)

But even if Reichenbach's aim of deductively vindicating induction has turned out to be an unattainable one, the discussion of Reichenbach's argument leads into profound depths. My own present stance is partly like Wittgenstein's. I agree with Wittgenstein that "The 'law of induction' can no more be *grounded* than certain particular propositions concerning the material of experience"[22], and I further agree that "[the language game] is not based on grounds. It is not reasonable [*vernünftig*] or unreasonable. It is there — like our life." Where I perhaps differ with Wittgenstein is in finding attempts like Reichenbach's of permanent value nonetheless.

---

[1] Published by the University of Chicago Press, 1938 and 1970. See 346ff.

[2] *Op. cit.*, 342.

[3] Cf. Herbert Feigl, "De Principiis Non Disputandum...? On the Meaning and the Limits of Justification," in M. Black (ed.), *Philosophical Analysis*, pp. 119-156, Cornell University Press, 1950. In another article (" Some Major Issues and Developments in the Philosophy of Science of Logical Empiricism," in *Minnesota Studies in the Philosophy of Science, vol. 1; The Foundations of Science and the Concepts of Psychology and Psychoanalysis*, University of Minnesota Press, 1956, 3-37) Feigl writes (p.29), "In some of my early papers I had been groping for this sort of solution of the problem of induction, and I think I came fairly close to a tenable formulation in the paper of 1934 [Feigl means "The Logical Character of the Principle of Induction," *Philosophy of Science*, 1:20-29 (1935)]. But with genuine appreciation I credit the late Hans Reichenbach with the independent discovery and the more elaborate presentation of this solution.".

[4] Reichenbach, it should be noted, did not use Feigl's terminology; he speaks of his purported demonstration of the conditional proposition simply as a "justification" of induction.

[5] Published by the University of California Press, 1951. See 245-246. Reichenbach employs the same analogy in *Experience and Prediction*, cf. 362-3.

[6] Actually, Reichenbach has a problem with *uniqueness*. As he himself points out (*Experience and Prediction*, 354) "Now it is easily seen not only that the inductive method will lead to success, but that every method will do the same if it determines as our wager the vale $h_n + c_n$, where $c_n$is a number which is a function of n, or also of h_n, but bound to the condition $\lim_{n\to\infty} c_n = 0$. ...all such methods must converge asymptotically. The inductive principle is the special case where $c_n = 0$". Reichenbach has two strategies for dealing with this problem; one is to argue that there is less "risk" in chosing the value $c_n = 0$ because "any other value may worsen the convergence" (p. 355) ; this is clearly fallacious, since the choice of $c_n = 0$ may also "worsen the convergence". The other strategy is to say that (for the immortal inquirer) the choice is "not of great relevance, as all these methods must lead to the same value of the limit if they are sufficiently continued."(P. 356).

[7] *Experience and Prediction*, 362.

[8] One of the few philosophers to devote sustained examination to Reichenbach's argument

was Max Black. See his *Language and Philosophy*, Cornell University Press, 1949.

[9] See, for example, *Experience and Prediction*, 358.

[10] I attempted to meet this objection to Reichenbach's vindication argument in my Ph.D. thesis, published as *The Meaning of the Concept of Probability in Application to Finite Sequences* (Garland, 1990, in the series *Harvard Dissertations in Philosophy*, ed. Robert Nozick). For reasons given in the "Introduction Some Years Later" to that work, and developed at greater length here, I no longer regard my attempt as successful.

[11] It is not necessary because we have learned to live with the fact that a good system may be incomplete, and indeed incompletable, for Gödelian reasons. But this point is not to my purpose here; my point here is that completeness, even when attainable, is no virtue unless the system is *consistent.*

[12] "Now let me introduce another predicate less familiar than 'green'. It is the predicate 'grue' and it applies to all things examined before t just in case they are green but to other things just in case they are blue. Then at time t we have, for each evidence statement asserting that a given emerald is green, a parallel evidence statement asserting that that emerald is grue. And the statements that emerald a is grue, that emerald b is grue, and so on, will each confirm the general hypothesis that all emeralds are grue." Nelson Goodman, *Fact, Fiction and Forecast*, 4th edition, Harvard 1983, 74-75.

[13] Published by the University of California Press, 1940.

[14] Here and in what follows, when I speak of "projecting a hypothesis using the Rule of Induction" what I mean is estimating the "reliability" of that hypothesis by using the Rule of Induction applied to an appropriate attribute **C**.

[15] Reprinted in my *Philosophical Papers, volume 1, Mathematics, Matter and Method*, 270-292 (Cambridge University Press, 1975).

[16] *loc. cit.* 283-4.

[17] These are given in the paper cited in n. 15, 279-80.

[18] Cf. Peter Kugel, "Induction, Pure and Simple," *Information and Control*, 33 (1977), 276-336; D. Osherson and S. Weinstein, "Identification in the Limit of First Order Structures,", *Journal of Philosophical Logic*, 15 (1986), 55-81; (by the same authors) "Paradigms of Truth Detection," *Journal of Philosophical Logic*, 18 (1989), 1-42; D. Osherson, M. Stob, and S. Weinstein *Systems that Learn*, M.I.T. Press (1986).

[19] Partial recursive functions are functions which are computable, but not necessarily defined on all integers. (In general, whether a partial recursive function is defined on a given integer is an unsolvable problem.) Partial recursive functions which *are* defined on all integers are called "general recursive"

[20] For the relevance of Gödel's and Church's work on undecidable problems to inductive logic see, in addition to the literature cited in n. 18, my " Trial and Error Predicates and a Solution to a Problem of Mostowski," *Journal of Symbolic Logic*, 30:1 (1965), 49-57, E.M. Gold, "Limiting Recursion," *Journal of Symbolic Logic*, 30:1 (1965), 27-48; the paper cited in note 15; and my "Reflexive Reflections," *Erkenntnis*, 22, 143-153 (1985).

[21] The reason for changing "the next individual" to "future individuals" is that, since we do not require the functions f to be total, an effective hypothesis may not in fact predict anything about the character of one or another individual. Also it may take an arbitrarily long time to compute what a hypothesis does predict about a given individual, if it does make a prediction; thus the fact that the user of this method *goes on forever* (i.e., is immortal), is essential.

[22] *On Certainty*, §499.

Logic, Methodology and Philosophy of Science IX
D. Prawitz, B. Skyrms and D. Westerståhl (Editors)

STIG KANGER IN MEMORIAM

DAGFINN FØLLESDAL

University of Oslo and Stanford University

The original invitation to hold this meeting in Uppsala came from Stig Kanger. Dag Prawitz and Dag Westerståhl had to take over the responsibility for the meeting when Kanger died on March 13, 1988, in his 64th year, on his way to Germany to recieve the Alexander von Humboldt prize for his research.

It is doubly appropriate to devote a symposium at this congress to Kanger's work. Not only was this to have been *his* congress. His work is also of the highest quality and deserves to be far better known. Later in this symposium Lars Lindahl, who worked together with Kanger for many years, will present and discuss Kanger's work on rights, and Amartya Sen will talk about Kanger's work on choice, preference and binariness.

However, Kanger made important contributions to many other areas of philosophy as well. He gave elegant formulations of various branches of elementary logic, many of them inspired by his teacher Anders Wedberg and by Tarski's work, which he admired and knew well. Kanger had new and interesting ideas on the theory of measurement and he even contributed to linguistics: in a paper on the notion of a phoneme and in unpublished work on time and tempus, where he presents an approach that is far superior to what was at that time available in the work of Reichenbach and Prior.

Kanger picked up new formal theories very easily, he quickly appreciated their basic point and their inherent difficulties. He spotted inconsistencies straightaway and came up with counterexamples promptly. As an example I can mention that some years ago he was in Oslo listening to Donald Davidson presenting a first version of his "Unified Theory of Language and Action." Kanger immediately saw that Davidson's basic idea of preferring a sentence to be true, led to contradictions, and he came up with a simple example to show this. Davidson was forced back to the drawing-board.

Kanger's little remark has been preserved for posterity in his "A Note on Preference-Logic." (In *Festschrift* for Thorild Dahlquist[1]).

Kanger's critical talent coupled with creativity, resourcefulness and great generosity made him an exceptional adviser, and through his advising and other contributions that have not found their way into writing, he has been of immeasurable importance for philosophy in Uppsala and among us others who had the opportunity to discuss philosophy with him.

I want, however, particularly to emphasize Kanger's development of a model theory for the modalities, in his 1957 dissertation, *Provability in Logic*,[2] which goes back to a course in logic that he gave at the University of Stockholm in the spring of 1955.

Most work done on the modalities until then was of a syntactic character. C. I. Lewis' varieties of propositional modal logic and Ruth Barcan Marcus' and Carnap's systems of quantified modal logic from 1946 all focus mainly on the syntactical. Carnap had proposed a quasi-semantic interpretation in terms of state descriptions, but a clear and fully semantic interpretation had never been proposed. Kanger, in his dissertation, proposed the first fully model theoretic interpretation of modal logic. Moreover, he introduced a fundamental new idea, which now is familiar to everybody working on modal logic, but which at that time was an innovation: While all earlier attempts to provide a semantics for modal logic, from Leibniz to Carnap, had started out from the idea of necessity as truth in all possible worlds, Kanger regarded the notion of a possible world as a *relative* notion. One world may be possible relative to some other worlds, and not possible relative to further worlds.

Kanger shows that by imposing various restrictions on the relation between possible worlds, for example requiring it be reflexive, symmetrical and/or transitive, one gets natural interpretations of the systems $S5$ and $S4$ of C. I. Lewis and of Feys' calculus t. These results are now among the first things a student of modal logic learns, as the basis of so-called Kripke semantics for modal logic. I think it would be appropriate for us who are present at this meeting to give credit where credit is due and to honor Kanger's memory by calling this semantics *Kanger-Kripke semantics*. I am sure that Saul Kripke, who is here, would not object. Neither will, I am sure, Jaakko Hintikka, who is also here, and who suggested

[1] *ThD 60. Philosophical essays dedicated to Thorild Dahlquist on his sixtieth birthday.* (Philosophical Studies published by the Philosophical Society and the Department of Philosophy, University of Uppsala, Sweden, No. 32) Uppsala 1980, pp. 37–38.

[2] Stig Kanger, *Provability in Logic* (Stockholm Studies in Philosophy 1), Almqvist & Wiksell, Stockholm, 1957.

the idea of worlds being possible relative to one another in two papers in 1957,[3] the same year as Kanger's dissertation, but without working the idea out in any detail. Kripke, by the way, has never proposed the label "Kripke semantics" for the semantics that he developed in his 1963 and '64 papers.[4] The label has come into use because Kripke's papers were very widely read and highly influential, while hardly anybody read Kanger's dissertation. Also, Kanger confined himself to stating his main ideas and results in a very condensed manner, while Kripke developed the theory much further and presented it in a pedagogical and very readable way. Kripke, of course, refers to Kanger in his articles. Thus in "Semantical Analysis of Modal Logic I" (page 69, footnote 2), he says: "The modeling for modal logic given in Kanger [Provability in Logic], though more complex, is similar to that in the present paper."

Kanger must himself bear the blame for the lack of attention given to his dissertation. Although he modestly says in his preface that "the essay may be regarded as having a kind of pedagogical aim", pedagogy is not its main virtue. Typically, shortly after Saul Kripke had published his 1959 paper in the *Journal of Symbolic Logic*,[5] where he gives a clear and simple model theoretic semantics for $S5$, but treats the notion of possible world in the old-fashioned way, as an absolute notion and not as a relative one, I mentioned to him that Kanger in his dissertation had introduced the idea of possible world as a relative notion and thereby got natural models also for weaker systems than $S5$. Kripke then told me that he was aware that this idea was in Kanger's dissertation, but that he had tried to read it and found it forbidding to read in its very dry and condensed formal style.[6] Kanger himself may have had a foreboding of this when he wrote in the preface to his dissertation that "It is hoped that this [pedagogical] aim is not overshadowed by the technical character of my exposition".

---

[3] Jaakko Hintikka, "Quantifiers in Deontic Logic", Societas Scientiarum Fennica, *Commentationes humanarum litterarum*, vol. 23, no. 4 (Helsinki 1957). "Modality as Referential Multiplicity", *Ajatus* 20 (1957), 49–64, esp. pp. 61–62.

[4] Saul Kripke, "Semantical Considerations on Modal Logic", *Acta Philosophica Fennica* 16 (1963), 83–94. "Semantical Analysis of Modal Logic I, Normal Propositional Calculi", *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik* 9 (1963), 67–96. "Semantical Analysis of Modal Logic II, Non-normal Modal Propositional Calculi", in *The Theory of Models*, edited by J. W. Addison, L. Henkin, A. Tarski (Amsterdam: North-Holland, 1965), pp. 206–220.

[5] Saul Kripke, "A Completeness Theorem in Modal Logic". *Journal of Symbolic Logic* 24 (1959), 1–14.

[6] Among Kanger's papers, that are kept in Uppsala University Library, is a letter from Saul Kripke to Kanger, dated January 24, 1958, where Kripke asks whether Kanger will send him a copy of his dissertation and of his article "A Note on Partial

From the perspective of our time Kanger's dissertation is not hard to read. Kanger gives ample credit for the basic ideas of his model theoretic approach, including the idea of bringing in a relation between model, to Bjarni Jónsson and Alfred Tarski's 1951–52 articles "Boolean algebras with operators", which he had studied very carefully.[7] However, there is no hint in Jónsson and Tarski that their ideas can be applied to modal logic. Stig Kanger was the first to do so, and he thereby laid the ground for what is in our time the dominant approach to the semantics for the modalities. Let us give him proper credit for having done this by including his name in the designation for this semantics.

Postulate Sets for Propositional Logic" (*Theoria* 21 (1955)), which had been brought to Kripke's attention by Haskell Curry. (I am grateful to Mrs. Dagmar Kanger for this information.)

[7]Strangely enough, in the footnote in Kripke's "Semantical analysis of modal logic I", which I quoted above, Kripke writes: "The most surprising anticipation of the present theory, discovered just as the paper was almost completed, is the algebraic analogue in Jónsson and Tarski ["Boolean algebras with operators"]. Independently and in ignorance of [Jónsson and Tarski's work] (though of course much later) the present writer derived its main theorem by an algebraic analogue of his semantical methods; the proof will appear elsewhere". Given how profusely Kanger gives credit to Jónsson and Tarski in his dissertation, this passage confirms what Kripke told me in 1960 about not having studied Kanger's dissertation thoroughly.

Logic, Methodology and Philosophy of Science IX
D. Prawitz, B. Skyrms and D. Westerståhl (Editors)

STIG KANGER'S THEORY OF RIGHTS *

LARS LINDAHL

I. Introduction

Stig Kanger regarded his theory of rights as one of his substantial contributions to philosophy; he worked on it, intermittently, for nearly thirty years. A starting-point was Kanger's interest in the classification of "fundamental jural relations" proposed by the American jurist W. N. Hohfeld, in the second decade of this century. Hohfeld's theory concerns an area which is mainly legal, and it belongs to the tradition of jurists such as Jeremy Bentham and John Austin. Hohfeld distinguished the relations *right, privilege, power, immunity*, and their "correlatives" *duty, no-right, liability, disability*; one of Hohfeld's tenets was that each of these relations is a relation between two parties with regard to an action by one of them.[1]

In his little book *New Foundations for Ethical Theory*, from 1957, Kanger presented his first explication of Hohfeld. He suggested that standard deontic logic, with only a deontic operator applied to sentences, is not adequate for expressing the Hohfeldian distinctions. The improvement he proposed was to combine a standard deontic operator with an action operator and to exploit the possibilities of external and internal negation of sentences where these operators are combined.

In Kanger's 1963 paper "The Concept of a Right", his explication of Hohfeld was restated as a system of so-called *simple* types of rights. In this paper, however, the simple types are the basis of a theory of *atomic* types of rights, which is more of a genuine typology. In the explication of atomic types, the combinatory method of "maxi-conjunctions" is used for providing an elegant logical typology of normative relations.

*The present essay is part of a project supported by *The Söderberg Foundations*.
[1]On the theories of Bentham, Austin and Hohfeld, see Lindahl (1977), Chapters 1 and 6, with further references.

During the last two decades of his life, Kanger was interested in the application of his theory of rights in connection with human rights and social justice; in particular, he turned to the problem of what, in the U. N. Declaration on Human Rights, is meant by having a right. In this connection, Kanger became aware of the distinction between a person's *having* a right and this right's being *realized* for the person. And so, in his last paper on rights, from 1985, Kanger dealt with the notion of realization of rights.

The first part of my paper contains a brief presentation of Kanger's typologies. After this, there follows a discussion of problematic points. The final part offers some suggestions for a positive solution to the most central problems.

Kanger's ideas about realization make use of a much enlarged logical framework, the treatment of which would lead too far in the present essay. His basic theory of rights, however, is independent of these ideas.[2]

II. Stig Kanger's theory of rights: a presentation

1. The language used by Kanger[3]

The sentences on rights that Kanger tries to explicate are taken from juristic usage or plain ordinary language. Moreover, Kanger's explications are not stated within a strictly formal language but only semi-formally. Only two kinds of entities are explicitly referred to, namely *agents*, on the one hand, and *states of affairs* or *conditions*, on the other. Agents are either persons, like Mr. Smith, or so-called collective agents, such as the Swedish Government or Smith & Co, Ltd. As illustrations of the second group of entities we have, for instance, the state of affairs (or condition) that Mr. Smith gets back the money lent by him to Mr. Black, or that Mr. Smith walks outside Mr. Black's shop. In Kanger's view, negation, conjunction, disjunction etc. can be applied to states of affairs (or conditions) in the same way as they are applied to sentences, and the notion of logical consequence is applicable to them by analogy as well.[4]

In order to state his explications in a general way, Kanger introduces letters for referring to agents or states of affairs that are chosen arbitrarily. He assumes that $x, y, z, \ldots$ are agents and that $F, G, H, \ldots$ are states of affairs. Moreover, $F(x, y), G(x, y), \ldots$ are assumed to be states of affairs

[2]See Kanger (1985). The enlarged logical apparatus is developed in Kanger (1977) and (1986).

[3]Kanger (1963), Kanger & Kanger (1966), and Kanger (1972).

[4]During the last years of his life, Kanger planned to develop a general theory of conditions based on cylindric algebra; unfortunately, however, the plan was never realized.

"involving" (as Kanger says) agents x and y.

To the Boolean connectives of negation, conjunction etc., Kanger adds the modal expressions "Shall" and "Do". Shall F is to be read "It shall be that F" and $\mathrm{Do}(x, F)$ should be read "x sees to it that F".

In his explication of rights, Kanger exploits the possibilities of combining the deontic operator Shall with the action operator Do. One example is Shall $\mathrm{Do}(x, F)$ which means that it shall be that x sees to it that F; another is $\neg$Shall $\mathrm{Do}(y, \neg F)$ which means that it is not the case that it shall be that y sees to it that not F.[5]

The logical postulates for Shall and Do assumed by Kanger are as follows (where $\longrightarrow$ is a relation of logical consequence, satisfying some resonable postulates[6]):

1. If $F \longrightarrow G$, then Shall $F \longrightarrow$ Shall G.
2. (Shall F & Shall G) $\longrightarrow$ Shall$(F\&G)$.
3. Shall $F \longrightarrow \neg$ Shall $\neg F$.
4. If $F \longrightarrow G$ and $G \longrightarrow F$, then $\mathrm{Do}(x, F) \longrightarrow \mathrm{Do}(x, G)$.
5. $\mathrm{Do}(x, F) \longrightarrow F$.

2. The simple types of rights[7]

In Kanger's theory, there are several *types* of rights. A type of rights is always a relation between two agents with respect to a state of affairs or a condition. For instance, if Mr. Smith has lent 100 dollars to Mr. Black, and, therefore, has a right to get back the money lent, then, according to Kanger, Smith has a right of the simple type Claim against Black with regard to the state of affairs (or condition) that Smith gets back the money he has lent to Black. In this example, Claim is the *type*, Smith is the *bearer*, Black is the *counter-party*, and the state of affairs (or condition) that Smith gets back the money lent is (what may be called) the "*object-matter*".

---

[5]The systematical use of "sees to it that" in combination with other operators is a characteristic feature in the work of Kanger's pupils within the Fenno-Scandian school of legal theory and social science. It is used in Pörn (1970), (1971), (1974), (1977), in Lindahl (1977), in H. Kanger (1984), in S. O. Hansson (1986), (1990-91), and in Holmström-Hintikka (1991). For some early suggestions, resembling Kanger's idea of combining Shall and Do, see Anderson (1962) and Fitch (1967).

[6]The principles assumed by Kanger for the relation of logical consequence are as follows:

(i) If F and $F \longrightarrow G$, then G;
(ii) If $F \longrightarrow G$, then $\neg G \longrightarrow \neg F$;
(iii) If $F \longrightarrow G$ and $G \longrightarrow H$, then $F \longrightarrow H$.

See, Kanger & Kanger (1966), at p. 88, note 3.

[7]See, concerning Kanger's typologies, Kanger (1963), Kanger & Kanger (1966), Lindahl (1977), and Makinson (1986).

In the history of the analysis of rights, there is a traditional distinction between, on the one hand, "passive rights", or rights to have something done, and on the other hand, "active rights" or rights to do something. In Kanger's theory of simple types of rights, the first group, henceforth called *O-rights*, is explicated by "counter-party obligatives", while the second, called *P-rights*, are explicated by "bearer permissives". In the first group, we have four simple types, explicated as follows:

Explicandum: *O-right*	*Explicans:* *Counter-party obligative*
Claim$(x, y, F(x, y))$	Shall Do$(y, F(x, y))$;
Counter-claim$(x, y, F(x, y))$	Shall Do$(y, \neg F(x, y))$;
Immunity$(x, y, F(x, y))$	Shall $\neg$ Do$(y, \neg F(x, y))$;
Counter-immunity$(x, y, F(x, y))$	Shall $\neg$ Do$(y, F(x, y))$.

For example, if Mr. Smith has an immunity against Mr. Black with regard to the condition that Mr. Smith walks outside Mr. Black's shop, this is explicated by: It shall be that Mr. Black does not see to it that Mr. Smith does not walk outside Mr. Black's shop. Each explicans satisfies the scheme,

$$\text{Shall } \pm \text{ Do}(y, \pm F(x, y)),$$

where $\pm$ stands for the two alternatives of affirmation or negation.

The four bearer permissive types are explicated in this way:

Explicandum: *P-right*	*Explicans:* *Bearer permissive:*
Power$(x, y, F(x, y))$	$\neg$ Shall $\neg$ Do$(x, F(x, y))$;
Counter-power$(x, y, F(x, y))$	$\neg$ Shall $\neg$ Do$(x, \neg F(x, y))$;
Freedom$(x, y, F(x, y))$	$\neg$ Shall Do$(x, \neg F(x, y))$;
Counter-fredom$(x, y, F(x, y))$	$\neg$ Shall Do$(x, F(x, y))$.

Here, each explicans satisfies the scheme,

$$\neg \text{ Shall } \pm \text{ Do}(x, \pm F(x, y)).$$

As an example, consider Mr. Smith's counter-freedom versus the police with regard to the condition that the police are informed about Mr. Smith's private life. In Kanger's explication, this would amount to: It is not the case that it shall be that Mr. Smith sees to it that the police are informed about Mr. Smith's private life.

Between the types of O-rights and the types of P-rights there exists a correspondence f such that if T is a type of O-right and T' is a type of P-right, then $T' = f(T)$ in case for any x, y, F it holds that x has a right of type T versus y with regard to $F(x, y)$ if and only if y has *not* a right of type T' versus x with regard to $F(x, y)$. For example, Claim is the counter-part of counter-freedom, in the sense, that x has a claim versus y with regard to $F(x, y)$ if and only if y has not a counter-freedom versus x with regard to $F(x, y)$.

According to the logical postulates, for some types it holds that membership of one type implies membership of another. For example, since $\mathrm{Do}(y, F(x, y))$, $\mathrm{Do}(x, \neg F(x, y))$, are inconsistent, $\mathrm{Shall\ Do}(y, F(x, y))$, $\mathrm{Shall\ Do}(x, \neg F(x, y))$, are inconsistent as well; therefore, according to Kanger's explication, $\mathrm{Claim}(x, y, F(x, y))$, not $\mathrm{Freedom}(x, y, F(x, y))$, are inconsistent.

2. The atomic types of rights

The construction of atomic types is as follows. We begin with the list,

$\mathrm{Claim}(x, y, F(x, y))$,
$\mathrm{Counter\text{-}claim}(x, y, F(x, y))$,
$\mathrm{Immunity}(x, y, F(x, y))$,
$\mathrm{Counter\text{-}immunity}(x, y, F(x, y))$,
$\mathrm{Power}(x, y, F(x, y))$,
$\mathrm{Counter\text{-}power}(x, y, F(x, y))$,
$\mathrm{Freedom}(x, y, F(x, y))$,
$\mathrm{Counter\text{-}freedom}(x, y, F(x, y))$.

Starting from this list, we form every new list that can be obtained by negating either $0, 1, 2, \ldots$, up to all 8 members of the list, while keeping the other members unnegated. Obviously, the number of all such lists will be 2^8, i.e., 256. (Each choice of negated members of the list corresponds to the choice of a subset of the original list; since the list has eight members, the number of its subsets is 2^8.) Of the 256 lists, however, all but 26 are inconsistent according to the logic of Shall and Do. Each of the remaining 26 lists, when regarded as a conjunction of its members, specifies an atomic type of rights.

As an example, we consider atomic type No. 5.

Name
"Power, immunity, counter-power, counter-immunity".

Explicans
$\{\langle x, y, F\rangle \mid \neg\mathrm{Shall}\neg\mathrm{Do}(x, F(x, y))\ \&\ \mathrm{Shall}\neg\mathrm{Do}(y, \neg F(x, y))\ \&$
$\neg\mathrm{Shall}\neg\mathrm{Do}(x, \neg F(x, y))\ \&\ \mathrm{Shall}\neg\mathrm{Do}(y, F(x, y))\}$.

We see that each conjunct in the explicans satisfies the scheme,

$$(*) \qquad \pm \text{ Shall } \pm \text{ Do } \left(\tfrac{x}{y}, \pm F(x,y)\right),$$

where $\pm$ and $\frac{x}{y}$ represent choices, as before. As suggested by David Makinson,[8] we can say that each atomic type is explicated by a "maxi-conjunction", i.e., a maximal and consistent conjunction such that each conjunct satisfies scheme $(*)$. Maximality means that if we add any further conjunct, satisfying $(*)$, then this new conjunct either is inconsistent with the original conjunction or redundant.

Given the underlying logic, the atomic types are mutually disjoint and their union is exhaustive.

Not all of Kanger's types of atomic rights are types of rights in any reasonable sense. Consider Kanger's atomic type No. 23. According to Kanger, x has a right of atomic type No. 23 versus y with regard to $F(x,y)$ if the following is true:

Not freedom$(x, y, F(x,y))$,
Not immunity$(x, y, F(x,y))$,
Not counter-claim$(x, y, F(x,y))$.

(Type 23 is specified by the list we obtain if all the lines of the original list of bearer permissives and counter-party obligatives are negated, and redundant members of the list have been dropped.) Since all members of the list are negated, x's relationship versus y with regard to $F(x,y)$ is one of *not* having a right on any kind, rather than one of *having* a right of a certain type. To say, in this case, that x has a right of a particular kind is like saying that poverty is a particular kind of opulence. Kanger's atomic typology, therefore, is a typology of normative relations from the "rights perspective" rather than a typology of rights.

III. Some aspects of Kanger's theory

In this section I will argue that Kanger's typology represents an improvement in the theory of *duties*; as a theory of rights, it suffers from a number of difficulties.

1. Kanger's theory as a theory of duties

Kanger's typologies are primarily typologies of *duties* and *non-duties*; x's O-rights versus y are explicated in terms of y's duties, i.e., in terms of counter-party obligatives; correspondingly, x's P-rights versus y are explicated in terms of x's non-duties, i.e., in terms of bearer permissives.

[8] See Makinson (1986), at pp. 405 f.

Thus, the counter-party obligative Shall $\mathrm{Do}(y, F(x, y))$ is an explication of "y has a duty to the effect that $\mathrm{Do}(y, F(x, y))$". Correspondingly, the bearer permissive $\neg$Shall $\mathrm{Do}(x, F(x, y))$ is an explication of "x has no duty to the effect that $\mathrm{Do}(x, F(x, y))$".

Other types of duty/non-duty are explicated if a negation sign is inserted before Do, before $F(x, y)$ or before both.

It follows that the atomic types are intersections of different types of duty/non-duty for two agents with regard to one and the same state of affairs.

If conceived of as typologies of duties/non-duties, Kanger's typologies represent a considerable improvement on earlier representations. In deontic logic, statements of duties are sometimes reproduced with the help of deontic operators carrying an index, like $\mathrm{O}_i, \mathrm{O}_j, \ldots$ where i, j are parameters or variables for agents; an expression of the form $\mathrm{O}_i F$ is read "F is obligatory for i".[9] Compared with this construction, Kanger's combinations of Shall and Do have greater expressive power; for example, instead of staying with "not $-$ F is obligatory for x", as expressed by $\mathrm{O}_x \neg F$, a distinction can be made between the cases Shall$\neg\mathrm{Do}(x, F)$, Shall $\mathrm{Do}(x, \neg F)$.

The idea of combining a non-relativized deontic operator with an agent-relative action operator has another advantage as well (though this was not exploited by Kanger himself). This advantage consists in the possibility of *iterating* operators in a meaningful way. It is controversial whether iterations of the kind $\mathrm{OO}F, \mathrm{O}\neg\mathrm{O}F$ etc., make sense; in any case it is not clear what is meant by statements of this form.[10] If we combine Shall and Do, however, new possibilities of iterations are opened. For example, in an organization, the boss is the superior of the clerk who is the superior of the errand-boy; it may well be the case that the boss is permitted to impose a duty on the errand-boy to work over-time, while the clerk is not permitted to impose such a duty on him. This distinction can be expressed by the two sentences

$$\neg\text{Shall}\neg\mathrm{Do}(x, \text{Shall } \mathrm{Do}(z, F));$$
$$\text{Shall}\neg\mathrm{Do}(y, \text{Shall } \mathrm{Do}(z, F));$$

where x is the boss, y is the clerk and z is the errand-boy.[11] It appears that in Kanger's language we can avoid iterations of the problematical kind; in the sentences just illustrated there is an instance of the Do operator

[9] See, for example, B. Hansson (1970).

[10] For a discussion of this problem, see Barcan Marcus (1966), v. Wright (1968), Szewak (1974) and Opfermann (1977).

[11] For a theory exploiting these possibilities, see Lindahl (1977), Part II (the theory of "ranges of legal action" or *Spielraum*). For a comment, see Talja (1980), where the tools of lattice theory are used.

between the two instances of the deontic operator.

2. Problems for Kanger's theory of rights

There are well-known problems connected with Kanger's theory conceived of as a theory of rights.

(*i*) Identification of bearer and counter-party. As remarked by J. S. Mill, the notion of a claim-right is connected with the idea that particular actions or omissions constitute cases of injustice committed against an assignable person (the bearer of the right); the injustice may consist in "depriving a person of a possession, or in breaking faith with him, or treating him worse than he deserves, or worse than other people who have no greater claims". The assumption that an injustice is committed, in turn, implies that the bearer of the right is *wronged*: "in each case the supposition implies two things — a wrong done, and some assignable person who is wronged".[12] In accordance with this suggestion, a criterion of appropriateness for the explication of a claim-right is as follows:

(1) x has a claim-right versus y to the effect that $F(x, y)$

only if it is true that,

(2) if $F(x, y)$ is not the case, then x is wronged,

(or x has a legitimate complaint). There are many interpretations of x, y, F such that Kanger's explicans for (1), i.e.,

(3) Shall $\mathrm{Do}(y, F(x, y))$,

holds, while (2) is false. The policeman has a duty to seize the murderer, who tries to get away. If we set x = the murderer, y = the policeman, and $F(x, y)$ for "x is seized by y", (3) is true. On the other hand, (2) is false in this case; the murderer is not wronged, and has no legitimate complaint, if the policeman does not succeed to seize him. The murderer has no right to the effect that he be seized.

Assume, on the other hand, that Creditor has lent 100 dollars to Debtor, and that, as a consequence, Debtor has a duty to pay this amount back. If we set x = Creditor, y = Debtor, and $F(x, y)$ for "x receives 100 dollars from y", the same counter-obligative formula (3) is true for this interpretation of the variables as well. In this case, however, (2) is true, and Creditor has a right to get his money back. Kanger's explicative formula (3) does not suffice to distinguish the two cases.[13]

[12] Mill (1910), p. 46.

[13] Cf. Lindahl (1977), pp. 45 f., and Makinson (1986).

One might try to defend Kanger's theory by going to the theory of *atomic* types of rights. But this does not help much since the same atomic type, viz. No. 6 (claim, power, counter-freedom) seems to be appropriate in both of the two examples illustrated. As applied to x versus y with regard to $F(x, y)$, type No. 6 is explicated as follows:

$$\text{Shall Do}(y, F(x, y)),$$
$$\neg\text{Shall } \neg\text{Do}(x, F(x, y)),$$
$$\neg\text{Shall Do}(x, F(x, y)).$$

The three sentences are true in the murderer case as well as in the Creditor case. (Observe that the third formula is true for the murderer, since he has no duty to see to it that he is seized by the particular policeman in view.)

The problem just illustrated for Claim-rights is that the explicandum is not entailed by the explicans. This problem can be shown to exist as well for the other types of O-right, i.e., counter-claim, immunity, counter-immunity.

If this objection is correct for O-rights, there will be a problem for P-rights as well. This time, however, the problem is that the explicans is not entailed by the explicandum. Let us remember that, in Kanger's construction, if T is a type of *P-right*, there is a type T^* of *O-right* such that $T(x, y, F(x, y))$ if and only if *not* $T^*(y, x, F(x, y))$. Furthermore, the types are constructed in such a way that φ is the explicans of $T(x, y, F(x, y))$ if and only if $\neg\varphi$ is the explicans of $T^*(y, x, F(x, y))$. By contraposition, therefore, if φ does not entail $T^*(y, x, F(x, y))$, then $T(x, y, F(x, y))$ does not entail $\neg\varphi$.

Let us illustrate the technical argument with an example. Suppose that y has a house in a suburban area. We may plausibly assume: y has no right that x does not walk around in the garden of y's neighbor (x's walking in that garden is no concern of y's). In Kanger's language, this means that

$$\text{(1)} \qquad \text{not Counter-immunity}(y, x, F(x, y))$$

where $F(x, y)$ expresses that x walks in the garden of y's neighbor. (1) is equivalent to

$$\text{(2)} \qquad \text{Power}(x, y, F(x, y)).$$

However, from (1) and (2) it ought not to follow, as in Kanger's theory,

$$\text{(3)} \qquad \neg\text{Shall}\neg\text{Do}(x, F(x, y)),$$

i.e., it should not follow that it is permitted that x walks in the garden of y's neighbor. For example, we may well suppose that x is a mortal enemy of y's neighbor, and that this neighbor has expressly forbidden x to walk in his garden; if so, the negation of (3) is true.

(*ii*) RIGHTS OF RECIPIENCE WITHOUT A COUNTER-PARTY. There are statements about "rights to receive", which do not imply statements about duties and which are not tractable in terms of Kanger's typologies. An example is as follows:

(1) Children have a right to be nurtured.

If x is a child, nothing follows from (1) about *who* has a duty to nurture it. Rather, it has been suggested, the acceptance of (1) is a first and basic point of departure from which further considerations can be made concerning duties for others (parents, guardians, authorities and so on).[14] Indeed, from (1) it does not even follow that for each child there *is* some y such that y has a duty to nurture it; i.e., if x is a child it does not follow that

(2) $$(\exists y)(\text{Shall Do}(y, F(x, y)))$$

where $F(x, y)$ means that x is nurtured by y. It may be suggested that (1) entails that if x is a child, then,

(3) $$\text{Shall } (\exists y)[\text{Do}(y, F(x, y))].$$

(2), however, does not satisfy the Kanger scheme for counter-party obligatives since a quantifier is embedded between Shall and Do. Since the quantifier is located after Shall, not before it, (2) does not say that anyone has a duty; rather (2) prescribes that there be someone who nurtures x.

(*iii*) LEGAL POWER. It is often maintained that so-called legal power is a type of right not tractable in terms of duties or non-duties. Suppose that F is a *legal* predicate; $F(x, y)$ signifies, for example, that the ownership of the Glenroy estate is transferred from x to y. Then (it is argued), the statement

(4) x has the legal power to see to it that $F(x, y)$,

[14] See N. MacCormick's essay "Children's rights: A Test-case for the Theories of Right", in MacCormick (1982).

cannot be analyzed as

(5) $$\neg\text{Shall}\neg\text{Do}(x, F(x, y)),$$

which is Kanger's general explication scheme for the simple type of right called "power": (5) expresses permission, while (4), it is usually held, expresses a capacitive dimension.[15]

On this point, I think that Kanger's analysis can be defended. It is true that, as "legal power" is usually understood, (4) and (5) are not synonymous, and Kanger's use of the term "power" is misleading. What Kanger wants to assert, however, is rather that (5) is an explication of a general notion of a right-to-do (what in German would be called *Befugnis*), i.e., of

(6) x has a right to see to it that $F(x, y)$.

(Apparently, Kanger did not find a suitable word in English corresponding to *Befugnis*.) Admittedly, in some circumstances, a thief is able to transfer the ownership of stolen goods to a purchaser who is in good faith (the sale will be legally valid). But, obviously, the thief has no *right* to do this. Perhaps (4) is true for this interpretation of F, x, y, but since (5) is false, (6) is false as well. In one sense of "legal power", the thief has the legal power to sell the stolen goods. But if so, "legal power" is not a type of *right*.

(*iv*) RELEVANCE OF CLAIM-HOLDER'S WILL. Suppose that Mr. Smith has a claim versus the community to receive medical care. If x = Mr. Smith, y = the community, and $F(x, y)$ is the condition that x receives medical care from y, then

(1) $$\text{Claim}(x, y, F(x, y))$$

is explicated by

(2) $$\text{Shall Do}(y, F(x, y)).$$

According to (2), the laws are disobeyed if y does not see to it that x receives medical care, even if this is due to x's refusing to receive it. However, all duties can be fulfilled even if x does not receive medical care, namely, in the case that he does not want to have it.

[15]See Lindahl (1977), p. 51 and pp. 194–211, with further references.

However, we might say that the "object-matter" of Smith's claim, expressed by $F(x, y)$, should appropriately be constructed in a different way, namely as the condition that medical care is made *available* to him by the community. The latter is another way of saying that Smith receives medical care, *if he wants* to have it. Of course, the expression $F(x, y)$ does not make it explicit that a conditional is involved, and it will be a problem how such a conditional should be expressed within the simple language presupposed by Kanger. However, this is a difficulty about expressing the "object-matter" of rights rather than an objection to the typology of rights itself.

A possible way out, in the specific example, is to replace $F(x, y)$ in (2) by the material equivalence $G(x, y) \leftrightarrow H(x, y)$, i.e., to substitute (2) by

$$(2') \qquad \text{Shall Do}(y, G(x, y) \leftrightarrow H(x, y)),$$

where $G(x, y)$ expresses that x (informs y that he) wants medical care and $H(x, y)$ that x receives medical care from y. This would keep the analysis within Kanger's basic framework; however, it remains an open question whether the construction is a good one.

As regards bearer-permissive rights, the problem is somewhat different. Mr. Brown has a right to walk in the municipal park, if he wants to, but need not walk there if he does not want to. In Kanger's typology, the relevance of Mr. Brown's will in this case can be expressed by the conjunction

$$\neg\text{Shall}\neg\text{Do}(x, F(x, y)) \ \& \ \neg\text{Shall}\neg\text{Do}(x, \neg F(x, y)),$$

where $F(x, y)$ expresses that x walks in y's park; the sentence says that x has both power (= *Befugnis*) and counter-power, as regards his walking in the park. Since, in this case, the power is "two-sided" (power and counter-power), it is sometimes described as *bilateral*.

Among theories of rights the so-called *will theory*, making relevance of the right-holder's will a conceptual characteristic of rights, has a respectable ancestry. A modern version of this theory has been developed by the Oxford legal philosopher Sir Herbert Hart. In Swedish philosophy, views similar to Hart's have been proposed by Sven Danielsson.[16]

However, there are claim-rights where the claim-holder's will is irrelevant, and there can be powers (in Kanger's sense) which are not bilateral. The statement that all children have a right to be given elementary education is compatible with the proposition that such eduction is compulsory,

[16]See Hart (1972), and S. Danielsson's essay "Fri- och rättigheter" in Danielsson (1986).

i.e., that refusal to partake in the education is inoperative. This shows that the object-matter of a claim-right should not always be construed by a conditional of the kind illustrated, where the claim-holder's will is made relevant: relevance of the claim-holder's will is not a general characteristic of claim-rights. Similarly, the statement that the policeman has a right to try to seize the thief is compatible with the statement that trying to do so is compulsory. The policeman's power is not bilateral, and it is not relevant what the policeman wants to do.

As is well-known, the notion of a right plays, and has played, an important part in many moral and political theories. Various theories emphasize different features of the notion of a right, or even define the notion in different ways, using it as a tool for an ideological message. This fact can be described in various ways: we might say that the notion of a right is "theory-dependent", or, that it is a "contested concept", or with Charles Stevenson, that there exist various persuasive definitions of the notion.[17] Those modern theories emphasizing relevance of the right-holder's will can be called *liberal* theories, in a wide sense. Since liberal theory occupies an important place in political thought, it is only to be expected that we are apt to regard cases where the right-holder's will is relevant as the central cases of rights. On the other hand, a general philosophical analysis of rights ought to avoid incorporating as definitional characteristics such features that are asserted by a specific moral or political theory.

(*v*) THE HETEROGENEITY OR HOMOGENEITY OF RIGHTS. Kanger never addresses the question whether the various types have anything in common which justifies calling all of them types of *rights*. He seems to hold that this problem is not worth pursuing, since the term "a right" is ambiguous; in fact, in the opening of his 1963 paper on rights, he says: "It is almost a commonplace that the idea of a right is vague and ambiguous ...".

The problem is whether there is any predicate φ such that, by analytical necessity,

(1) x has a right to (the effect that) A if and only if $\varphi(x, A)$,

where A is any condition, and $\varphi(x, A)$ expresses, in a non-trivial way, the point made when we ascribe a right to A to the agent x.

In the theory of rights, there are two basic attitudes to this question. One is that the term "a right" is used in such different ways that it is no use to look for a predicate of the kind referred to. According to this view, there are different explications $\varphi_1(x, A)$, $\varphi_2(x, A), \ldots, \varphi_n(x, A)$,

[17] See Stevenson (1944).

appropriate for different sentences of the form "x has a right to A"; the only way of explicating this formula according to (1) is the trivial one of interpreting $\varphi(x, A)$ as the disjunction $\varphi_1(x, A) \vee \varphi_2(x, A) \vee \ldots \vee \varphi_n(x, A)$.

The second basic attitude is that there exists a predicate φ appropriate for the explication of all rights. There is no agreement, however, as to which one of several explications is the appropriate one. In fact, as shown by Richard Tuck, the issue has been a bone of contention from the Middle Ages and onwards; various proposals are closely tied to specific theological, moral and political theories (cf. above, about "theory-dependence").[18] In a recent work, Alan White maintains that "x has a right" expresses that x is entitled to, has a title to, something. Basically, White's observation seems correct. However, White has not developed his suggestion, and, as White admits, the idea of being entitled and having a title is not more helpful than the information we can get from an ordinary dictionary.[19] This result is not surprising: if a theory of a common feature of rights is not to be tied to a specific legal, moral, or political theory it has to be exceedingly minimalistic and expressed in terms (like "being entitled") which are highly unprecise.

IV. A positive proposal

1. "Being wronged" and a new start

After the foregoing survey of problems, some positive proposals will now be made. A central subject is the identification of bearers and counterparties of rights within a minimalistic theory. In the suggestions that follow, the notion of *being wronged*, introduced in the previous section, features prominently.

As is well-known, in 1956 A. R. Anderson suggested an interpretation of the deontic operator O (for "obligatory") in terms of alethic modal logic.[20] Applied to Kanger's expression Shall, this interpretation amounts to the following:

$$\text{Shall } F \leftrightarrow \text{N}(\neg F \rightarrow \text{S}).$$

In the expression to the right, N stands for "necessary", $\rightarrow$, as usual, is the symbol for material implication, and S is a propositional constant. S can be understood as "deontic", expressing that the Bad Thing occurs. N

[18] Tuck (1979). Cf. M. Golding (1990), at p. 55.

[19] White (1984), especially at p. 114. Of course, the idea of unambiguity is compatible with holding that there are, nevertheless, several types of rights. To make an analogy, the unambiguity of the term "bird" in zoology is perfectly compatible with assuming that there are various kinds of birds.

[20] See Anderson (1956), reprinted in Rescher (1967).

is supposed to satisfy what Anderson calls the minimal requirements of a normal alethic modal logic. For S, Anderson assumes the axiom $\neg N(S)$.

The so-called *system T* for alethic propositional modal logic has the following rule of inference and axioms:[21]

If A is a theorem, then $N(A)$ is a theorem.
$N(A \rightarrow B)\ \&\ N(A) \rightarrow N(B)$.
$N(A) \rightarrow A$.

With system T for N, and with $\neg N(S)$, the Anderson interpretation of Shall yields the theorems of standard deontic logic, as formulated by:

If A is a theorem, then $Shall(A)$ is a theorem.
$Shall(A \rightarrow B)\ \&\ Shall(A) \rightarrow Shall(B)$.
$Shall(A) \rightarrow \neg Shall(\neg A)$.

Therefore, this interpretation remains basically within Kanger's framework, which is also the framework accepted in this essay.[22]

A useful tool for the explication of rights is obtained if we substitute Anderson's propositional constant S by a two-place predicate constant W for "is wronged by". The notion of an agent's being wronged, introduced above with reference to J. S. Mill, is important in criminal, private, and procedural law; moreover, it plays a prominent part in moral theory. (See, for instance, G. E. M. Anscombe's essay "Who is wronged?"[23]) In what follows, $W(x, y)$ is to be read "x is wronged by y".

2. The logic of rights-proper

We start with system T for N, the constant S and Anderson's axiom

(I) $$\neg N(S).$$

We add the predicate constant W and the axiom

(II) $$W(x, y) \rightarrow S,$$

expressing that if x is wronged by y, then the Code is violated etc. In passing, we observe that a weaker logic is obtained if we drop S together

[21] See, for example, Hughes & Cresswell (1968).

[22] The Anderson construction is, of course, connected with the problem of how to express "If ..., then" in a satisfactory way within a logically well-written language. Our reason for not discussing this problem is that, even if $N(. \rightarrow .)$ is questionable in the context at hand, it will keep us close to the Kanger typologies and logical framework. A recent attept to solve the problem of conditionals in deontic logic is made in Alchourrón (1991), where the relation of strict implication plays an important part.

[23] Anscombe (1967).

with (I) and (II), and rather stay with the axiom,

$$(1) \qquad \neg \mathrm{N}[\mathrm{W}(x, y)],$$

expressing that it is not necessary that x is wronged by y. Thus, from (I) and (II) we can derive

$$(2) \qquad \neg \mathrm{N}[textW(x, y) \vee \mathrm{W}(z, w)]$$

but (2) cannot be derived from (1). As will appear, the stronger logic resulting from (I) and (II) will yield typologies closer to those proposed by Kanger.

Next, we introduce the notion of a *right-proper*. If A is a condition, $\mathrm{R}(x, y, A)$ is read "x has a right-proper versus y to the effect that A", and is interpreted as follows:

$$\mathrm{R}(x, y, A) \leftrightarrow \mathrm{N}(\neg A \rightarrow \mathrm{W}(x, y)).$$

If, for fixed x, y, $\mathrm{R}(x, y, .)$ is regarded as an operator with x, y as parameters, the logic of $\mathrm{R}(x, y, .)$ will be standard deontic logic:

If A is a theorem, then $\mathrm{R}(x, y, A)$ is a theorem;
$\mathrm{R}(x, y, A)\ \&\ \mathrm{R}(x, y, A \longrightarrow B) \rightarrow \mathrm{R}(x, y, B)$;
$\mathrm{R}(x, y, A) \rightarrow \neg \mathrm{R}(x, y, \neg A)$.

Since we have (I) and (II) among the axioms, we obtain, as well, further theorems for cases where x, y are not kept fixed; in particular, we have,

$$\mathrm{R}(x, y, A) \rightarrow \neg \mathrm{R}(z, w, \neg A).$$

3. Simple and atomic types of rights

As will be remembered from section II, Kanger's explicans-formulae for simple types of O-rights (x versus y with regard to F) all satisfy the scheme,

$$\mathrm{Shall} \pm \mathrm{Do}(y, \pm F).$$

In this scheme, let us substitute Shall(.) by $\mathrm{R}(x, y, .)$ and we obtain,

$$\mathrm{R}(x, y, \pm \mathrm{Do}(y, \pm F)).$$

This way we can reconstruct all of Kanger's simple types of O-rights: claim, counter-claim, immunity, and counter-immunity. Due to the introduction of the notion of being wronged, however, their explication will differ from Kanger's and the problem of identifying the bearer does not occur. For example, Claim(x, y, F), i.e., R$(x, y,$ Do$(y, F))$, is explicated by,

$$\mathrm{N}[\neg \mathrm{Do}(y, F) \rightarrow \mathrm{W}(x, y)].$$

We will no longer have to say, as in the example discussed in section III, that the murderer has a claim versus the policeman to the effect that he is arrested by the policeman.

In a similar way, all of Kanger's simple types of P-rights can be reconstructed within the new system. To simplify the exposition, let us introduce the notions May and R* by the following conventions:

$$\mathrm{May}\ A \leftrightarrow \neg \mathrm{Shall} \neg A;$$
$$\mathrm{R}^*(x, y, A) \leftrightarrow \neg \mathrm{R}(y, x, \neg A).$$

May expresses permission (in a weak sense), and R*$(x, y, .)$ expresses a weak permissive right, which, with Hohfeld, we might call *privilege*, x versus y. Kanger's explicans-sentences for P-rights (x versus y with regard to F) all satisfy the scheme

$$\mathrm{May} \pm \mathrm{Do}(x, \pm F).$$

If we substitute May(.) by R*$(x, y, .)$, we get the scheme,

$$\mathrm{R}^*(x, y, \pm \mathrm{Do}(x, \pm F)).$$

Using this scheme, all of Kanger's types of P-rights can be reconstructed: power (= *Befugnis*), counter-power, freedom, counter-freedom. For example, Power(x, y, F) becomes R*$(x, y,$ Do$(x, F))$, and is explicated by

$$\neg \mathrm{N}[\mathrm{Do}(x, F) \rightarrow \mathrm{W}(y, x)].$$

We avoid the problem about counter-parties that is connected with Kanger's explication. In the example from section III, of x's walking in the garden of y's neighbor z, we can make the two statements,

$$\mathrm{Power}(x, y, F),\ \text{not}\ \mathrm{Power}(x, z, F)$$

i.e., x has a power (= *Befugnis*) versus y with regard to walking in z's garden, but x does not have this power versus z himself. The distinction is accomplished, since we have the respective explications,

$$\neg \mathrm{N}[\mathrm{Do}(x, F) \rightarrow \mathrm{W}(y, x)],$$
$$\mathrm{N}[\mathrm{Do}(x, F) \rightarrow \mathrm{W}(z, x)].$$

Thus it appears that, using W, we can reconstruct the complete lists of four types of O-rights and four types of P-rights. Given the list of the eight simple types, we can, of course, reconstruct a theory of *atomic* types of rights by the method of "maxi-conjunctions". The number of atomic types, however, will be greater than Kanger admitted. This is due to the fact that while

(1) $\quad \text{Shall Do}(y, F) \rightarrow \text{ Shall}\neg \text{ Do}(x, \neg F)$

is a theorem in Kanger's theory (since $\text{Do}(y, F)$, $\text{Do}(x, \neg F)$ are inconsistent), the corresponding reconstructed formula

(2) $\quad \text{R}(x, y, \text{Do}(y, F)) \rightarrow \text{ R}(y, x, \neg \text{ Do}(x, \neg F))$

does *not* follow from the axioms hitherto assumed in the reconstructed theory. (If (2) were a theorem, we would get 26 atomic types, as does Kanger.) It would lead too far afield to discuss in any detail the merits of (2). If, however, we want to have (2) as a theorem, while keeping the former basis of the reconstructed theory untouched, the question arises which further axiom or axioms should be added. There may be various possibilities. Among these are the following additions:

III. $\quad \text{N}(F \longrightarrow G) \rightarrow \text{ N}[\text{Do}(x, F) \rightarrow \text{ Do}(x, G)]$;

IV. $\quad \text{N}[\text{Do}(x, \text{W}(x, y)) \rightarrow \text{ W}(y, x)]$

(If these are added, (2) can be derived.[24]) III is easily understood; but IV needs some comment. It says that, necessarily, if x himself sees to it that he is wronged by y, then it follows that y is wronged by x. (This seems, in fact, to be the rationale behind the Kanger theorem (1).) For example, suppose that a child, by escaping from school, sees itself to it that it is wronged insofar as it does not receive the education that is due to it. Then it follows that those who have the duty to give the child its education (teachers, schoolmasters etc.) are wronged by the child's escaping, which prevents them from fulfilling their duty.

The acceptability of III and IV as logically valid may well be questioned. But if so, the Kanger theorem (1) can be questioned with as much justification.

[24]The antecedent of (2) is equivalent to $\text{N}(\neg\text{Do}(y, F) \rightarrow \text{ W}(x, y))$ which implies $\text{N}(\neg F \rightarrow \text{ W}(x, y))$. From this formula and III we get $N[Do(x, \neg F) \rightarrow Do(x, W(x, y))]$; using IV we get $\text{N}(\text{Do}(x, \neg F) \rightarrow \text{ W}(y, x))$, which is equivalent to the consequent of (2).

4. RIGHTS WITHOUT A COUNTER-PARTY

We often use statements of the kind "x has a right to ..." without mentioning any counter-party. Is it possible to explicate such statements using our two-place predicate W? Three examples will be discussed. The first one concerns the colloquial use of "having a right", emphasized by Alan White. Suppose we say to x: "You have the right to feel proud." Such a statement is somewhat ambiguous. One plausible interpretation, however, might go as follows. If x does not see to it that he feels proud, then he is wronged by himself; furthermore, for any y other than x, if y sees to it that x does not feel proud, then x is wronged by y. This way, counter-parties are seen as implicitly referred to, and the statement can be explicated in terms of the reconstructed notion.

The next example is adapted after one proposed by Bengt Hansson.[25] Petaluma is an area of private property, where different parts are owned by different people; we assume that for each land-owner y, y is wronged if x walks on his land. If $F(x)$ expresses that x walks on Petaluma land, we have

$$\neg(\exists y)(\mathrm{N}[F(x) \rightarrow \mathrm{W}(y, x)]),$$

since no land-owner y is wronged if x walks on Petaluma land belonging to another land-owner z (cf. the example, above, concerning x walking in the garden of y's neighbor). On the other hand, in the example,

$$\mathrm{N}[F(x) \rightarrow (\exists y)(\mathrm{W}(y, x))].$$

This sentence expresses, simpliciter, that x has no right to walk on Petaluma land.

The third example is the one referred to in section III, that all children have the right to receive nutrition. We suppose that x is a child and that $F(x)$ expresses that x receives nutrition; we want to express that x has the right to receive nutrition. This sentence is compatible with

$$\neg(\exists y)\mathrm{N}[\neg\mathrm{Do}(y, F(x)) \rightarrow \mathrm{W}(x, y)],$$

i.e., there need not be any particular agent by whom the child is wronged if that agent does not see to it that the child receives nutrition. On the other hand, we might suggest the following as an improved interpretation:

$$\mathrm{N}[\neg(\exists y)(\mathrm{Do}(y, F(x))) \rightarrow (\exists y)(\mathrm{W}(x, y))].$$

That is: if no-one sees to it that x receives nutrition, then there is someone by whom x is wronged.

[25] B. Hansson (1970), at pp. 245 f.

The last two examples illustrate how predicate W can be used in a flexible way to explicate sentences that cannot be well interpreted even in terms of the reconstructed notions of rights against a counter-party. In the last of the three examples, however, the explication given may be questionable. Indeed, the example may suggest that, in addition to the two-place predicate W, we can be in need of a one-place predicate W for "is wronged", such that $\mathrm{W}(x)$ expresses that x is wronged, simpliciter. If we introduce such a notion, we should assume that $\mathrm{W}(x, y)$ implies $\mathrm{W}(x)$ but not that $\mathrm{W}(x)$ implies $(\exists y)(\mathrm{W}(x, y))$.

The purpose of introducing a one-place predicate W would be to use it for interpreting a notion $\mathrm{R}(x, .)$, i.e., a right proper where there is no counter-party, according to the formula:

$$\mathrm{R}(x, A) \leftrightarrow \mathrm{N}(\neg A \rightarrow \mathrm{W}(x)).$$

With axiom (I), as well as $\mathrm{W}(x, y) \rightarrow \mathrm{W}(x)$ and $\mathrm{W}(x) \rightarrow \mathrm{S}$, we would get standard deontic logic for $\mathrm{R}(x, .)$, as well as further theorems like

$$\mathrm{R}(x, A) \rightarrow \neg \mathrm{R}(y, \neg A);$$
$$\mathrm{R}(x, y, A) \rightarrow \mathrm{R}(x, A);$$

and so on. The question whether there is a need for introducing the one-place predicate W, however, is left open here.

5. The impersonal operator Shall and the reconstructed notion of a right

A typology of rights, based on the notion of "being wronged by", as developed in the foregoing, is more akin to Hohfeld's original idea of jural relations between parties than is the Kanger typology, based on the impersonal operator Shall.[26] By the axiom $\mathrm{W}(x, y) \rightarrow \mathrm{S}$, we established a connection since, from our assumptions, it follows that $\mathrm{R}(x, y, F) \rightarrow$ Shall F.

The suitability of establishing this connection may be questioned. In any case, however, we ought *not* to assume any of

$$\mathrm{S} \rightarrow (\exists x)(\exists y)(\mathrm{W}(x, y));$$
$$\text{Shall } F \rightarrow (\exists x)(\exists y)(\mathrm{R}(x, y, F)).$$

That is, we should not assume that if the Code is violated, then someone is wronged by someone, or that if something is prescribed, then someone

[26] For an approach closer to Hohfeld's than Kanger's, see, as well, B. Hansson (1970); cf. also Makinson (1986), at pp. 48 ff.

has a right versus someone as regards the fulfillment of what is prescribed. There are many prescriptions (administrative regulations, traffic prescriptions etc.) which do not imply rights for particular agents; the contrary assumption would lead to an inflation of rights where the group of right-holders is very diffuse. This shows that there is room for the reconstructed typologies of rights that are genuine relations of rights between parties, alongside with typologies of normative positions expressed in terms of the operator Shall. For the latter kind of typologies, Stig Kanger's idea of combining Shall and Do is very useful. As suggested in the foregoing, typologies satisfying the Kanger schemes can be seen as typologies of positions of duty or non-duty.

REFERENCES

ALCHOURRÓN, C. E., *Philosophical Foundations of Deontic Logic and Its Practical Application in Computational Contexts*, Paper presented at Deon'91, Vrije Universiteet Amsterdam, December 1991. (Mimeographed).

ANDERSON, A. R., *The Formal Analysis of Normative systems*, Technical report N:o 2, contract N:o SAR/nonr-609 (16), Office of Naval Research, Group Psychology Branch, New Haven, Conn. 1956. (Reprinted in Rescher (1967), at pp. 147–213).

———————, "Logic, Norms and Roles", 4 *Ratio* (1962), pp. 36–49.

ANSCOMBE, G. E. M., "Who is Wronged?", *The Oxford Review* (1967).

BARCAN MARCUS, R., "Iterated deontic modalities", 75 *Mind* (1966), pp. 580–582.

DANIELSSON, S., *Filosofiska invändningar*, Stockholm 1986.

FITCH, F. B., "A Revision of Hohfeld's Theory of Legal Concepts", 10 *Logique et Analyse* (1967), pp. 269–276.

GOLDING, M. P., "The Significance of Rights Language", 18 *Philosophical Topics* (1990), pp. 53–64.

HANSSON, B., "Deontic Logic and Different Levels of Generality", 36 *Theoria* (1970), pp. 241–248.

HANSSON, S. O., "Individuals and collective actions", 52 *Theoria* (1986), pp. 87–97.

———————, "A formal representation of declaration-related legal relations", 9 *Law and Philosophy* (1990–91), pp. 399–416.

HART, H., "Bentham on Legal Rights", in *Oxford Essays in Jurisprudence*, Second Series, ed. A. W. B. Simpson, Oxford 1973, pp. 171-201.

HILPINEN, R., ed., *Deontic Logic: Introductory and Systematic Readings*, Dordrecht 1971.

HOHFELD, W. N., *Fundamental Legal Conceptions As Applied in Judicial Reasoning and Other Legal Essays*, ed. W. W. Cook, New Haven 1923.

HOLMSTRÖM-HINTIKKA, G., *Action, Purpose, and Will: A Formal Theory*, Helsinki 1991.

HUGHES, G. E. and CRESSWELL, M. J., *An Introduction to Modal Logic*, London 1968.

KANGER, H., *Human Rights in the U. N. Declaration*, Acta Universitatis Upsaliensis, Uppsala 1984.

KANGER, S., *New Foundations for Ethical Theory*, Part I, Stockholm 1957. (Reprinted, with minor changes, in Hilpinen (1971)).

———————, "Rättighetsbegreppet" ("The Concept of a Right"), in *Sju Filosofiska Studier tillägnade Anders Wedberg den 30 mars 1963*, Philosophical Studies published by the Department of Philosophy, University of Stockholm, N:o 9, Stockholm 1963. (Reprinted in English translation, as the first part of Kanger & Kanger (1966)).

———————, "Law and Logic", 38 *Theoria* (1972) pp. 105–132.

———————, "Några synpunkter på begreppet inflytande" ("Some aspects of the concept of influence"), in *Filosofiska Smulor tillägnade Konrad Marc-Wogau*, Filosofiska Studier utgivna av Filosofiska Föreningen och Filosofiska Institutionen vid Uppsala Universitet, Uppsala 1977.

———, "On Realization of Human Rights", in *Action, Logic, and Social Theory*, ed. by G. Holmström and A. J. I. Jones, 38 *Acta Philosophica Fennica* (1985), pp. 71–78.

———, "Unavoidability", in *Logic and Abstraction. Essays Dedicated to Per Lindström on His Fiftieth Birthday*, ed. M. Furberg et al., Göteborg 1986.

KANGER, S. & KANGER, H. "Rights and Parliamentarism", 32 *Theoria* (1966), pp. 85–116.

LINDAHL, L., *Position and Change: A Study in Law and Logic*, Dordrecht 1977.

MACCORMICK, N., *Legal Right and Social Democracy*, Oxford 1982.

MAKINSON, D., "On the Formal Representation of Rights Relations. Remarks on the Work of Stig Kanger and Lars Lindahl", 15 *Journal of Philosophical Logic* (1986), pp. 403–425.

MILL, J. S., *Utilitarianism, Liberty, Representative Government*, Everyman's Library, London 1910 (repr. 1964).

OPFERMANN, W., "Zur Deutung normlogischer Metaoperatoren", in *Deontische Logik und Semantik*, ed. by A. G. Conte et al., Wiesbaden 1977.

PÖRN, I., *The Logic of Power*, Oxford 1970.

———, *Elements of Social Analysis*, Filosofiska Studier utgivna av Filosofiska Föreningen och Filosofiska Institutionen vid Uppsala Universitet, Uppsala 1971.

———, "Some basic concepts of action", in *Logical Theory and Semantic Analysis*, ed. by S. Stenlund, Dordrecht 1974, pp. 93–101.

———, *Action Theory and Social Science*, Dordrecht 1977.

RESCHER, N., ed., *The Logic of Decision and Action*, Pittsburgh 1967.

STEVENSON, Ch. L., *Ethics and Language*, New Haven 1944.

SZEWAK, E. J., "Iterated modalities and the parallel between deontic and modal logic", 67–68 *Logique et Analyse* (1974), pp. 323–333.

TALJA, J., "A technical note on Lars Lindahl's Position and Change", 9 *Journal of Philosophical Logic* (1980), pp. 167–183.

TUCK, R., *Natural Rights Theories: Their Origin and Development*, Cambridge 1979.

WHITE, A. R., *Rights*, Oxford 1984.

VON WRIGHT, G. H., *An Essay in Deontic Logic and the General Theory of Action*, Amsterdam 1968.

Logic, Methodology and Philosophy of Science IX
D. Prawitz, B. Skyrms and D. Westerståhl (Editors)

NON-BINARY CHOICE AND PREFERENCE: A TRIBUTE TO STIG KANGER

AMARTYA SEN*

Depts. of Economics and Philosophy, Harvard Univ., Cambridge, Ma. 02138, USA

1. Introduction

Stig Kanger was a philosopher of extraordinary power and creativity. In logic, in choice theory, in the therory of rights, and in many other fields, Kanger made far-reaching contributions which were profoundly important for the respective subjects. But he was not invariably a person of the greatest perseverance. He would often make an extremely innovative departure from the received tradition, but then move on to something else without staying on to finish the work he had started.

This is especially the case with his deep and penetrating contributions to choice thory. His slender paper "Choice Based on Preference"—a thoroughly original contribution—was written some time in the middle 1970s (it will be called here Kanger I). It was seriously incomplete when it was first presented (with two sections of the text and the entire reference list missing), and it remained incomplete even at the time of his death more than a decade later. A subsequent paper "Choice and Modality" (to be called Kanger II) seemed like an attempt at completing the exercise, and it did extend the analysis, but it too needed more work which never came.[1]

In this paper, I want to talk about some specific aspects of choice theory that emerge forcefully from Kanger's ingenious conributions in this field.

*For helpful discussions on this and related topics, I am most grateful to Nick Baigent, Ben Fine, Dagfinn Follesdal, Wlodzimierz Rabinowicz, Ryszard Sliwinski, and of course—over many years—to Stig Kanger himself

[1]Both the papers contained, in fact, a small error, which was detected and sorted out by Stig Kangers's associates, Wlodzimierz Rabinowicz and Ryszard Sliwinski, in a forthcoming volume of Scandinavian texts on decision theory and ethics, which will include Kanger's unpublished — and unfinished—paper "Choice Based on Preference"; Pörn *et al.* (1992). The "Introduction" also comments generally and illuminatingly on the nature of Kanger's contributions to decision theory.

But given the incompleteness of the papers, this exercise must involve some speculation on what Kanger was really after. I am helped in this exercise by the discussions I had with him, first, at the London School of Economics in the mid-seventies, and later on, during my two visits to Uppsala in 1978 and 1987 respectively.

In the next section, the standard models of binary and non-binary choice theory are briefly discussed, followed—in section 3—by some reformulations reflecting Stig Kanger's ideas and suggestions. In section 4, the motivation underlying the reformulations are examined, and the importance of these departures is illustrated with paricular substantive examples. The essay ends with a concluding remark on the over-all significance of Kanger's proposals.

2. Choice functions and binariness

At the risk of some over-simplification, the literature in choice theory can be divided into two categories in terms of what is taken to be "the primitive", viz, (1) some *binary relation* R (interpreted as "preference", or "value", or "objective", or "the utility relation" — something seen as *prior* to choice), *or* (2) the *choice function* $C(.)$ itself.[2] These two standard approaches can serve as the background against which we see Kanger's departures.

2.1. Binary relation as the primitive

Consider, first, the traditional view of "relational choice", basing choice on the primitive relation R in the standard way. A binary relation R ranks the set of available alternatives X from which a non-empty "menu" S is offered for choice, $S \subseteq X$, and from this S an "optimal set" $C(S, R)$ is chosen on the basis of the binary relation R. In fact, only one element of the optimal set must ultimately be picked, but the optimal set reflects the set of "chooseable" elements of S.

$$C(S, R) = \{x \mid x \in S \ \& \ \forall y \in S : xRy\} \tag{1}$$

$C(S, R)$ is sometimes called the "choice set" of S with respect to the binary relation R. The interpretation of $C(S, R)$ depends on the content of the binary relation R. If, for example, R stands for the relation "at least as good as", then $C(S, R)$ is the set of "best" elements in S.

Here we move from a binary relation, taken as the primitive, to the derived choices. Within this general structure, the approach can vary with the

[2]The distinction applies to choice under uncertainty as well as certainty. However, in this paper I shall not go into the former, since neither of Kanger's essays deals with uncertainty.

characteristics of R, which may or may not be complete, may or may not be transitive, and so forth.

The symmetric and asymmetric factors of R partition the different cases in which xRy holds into xPy and xIy.

$$xPy \iff [xRy \ \& \ not \ yRx] \tag{2}$$

$$xIy \iff [xRy \ \& \ yRx] \tag{3}$$

If R is interpreted as at least as good as, then P can be seen as the relation "better than" and I as the relation "indifferent to".

In another variant of this approach of relational choice, the elements to be chosen may be specified as the set of "maximal" elements, rather than as the "optimal elements".[3] In the case of choosing from the "maximal element" set, to qualify for choice, and element x has to be undominated by any other element (that is, for no y should it be true that yPx), even though xRy need not hold either.

$$M(S,P) = \{x \mid x \in S \ \& \ not \ \exists y \in S : yPx\} \tag{4}$$

The distinction between the maximal set $M(S,P)$ and the optimal set $C(S,R)$ is helpful for relational choice for various reasons, but perhaps most of all because the optimal set $C(S,R)$ might well be empty when R is incomplete. While reflexivity (requiring xRx for all x) may be trivial in the context of many cases in choice theory (it is, for example, hard to dispute that x is "at least as good as" itself), completeness certainly can be a really exacting demand. Even with incompleteness, the maximal set can sometimes exist even though the optimal set is empty. For example, if neither xRy, nor yRx, then $C(\{x,y\},R) = \emptyset$, whereas $M(\{x,y\},R) = \{x,y\}$.

One type of preference relation much studied in choice theory is a "quasi-ordering", in which R is transitive but not necessarily complete. Kanger too has tended to take that type of relation as a good starting point of his analysis of "choice based on preference". For a quasi-ordering, an "optimal set" may well be empty even when a "maximal set" is clearly non-empty. Indeed, over a finite set S, a maximal set $M(S,R)$ will always exist for a quasi-ordering R (Sen 1970, Lemma 1*b). However, the following theorem holds (for a proof see Sen 1970, Lemma 1*d, pp. 11-2).

(T. 1) *For quasi-ordering R, if $C(S,R)$ is non-empty, then $M(S,R) = C(S,R)$.*

[3] On the distinction between "optimal" and "maximal" see Debreu (1959), Chapter 1, and Sen (1970).

The interest in the maximal set—as opposed to the optimal set — particulary arises when the optimal set does not exist.

2.2. Choice function as the primitive

In the alternative traditional approach, the primitive is taken to be the choice function $C(.)$ itself, which is a functional relationship that specifies for any non-empty subset S of the universal set X, a "choice set" $C(S)$, a subset of S. It is possible to obtain binary relations of "revealed" or "underlying" preference, from such a choice function (by making some standard assumptions), and indeed there is a quite a literature on this. For example, x is weakly "revealed preferred" to y if and only if from some set of which y is a member, x is actually chosen (whether or not y is also chosen).[4] Further, x is weakly "base relation preferred" to y if and only if x is picked precisely from the pair $\{x, y\}$.[5]

Weak revealed preference:

$$xR_cy \iff [\exists S : x \in C(S) \;\&\; y \in S] \tag{5}$$

Weak base relation:

$$x\bar{R}_cy \iff [x \in C(\{x, y\})] \tag{6}$$

The asymmetric and symmetric factors of R_c (denoted, P_c and I_c respectively) can be obtained in the usual way, following (2) and (3) applied to R_c. Similarly, with $\bar{R}_c$.

It is, in fact, also possible to define a *strong* revealed preference relation P^c directly, in terms of x being chosen from a set that contains y but from which y is not chosen (that is, x is chosen and y rejected).[6]

Strong revealed preference:

$$xP^cy \iff [\exists S : x \in C(S) \;\&\; y \in (S - C(S))] \tag{7}$$

2.3. Binary choice

A choice function is binary if and only if the revealed preference relation R_c generated by that choice function would generate back the same choice function if R_c is used as the basis of relational choice. Invoking (1) and (5), binariness is defined thus.

[4]See Samuelson (1938), Arrow (1959), Hansson (1968), Herzberger (1973).
[5]See Uzawa (1956), Herzberger (1973), Suzumura (1983).
[6]See Arrow (1959), Suzumura (1983).

Binariness of a choice function: A choice function is binary if and only if, for all $S \subseteq X$:

$$C(S) = C(S, R_c) \tag{8}$$

Various consistency conditions have been proposed for choice functions, such as the weak axiom of revealed preference, path independence, and so on. The following two elementary conditions are central for the binariness of a choice function.

Property α (basic contraction consistency): For all x in X and all $S, T \subseteq X$,

$$[x \in C(X) \ \& \ x \in T \subseteq S] \Longrightarrow [x \in C(T)] \tag{9}$$

Property γ (basic expansion consistency): For all x in X and any class of sets $S_j \subseteq X$:

$$[x \in \bigcap_j C(S_j)] \Longrightarrow [x \in C(\bigcup_j S_j)] \tag{10}$$

Property α demands that if a chosen element x from a set S belongs to a subset T of S, then x would be chosen from T as well. *Property γ* requires that if some x is chosen from every set S_j in a class, then it would be chosen also from the union of all such S_j.

The following result is easily established linking *Properties α* and γ to binariness of choice for a complete choice function, that is, for choice functions such that $C(S)$ is non-empty for any non-empty S (see Sen 1971 and Herzberger 1973).

(T. 2) *A complete choice function is binary if and only if it satisfies Properties α and γ.*

Binariness can also be defined in terms of the base relation $\bar{R}_c$, rather that the revealed preference relation R_c, in exactly the same way, and it can be shown that "basic binariness" thus defined is equivalent to binariness with respect to the revealed preference relation and thus equivalent to the combination of *Properties α* and γ (on this and related matters, see Herzberger 1973). By varying the required properties, the choice function can be made less or more demanding than binariness.[7]

3. Kanger's departures

The basic variation that Kanger introduces in this standard structure is the possibility of choosing according to a binary relation of preference R^V that

[7]For the main results, see Arrow (1959), Hansson (1968), Sen (1971), Herzberger (1973), Suzumura (1983).

depends on the "background" set V rather than being independent of the set of alternatives (as assumed in the case of R considered in the last section). While the choices are seen as being based firmly on binary relations, the particular binary relation to be used in the Kanger system varies with the background set V. The far-reaching significance of this variation will be considered in the next section.

The present section is concerned mainly with sorting out the formalities in Kanger's formulation, which is rather complex and in some ways quite hard to follow.[8] I shall first present the logical sequence in Kanger's own presentation, but it will emerge that the main differences introduced by him can be stated in another—rather simpler—way in terms of the standard format of choice theory. So if the reader is disinclined to go through a lot of formalities, he or she could move straight on to equations (15) and (16) below.

Kanger proceeds from a "primitive" notion of a decision function D, from which a choice function C is obtained. We shall call them D^K and C^K respectively, in honour of Kanger. The different concepts can be perhaps more easily understood by invoking a diagram of intersecting sets V and X (at the cost of some loss of generality, which will not however affect the formal definitions presented here). We take $S = V \cap X$.

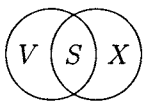


Figure 1:

$D^K(V, X)$ are the elements of V that are no worse than any element of $V - X$ (equivalently, $V - S$) according to the strict binary relation P^V with respect to the background set V.

$$D^K(V, X) = \{x \mid x \in V \ \& \ not \ \exists y \in V - X : yP^V x\} \tag{11}$$

It is easily checked that the following relations hold:

$$D^K(V, X) \ = \ D^K(V, S) \tag{12}$$

[8]Rabinowicz and Sliwinski point out in their introduction in Pörn *et al.* (1992) that Kanger's "reason for choosing such an artificial concept as D as his primitive" relates to "the close formal connection between D and modal operators studied in modal logic". Rabinowicz and Sliwinski discuss these connections, and they are indeed important for the formal side of Kanger's reformulation of the choice problem (see Kanger I and Kanger II). In this paper, however, I am mainly concerned with the substantive differences pursued by Kanger. See also Danielsson (1974) on related issues.

$$D^K(V, V - X) \;=\; D^K(V, V - S) \tag{13}$$

The choice function C^K is defined in terms of D^K thus:

$$C^K(V, X) = D^K(V, V - X) \cap X \tag{14}$$

With the choice function C^K thus established, Kanger proceeds to introduce more structure into the background-dependent preference relation P^V: first the elementary need for this notationally "strict" P^V to be irreflexive; then the requirement that P^V be a strict partial ordering with no infinitely ascending chain; then it be also a semi-ordering; and finally that it be a strict weak ordering. He examines their various properties and relates them to the consistency conditions used in the standard literature (such as *Properties* α and γ).

The basic idea behind the choice function C^K can be understood in more direct terms in the following way. Consider the maximal set $M(S, P)$, defined earlier, in equation (4). The strict preference relation P invoked there did not depend on any background set V. Now make it dependent on a selected background V, and call it P^V. Define $C^*(S, V)$ simply as $M(S, P^V)$, exactly like a traditional maximal set, except for using P^V rather than P.

$$C^*(S, V) = M(S, P^V) = \{x \mid x \in S \;\&\; \textit{not}\; \exists y \in S : yP^Vx\} \tag{15}$$

Now bearing in mind that S is the intersection of V and X, it can be easily established that Kanger's Choice function C^K relates to C^* (and thus to the standard maximal function M) in the following way:

(T. 3)

$$C^K(V, X) = C^*(S, V) = M(S, P^V) \tag{16}$$

The result is easily checked by comparing (15) with the characterization of $C^K(V, X)$ in the Kanger system, given by (17), based on (14):

$$C^K(V, X) = \{x \mid x \in V \cap X \;\&\; \textit{not}\; \exists y \in V \cap X : yP^Vx\} \tag{17}$$

Thus, we are essentially in the same territory as the traditional maximal function $M(.)$, with the added proviso that the strict preference relation P is now a *background dependent* P^V. And bearing in mind the old result (T.1) that the traditional maximal set $M(S, P)$ is the same as the traditional choice set $C(S, R)$ whenever the latter is non-empty and R is a quasi-ordering (Sen 1971), we have a clear relationship between Kanger's choice system and the standard system of choice sets and maximal sets.

The Kanger system opts for the idea of maximality rather than that of optimality (underlying the traditional binary choice function), and furthermore makes the binary relation of preference P^V (on the basis of which maximality is defined) dependent on the specification of the background set V. The latter is a truly substantial departure, and in the next section, the motivation underlying this change and its extensive importance are discussed and exemplified. But as far as formalities are concerned, we lose nothing substantial by using the simpler notion of a background-dependent maximal functions $M(S, P^V)$, rather than $C^K(V, X)$, as in the Kanger system.

The discussion that follows will be conducted entirely in these less specialized terms, using the older notion of maximality coupled with Kanger's ideal of a background-dependent preference relation P^V.

4. Why background dependence?

At the substantive level, the idea behind a background-dependent maximal choice $M(S, P^V)$, equivalent to Kanger's differently formulated choice structure, can be seen in terms of two distinct departures from the standard maximal choice $M(S, P)$: (1) the preference relation P is taken to be dependent on a background set V in terms of which it is defined, and (2) the background set V need not be the set S (the menu) from which choice is being made. I shall briefly consider different types of motivations that can justify the broader conception of choice behaviour proposed by Kanger. Since Kanger himself has tended to shy away from motivational discussions in general, I cannot claim that these motivations explain *why* Kanger made his proposals. But nevertheless these motivational arguments help us understand some of the advantages of the Kanger formulation over more traditional models of choice behaviour.

Let us first consider the former departure without the second (i.e., background-dependence of preference when the background is required to be the menu itself). Take the preference relation P^S to be dependent on the set S from which choice is being made: $M(S, P^S)$. This is already a considerable departure from the standard model of choice, given by $C(S, R)$ or $M(S, P)$, in which the preference relations R and P are taken to be menu-independent (and of course, more generally, background-independent). This relaxed requirement can deal with cases in which the nature of the menu from which choice is being made can affect the ranking of alternative elements. The reasons for such menu-dependence of rankings can be diverse and they tend to be comprehensively ignored in the traditional models of binary choice.

I present here briefly three quite different—and essentially independent — reasons for menu-dependence of preference, which I have discussed more

extensively elsewhere (Sen 1992).[9]

Positional choice: The ranking of alternatives may depend on the position of the respective alternatives vis-a-vis the others in the menu. For example, when picking a slice of cake from a set of slices, a cake-loving person who nevertheless does not want to be taken to be greedy may decide not to pick the largest slice, but choose instead one that is as large as possible subject to its not being the largest, to wit, she may choose the *second* largest slice.[10] This type of choice would violate binariness and even the elementary condition of *Property* α (basic contraction consistency). If, for example, three slices of cakes are ranked in decreasing order of size as a over b and that over c, then from the menu (a, b, c), the person may pick b, and from (b, c) may choose c.

There is nothing particularly "irrational" in such behaviour, even though these choices violate *Property* α and binariness. Similarly, a person may decide not to pick the last apple from an after-dinner fruit basket, having one of the pears instead, even though she may pick an apple from a larger basket containing many apples and many pears.

Epistemic value of the menu: A person may accept the invitation to tea from an acquaintance she does not know well, but refuse that invitation to tea if the acquaintance were also to invite this person to have some cocaine with him. The addition of the latter invitation may give her some extra information about him which might make her more skeptical of the idea of having tea with him. The menu offered has informational value in ranking the individual courses of action. Again, we see here a violation of *Property* α and of binariness, but the reasoning is canny enough.

Valuation of freedom: The freedom a person enjoys depends on the nature of the menu open to her. The choice of courses of action may be influenced by the extent of freedom. For example, a person may choose to read a particular newspaper when she could read any one she chooses (or none), and yet decide to protest and read none if she is forced to read *that* particular newspaper and no others.

Contraction consistency and binariness are violated in all these cases, but there is no difficulty in explaining and rationalizing the choices in terms

[9]See also Sen (1982, 1992), Elster (1983), Levi (1986), Fine (1990), among others, for different types of reasons for menu-independence.

[10]Positional valuation has been extensively investigated in the context of social choice by Gärdenfors (1973) and Fine and Fine (1974).

of "choice based on preference" when the preference relation P^S depends on the menu from which choice is being made. These and other examples have been discussed and scrutinized elsewhere in terms of the particular properties of menu-dependent preference P^S, but they are covered *inter alia* by the more general case of background-dependent preference P^V proposed by Stig Kanger.

Now we can turn to the case in which the background set V need not coincide with the menu set S. This is a particularly Kanger territory. What can be the reason for choosing a background set that is different from the menu from which choice is being made? While Kanger himself has not discussed the motivational issues in his papers, possible reasons for the additional departure are not hard to seek. The menu tells us what we can choose from. The ranking of the alternatives may depend, however, on the role of the chosen alternatives *after* the choice has been made.

For example, consider the problem of selecting tennis players to represent a country in the Davis Cup—an international tournament. What the selectors have to seek are not the best players in the country in terms of playing against each other, but the best players in terms of playing against tennis players from other nations. Consider a case in which players A and B can defeat players C, D, E and F individually and in pairs. That is a good reason for declaring them to be champion players within the nation. But it is still possible—given differences in the style of playing—that players C and D can defeat the Davis Cup team from the United States while the others cannot do that, and players E and F can defeat the Davis Cup players from Sweden, while the others cannot perform that feat. In that case, in picking Davis Cup players, there would be a good argument for picking C and D if it looks that this country will have to play against the United States, and for picking E and F if it appears that the contest will be against Sweden. The ranking relation P^V must, thus, take note of the ranking of the domestic players not vis-a-vis each other, but of their abilities to play against the likely international competitors—the appropriate "background" in this case.

Similarly, in selecting a poet laureate, the selectors may be guided not just by the merits of the likely candidates seen in terms of internal comparisons, but by the respective standings and comparative standards of these candidates vis-a-vis other well-known poets—including dead poets and lyricists from other nations. To take another type of example, in making admissions decisions, a college may be guided not just by comparisons of the applicants against each other seen in purely internal terms, but also by comparing them to general categories of students whether or not applicants to this particular college. Many other types of examples can be easily presented.

The common factor in all this is the need for external reference—external

to the menu—in comparing the alternatives in the menu. It is that general possibility that the Kanger formulation of choice can capture in a neat and elegant way by explicitly bringing in the reference to a background set V that may or may not coincide with the menu S.

5. A final remark

In this essay I have briefly presented the special features of Stig Kanger's model of "choice based on preference". By presenting his formulation in a slightly different way, we can see it as an extension of the standard model of binary choice in terms of maximal sets with the binary relation of choice P^V made dependent on a background set V which may or may not coincide with the menu S. The departures, thus, involve three distinct elements: (1) use of maximality rather than optimality, (2) admitting menu dependence of preference, and (3) admitting dependence of preference on a set different from the menu itself. I have discussed the case for each of these departures, of which the last is most specific to Kanger's own work.

I end with a final remark that while Kanger's formulation takes choice theory well beyond the limited framework of binary choice as it is standardly defined, the primitive notion that Kanger invokes is still a binary relation P^V defined in terms of a specified background set. In this sense, Kanger's model can be seen as a generalized formulation of binary choice (as he calls it, "choice based on preference").

One of the implications of Kanger's analysis is the need to rethink on the requirements of maximization as the basis of decisions and choice. The Kanger framework violates the standard conditions of maximal choice quite robustly, but the differences arise not from rejecting any intrinsic feature of maximization as such, but from dropping the implicit presumption in the standard literature that the preference relation be background independent. In effect, Stig Kanger has shown that maximization is a much more general discipline than theorists of maximization have tended to assume. That is the key to a different world of choice through maximization.

References

ARROW K.J. (1959), *Rational Choice Functions and Orderings*, Economica 26.

DANIELSSON, S.(1974), *Two Papers on Rationality and Group Preference*, Uppsala: Philosophy Department, Uppsala University.

DEBREU, G. (1959), *Theory of Value*, New York: Wiley.

ELSTER, J. (1983), *Sour Grapes*, Cambridge: Cambridge University Press.

FINE, B. (1990), *On the Relationship between True Preference and Actual Choice*, mimeographed, Birkbeck College, London.

FINE, B. AND FINE, K. (1974), *Social Choice and Individual Ranking*, Review of Economic Studies, 41.

GÄRDENFORS, P. (1973), *Positional Voting Functions*, Theory and Decision 4.

HANSSON, B. (1968), *Choice Structures and Preference Relations*, Synthese 18.

HERZBERGER, H.G. (1973), *Ordinal Preference and Rational Choice*, Econometrica 41.

KANGER, STIG (1970S), *Choice Based on Preference*, mimeographed, University of Uppsala (cited here as Kanger I).

KANGER, STIG (1980S), *Choice and Modality*, mimeographed, University of Uppsala (cited here as Kanger II).

LEVI, I. (1986), *Hard Choices*, Cambridge: Cambridge University Press.

PÖRN, I. *et al.* (1992), *Choices, Actions and Norms. Conceptual Models in Practical Philosophy—Scandinavian Contributions*, to appear.

RABINOWICZ, W., AND SLIWINSKI, R. (1991), *Introduction*, Pörn et al. (1992).

SAMUELSON, P.A. (1938), *A Note on the Pure Theory of Consumers' Behaviour*, Economica 5.

SEN, A.K. (1970), *Collective Choice and Social Welfare*, San Francisco: Holden-Day; republished, Amsterdam: North-Holland, (1979).

SEN, A.K. (1971), *Choice Functions and Revealed Preference*, Review of Economic Studies 38; reprinted in Sen (1982).

SEN, A.K. (1982), *Choice, Welfare and Measurement*, Cambridge, MA: MIT Press, and Oxford: Blackwell.

SEN, A.K. (1992), *Internal Consistency of Choice*, 1984 Presidential Address to the Econometric Society, forthcoming in Econometrica 1993.

SUZUMURA, K. (1983), *Rational Choice, Collective Desicions, and Social Welfare*, Cambridge: Cambridge University Press.

UZAWA, H. (1956), *A Note on Preference and Axioms of Choice*, Annals of the Institute of Statistical Mathematics 8.

Logic, Methodology and Philosophy of Science IX
D. Prawitz, B. Skyrms and D. Westerståhl (Editors)

DeBAYESING GAME THEORY

KEN BINMORE

University College London, University of Michigan

> The *look before you leap* principle is preposterous if carried to extremes ...
> Leonard Savage, *Foundations of Statistics*

1. Bayesianism

Debasing the coinage is a serious offence. DeBayesing game theory would be even worse if it meant denying game theorists the use of Bayes' rule. How would we make a living if deprived of the most fundamental of the tools of our trade? It therefore needs to be explained that this lecture is not an attack on Bayesian decision theory as commonly used in analyzing particular games. I am a Bayesian myself in such a context. The paper is an attack on *Bayesianism*, which I take to be the philosophical principle that Bayesian methods are always appropriate in all decision problems. I want to argue in particular that Bayesianism is an inappropriate standpoint from which to view the foundations of game theory. My own hopes for progress on this front depend on importing evolutionary ideas into game theory. However, I shall have nothing to say about such alternative approaches.

The ugly word Bayesianismist will be used to describe an adherent of the creed of Bayesianism. I freely admit that few serious researchers would react with pride if such a label were pinned on them. But I do not think I am merely attacking a straw man. What matters for this purpose is not so much what people say about their philosophical attitudes, but what models they choose to construct. As Robert Aumann likes to say of game-theoretic concepts in general: By their fruits shall ye know them.

There is an exception to the rule that Bayesianism is an underground creed. This is provided by the economics profession. For many young economists just out of graduate school, it is almost a heresy to argue that alternatives to Bayesian decision theory might ever make any sense. The defence against charges of heresy is to refer to the scriptures. In the case of

Bayesianism, the appropriate text is Savage's. *Foundations of Statistics.* Savage is very clear that his is a *small world* theory.[1] Others speak of a *closed universe*, but for reasons that will emerge, I prefer to refer instead to a *completable universe.*

Savage makes the distinction between a small and a large world in a folksy way by quoting the proverbs "look before you leap" and "cross that bridge when you come to it". You are in a small world if it is feasible always to look before you leap. You are in a large world if there are some bridges that you cannot cross before you come to them. As Savage comments, when proverbs conflict, it is proverbially true that there is some truth in both. The words of the prophet therefore seem quite clear. Some decision situations are best modeled in terms of a completable universe; others are not. Savage rejects the idea that *all* universes are completable as both "ridiculous" and "preposterous".

My view is that the foundational problems of game theory are not completable universe problems, and hence are not amenable to a Bayesianismist methodology along the lines proposed by Robert Aumann [4,5] and others. On the contrary, I see one of the major purposes of studying foundational questions as being that of finding appropriate ways of closing the universe of discourse so as to *legitimize* the use of Bayesian methods in analyzing particular games.

I am well aware that a formal theory of rational decision-making in an incompletable universe seems likely to remain as elusive in the near future as it always has in the past. To maintain otherwise would be to maintain that the problem of scientific induction is on the point of being solved. However, I would prefer to work with a game theory that has no foundations at all, than to operate using foundational principles based on a flawed methodology.

2. Using Savage's theory

This section reiterates the reasons given in Binmore [10] for rejecting unqualified Bayesianism as naive. Savage's theory is entirely and exclusively a *consistency* theory. It says nothing about how decision-makers come to have the beliefs ascribed to them; it asserts only that, if the decisions taken are consistent (in a sense made precise by a list of axioms), then

[1] There is room for confusion here for those who are well-read in the scriptures. I do not intend when speaking of small worlds to refer to Savage's attempt to explain how a small world, which he calls a *microcosm* in this context, may be embedded in a grand world. This attempt does not seem to me to be very successful. I intend the concept of a small world to be interpreted in the wider, non-technical sense of the earlier portion of his book.

they act *as though* maximizing expected utility relative to a subjective probability distribution. Objections to the axiom system can be made, although it is no objection when discussing rational behavior to argue, along with Allais [1] and numerous others, that real people often contravene the axioms. People also often get their sums wrong, but this is no good reason for advocating a change in the axiomatic foundations of arithmetic! In any case, it is not Savage's consistency axioms that are to be attacked here.

What is to be denied is that Savage's passive *descriptive* theory can be reinterpreted as an active *prescriptive* theory at negligible cost. Obviously, a reasonable decision-maker will wish to avoid inconsistencies. A Bayesianismist therefore assumes that it is enough to assign prior beliefs to a decision-maker and then forget the problem of where beliefs come from. Consistency then forces any new data that may appear to be incorporated into the system via Bayesian updating. That is, a posterior distribution is obtained from the prior distribution using Bayes' rule. The naiveté of this approach does not consist in using Bayes' rule, whose validity as a piece of algebra is not in question. It lies in supposing that the problem of where the priors came from can be quietly shelved. Some authors even explicitly assert that rationality somehow *endows* decision-makers with priors, and hence that the problem does not exist at all.

Savage did argue that his descriptive theory of rational decision-making could be of practical assistance in helping decision-makers form their beliefs, but he did not argue that the decision-maker's problem was simply that selecting a prior from a limited stock of standard distributions with little or nothing in the way of soul-searching. His position was rather that one comes to a decision problem with a whole set of subjective beliefs derived from one's previous experience. This belief system may or may not be consistent. In a famous encounter with Allais, Savage himself was trapped into expressing inconsistent beliefs about a set of simple decision problems. The response he made is very instructive. He used his theory to adjust his beliefs until they became consistent. Luce and Raiffa [16, p 302] explain the process by means of which such a consistent set of final beliefs is obtained as follows:

> Once confronted with inconsistencies, one should, so the argument goes, modify one's initial decisions so as to be consistent. Let us assume that this jockeying — making snap judgments, checking on their consistency, modifying them, again checking on consistency, etc — leads ultimately to a bona fide, *a priori* distribution.

For Savage therefore, forming beliefs was more than a question of attending to gut-feelings. It was a matter for *calculation* — just as the question of whether you or I prefer $\$17 \times 29$ to $\$19 \times 23$ is a matter for calculation.

But why should we wish to adjust our gut-feelings using Savage's methodology? In particular, why should a rational decision-maker wish to be consistent? After all, scientists are not consistent, on the grounds that it is not clever to be consistently wrong. When surprised by data that shows current theories to be in error, they seek new theories that are inconsistent with the old theories. Consistency, from this point of view, is only a virtue if the possibility of being surprised can somehow be eliminated. This is the reason for distinguishing between incompletable and completable universes. Only in the latter is consistency an unqualified virtue.

One might envisage the process by means of which a decision-maker achieves a consistent set of subjective beliefs in a completable universe as follows. The decision-maker knows that subjective judgments need to be made, but prefers to make such judgments when his information is maximal rather than minimal. He therefore asks himself, for every conceivable possible course of future events: what would my beliefs be *after* experiencing these events? Such an approach automatically discounts the impact that new knowledge will have on the basic model that the decision-maker uses in determining his beliefs — that is, it eliminates the possibility that the decision-maker will feel the need to alter his basic model after being surprised by a chain of events whose implications he had not previously considered. Next comes the question: is this system of *contingent* beliefs consistent? If not, then the decision-maker may examine the relative *confidence* that he has in the "snap judgments" he has made, and then adjust the corresponding beliefs until they *are* consistent.[2] With Savage's definition of consistency, this is equivalent to asserting that the adjusted system of contingent beliefs can be deduced, using Bayes' rule, from a single prior.

At the end of the story, the situation is as envisaged by Bayesianismists: the final "massaged" posteriors can indeed be formally deduced from a final "massaged" prior using Bayes' rule. This conclusion is guaranteed by the use of a complex adjustment process that operates until consistency is achieved. As far as the massaged beliefs are concerned, Bayes' rule has the status of a *tautology* — like 2+2=4. Together with the massaged prior, it serves essentially as an indexing system that keeps track of the library of massaged posteriors. However, what is certainly false in this story, is the Bayesianismist view that one is *learning* when the massaged prior is updated to yield a massaged prior. On the contrary, Bayesian updating only takes place *after* all learning is over. The actual learning takes place while the decision-maker is discounting the effect that possible

[2]Gärdenfors' [13] *Knowledge in Flux* assesses the considerations that will control how such adjustments are made.

future surprises may have on the basic model that he uses to construct his beliefs, and continues as he refines his beliefs during the massaging process. Bayesianismists therefore have the cart before the horse. Insofar as learning consists of deducing one set of beliefs from another, it is the massaged *prior* that is deduced from the unmassaged *posteriors*.

A *caveat* is necessary before proceeding. When the word "learning" is used in the preceding paragraph, it is intended in the sense of "adding to one's understanding" rather than simply "observing what happens". Obviously, a person with perfect recall will have more facts at his disposal at later times than at earlier times, and it is certainly true that there is a colloquial sense in which he can be said to "learn" these facts as time goes by. However, it seems to me that this colloquial usage takes for granted that whoever is "learning" the facts is also sorting and classifying them into some sort of orderly system with a view to possibly making use of his knowledge in the future. Otherwise it would not seem absurd to say that a video camera is "learning" the images it records. In any case, it is not the simple recording of facts that is intended when "Bayesian learning" is discussed. Any proposal for a rational learning scheme will presumably include recording the facts (if the cost of so doing is negligible). What distinguishes "Bayesian learning" from its alternatives must therefore be something else.

In spite of this *caveat* about what I intend when speaking of learning, the suggestion that Bayesian updating in a completed universe involves no learning at all commonly provokes expressions of incredulity. Is it being said that we can only learn when deliberating about the future, and never directly from experience? The brief answer is *no*, but I have learned directly from experience that a longer answer is necessary.

In the first place, the manner in which you and I (and off-duty Bayesianismists) learn things about the real world is not necessarily relevant to the way a Bayesian learns. Still less is it relevant to the way in which a Bayesianismist learns when on duty. Experimental evidence offers very little evidence in favor of the proposition that we are natural Bayesians of any kind. Indeed, what evidence there is seems to suggest that, without training, even clever people are quite remarkably inept in dealing with simple statistical problems. In my own game theory experiments, no subject has ever given a Bayesian answer to the question "Why did you do what you did?" when surveyed after the experiment — even though, in most cases, the populations from which the subjects were drawn consisted entirely of students who had received training in Bayesian statistics. I therefore think introspection is unlikely to be a reliable guide when considering what learning for a Bayesian may or may not be.

The fact that real people actually learn from experience is therefore not relevant to whether Bayesian updating in a completed universe should count as genuine learning. The universes about which real people learn are almost always incomplete and, even when they are confronted with a completable universe, they almost never use Bayesian updating. Of course, Bayesian statisticians are an exception to this generalization. They use Bayesian updating all the time, but, just like real people, they are almost never working in a completed universe. That is to say, they have not asked themselves why a knee-jerk adherence to consistency requirements is appropriate, but simply update from a prior distribution chosen on *a priori* grounds. I do not argue that such a procedure is necessarily nonsensical. On the contrary, it often leads to descriptions of the data that provide much insight. Nor do I argue that a Bayesian statistician who updates from a prior distribution chosen on *a priori* grounds is not learning. All I have to say to such a Bayesian statistician is that I see no grounds for him to claim that he is learning *optimally*, or that his methodology is *necessarily* superior to those of classical statistics.[3] The problem of how "best" to learn in an incompletable universe is unsolved. Probably it is one of those problems that has no definitive solution. But, until the problem of scientific induction is solved, any learning procedures that we employ in the context of an incompletable universe will necessarily remain arbitrary to some extent.

Recall that we are still not through with the question of whether Bayesian updating in a completed universe can properly count as learning. So far, it has been argued that the fact that real people clearly learn from experience is irrelevant to this question. The same is true of Bayesian statisticians operating in a universe that is incompletable, or which they have not chosen to complete. This leaves us free to focus on what is genuinely at issue. For this purpose, I want to draw an analogy between how a Bayesian using the massaging methodology I have attributed to Savage learns, and how a child learns arithmetic. It is true that the Bayesian is envisaged as teaching himself, but I do not think this invalidates the comparison.

When a child learns arithmetic at school, his teacher does not know what computations life will call upon him to make. Amongst other things, she therefore teaches him an algorithm for adding numbers. This algorithm requires that the child memorize some addition tables. In particular, he must memorize the answer to $2 + 3 = ?$. If the teacher is good at her job, she will explain *why* $2 + 3 = 5$. If the child is an apt pupil, he will

[3]Which is not the same as saying that there may not be *empirical* grounds for preferring Bayesian methods to classical methods.

understand her explanation. One may then reasonably say that the child has learned that, should he ever need to compute $2 + 3$, then the answer will be 5. Now consider the child in his maturity trying to complete an income tax form. In filling the form, he finds himself faced with the problem of computing $2 + 3$, and so he writes down the answer 5. Did he just learn that the answer to this problem is 5? Obviously not. He learned this in school. All that one can reasonably say that he "learned" in filling the form is that filling the form requires computing $2 + 3$. But such simple registering of undigested facts is excluded by the *caveat* that identifies learning with "adding to one's understanding". Of course, there may be children who are such poor students that they grow to maturity without learning their addition tables. Such a person might perhaps use his fingers to reckon with and thereby discover or rediscover that $2 + 3 = 5$ while filling the tax form. He would then undoubtedly have learned something. But he would not be operating in a completed universe within which all potential surprises have been predicted and evaluated in advance of their occurrence.

How is it that Bayesianismists succeed in convincing themselves that rational learning consists of no more than the trivial algebraic manipulations required for the use of Bayes' rule? My guess is that their blindness is only a symptom of a more serious disease that manifests itself as a worship of mathematical formalism. A definition-axiom-theorem-proof format is designed to close the mind to irrelevant distractions. But the aspects of the learning process that are neglected by Savage's formalism are not irrelevant. How decision-makers form and refine their subjective judgments really does matter. But the fact that Savage's theory leaves these aspects of the learning process utterly umodeled creates a trap into which Bayesianismists are only too ready to fall. The trap is to proceed as though anything that is not expressed in the formalism to which one is accustomed does not exist at all.

In game theory, however, the question of where beliefs come from cannot sensibly be ignored. Bayesianismist decision theory provides an adequate account of why we should study equilibria, but fails to make any headway at all with the problem of equilibrium *selection*. Game theorists therefore cannot afford to fall victim to Bayesianismist *newspeak*[4] if they hope to break out of the bridgehead they currently occupy.

[4]Recall from George Orwell's *1984* that *newspeak* is an invented language in which politically incorrect statements cannot be made.

3. Bayesianism in game theory

This section looks very briefly at two approaches to the problem of founding game theory on Bayesian principles. The second approach, due to Robert Aumann [4], hangs together very much better than the first. But this is because Aumann's approach does not attempt to do more than justify game theorists' obsession with the notion of an equilibrium. However, the first approach aims to say things about which equilibrium should be selected.

Harsanyi and Selten's [14] theory is without doubt the best known of the avowedly Bayesian approaches to the problem of equilibrium selection. However, it is too baroque a theory to lend itself to easy discussion in a paper like this. In brief, the notion of a *tracing procedure* lies at the heart of their model. Their procedure seeks to trace the manner in which Bayesian players will reason their way to an equilibrium. Other authors offer alternative accounts of how such reasoning might proceed. Skyrms [20] gives a particularly clean description of how he sees the deliberative process operating inside the head of a Bayesian player.

Skyrms [20] follows Harsanyi and Selten and others in supposing that, while deliberating, the players assign interim subjective probabilities to the actions available to their opponents. If these subjective probabilities are common knowledge,[5] along with the fact that everyone is a maximizer of expected utility, then an inconsistency will arise — unless the players' beliefs happen to be in equilibrium. When such an inconsistency arises, the players are assumed to update their subjective probabilities using Bayes' rule. Various candidates for the likelihood function can be considered (of which Skyrms offers a small sample). However, the modeling judgment made at this level is irrelevant to the point I want to make.

My criticism of this and similar models will be clear. By hypothesis, the players have *not* looked ahead to preview all possible lines of reasoning they might find themselves following in the future. They are therefore operating in a universe that is definitely incomplete. In such a universe, no special justification for the use of Bayesian updating exists. One might seek to rescue the special status of Bayesian updating by departing from Skyrms' story and postulating that the players have indeed previewed all the possible lines of reasoning open to them. But, after the previewing is over, there would be no scope for Bayesian updating because no there would then be no new information to incorporate into the system when the player began to reason for real. In summary, one might say that the conditions that justify the use of Bayes' rule in this context are satisfied

[5] As a consequence of the players' duplicating the reasoning processes of their opponents.

if and only if there is nothing for Bayes' rule to update.

One cannot make the same criticism of a recent paper by Kalai and Lehrer [15]. They envisage a game being played repeatedly in real time. The circumstances under which the repetition takes place need not concern us. For our purposes, it is enough that the players use Bayes' rule to update their beliefs after each repetition, and that Kalai and Lehrer give conditions under which there is convergence on a Nash equilibrium. What does such a conclusion mean? It is certainly a very reassuring consistency result for those like myself who regard Nash equilibrium as the basic tool of game theory. But is the result also a contribution to equilibrium selection theory? It is certainly true, as Kalai and Lehrer remark, that the limit equilibrium is a function of the players' prior beliefs,[6] but it seems to me that much care is necessary in interpreting this piece of mathematics. If we take seriously the notion that a players' prior beliefs are simply a summary of a set of massaged posterior beliefs, we have to abandon the idea that the players in Kalai and Lehrer's model are *learning* which equilibrium to play as the future unfolds. The players' *already knew* what equilibrium would be played under all possible future contingencies. Their initial snap judgements necessarily incorporate *preconceptions* on this subject that the model leaves unexplained. Any learning took place during the unmodeled introspection period before the play of the game when the players previewed all possible courses the game might take and predicted how the game would end up being played after each of these possible sets of common experience.

It should be emphasized that the last thing I wish to do is to criticize anyone for seeking to model the *process* by means of which equilibrium is achieved. Indeed, I have contributed to this literature myself (Binmore [11]). Far from decrying such work, I believe that the reason game theorists have made so little progress with the equilibrium selection problem is because of their reluctance to confront such questions. I do not even object to Bayesian updating being used as a learning rule in this context, *provided* that nobody is claiming any special status for it beyond the fact that it possesses some pleasant mathematical properties. However, other learning rules also have virtues, and the decision to use Bayes' rule in the context of an incomplete universe is no less *ad hoc* than the decision to use one of the rival rules. My own preferred research strategy on this subject is not to make any *a priori* choice at all of a learning rule, but to let one emerge endogenously as a consequence of the operation of evolutionary pressures. However, this is an approach fraught with many difficulties.

[6] In general, the limit equilibrium will also depend on random events that occur during play.

Aumann's [4,5] attempt to provide Bayesian foundations for game theory is very different in character from the work discussed so far in this section. Nobody learns anything or even decides anything in his very static model. Things are "just the way they are", and we are offered the role of a passive observer who sits on the sidelines soliloquizing on the nature of things. Such a model is not well-adapted to the equilibrium selection problem. Its purpose is to clarify what kinds of equilibria should lie in the set from which a selection needs to be made.

In brief, Aumannn postulates a universe of discourse whose states are *all-inclusive*. A description of such a state includes not only what players know and believe about the world and the knowledge and beliefs of other players, but also what all the players will *do* in that state. In such a framework, it becomes almost tautological that players whom fate has decreed will be Bayesian-rational in every state will necessarily operate some kind of equilibrium. Aumann then notes that, if what the players know always includes what strategy they find themselves using, then they will necessarily be frozen into what he calls a "subjective correlated equilibrium".[7]

The preceding paragraph is a sorry excuse for an assessment of how Aumann proceeds. A longer and more detailed account appears in Binmore [8]. However, what has been said is perhaps enough to make it clear that Aumann's universe is definitely not a small world. Indeed, his universe is as large as a universe could possibly be, since its states encompass everything that matters. However, Aumann evades the traps that await the Bayesianismist by refusing to classify his theory either as descriptive or as prescriptive. He describes his model as "analytic" to indicate that all the action takes place in the head of an otherwise passive observer. The model certainly cannot be prescriptive because there is no point in offering advice to players who "just do what they happen to do" and "believe what they happen to believe". Nor can the model be descriptive of a world in which people make conscious choices after transferring their experience into subjective judgments about the way things are. However, it seems to me that the latter is precisely the kind of world with which game theory needs to grapple.

I want to argue now that such a world is *necessarily* large in Savage's sense. The case for this is even stronger than the standard claim that the universe within which physics is discussed is incompletable. Or, to say the same thing more flamboyantly, inner space is necessarily even more

[7]This is not such an innocent assumption as it may appear. When the players are modeled as self-correcting computing machines, it becomes more than a little problematic (Binmore [10]).

mysterious than outer space. The reason is that, if the thinking processes of a player are to be modeled, then we are no longer free to envisage that all possible mental processes have been completed. A player *cannot* exhaustively evaluate all contingencies in a universe that includes his own internal deliberations and those of other players like himself. The issue is more fundamental than whether Bayesianism is applicable or not, since one cannot even rely on the *epistimology* that Bayesianismists take for granted.

Bayesians usually work with possibility sets[8] in specifying what a person knows. The possibility set $P(\omega)$ consists of the set of all states that the decision-maker thinks possible when the true state is ω. Equivalently, it is the event that he perceives in state ω. But suppose that we model players as Turing machines[9] — i.e. as programs that run on computers which have no storage constraints. Then we have to take on board the fact that possibility questions must be settled algorithmically.

To explore this issue, imagine that, for each *all-inclusive* state ω, possibility questions are resolved by a Turing machine $S = S(\omega)$ that sometimes answers NO to questions that begin: *Is it possible that ...?* Unless the answer is NO, possibility is conceded. (Timing issues are neglected.)

Consider a specific question concerning the Turing machine N. Let the computer code for this question be $\lceil N \rceil$. Let $\lfloor M \rfloor$ be the computer code for the question: *Is it possible that M will answer* NO *to* $\lceil M \rceil$*?* Finally, let T be a Turing machine that outputs $\lfloor x \rfloor$ when its input is $\lceil x \rceil$. Then the program $R = ST$ that consists of first operating T and then operating S, responds to $\lceil M \rceil$ as S responds to $\lfloor M \rfloor$.

Suppose that R responds to $\lceil R \rceil$ with NO. Then S reports that it is *impossible* that R responds to $\lceil R \rceil$ with NO. If what I know is true, it must therefore be that R never responds to $\lceil R \rceil$ with NO. But, if we as observers know this, why don't we replace S with a better program: one that accurately reflects our knowledge? Either our algorithm for determining what is possible is "incomplete" in that it allows as possible events we know to be false, or it is "inconsistent" in that it rejects as impossible events we know to be true.

This echoing of Gödel is no accident. The halting problem for Turing machines, from which the preceding example is adapted, is closely related

[8]Game theorists refer to an elaboration of the idea of a possibility set as an *information set* (Binmore [9]).

[9]The Church-Turing hypothesis asserts that any formal calculation possible for a human mathematician can be aped by a Turing machine. Penrose [17] bravely puts the case for humans being able to transcend the limitations of such machines. Those who are consitutionally inclined to this view should read his book to find out what they are letting themselves in for in the way of assumptions about how the human mind works.

to part of Gödel's reasoning. Note, in particular, the self-reference involved in asking a machine how it will respond to a question about how it responds to questions.

If the implications of taking an algorithmic view of knowledge acquisition are taken seriously, then the consequences for Bayesian epistomology run very deep. Binmore and Shin [6] give some (not very profound) arguments why the modal logic (S5) that characterizes knowledge for Bayesians would need to be replaced by the modal logic (G) that Solovay [21] showed to represent the "provable principles of provability" in Peano Arithmetic. (See also Shin [19] and Artemov [3].)

One escape from such difficulties is to abandon the requirement that states be all-inclusive, so that self-referential questions that trouble knowledge algorithms can be disbarred. That is, one can seek to *complete* the universe of discourse. But self-reference is intrinsic to game theory, which is *about* chains of reasoning that go, "If I think that he thinks that I think ...". Papers that exploit the self-referential difficulties that arise in this specific context are Binmore [10], Anderlini [2] and Canning [12].

Anderlini offers a particularly insightful observation for those Bayesians who like to argue that "game theory is based on the assumption that it is common knowledge that the players are rational". Such observations are thrown into the ring with no thought as to the nature of the universe of discourse. We are not even told what sort of entity a player is.[10] However, if a player is a Turing machine and "rationality" is defined in a natural way, Anderlini notes that the latter is not an effectively computable concept. That is, one can know every instruction in the computer program of the opponent and still not be able to tell whether the opponent is "rational".

4. Modeling players

In this section the need for modeling the players in game theory will be taken for granted. Some reasons are given in Binmore [7,8], but perhaps the most persuasive reason is the manner in which game theorists of all stripes have been driven, almost in spite of themselves, to the study of "bounded rationality".

Once a player has been modeled, one can say things about his complexity. In particular, one can compare his complexity to that of his environment. One might summarize this lecture so far by saying that, if his environment is sufficiently complex compared with the complexity of

[10]Whatever the definition of a player may be, it must certainly be rich enough in structure to admit the possibility of a player being irrational. Otherwise the statement would be empty.

his mental apparatus, then a Bayesianismist view of his predicament is untenable.

I do not have any Ism to offer as an alternative to Bayesianism for decision-making in incompletable universes. I want only to make a plea for the issue to be returned to the research agenda from which it was displaced by the triumph of Bayesianism in the economics profession. It is worth noting that those who knew and worked with Savage in the fifties were under no illusions about the importance of the problem. Luce and Raiffa [16], for example, list a number of systems for making decisions "under complete ignorance" in which the incompleteness of the universe of discourse is explicitly acknowledged. The existence of such systems indicates that the problem of decision-making in an incompletable universe is not a featureless desert about which one can hope to say nothing at all. I do not feel able to endorse any of these systems, since they all appeal to axioms that I find it hard to evaluate in the abstract. Instead, I plan to describe some simple structural observations that seem to me to follow from little more than the requirement that a decision-maker be modeled as a computing machine.

4.1 A belief machine

Let us simplify the problem to be considered by allowing only two consequences, winning and losing. Which of these will occur depends on some process about which the decision-maker is only partially informed. A Turing machine M will be used to model the manner in which the decision-maker evaluates his partial information. The input to M is therefore the data D available to the decison-maker about the unknown process. Since this lecture is a piece of rhetoric directed at Bayesianismists who believe that all ignorance can and should be quantified using subjective probabilities, let us restrict the output of the machine M to probabilistic statements. More specially, imagine that the machine M has k output devices $H_1, H_2, \ldots H_k$, each of which corresponds one of the intervals $I_1, I_2, \ldots I_k$ in a partition of $[0, 1]$. Each output device may or may not eventually type NO. When the output device H_j types NO, the understanding is that this answers the question: *Is it possible that the notional probability π of winning lies in the interval I_j?* It must be remembered that nothing guarantees that a Turing machine will stop calculating at all.

We presumably wish to exclude the possibility that all output devices will eventually print NO. But, if we are to take incompletable universes seriously, it must be recognized that we cannot simultaneously insist that only *one* output device will fail to output NO. If the set of admissible inputs D is not artificially restricted, then sometimes M will calculate

forever without succeeding in tying π down to a single interval I_j.[11]

4.2 Upper and lower probabilities

Such considerations lead very naturally to the notion of upper and lower probabilities with which many decision-theorists have toyed. All that is needed, in addition to what has already been assumed, is the assumption that the subset S of $[0,1]$ that remains after all the intervals I_j that are going to be excluded have been excluded should necessarily be convex. One may then argue that all that is known about the notional probability π is that it lies between the upper and lower limits of the interval S. The idea of a probability π therefore has necessarily to be supplemented by allowing intervals $[\underline{\pi}, \overline{\pi}]$ in which $\overline{\pi}$ is an *upper probability* and $\underline{\pi}$ is a *lower probability*.[12]

No-nonsense subjectivists like to debunk their critics by insisting that the critics compare bets on events to which the critics are reluctant to assign subjective probabilities with bets on events for which the appropriate probabilities are uncontroversial. In the case of a decision-maker who uses the machine M, they would therefore seek two situations between which the decision-maker is indifferent — one in which M outputs π and one in which M outputs $[\underline{\pi}, \overline{\pi}]$. However, even if the decision-maker expresses such an indifference, it does not follow that he is saying that he regards the output $[\underline{\pi}, \overline{\pi}]$ from M as being equivalent to the output π. He will be expressing a *preference* not a *belief*. It is true that, with Savage's consistency axioms, these ideas merge. But Savage's consistency axioms are not designed for application in an incompletable universe, and it is therefore no longer possible to take for granted that a person's von Neumann and Morgenstern *utility* for a process that can lead only to winning or losing may be identified with the person's subjective *probability* of winning.

Von Neumann and Morgenstern utilities are mentioned because there seems no particular reason why one should not ask that the preferences a decision-maker has over lotteries in which the prizes are objects of the form $[\underline{\pi}, \overline{\pi}]$ should not satisfy the von Neumann and Morgenstern rationality axioms. If so, it will make sense to speak of the von Neumann and

[11] One might argue that the decision-maker might be able to tie things down further if he were allowed to examine a transcript of the calculations made by M. But the ground rules are that any such examination would need to be expressible algorithmically. We could then construct a Turing machine that does the same thing as the decision-maker and run this along with M. We would then be making our judgments with a Turing machine again, albeit a larger Turing machine than M.

[12] I am ignoring two issues. The first is that only approximations to probabilities can emerge from such a procedure. The second is that one cannot wait for ever to learn for sure which output devices are going to fail to print NO.

Morgenstern utility $u[\underline{\pi}, \overline{\pi}]$ of a process. One will presumably wish to insist that

$$u[\underline{\pi}, \underline{\pi}] \leq u[\underline{\pi}, \overline{\pi}] \leq u[\overline{\pi}, \overline{\pi}],$$

and to normalize so that $u[p, p] = p$, but it is not clear what further rationality requirements are appropriate. One possibility is to ask that the decision-maker evaluates $\overline{\pi}$ and π "separately".[13] An argument of Keeney and Raiffa then shows that $u[\underline{\pi}, \overline{\pi}]$ must take one of the two forms:

$$\underline{u}(\underline{\pi}) + \overline{u}(\overline{\pi}) \qquad or \qquad \underline{u}(\underline{\pi}) \times \overline{u}(\overline{\pi}).$$

Thus, for example, it could be that

$$u[\overline{\pi}, \overline{\pi}] = \alpha\underline{\pi} + \beta\overline{\pi} \qquad or \qquad u[\underline{\pi}, \overline{\pi}] = \pi^{\alpha}\overline{\pi}^{\beta},$$

where $\alpha \geq 0$, $\beta \geq 0$ and $\alpha + \beta = 1$.

4.3 Updating upper and lower probabilities

Consider now three processes with respective data D_1, D_2 and D_3. Let D_3 be the process in which the decision-maker wins if and only if he wins in both D_1 and D_2. It is then natural to say that D_1 and D_2 should be regarded as independent processes if $\underline{\pi}_3 = \underline{\pi}_1\underline{\pi}_2$ and $\overline{\pi}_3 = \overline{\pi}_1\overline{\pi}_2$. One can then deal with conditioning by writing $D_2|D_1$ instead of D_2. It then seems that, although we may not be able to assign probabilities to all events in an incompletable universe, nevertheless Bayes' rule is still with us as the appropriate method for updating upper and lower probabilities.

I think this conclusion is correct, provided that one is not naive about the circumstances in which the procedure is used. In reaching the conclusion that upper and lower probabilities should be updated using Bayes' rule, I implicitly made use of *consistency* assumptions. However, earlier in the lecture, it was argued that such consistency assumptions only make good sense in a completable universe. If upper and lower probabilities are to be updated by Bayes' rule, we therefore need to be able to argue that the relevant universe for this particular operation is completable. Among other things, we need to be confident that the machine M would not respond differently to the input D_1 after being asked to evaluate $D_2|D_1$ than it did before. If the machine were to operate like a human decision-maker following Savage's methodology, this would be assured if the machine's massaging activities while originally assessing the data D_1

[13]This requires, for example, that $[\underline{\pi}, \overline{\pi}] \preceq [\underline{\pi}, \overline{q}] \Leftrightarrow [\underline{q}, \overline{\pi}] \preceq [\underline{q}, \overline{q}]$ whenever the expressions are meaningful. Not only this, the relationship must survive when lotteries are taken over $\overline{\pi}$ and $\overline{q}$. Moreover, everything must be the same when it is the second argument that is held constant in comparisons rather than the first.

were sufficiently wide ranging as to include the possibility that it might later be offered the data $D_2|D_1$. However, in an incompletable universe, it will not be possible for the machine to anticipate what the effect of *all* possible future data will be on the manner in which it processes data. Like ourselves, the machine will not only learn, it will learn how to learn as it gathers experience, and it is impossible for the machine to predict how it might possibly reprogram itself under all future contingencies.

Upper and lower probabilities in games?

It is necessary to round off this section be indicating how the ideas it presents would work in a game-theoretic setting. I hope, however, that what comes next will not be regarded as an attempt to construct a new theory of games. It is merely a piece of rhetoric whose aim is to discomfort Bayesianismists by bringing to their attention what they would have to believe if they genuinely sought to implement their ideas algorithmically.

Figure 1 shows a payoff matrix for a version of a well-known "toy" game called the Battle of the Sexes. It comes with a silly story about a husband (player I) and a wife (player II) who did not agree at breakfast whether to go to a boxing match or a ballet performance in the evening. Later in the day they get separated and hence have to make the decision of where to go in the evening independently.

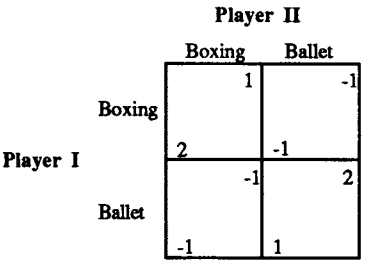


Figure 1: The Battle of the sexes game.

According to a traditional analysis, the game has three Nash equilibria.[14] There are two Nash equilibria in pure strategies: namely (*Boxing, Boxing*) and (*Ballet, Ballet*). However, unless some way can be found to break the symmetry, neither can be the "solution" of the game, since any argument in favor of one of the pure strategy equilibria is equally an argument in favor of the other. The third Nash equilibrium calls for both

[14]In a Nash equilibrium, each player's strategy choice is optimal given the strategy choices made by the other players.

players to use mixed strategies. That is, each player randomizes over his or her pure strategies. To be precise, the husband and wife each choose *boxing* independently with respective probabilities $\frac{3}{5}$ and $\frac{2}{5}$. However, this third Nash equilibrium is not a very attractive candidate for the "solution" of the game, because each player's solution payoff would then be no more than his or her security level.[15]

However, we perhaps ought to ask ourselves whether we have really exhausted all the Nash equilibria. Is it not possible, for example, that the wife might employ a decision process in deciding what action to take whose data, when taken as input for the husband's assessment machine, leads it to produce the output $[\underline{\pi}_2, \overline{\pi}_2]$ when questioned about the notional probability with which she will use *Boxing*?

To simplify the situation, imagine that the husband's ultimate aim is to win a prize, and the wife's is to win a second and separate prize. One may then take the entries in the payoff matrix of Figure 1 to be the probabilities with which the players will win their respective prizes for each the four possible pure strategy combinations. If player I now uses a decision process with data D_1 and player II independently uses a decision process with data D_2, then we can symbolically represent the process that decides whether player I wins his ultimate prize as

$$(D_1 \wedge D_2) \vee (\neg D_1 \wedge \neg D_2 \wedge D),$$

where D is the data for a process, independent of D_1 and D_2, that is assessed at $\frac{2}{3}$. If we make Bayesian assumptions about how such combinations of processes should be manipulated, the husband's machine will assess the combination as

$$[\underline{\pi}_1\underline{\pi}_2 + \frac{2}{3}(1-\overline{\pi}_1)(1-\overline{\pi}_2), \overline{\pi}_1\overline{\pi}_2 + \frac{2}{3}(1-\underline{\pi}_1)(1-\underline{\pi}_2)].$$

To proceed further it is necessary to make assumptions about the utility functions with which the players evaluate such assessments. I want only to observe that if both players have utility functions defined by

$$u[\underline{\pi}, \overline{\pi}] = \{\underline{\pi}\}^{\frac{1}{2}}\{\overline{\pi}\}^{\frac{1}{2}},$$

then the Battle of the Sexes not only has Nash equilibria other than those traditionally considered, it has Nash equilibria that generate payoffs for

[15]Player I's security level is his expected payoff if he acts on the assumption his opponent will guess his choice of mixed strategy in advance and respond by choosing a strategy herself that minimizes his payoff.

both players that are better than the $\frac{1}{5}$ the players get when the traditional mixed equilibrium is used.

5. Summary

This lecture has been an attack on Bayesianism, which I see as a metaphysical doctrine that hinders advances in the foundations of game theory. The lecture began with an appeal to the authority of Savage. It continued with an attempt to explain how it can be shown that certain universes of discourse *cannot* be completable in the sense required to legitimize a Bayesianismist methodology. It concluded with a brief discussion of some of the implications of looking seriously at the idea that decision-making should be described in terms of algorithms.

REFERENCES

[1] M. ALLAIS. Le comportement de l'homme rationnel devant le risque: critiqye des postulants et axiomes de l'ecole Americaine. *Econometrica*, 21:503–546, 1953.

[2] L. ANDERLINI. *Some Notes on Church's Thesis and the Theory of Games.* Cambridge Economic Theory Discussion Paper, 1988.

[3] S. ARTEMOV. Kolmogorov's logic of problems and a provability interpretation of intuitionistic logic. In R. Parikh, editor, *Theoretical Aspects of Reasoning About Knowledge*, Morgan Kaufmann, San Mateo, CA, 1990.

[4] R. AUMANN. Correlated equilibrium as an expression of Bayesian rationality. *Econometrica*, 55:1–18, 1987.

[5] R. AUMANN. Interactive epistemology. 1989. Working Paper, Cowles Foundation, Yale University.

[6] K. BINMORE and H. SHIN. Algorithmic knowledge and game theory. 1990. University of Michigan Discussion Paper.

[7] K. BINMORE. *Essays on Foundations of Game Theory.* Basil Blackwell, Oxford, 1990.

[8] K. BINMORE. Foundations of game theory. In J-J. Laffont, editor, *Advances in Economic Theory: Proceedings of the Sixth World Congress of the Econometric Society*, Cambridge University Press, Cambridge, 1991.

[9] K. BINMORE. *Fun and Games.* D. C. Heath, Lexington, Mass., 1991.

[10] K. BINMORE. Modeling rational players I. *Economics and Philosophy*, 3:9–55, 1987.

[11] K. BINMORE. Modeling rational players II. *Economics and Philosophy*, 4:179–214, 1987.

[12] D. CANNING. Rationality and game theory when players are turing machines. 1988. London School of Economics ST/ICERD Discussion Paper 88/183.

[13] P. GÄRDENFORS. *Knowledge in Flux.* MIT Press, Cambridge, Mass., 1988.

[14] J. HARSANYI and R. SELTEN. *A General Theory of Equilibrium Selection in Games.* MIT Press, Cambridge, 1988.

[15] E. KALAI and E. LEHRER. Rational learning leads to Nash equilibrium. 1990. Discussion Paper 895, Northwestern University.

[16] R. LUCE and H. RAIFFA. *Games and Decisions.* Wiley, New York, 1957.

[17] R. PENROSE. *The Emperor's New Mind.* Oxford University Press, Oxford, 1989.

[18] L. SAVAGE. *The Foundations of Statistics.* Wiley, New York, 1951.

[19] H. SHIN. *Logical Structure of Common Knowledge, I and II.* Nuffield College, Oxford, 1987.

[20] B. SKYRMS. Dynamic models of deliberation and the theory of games. In R. Parikh, editor, *Reasoning About Knowledge*, Morgan Kaufman, San Mateo, CA, 1990.

[21] R. SOLOVAY. Provability interpretation of modal logic. *Israel Journal of Mathematics*, 25:287–304, 1976.

Logic, Methodology and Philosophy of Science IX
D. Prawitz, B. Skyrms and D. Westerståhl (Editors)

NORMATIVE VALIDITY AND MEANING OF VON NEUMANN-MORGENSTERN UTILITIES

JOHN C. HARSANYI

University of California at Berkeley

1. The problem

Payoffs in game theory are usually expressed in von Neumann-Morgenstern (vNM) utilities. Yet, there is a lot of experimental evidence that people's behavior often fails to conform to the vNM axioms. This empirical fact is part of the more general observation that people do not consistently follow *any one* of the rationality requirements of economic theory, and deviate even from such very basic ones as extensionality[1] and transitivity (see Tversky and Kahneman, 1981; Arrow, 1982; and Schoemaker, 1982).

The observed deviations from the vNM axioms pose two different problems. One concerns the *predictive* value of these axioms and of economic models based on these axioms. The other concerns their *normative* validity as rationality requirements. In this paper I shall restrict my discussion to this latter problem. These are two different problems. For even if we decided that the vNM axioms had full normative validity, we should not be surprised if natural selection failed to provide us with an instinctive ability to make rational and efficient choices between often quite complicated lotteries in accordance with the vNM axioms, since our animal and early human ancestors were never confronted with such problems, and obviously did not suffer any evolutionary disadvantage by lacking the instinctive ability to make such choices in a proficient manner. Nobody

[1] I shall follow Arrow (1982) in describing as *extensionality* the requirement that people's choices between two alternatives should not depend on the way these are described to them as long as these descriptions are logically clearly equivalent. As we all know, in actual fact people do not satisfy this requirement. For instance, their willingness to undergo an operation may be quite different if they are told that this operation has a survival rate of 95 per cent than if they are told that it has a fatality rate of 5 per cent.

doubts the normative validity of arithmetic, yet most children cannot solve arithmetic problems of any complexity without special training.

I shall argue that the vNM axioms do have full normative validity as rationality axioms — but that this is true only under some very specific *motivational* assumptions.

2. Notations

I shall distinguish between *pure* and *mixed* alternatives depending on whether they do, or do not, include risk and/or uncertainty. Pure alternatives will be regarded as degenerate special cases of mixed alternatives. I shall use also the term *lotteries* to describe mixed alternatives.

Strict preference, equivalence and *nonstrict preference* will be denoted as $\succ$, $\sim$, and $\succsim$, respectively.

Let L be a lottery yielding alternative A_k if event e_k occurs, with $k = 1, \ldots, n$. Then I shall write

$$L = (A_1 \,|e_1; \ldots; A_n|\, e_n)\,. \tag{1}$$

The alternatives A_k will also be called *prizes* or *outcomes*. The events e_k will be called *conditioning events*. It will be assumed that these events are mutually exclusive and exhaust all possibilities.

Suppose the decision maker *knows* the objective probabilities $p_1, \ldots, p_n$ associated with the events $e_1, \ldots, e_n$. Then I shall write

$$L + (A_1, p_1; \ldots; A_n, p_n) \tag{2}$$

and shall call L a *risky* lottery. Of course, all these probabilities must be non-negative and must add up to unity. On the other hand, a lottery will be called an *uncertain* lottery if it is *not* a risky lottery under this definition.

3. A simplified and generalized version of the von Neumann-Morgenstern axioms

Von Neumann and Morgenstern's (1947) theory was restricted to the case where all probabilities were *objective* probabilities *known* to the decision maker, i.e., to the case of *risky* lotteries. But most economists use Savage's (1954) theory that employs only *subjective* probabilities (which he called *personal* probabilities). Therefore, it does cover also the general case where objective probabilities may be unknown or may be even undefined. Anscombe and Aumann (1963) proposed still another theory,

likewise covering the general case, but using much simpler axioms. Their theory, however, has to use *both* objective and subjective probabilities. (But all the objective probabilities they need may be generated by *one* random device whose statistical behavior is known to the decision maker.)

What makes the vNM axioms rather complicated is the fact that some of them simply *restate* certain propositions of probability theory. In order to simplify my axioms — and in order to make it clear what the logical status of each axiom is — I have separated my four rationality postulates, to be called simply *axioms*, from my two *background assumptions*, whose only purpose is to enable us to use the theorems of the Propositional Calculus and of the Probability Calculus in any mathematical proof. My two background assumptions are:

ASSUMPTION I. *The conditional statements defining a lottery [as stated in the sentence preceding (1)] follow the laws of the Propositional Calculus.*

ASSUMPTION II. *The objective probabilities defining a risky lottery [as in (2)] follow the laws of the Probability Calculus.*

I need Assumption I because I want to use Anscombe and Aumann's "Reversal of order" postulate without making it into a separate axiom. Their postulate assumes that their "roulette lottery" and their "horse lottery" will be conducted *consecutively* but that it makes no difference if their time order is *reversed*. But we can just as well assume that the two lotteries will be conducted *simultaneously*. Once this assumption is made, their postulate becomes a corollary of a well-known theorem of the Propositional Calculus. If we write $p \rightarrow q$ for the statement "If p then q", and write $=$ for logical equivalence, then the relevant theorem can be written as

$$p \rightarrow (q \rightarrow r) = q \rightarrow (p \rightarrow r)\,. \tag{3}$$

I need Assumption II because, in order to compute the final probability of each outcome in a two-stage lottery, I want to use the addition and the multiplication laws of the Probability Calculus, without making them into separate axioms.

I also need the following four rationality axioms.

AXIOM 1 (COMPLETE PREORDERING). The relation $\succsim$ (non-strict preference) is a complete preordering over the set of all lotteries. (That is to say, $\succsim$ is both transitive and complete.)

AXIOM 2 CONTINUITY). Suppose that $A \succ B \succ C$. Then there exists some probability mixture

$$L(p) = (A, p; C, 1-p) \tag{4}$$

of A and C with $0 \leq p \leq 1$ such that $B \sim L(p)$.

AXIOM 3 (MONOTONITY IN PRIZES). Suppose that $A_k^* \succsim A_k$ for $k = 1, \ldots, n$. Then also $L^* \succsim L$, where

$$L^* = (A_1^* | e_1; \ldots; A_n^* | e_n) \text{ and } L = (A_1 | e_1; \ldots; A_n | e_n). \tag{5}$$

(This axiom is a version of the sure-thing principle.)

AXIOM 4 (PROBABILISTIC EQUIVALENCE). Let *Prob* denote objective probability. Define the lotteries L and L' as

$$L = (A_1 | e_1; \ldots; A_n | e_n) \text{ and } L' = (A_1 | f_1; \ldots; A_n | f_n). \tag{6}$$

Suppose the decision maker *knows* that

$$Prob\ (e_k) = Prob\ (f_k) \text{ for } k = 1, \ldots, n. \tag{7}$$

Then, for this decision maker, we must have

$$L \sim L'. \tag{8}$$

In other words, a rational decision maker will be indifferent between two lotteries yielding the *same prizes*, and yielding each prize with the *same objective probability* — regardless of the physical mechanisms the two lotteries use to generate these probabilities. In particular, he or she[2] must be indifferent between a one-stage and a two-stage lottery yielding the same prizes with the same probabilities.

We can now state the following theorem.

THEOREM. *Given Assumptions I and II, an individual i whose preferences satisfy Axioms 1 to 4 will have a utility function U_i that equates the utility $U_i(L)$ of any lottery L to the* expected utility *of this lottery so that*

$$U_i(L) = \sum_{k1}^{n} p_k U_i(A_k), \tag{9}$$

where $p_1, \ldots, p_n$ are either the objective *probabilities of the conditioning events $e_1, \ldots, e_n$* known *to the decision maker,* or *are his own* subjective *probabilities for these events.*

If a utility function U_i satisfies (9) then it is said to possess the *expected-utility property*, and is called a *von Neumann-Morgenstern utility function.* Given Assumption II, our axioms are equivalent to the vNM axioms,

[2] In what follows, for stylistic reasons, in similar phrases I shall omit the female pronoun.

which means that they can be used to prove the above theorem for *risky* lotteries. On the other hand, as Anscombe and Aumann have shown, we can extend the theorem to *all* lotteries by using our axioms, after adding as new axioms the theorem itself, restricted to risky lotteries, and Anscombe and Aumann's "reversal of order" postulate, derived from equation (3) by means of Assumption I (as explained above).

In view of the above theorem, we can now extend the notation used in (2) also to *uncertain* lotteries if we reinterpret the probabilities $p_1, \ldots, p_n$ as the decision maker's *subjective* probabilities for the conditioning events $e_1, \ldots, e_n$.

It is easy to verify that the *converse of the theorem* is likewise true: If a person's choices among lotteries consistently maximize the expected value of some utility function then his behavior will satisfy Axioms 1 to 4.

4. Need for an outcome-oriented attitude

I shall now consider the *normative validity* of our four axioms as rationality requirements. To start with axiom 1 (complete preordering), this is a rationality axiom used in all parts of economic theory. Its normative validity is rather uncontroversial.[3]

On the other hand, Axiom 2 (continuity) is basically a regularity assumption, rather than a rationality requirement. Even in the absence of Axiom 2, we can show the existence of a utility indicator with the expected-utility property by using our other three axioms. But this utility indicator will not be a real-valued (scalar-valued) utility function, but rather will be a *utility vector* with two or more lexicographically ordered components (see Hausner, 1954). We need a continuity axiom (like Axiom 2) only to ensure the existence of a *scalar-valued* utility function.

Thus, the real question is how much normative validity our axioms 3 and 4 have. To answer this question, I propose to divide the utilities associated with choices involving risk and/or uncertainty into *outcome utilities* and *process utilities*. The former are the (positive and negative) utilities the chooser derives from the various possible *outcomes* (or prizes) of each lottery. The latter are the (positive and negative) utilities he derives from his *psychological experiences* before, during, and after the act of gambling itself. These experiences include the nervous tension produced by gambling; the joy of winning and the sorrow of losing; the pride or the regret of having made what has turned out to be the right choice or the

[3] Yet, we know from experimental studies that, in choices of some complexity, people's behavior often fails to conform to this axiom (mainly because they make *intransitive* choices; but see May, 1954, for an interesting discussion).

wrong choice; the favorable or unfavorable reactions by other people to the final outcome and to the decision maker's purported responsibility for this outcome; and so on.

When people gamble for entertainment, they tend to do so both in the hope of winning valuable prizes and also in the hope of having a good time, which means that they are guided *both* by their outcome utilities and by their process utilities. But this may not be true when the stakes are *very high*, or when the participants are business executives or political leaders gambling with *other people's* money (or even with other people's lives). In such cases, the decision makers will have very good reasons, based both on self-interest and on moral considerations, to concentrate on trying to achieve the best possible *outcomes* both for themselves and for their constituents, without being diverted from this objective by the pleasant or unpleasant subjective experiences they derive from the process of gambling itself.

This suggests the following definition. I shall say that a decision maker takes a strictly *outcome-oriented* attitude if he is guided solely by his *outcome utilities*, i.e., by the utilities he assigns to the various possible outcomes of each lottery, and by his *outcome probabilities*, i.e., by the probabilities he associates with these outcomes. I shall argue that our Axioms 3 and 4 are perfectly *valid normative rationality requirements* for decision makers with strictly *outcome-oriented* attitudes but are *not* valid rationality requirements for *other* decision makers, who are wholly or partly guided by their *process utilities* derived from the process of gambling itself.

Let me first discuss Axiom 4 (probabilistic equivalence). As we have seen, this axiom implies that a rational decision maker will be indifferent between a one-stage and a two-stage lottery if both yield the same prizes with the same probabilities. Since, by definition, a strictly outcome-oriented person will be interested only in the possible outcomes and their probabilities, and these two pieces of information will be the *same* for both lotteries, he will have to be indifferent between the latter, as required by our axiom.

On the other hand, Axiom 4 is *not* a valid rationality requirement for a person guided wholly or partly by his process utilities derived from the process of gambling itself. For a one-stage and a two-stage lottery may generate very *different* psychological experiences and therefore also very *different* process utilities, even if they both yield the same possible prizes with the same probabilities. For instance, a one-stage lottery will give rise only to *one* period of nervous tension whereas a two-stage lottery will give rise to *two* such periods. As a result, other things being equal, among

people who give some weight to their process utilities, some will tend to prefer *one-stage* lotteries because they derive higher process utilities from them, while others will tend to prefer *two-stage* lotteries because they derive higher process utilities from the latter. In both cases, this will be a *rational* preference — even though it will violate our Axiom 4.

Similar considerations apply to our Axiom 3 (monotonity in prizes). The axiom considers two lotteries L and L^*. Lottery L^* is obtained by replacing each prize A_k of L by a prize A_k^* *at least as desirable* as A_k itself from the decision maker's point of view. The axiom asserts that, under these assumptions, the decision maker will find the new lottery L^* itself also *at least as desirable* as the original lottery L was. The reason is that participation in lottery L^* will always yield the decision maker an outcome as good as or better than participation in lottery L would have yielded him. As this argument is based on comparing the possible *outcomes* of lottery L^* with those of lottery L, it shows that Axiom 3 is in fact a valid rationality requirement for any decision maker with a strictly *outcome-oriented* attitude.

Yet, it is *not* a valid rationality requirement for a person wholly or partly guided by his process utilities. To verify this, let me assume that the prizes $a_1, \ldots, A_n$ of lottery L are *pure alternatives* whereas the prizes $A_1^*, \ldots, A_n^*$ are themselves *lotteries* (lottery tickets). Under this assumption, L will be always a *one-stage* lottery because it will end as soon as one of its possible prizes, say, A_k, has been selected as outcome. In contrast, L^* will be a *two-stage* lottery, where at stage 1 one of the prizes A_k^* will be selected whereas at stage 2 the outcome of the lottery A_k^* itself will be decided. This means that a person who derives higher *process utilities* from one-stage than from two-stage lotteries may actually *prefer* L to L^*, even though he prefers the *prizes* of the latter to the corresponding prizes of the former — and this may be a perfectly rational preference from his point of view. In other words, a person who gives some weight to his process utilities may reasonably assign a *different utility* to a lottery A_k^* when it *is* embedded in a larger lottery L^* than when it is *not* so embedded.

5. Von Neumann-Morgenstern utility functions, outcome utilities, and process utilities

As we have seen, only people with a strictly outcome-oriented attitude will act consistently in accordance with the vNM axioms. By definition, these will be people guided only by their outcome utilities (and by the probabilities they assign to various outcomes) but paying no attention to

their process utilities. Yet, by the Theorem and its converse, both stated in section 3, only people with these characteristics will have vNM utility functions. This in turn implies that a person's vNM utility function, if he has any, *can express only his outcome utilities* and cannot but disregard his process utilities, which will have no influence on his choice behavior.

An even simpler way of verifying this is by inspection of equation (9) in section 3, which defined the utility $U_i(L)$ of a lottery L solely in terms of its *outcome utilities* $U_i(A_k)$ and the outcome probabilities p_k, without reference to any process utilities.

Let me add that von Neumann and Morgenstern (1947, esp. pp. 28 and 632) were perfectly aware of the fact that their axioms *excluded* what they called the "utility of gambling", and what I am calling "process utilities". But they apparently felt that exclusion of these utilities was simply a *shortcoming* of their theory, one to be removed eventually by devising a set of more powerful axioms. Of course, a formal theory covering also process utilities would be an important advance. Yet, in my own view, even though von Neumann and Morgenstern's original theory does not cover process utilities, it is an analytically very valuable theory because the vNM utility functions defined by it have very attractive mathematical properties, including that of being *cardinal* utility functions (see section 7 below).

6. Von Neumann-Morgenstern utility functions and attitudes toward risk

We often read in the literature that a person's vNM utility function expresses his attitude toward *risk taking*, i.e., toward *gambling*.[4] Yet, without proper qualifications, this is a very *misleading* statement. If we do not assume strictly outcome-oriented attitudes, then a person's willingness to take risks will depend on two factors:

(i) On his like or dislike for *gambling* as such, as determined by the positive and negative *process utilities* he associates with gambling.
(ii) On the utilities and probabilities he assigns to various possible *outcomes*.

For the sake of simplicity, I shall describe factor (i) as this person's *intrinsic attitude* toward gambling while describing factor (ii) as his *instrumental attitude* toward the latter. (In the latter case I shall speak of an *instrumental* attitude because it is not based on this person's like or

[4] In what follows, for convenience I shall follow colloquial usage and use the term "risk" so as to cover both "risk" and "uncertainty".

dislike for gambling *as such* but rather refers to his willingness to gamble for the sake of the various possible outcomes.)

As we have seen, in the case of people who have vNM utility functions at all, factor (i) will be completely inoperative, so that their only reason for gambling will be *instrumental*, based on their desire to achieve some specific outcomes.

Yet, when it is claimed that vNM utility functions express people's attitudes toward *gambling* without any qualification, it is natural to assume that their *intrinsic* attitude toward gambling — i.e., their *intrinsic* like or dislike for gambling — is being meant, even though, as we have seen, people's vNM utility functions cannot be affected by this attitude at all.

7. Von Neumann-Morgenstern utilities as *cardinal* utilities

I now propose to argue that vNM utility functions are cardinal utility functions. There are two basic differences between merely *ordinal* and *cardinal* utility functions. One is that the former allow meaningful comparisons only between the relevant individual's *utility levels* but not between his *utility differences*, whereas the latter allow both kinds of comparisons in a meaningful way. Thus, regardless of whether U_i is an ordinal or a cardinal utility function of individual i, the preference statement $A \succ B$ will be represented by the inequality $U_i(A) > U_i(B)$ whereas the indifference statement $A \sim B$ will be represented by the equation $U_i(A) = U_i(B)$.

On the other hand, if U_i is merely an *ordinal* utility function then inequalities and equalities between utility differences such as

$$\Delta U_i(A, B) = U_i(A) - U_i(B) \text{ and } \Delta U_i(C, D) = U_i(C) - U_i(D) \tag{10}$$

will have no introspective or behavioral meaning. In contrast, if U_i is a *cardinal* utility function then such inequalities and equalities will be meaningful. (As we shall see, in the special case where U_i is a vNM utility function, such inequalities and equalities will tell us something about i's preferences and indifferences between certain lotteries.)

The other difference is that an ordinal utility function U_i tells us only *what* i's preferences are whereas, if U_i is a cardinal utility function, then it will also permit us to *compare* i's different preferences as to their *intensities* or, equivalently, as to their *relative importance* for i.

The relevant mathematical facts will be stated in the form of the following:

LEMMA. *Consider the inequality*

$$\Delta U_i(A, B) > \Delta U_i(C, D). \tag{11}$$

This inequality will hold if and only if

$$L_1 = \left(A, \frac{1}{2}; D, \frac{1}{2}\right) \succ L_2 = \left(B, \frac{1}{2}; C, \frac{1}{2}\right). \tag{12}$$

Moreover, the Lemma remains true even if in (11) and in (12) the signs $>$ *and* $\succ$ *are replaced by the signs* $=$ *and* $\sim$*, respectively.*

To verify the first two sentences of the Lemma, note that, in view of (10), inequality (11) can be written also in the form

$$\frac{1}{2}U_i(A) + \frac{1}{2}U_i(D) > \frac{1}{2}U_i(B) + \frac{1}{2}U_i(C). \tag{13}$$

Yet, (13) implies, and is also implied by, statement (12). The last sentence of the Lemma can be verified in a similar way.

The Lemma shows how statements about one utility difference $\Delta U_i(A, B)$ being *larger than*, or being *equal to*, another utility difference $\Delta U_i(C, D)$ can be reduced to statements about i's *preference* for some lottery L_1 over some lottery L_2, or about i's *indifference* between the two lotteries. It also shows how, conversely, statements about i's preferences and indifferences can be reduced to inequalities and equalities between utility differences.

I now propose to show that, in view of our Lemma, if i prefers A to B but prefers C to D, then the utility differences $U_i(A, B)$ and $U_i(C, D)$ can be used to measure the *intensities* of these two preferences by i, or, equivalently, the *relative importance* of these two preferences for him.

Again consider the two lotteries

$$L_1 = \left(A, \frac{1}{2}; D, \frac{1}{2}\right) \quad \text{and} \quad L_2 = \left(B, \frac{1}{2}; C, \frac{1}{2}\right).$$

We can obtain L_1 from L_2 by making two moves: *Move I* will consist in replacing prize B by prize A in lottery L_2 whereas *Move II* will consist in replacing prize C by prize D. Since by assumption we have $A \succ B$ but $C \succ D$, Move I will amount to replacing a given prize by a *preferred* prize while Move II will amount to replacing a given prize by a *less preferred* prize. It is natural to assume that i will prefer lottery L_1 to lottery L_2 if and only if his preference for A over B has *greater intensity* or, equivalently, if it has *greater importance* for him, than his preference for C over D.

Yet, by our Lemma, i will prefer L_1 over L_2 if and only if $\Delta U_i(A, B)$ is *larger* than $AU_i(C, D)$. This means that i's preference for A over B

will have *greater intensity* and will have *greater importance* for him if and only if $\Delta U_i(A, B)$ is *larger* than $\Delta U_i(C, D)$. In other words, the two utility differences $\Delta U_i(A, B) = U_i(A) - U_i(B)$ and $\Delta U_i(C, D) = U_i(C) - U_i(D)$ can be used as *measures* for the *intensities* and for the relative *importance* of i's preference for A over B, and of his preference for C over D. This is of course an intuitively very plausible result: The mere fact that i prefers A to B is indicated by the piece of information that the utility difference $\Delta U_i(A, B)$ is *positive*. Thus, it is not surprising to find that the *magnitude* of this utility difference indicates the *intensity* of this preference and its *importance* for him.

8. Marginal utilities, complementarity, and substitution

Economists use vNM utilities primarily in analyzing choices involving risk and uncertainty. Other things being equal, the *more concave* a person's vNM utility function for money, i.e., the more strongly it displays *decreasing marginal utilities*, the less willing he will be to take risks; and the *more convex* his vNM utility function for money, i.e., the more strongly it displays *increasing marginal utilities*, the more willing he will be to take risks (cf. Friedman and Savage, 1948).

Yet, once vNM utility functions are available, they can be used also in other branches of economic theory. For instance, they can be used to replace the well-known Hicks-Allen definitions for complements and for substitutes (Hicks, 1939) by much simpler definitions. Let A and B denote specific amounts of commodities α and β. Let U_i be i's vNM utility function. Let $U_i(A\&B)$ denote the utility that i derives from consuming A and B *together*, and let $U_i(A)$ and $U_i(B)$ denote the utilities he derives from consuming A and B *separately*.

Then, A and B will be *complements* if

$$U_i(A\&B) > U_i(A) + U_i(B); \tag{14}$$

and they will be *substitutes* if

$$U_i(A\&B) < U_i(A) + U_i(B). \tag{15}$$

Under these definitions, i's vNM utility function for money will display *concavity*, i.e., *decreasing marginal utilities*, in those income ranges where among the commodities consumed by i *substitution* relations predominate. The opposite will be true in those income ranges where among

these commodities *complementarity* relations predominate. (For this purpose, indivisibilities must be considered to be special cases of complementarities.)

These conclusions usefully supplement those we reached in sections 6 and 7. There we concluded that a person's vNM utility function has nothing to do with his *intrinsic* like or dislike for gambling. Rather, it expresses his *instrumental* attitude toward risk taking and is itself determined by his *cardinal utilities* (outcome utilities) for various alternatives (such as alternative commodity baskets). Now we have found that these cardinal outcome utilities themselves depend on the *substitution* and *complementarity* relations existing among the commodities consumed by the relevant individual.

In any case, it is the decision maker's *cardinal utilities* (outcome utilities) for various alternatives that determine his (instrumental) *willingness to take risks* in order to obtain some desirable alternatives. These cardinal utilities determine his attitude toward risk taking, rather than the other way around.

9. Conclusion

In game theory, payoffs are usually expressed in vNM utilities. Yet, experiments show that many people repeatedly deviate from the vNM axioms as well as from other rationality axioms. This raises the question whether the vNM axioms have even *normative validity* as rationality requirements. To make it easier to answer this question, I proposed a simplified form of the vNM axioms, based on the Anscombe-Aumann (1963) approach to decision theory. Then, I proposed to divide the (positive and negative) utilities people derive from risky choices into *outcome utilities* and *process utilities*. The former are the utilities people assign to the various possible outcomes of any lottery whereas the latter are the utilities they derive from the process of gambling itself.

I argued that, in many choice situations involving risk, some people will have good reasons to *disregard* their process utilities and to be guided solely by the utilities and the probabilities they assign to the various possible outcomes. This attitude I described as a strictly *outcome-oriented* attitude.

I suggested that the vNM axioms have full normative validity as rationality requirements — but only for people taking this particular attitude. This, however, means that, if a person has a vNM utility function at all, this utility function can express only his *outcome utilities* and cannot but disregard his *process utilities*. Already von Neumann and Morgenstern re-

alized this fact (though they spoke of "the utility gambling" rather than of "process utilities").

It is often claimed that vNM utility functions express people's attitudes toward *gambling*. But the truth is that these utility functions have nothing to do with people's *intrinsic* attitudes toward gambling, i.e., with their intrinsic like or dislike for gambling as such. What they do express is people's *instrumental* attitudes toward risk taking, i.e., their willingness to take risks in order to obtain some desirable outcomes.

Then I tried to show that vNM utility functions are *cardinal* utility functions, which permit us to make meaningful comparisons, not only between *utility levels* but also between *utility differences*, and which also permit us to compare a person's different preferences as to their *intensities* or, equivalently, as to their relative *importance* for the individual in question.

Finally, I proposed definitions for *complementarity* and for *substitution* in terms of a person's vNM utility function, and argued that the convexity or the concavity of a person's vNM utility function for money in any given income range depends on whether complementarity or substitution relations predominate among the commodities consumed by him.

REFERENCES

ANSCOMBE, F. J. and R. J. AUMANN (1963), *"A definition of subjective probability"*, Annals of mathematical Statistics, 34, pp. 199–205.

ARROW, K. J. (1982), *"Risk perception in psychology and economics"*, Economic Inquiry, 20, pp. 1–9.

FRIEDMAN, M., and L. J. SAVAGE (1948), *"The utility analysis of choices involving risk"*, Journal of Political Economy, 56, pp. 279–304.

HAUSNER, M. (1954), *"Multidimensional utilities"*, in Thrall et al. (eds.), Decision Processes. Wiley: New York, N.Y.; pp. 167–180.

HICKS, J. R. (1939), *Value and Capital*, Oxford University Press: Oxford, England.

MAY, K. O. (1954), *"Intransitivity, utility, and the aggregation of preference patterns"*, Econometrica, 22, pp. 1–13.

SAVAGE, L. J. (1954), *The Foundations of Statistics*, Wiley: New York, N.Y.

SCHOEMAKER, P. (1982), *"The expected utility model: its variants, purposes, evidence, and limitations"*, Journal of Economic Literature, 20, pp. 529–563.

TVERSKY, A. and D. KAHNEMAN (1981), *"The framing of decisions and the psychology of choice"*, Science, 211, pp. 453–458.

VON NEUMANN, J., and O. MORGENSTERN (1947), *Theory of Games and Economic Behavior*, Princeton University Press: Princeton, N.J.

CONTRIBUTED PAPERS

1. Proof Theory and Categorial Logic

G. BELLIN, Natural deduction for intuitionistic linear logic
M. BENEVIDES, Axioms and assumptions
V. DE PAIVA, M. HYLAND, Full intutionistic linear logic
J. DEGEN, Takeuti's conjecture for certain infinitary type theories
Y. GAUTHIER, Infinite descent in a fermatian set theory
E. HAEUSLER, L.C. PEREIRA, A denotational semantics for arbitrary level typed λ-calculus
L. HALLNÄS, Induction and bar induction
T. HOSOI, I. MASUDA, A classification of the intermediate logics on the third slice
H. JERVELL, Interpolation and completeness in linear logic
SHIH-CHAO LIU, Constructive proof in number theory N
E. MONTEIRO, Linear logic as CSP
E. PALMGREN, Type-theoretic interpretation of inductive definitions
C. MASSI, L.C. PEREIRA, Strong normalization via the worst reduction sequence
H. PFEIFFER, A notation system for ordinals using α-Mahlo-β-inaccessible ordinals
A. PITTS, On an interpretation of second order quantification in first order intuitionistic propositional logic
A. PRELLER, M. FOUDA, M. RIBEAUD-DUPASQUIER, Categories with frame and intuitionistic type theory
W. RUITENBURG, A new constructive logic and category theory
G. SAMBIN, An intuitive interpretation of intuitionistic logic without contraction
K. SASAKI, The simple substitution property of the intermediate propositional logics
J. SMITH, An interpretation of Kleene's slash in type theory
R. SOMMER, Ordinals in bounded arithmetic

2. Model Theory, Set Theory and Formal Systems

M. AMER, Cylindric algebras of sentences
S. ARTEMOV, On propositional quantifiers in provability logic
A.S. DENISOV, The structural property of the nuclear theory
C. GONZÁLEZ, A note on Zermelo set theory
P. HÁJEK, Interpretability and fragments of arithmetic
I. JOKISZ, Many-sorted theories: A characterization of interpretations between them
M. KRYNICKI, Quantifiers determined by classes of binary relations
D. KUREPA, The scope of the fixpoint approach
W. LENSKI, Order indiscernibles for formulas and types in classes of ordered abelian groups
D. MILLER, Infringements of classical laws in general metamathematics

M. MOSTOWSKI, Divisibility as formalized concept
SIU-AH NG, Locally pure measures in forking theory
E. NOGINA, Some properties of Gödel-Löb lattice
J. OIKKONEN, Chain models and infinitely deep logics
D.E. PAL'CHUNOV, Axiomatic complexity of theories
P. PAZDYKA, Universality of the language with a single binary relation symbol
N. PRATI, A partial model of NF with E
J. SÁNCHEZ-POZOS, Boolean formulae in relevant logic and Dedekind's problem
G. SANDU, A separation theorem for the logic of informational independence
Z. SOKOLOVIC (with A. PILLAY) On semisimple differential algebraic groups
S. TITANI, Completeness of global intuitionistic set theory
H. TUURI, A structure-nonstructure dichotomy for first-order theories
P. VELOSO, A. HAEBERER, A new algebra of first-order logic: binary relations resumed

3. Recursion Theory and Constructivism

K. AMBOS-SPIES, A. NIES, R. SHORE, The theory of the recursively enumerable weak truth-table degrees is undecidable
W. CALHOUN, Incomparable prime ideals of recursively enumerable degrees
A. DËGTEV, On some reducibilities of numerations,
M. DOMBROVSKY, On an approach to subrecursiveness
P. FEJER, K. AMBOS-SPIES, DING DECHENG, Embedding distributive lattices preserving 1 below a nonzero recursively enumerable turing degree
D. HARTZ, Many-one degrees of the stages of inductive definitions
K. KUCERA, Randomness and turing degrees
M. KUMMER, F. STEPHAN, Weakly semirecursive sets and r.e. orderings
V.L. SELIVANOV, On definable index sets
R. SHORE, K. AMBOS-SPIES, 1-types and undecidability in the r.e. degrees
R. SOARE, Definable properties and automorphisms of recursively enumerable sets
YU. G. VENTSOV, A problem of effective choice of constructivizations
S. WAINER, Ordinal complexity of recursive definitions

4. Logic and Computer Science

A. CHAUVIN, Untyped λ-calculus extensions by two unclassical logics: Markov-Myhill typed logic, partial logic
R. COWEN, Run-time bounds for Iwama's IS algorithm
M. D'AGOSTINO, M. MONDADORI, Classical analytic deduction and complexity
R. DE QUEIROZ, D. GABBAY, The functional interpretation of the existential quantifier
H. DE SWART, W. OPHELDERS, Tableaux, resolution and complexity of formulas
J. DIX, Normal logic programs and cumulativity
W. FARMER, J. GUTTMAN, A simple type theory with partial functions and subtypes
D. GHEORGHIU, Substitution rule and the decision problem in propositional logic
A. GOMOLINSKA, A non-standard approach to autoepistemic logic
A. HAEBERER, P. VELOSO, Program derivation calculi cannot preserve termination
E. HAEUSLER, A general and mechanical proof procedure for natural decuction systems

A. HUERTAS, M. MANZANO, Many-sorted logic as a unifying framework
XUEFENG JIANG, Resolution method based on Lin's entailment logic
M. KANOVICH, The intuitionistic logic as a logic of computational problems
E. KOUNALIS, On the equational bases of algebras
W. LABUSCHAGNE, J. HEIDEMA, The unit of information in deductive databases
S. LINDSTRÖM, A semantic approach to nonmonotonic reasoning: inference operations and choice
W. MACCAULL, Tableau method for residuated logic
G. MARCO, Non-monotonic systems of information
G. MASCARI, A. VINCENZI, Model-theoretic specifications and back-and-forth equivalences
I. MAUNG, Possibilistic inference processes
M. MIRCHEVA, Default reasoning—a way of using contradictions and ignorance
Y. NAKAYAMA, A treatment of counterfactuals within a framework for default reasoning
Z. STACHNIAK, The resolution rule and the meaning of verifiers
V.N. STEBLETSOVA, Branching time temporal logic as a tool for reasoning about concurrent programs
F. WIDEBÄCK, Assuredly hard tautologies
XIAN MING, Some results of handling metalogical problem of Lin's entailment system Cm by computer
J. YAMAGUCHI, Phases, informations and the permitting relation—A Boolean-valued representation of knowledge in AI
G. ZAVERUCHA, A goal directed theorem prover for a Modal Defeasible Relevant logic

5. Philosophical Logic

T. AHO, On the interpretations of attitude logics
M. ASTROH, On the logical form of epistemic predicates
G. BOOLOS, The advantages of honest toil over theft (Russell vs. Frege)
B. BORICIC, A syntactic concept of fuzzy logic
S. BRKIC, Is epistemology the basis for logic?
M. BROWN, A logic of de se knowledge and belief
A. BUCHSBAUM, T. PEQUENO, Ten paraconsistent and/or paracomplete calculi with recursive semantics
J. BUTTERFIELD, Probabilities and conditionals: distinctions by example
B. CHELLAS, Some problems in the logic of agency
J. CHENG, K. USHIJIMA, C—an entailment logic of conditionals
G. CORSI, A proof of Bull's theorem using the method of diagrams
M. DALLA CHIARA, G. TORALDO DI FRANCIA, Identity questions from quantum theory
M. DUMITRU, Intensional presuppositions of the realism-antirealism debate
M. FARTOS MARTÍNEZ, Logical consideration of metalogic
J. FRIEDMAN, On an adequate definition of distribution for first-order logic
S. GHILEZAN, Lambda calculi with intersection types
A. GIUCULESCU, The construction of a logical theory of action
GONG QIRONG, Lin's entailment logic can logically represent all knowledge
V. GORANKO, Applications of quasi-structural rules to axiomatizations in modal logic
C. GRECU, The logic of perspectival time

A. GUPTA, A calculus for definitions
T. HAILPERIN, Herbrand semantics, the potential infinite, and ontology-free logic
I. HALONEN, Critical comments on paraconsistent logics
S.O. HANSSON, Belief base dynamics
J. HAWTHORNE, Defeasible entailments and Popper functions: an essential connection
J. HEIDEMA, F. WIID, Quantum logics as semilattice retracts of heyting and boolean algebras
G. HOLMSTRÖM-HINTIKKA, Towards a semantics of action
T. INOUÉ, On some topological properties of some classes of rejected formulas and satisfiable formulas
D. JACQUETTE, The validity paradox in modal S5
A. JOHANSON, Imperative logic as based on a galois connection
A.S. KARPENKO, Classification of implicational logics
A. KAZMI, Indiscernibility
D. KHLENTZOS, What in the world could correspond to thruth?—Deflationism, correspondence, realism: prospects for naturalism
P. KOLÁR, Logical analysis of action sentences
P. KOT'ÁTKO, The nature of predication: Strawsonian and an alternative point of view
K.-H. KRAMPITZ, Predication theory and the concept of existence
A. KRON, A decidable first-order relevance logic
S. LAPIERRE, Montague-Gallin's intensional logic, structured meanings and Scott's domains
F. LEPAGE, Strong isomorphisms between some sets of partial functions and the set of standard functions in type theory
LIN BANGJIN, The notional calculus Cn of Lin's entailment logic
H. LINNEWEBER-LAMMERSKITTEN, Are principles of simplification necessary?
L. LISMONT, PH. MONGIN, Knowledge structures: completeness properties and applications to game theory
LIU BAOCHANG, Liu's universal comparison
SHANG-CHIANG LU, An odd thesis on physical existence, its deduction and the analysis of related semantic problems
L. MAKSIMOVA, On the Beth definability properties in modal logics
E. MARES, Classically complete modal relevant logics
D. MIEVILLE, Some aspects of a free, universal and developmental logic
T. MIHÁLYDEÁK, Semantic partiality and consequence relations
K. MISIUNA, An argument for modern nominalism
PH. MONGIN, The logic of belief change and nonadditive probability
G. ODDIE, Universalizability and supervenience: by degrees
M. OMYLA, The theories of possible worlds in the language of formal logic
E. ORLOWSKA, Semantics of relevant logics based on relation algebras
T. PARSONS, Truth and meaning in fregean semantics
J. PASNICZEK, Meinongian logic, generalised quantifiers, and classical logic
T. PEQUENO, A. BUCHSBAUM, The notion of epistemic inconsistency and its logic
W. RABINOWICZ, S. LINDSTRÖM, Epistemic entrenchment with incomparabilities and relational belief revision
H. ROTT, Preferential belief change using generalized epistemic entrenchment
V. RYBAKOV, Poly-modal logic as metatheory of pure predicate calculus
P. RÖPER, The topology of continua - without points
L. SAVION, Semantics for belief attributions

U. SCHEFFLER, How to talk about events
G. SCHURZ, Relevant deduction
J. SEREMBUS, Generalization and comparative probabilistic semantics
M. SEYMOUR, On rehabilitating redundant truth
YA. SHRAMKO, New semantics for minimal and intuitionistic logics
T. SKURA, Pure refutation formulations of propositional logics
D.P. SKVORTSOV, V.B. SHEHTMAN, Maximal kripke-type semantics for non-classical predicate logics
V.A. SMIRNOV, Multidimensional logics
W. STELZNER, Semantic relevance and conceptivism
K. SWIRYDOWICZ, A system of dyadic deontic logic
E. TATIEVSKAYA, Russell on word meaning
S. URAL, Temporal interpretation of the connectives
M. URCHS, Paraconsistency: "stric" is too strong
D. VAKARELOV, Modal logics for reasoning about arrows: arrow logics
A. VASILCHENKO, Hierarchical account in the logical theory of action
R. VERGAUWEN, Entia non sunt deminuenda praeter necessitatem
A. VOIZARD, What can, and what should truth be?
B. WALLISER, Updating of beliefs with non additive or multilevel probabilities
WANG SHIAN, ZHANG JUSHENG, The unsteady logic
M. WEGENER, St Anselm's proof of God: discussion and formal reconstruction
P. WEINGARTNER, Types of redundant and irrelevant components in logical deduction
T. WILLIAMSON, Some admissible rules in modal systems
A. WISNIEWSKI, Multiple-conclusion arguments and erotetic arguments
XIANG RONGXIAN, Logical law is the law of the objective world, not one of thinking
ZHANG DEWENG, On logical object
ZHOU GANHUA, Intensional analysis on n-ary relation by Lin's entailment logic
ZHOU XUNWEI, Why is Chinese modal logic different from its Western counterpart

6. Methodology

A. AKKERMAN, Are demographic predicates attributes of cohorts? Some issues in the methodology of age cohorts
A. AMBROGI, Defences of realism
M. ARTIGAS, Reliability and fallibilism
W. BALZER, M. BURGIN, V. KUZNETSOV, Reduction and the structure-nominative view of theories
B. BERCIC, The bricks of the universe
G. BOTTERILL, Effects of a common cause
H. BROWN, Observational evidence
CHEN KUN, New thinking mode in new age
T. CLEVELAND, Empiricism without observation sentences
D. COSTANTINI, U. GARIBALDI, A probabilistic foundation of statistical mechanics
J. DERDEN,Why good scientific theories cannot be falsified and cannot explain particular facts
H. DE REGT, The heuristic role of philosophy in the development of science
M. EDMAN, Innate knowledge and scientific rationality
A. ESTANY, The role of methodology in models of scientific change

GUANHUA ZHOW, On the relations between the experiencial and rational functions of scientific method

7. Probability, Induction and Decision Theory

M. BACHARACH, Description and coordination
T. BURNS, E. GRIFFOR, A mathematical theory of human interaction and games: an extension of classical game theory
A. FUHRMANN, Nonmonotonic reasoning and cautious operations on premisses
I. GILBOA, Philosophical applications of Kolmogorov's complexity measure
B. HAGLUND, Inductive inference strategies
T. HAVRÁNEK, On connection of subjective probability and uncertainty processing in knowledge systems
JU SHIER, XIA YAN, The analyses of non-pascalian probability and its application
R. KOONS, Liar-like paradoxes and reputation effects in game theory
A. KRON, D. PAVLICIC, Principles and preferences
P-E. MALMNÄS, Towards a mechanization of real-life decisions
P. MENZIES, What are conditional propensities?
G. ODDIE, P. MENZIES, An objectivist's guide to subjective value
P. RAWLING, Choice and conditional expected utility
C.L. SHENG, An explanation of the Ellsberg paradox
F.P. SIMION, G. SIMION, Probability with applications to technical diagnosis
J.H. SOBEL, Backward induction arguments in infinitely repeated prisoner's dilemmas
P. WEIRICH, The hypothesis of Nash equilibrium and its bayesian justification
H. WESSEL, Eine Lösung der Rabenparadoxie

8. History of Logic, Methodology and Philosophy of Science

A. ADAM, Einstein's philosophy of science as a revolution in the philosophy of science
M. ALY, Philosophy of science in the Arab world: a critical study of its main issues
J. BERG, The ontological foundations of Bolzano's philosophy of mathematics
A. BOTEZ, Edinburgh school and epistemological relativism
H. BURKHARDT, Monadology and mereological structure
N. CARTWRIGHT, J. CAT, Otto Neurath: materialism, socialism, and the demise of scientific method
J. COVER, Non-basic time and reductive strategies: Leibnizian foundations
R. CREATH, Carnap's engineering conception of philosophy
J. DAWSON, The compactness of first-order logic: from Gödel to Lindström
A. DRAGO, The prosecution of Leibniz' program in Lazare Carnot's works
D. FELIPE, Burden of proof in the post-medieval *ars disputandi*
H. FESTINI, Verificationism: neoverificationism and verificational games
P. GAIDENKO, Paradoxe des Unendlichen in Galilei's Mechanik
J. GASSER, Frege and Lukasiewicz on reasoning from false premises
K. GOODMAN, Anticipations of truth: historical evidence for a realist account of scientific progress
B. GOWER, Realism and empiricism in Schlick's philosophy of science
L. HAAPARANTA, The role of judgements in Frege's and Peirce's logic

9. Ethics of Science and Technology

10. Foundations of Logic, Mathematics and Computer Science

D. ISAACSON, On a criticism by Wittgenstein of the Frege-Russell definition of natural number
K. MANDERS, Hilbert's method of ideal elements
C. MCLARTY, Why category theory is not a foundation for mathematics
P. MÄENPÄÄ, Extending natural deduction into a general method of analysis
A. OLIVER, The metaphysics of singletons
J. PEREGRIN, General notion of model theory
M. PICARD, Impredicativity and monism
M. RESNIK, Aristotelian and Platonic structuralism
P. SCHROEDER-HEISTER, An asymmetry between introduction and elimination inferences
S. SHANKER, The resurgence of (cognitive) psychologism
P. SHIMAN, Arithmetizing the transfinite
S. STENLUND, The limits of formalization
G. SUNDHOLM, Ontologic vs. epistemologic
WUIJA ZHU, XIAN XIAO, The foundation of logic and set theory for uncertain mathematics

11. Foundations of Physical Sciences

M. AKEROYD, Laudan's model criticised
T. ANGELIDIS, On the problem of a local extension of the quantum formalism
D. BOZIN, Non-standard scales for fundamental measurement
M. BRODOWSKI, Structure of mathematicized empirical theories
J. BUB, Solving the measurement problem of quantum mechanics
F. BUNCHAFT, T. LOBAO, An axiomatic reconstruction of the 'experimental propositions' of Mach
M. BURGIN, V. KUZNETSOV, Laws and forms of their representation in physical theories
J. BUTTERFIELD, Causation and counterfactuals in the Bell experiment
D. CARTIANU, The history of the idea of synchronization from Huygens until today
G. CASTAGNOLI, A. VINCENZI, Quantum simultaneous machines and computability
R. CLIFTON, Reviving the counterfactual approach to quantum nonlocality
R. COLEMAN, H. KORTÉ, A refutation of Reichenbach's thesis that geometry is empirically underdetermined
R. COLEMAN, H. KORTÉ, Conformal causality and the universality of free fall entail the existence of a unique Weyl structure on spacetime
A. CORDERO, Practical reasoning in the foundations of quantum theory
J. CUSHING, Quantum mechanics, historical contingency and the Copenhagen interpretation
P. DIAS, A path from Watt's engine to the principle of heat transfer
D. DIEKS, A realistic interpretation of quantum mechanics in terms of contextual properties
I. DOBRONRAROVA, Non-linear style of reasoning
J. FAYE, Non-separability or non-locality?
H. GROENEWOLD, Field or print
M. HOGARTH, Predictable space-times
HONG DINGGUO, Philosophical analysis of physical conceptions

C. HOOKER, Intelligibility, objectivity and completeness in physics: The divergent ideals of Bohr and Einstein
W. JONES, Concerning the question: "Is the quantum theory complete?"
F. KRONZ, The projection postulate and the time-energy uncertainty relation
J. LEROUX, Helmholtz's scientific realism: not so clear a case
A. MAIDENS, Active and passive transformations in general relativity
K. MAINZER, Symmetries in the physical sciences
S. MARTINEZ, From mechanisms to non-separability
B. MUKHERJEE, Two problems of quantum field theory
J. MURGAS, The role of invariancy in physics
F.G. NAGASAKA, Explanation in physical science
G. NERLICH, Holes in the hole argument
R. NUGAYEV, Will quantum field theory be reconciled with general relativity?
C. PAGONIS, M. REDHEAD, R. CLIFTON, Nonlocality and the classical limit of quantum mechanics
I. PÂRVU, The completeness of quantum mechanics and the philosophical theory of possibility
A. POLIKAROV, A. KUNAEVA, Stable structures in the dynamics of physical theories
H. PRICE, Agency and causal asymmetry
J. RAMSEY, Ideal reaction types and the reactions of real alloys
M. RÉDEI, Proof and significance of the fact that quantum field theory is a stochastic Einstein local theory
V. RODRIGUEZ, On physical constants
L. ROPOYI, On the nature of thermodynamics
K. RUTHENBERG, Is there a philosophy of chemistry?
D. SAPIRE, General causal propensities and quantum probabilities
P. SIBELIUS, Discrete dynamics and first order logic
S. SIT'KO, I. DOBRONRAROVA, Physics of the alive
A.V. SOLDATOV, Structure of the scientific picture of the world and the main determinant of its development
A. STODDART, Possible worlds and indistinguishable particles
L. SZABÓ, Spacetime structure on quantum lattice
P. SZEGEDI, Correspondence or incommensurability?
GUOZHENG WANG, Philosophical and methodological significance of black hole physics
E. ZIELONACKA-LIS, The cognitive status of the reconstruction of mechanisms in modern organic chemistry
C. ÅBERG, On the conceptual foundations of quantum field theory

12. Foundations of Biological Sciences

W. BECHTEL, Levels of organization and research instruments in biology
M. BRADIE, Expanding the circle: some moral implications of darwinian theory
P. DUMOUCHEL, Natural selection and selection type theories
S. FISHER, Mayr's formulation of an ultimate/proximate distinction
E. LLOYD, An anatomy of the units of selection debates
A. MARCOS, A measure of functional information relevant to biology
J. MOSTERIN, Mereology, set theory and biological ontology
L. NISSEN, Teleology and natural selection

13. Foundations of Cognitive Science and AI

14. Foundations of Linguistics

J. PASEK, On formalization of maxims of conversation
F.J. PELLETIER, On an argument against semantic compositionality
S. PETERS, What is a translation?
A. RANTA, The structure of a formal grammar
V. RIIKONEN, Propositional models or images?
P. SEGERDAHL, The theoretical approach to language
G. SHER, Towards a general definition of partially-ordered (branching) generalized quantifiers
A. TER MEULEN, The quantificational force of static and dynamic predication
R. TURNER, Properties and types
S. ÖHMAN, Metaphysics in contemporary linguistics

15. Foundations of Social Sciences

L. ADDIS, Human action and the Humean universe
A. ANDRADE-CARRENO, Rationality and multidimensional theories
W. BALZER, J. SANDER, M. KUOKKANEN, Theory and data in social science: an example
C. BICCHIERI, Learning in games
M. BLEGVAD, Karl Popper's philosophy of social science
V. BLUM, The nature of anticultural attitudes
T. BOYLAN, Kaldor's methodology of economics: a constructive empiricist interpretation
M. COLBERG, M. NESTER, S. REILLY, L. NORTHROP, B. RHEINSTEIN, The use of illogical biases in the measurement of reasoning skills
F. COLLIN, Constructivism and social science
A. CUDD, Modeling rationality
E. ERWIN, Recent philosophical work on the foundations of psychoanalysis
F. FORGES, Three legitimate definitions of correlated equilibrium in games with incomplete information
D. GINEV, Ontological assumptions of interpretive social science
W. GONZÁLEZ, Prediction in economics. On theoretical basis of its methodological development
M. JANSSEN, Methodological individualism and the role of aggregation
E. KALAI, E. LEHRER, Bayesian learning and Nash equilibrium
M. KEBEDE, Science or ethics of development?
V. KEMEROV, Dynamics of methodology
E. KRAUSZ, The use of aggregated data in sociological analysis
L. KULTTI, Holistic concepts as approximations in social sciences
M. LAGUEUX, Rationality and mechanisms in economics
B. LAHOTI, Methodological problems pertaining to analysis and understanding of traditional societies (with special reference to Indian society)
H. LARSSON, Game theory and rationality. A study in modelling
H. LAUER, Causal facts as logical entailments of action statements
H. LIND, Does there exist a Lakatiosian hard core in "neo-classical" economics?
LIU HUA, Economic philosophy is faced with the basic challenge
S.W. MAN, Epistemological questions in the incremental-political approach to decision-making in public administration

Name index