KNOWLEDGE AND REALITY

PHILOSOPHICAL STUDIES SERIES

VOLUME 103

Founded by Wilfrid S. Sellars and Keith Lehrer

The titles published in this series are listed at the end of this volume.

KNOWLEDGE AND REALITY

Essays in Honor of Alvin Plantinga

Edited by

THOMAS M. CRISP
Biola University, La Mirada, CA, U.S.A.

MATTHEW DAVIDSON
California State University, San Bernardino, CA, U.S.A.

and

DAVID VANDER LAAN
Westmont College, Santa Barbara, CA, U.S.A.

Springer

A C.I.P. Catalogue record for this book is available from the Library of Congress.

ISBN-10 1-4020-4732-0 (HB)
ISBN-13 978-1-4020-4732-9 (HB)
ISBN-10 1-4020-4733-9 (e-book)
ISBN-13 978-1-4020-4733-6 (e-book)

Published by Springer,
P.O. Box 17, 3300 AA Dordrecht, The Netherlands.

www.springer.com

Cover art: After the Ascent, Acrylic Painting for Plantinga, Keith Lehrer, 2004

Printed on acid-free paper

Printed in the Netherlands.

For Al

Contents

Contributing Authors

William P. Alston
Syracuse University

Michael Bergmann
Purdue University

Richard Fumerton
University of Iowa

Jenann Ismael
University of Arizona

Jonathan Kvanvig
University of Missouri

Trenton Merricks
University of Virginia

Richard Otte
University of California, Santa Cruz

John L. Pollock
University of Arizona

Michael C. Rea
University of Notre Dame

Eleonore Stump
St. Louis University

James E. Tomberlin
California State University, Northridge

Peter van Inwagen
University of Notre Dame

Nicholas Wolterstorff
Yale University

Keith E. Yandell
University of Wisconsin, Madison

Preface

This volume was conceived as a *Festschrift* to surprise Alvin Plantinga on his 70th birthday. That original plan was not entirely successful. For one thing, the day came and went well before the work was complete. For another, the project wasn't quite a surprise: Plantinga caught wind of it (though not of its details) before the unveiling.

The occasion was marked, however, by a presentation of a projected table of contents and an early draft of the cover painting by Keith Lehrer. Plantinga then saw the details, and was quite taken aback that the editors and a few contributors had come to South Bend from as far Virginia, Florida, and California to celebrate.

Now we are pleased to offer the essays themselves. The collection ranges widely over metaphysics and epistemology. Its wingspan testifies to the breadth both of Plantinga's own work and of the audience that has valued it. Some of the essays deal with ontology, examining actualism, presentism, antirealism, properties, and artifacts. Several essays in epistemology raise skeptical questions, work through the implications of naturalism or internalism, and engage Plantinga's own Reidian account of warrant. Other contributions consider the bearing of philosophical ideas on the Christian faith–of split brain cases on the doctrine of the Trinity, for example, and of materialism on the afterlife.

The contributors are friends, colleagues, and former students of Plantinga. The editors thank all of them for their eager participation in this project.

There is little we can add to the expressions of praise and gratitude the contributors and others have voiced. Plantinga's seminal work on modality, the problem of evil, and the rationality of religious belief has long been

celebrated for its rigor, depth, and clarity, and his more recent work in epistemology has been eagerly received as an important, stimulating contribution to the field.

Beyond that, many have had occasion to thank Plantinga for his availability and encouragement, among both his own students and others. As James Sennett has recently observed, Plantinga follows through on his expressed concern for the development of younger philosophers, and in particular members of the Christian philosophical community. The editors add their testimony that Plantinga has been consistently generous with his time and attention.

But perhaps one facet of Plantinga's work deserves greater appreciation: the scope of his vision, the big picture which animates his thought. Plantinga excels not only in analysis but also in synthesis. He aims always to place his ideas in their context, and he has encouraged his students to do the same. His work thus contributes not merely to the development of a handful of philosophical problems, but to a worldview in which all of knowledge, freedom, possibility, the character of propositions, the origins of humanity, and the nature and purposes of God twine each other and cannot be understood in isolation.

Here Plantinga owes something to his own teacher, William Harry Jellema, who conceived of philosophy's history as an arena of competing commitments which are ultimately religious. Plantinga is quick to credit Jellema as a model of historically informed, insightful, and subtle thinking that appreciates criticism without being overawed by intellectual fads–nor even by entrenched errors. Jellema is himself indebted to Abraham Kuyper's conviction that every human enterprise (not least philosophy) is claimed by the sovereignty of Christ, Kuyper in turn to John Calvin's understanding of the original goodness of creation and humans' tasks in it, and Calvin to Augustine's longing for rest in God, the human end.

At Jellema's hands Plantinga saw the vision. We hope to honor them both by applying the proceeds of sales of this book toward a chair in Jellema's name at Calvin College. We are glad to have caught the vision from you, Al.

Thomas M. Crisp
Matthew Davidson
David Vander Laan

Acknowledgments

We would like to thank Keith Lehrer for commissioning the volume, as well as for the painting for the cover; Ann Hickman for her patience and hard work in typesetting the pages; and our terrific (and understanding) editors—Ingrid van Laarhoven and Floor Oosting. We are grateful to each of the contributors for the essays, and to Hilda Tomberlin and Greg Fitch for help in preparing James Tomberlin's essay.

Chapter 1

ACTUALISM AND PRESENTISM*

James E. Tomberlin
California State University, Northridge

In the metaphysics of time and tense, presentism is the view that there are no objects that do not presently exist.[1] According to the presentist, there are no philosophical problems whose solution calls for or requires an ontological commitment to non-presently existing individuals. In the metaphysics of modality, actualism is minimally the view that there are no objects that do not actually exist.[2] By the lights of actualism, there are no philosophical problems whose proper treatment demands an ontological commitment to nonactual objects. Now I harbor a deep skepticism as regards both of these ontological stances. In what follows, accordingly, I aim to extend and sharpen the skeptical concerns previously voiced in my 1993, 1996a, 1996b, forthcoming a, and Tomberlin and McGuinness 1994.

1. A DEONTIC CASE[3]

Jones, as it happens, has taken up nouvelle cuisine with its laudable emphasis on fresh and unusual ingredients. One weekend, in seclusion, he opts to prepare for himself the remarkable ragoût of wild mushrooms with veal stock and red wine concocted by Alice Waters for her renowned restaurant, *Chez Panisse*. For the preparation, Jones decides, why not utilize wild mushrooms he gathered from the nearby woods just yesterday? A splendid dish indeed, he observes upon dining. But alas, some time later, Jones, still home alone and miles from the nearest person, is rendered comatose. Several of the wild mushrooms were highly toxic and Jones, alone and physically incapable of conveying his plight, faces certain death.

T. M. Crisp, M. Davidson and D. Vander Laan (eds.), Knowledge and Reality, 1-14.

In this situation, I take it, (1) and (2) are true but (3) is false:

(1) For any individual *x*, if *x* is a moral agent and *x* is available and able to come to Jones's assistance, *x* ought prima facie to provide Jones with aid.
(2) No actual moral agent is available and able to come to Jones's assistance.
(3) For any individual *x*, if *x* is a moral agent who is available and able to come to Jones's assistance, *x* ought not prima facie to provide Jones with aid.

If so, however, the actualist cannot read (1) and (3) as universally quantified material conditionals. For suppose otherwise. Since no actual individual satisfies the open sentence

x is a moral agent who is available and able to come to Jones's assistance,

every actual individual satisfies both (1a) and (3a):

(1a) (*x* is a moral agent who is available and able to come to Jones's assistance) ⊃ (*x* ought prima facie to provide Jones with aid).

(3a) (*x* is a moral agent who is available and able to come to Jones's assistance) ⊃ (*x* ought not prima facie to provide Jones with aid).

But then (1) and (3) are both true, after all.

These considerations lead directly to a serious challenge for actualism:

Challenge One. With objectual quantification,[4] provide an interpretation of (1) and (3) satisfying these conditions: (1) and (2) are true, (3) is false, and the quantifiers range over actual individuals only.

The above deontic case likewise poses a serious threat to presentism. For in the scenario around Jones it is also the case that whereas (1_1) and (2_1) are true, (3_1) is false:

(1_1) For any individual *x*, if *x* is a moral agent and *x* is available and able to come to Jones's assistance, *x* ought prima facie to provide Jones with aid.

(2_1) No presently existing moral agent is available and able to assist Jones.

(3_1) For any individual x, if x is a moral agent who is available and able to come to Jones's assistance, x ought not prima facie to aid Jones.

If so, the presentist must not read (1_1) and (3_1) as universally quantified material conditionals. For suppose otherwise. Because no presently existing individual satisfies the open sentence 'x is a moral agent who is available and able to assist Jones', every presently existing individual satisfies both (1a) and (3a):

(1a) (x is a moral agent who is available and able to assist Jones) $\supset$ (x ought prima facie to provide Jones with aid).

(3a) (x is a moral agent who is available and able to assist Jones) $\supset$ (x ought not prima facie to provide Jones with aid).

But then (1_1) and (3_1) are both true, after all.

As with actualism before, these considerations yield the following challenge for presentism:

Challenge One. With objectual quantification, provide an interpretation of (1_1) and (3_1) satisfying these conditions: (1_1) and (2_1) are true, (3_1) is false, and the quantifiers range over presently existing individuals only.

2. HOW NOT TO MEET THE CHALLENGE

The above trouble with treating (1_1) and (3_1) as universally quantified material conditionals naturally suggests that the presentist construe the notion of conditionality at work in (1_1) and (3_1) in such a way that a conditional of the sort in question does not come out true just because its antecedent is (merely) in fact not satisfied. This in turn suggests that the presentist entertain one of the following proposals.

Strict Conditionals. By this alternative, (1_1) and (3_1) are to be construed as universally quantified strict conditionals, where a strict conditional $\Box(A \supset B)$ is true (at a world w) if and only if B is true in every logically and/or metaphysically possible world (relative to w) where A is true. So understood, however, this proposal scarcely meets our challenge: since it is *logically* possible that some presently existing moral agent who is available and able to assist Jones does not have a prima facie obligation to aid Jones, (1_1) turns out false under this interpretation.

Nomic Conditionals. According to this view, (1_1) and (3_1) are to be taken as universally quantified conditionals of nomic necessity, where a conditional of nomic (= physical) necessity $\Box_P(A \supset B)$ is true (at a world w)

exactly on the condition that *B* is true in every physically possible world (relative to *w*) in which *A* is true. While more modest than the previous interpretation owing to a switch from logical to physical necessity, it should nevertheless be clear that the present alternative fails: insofar as no (actual) law of nature or statement of nomic necessity is violated under the assumption that some presently existing moral agent is not prima facie obligated to aid Jones even though he or she is available and able to provide assistance, true (1_1) won't be true, after all.

Soft (or Hedged) Laws. Under the present view, (1_1) and (3_1) are deemed seriously incomplete owing to the fact that each implicitly contains a *ceteris paribus* clause. Bringing this clause into the open, (1_1) becomes the allegedly true *soft law*: For any individual *x*, if *x* is a moral agent who is available and able to assist Jones, *all things being equal*, *x* ought prima facie to provide Jones with aid. For soft laws, we are told, the consequent holds in any physically possible situation in which the antecedent and the *ceteris paribus* condition are jointly satisfied.[5] If so, however, this interpretation likewise fails the challenge: because no law is violated under the condition that some presently existing moral agent is not prima facie obligated to aid Jones, even though this agent is available and able to provide assistance and all other things are equal, (1_1) comes out false under this interpretation. No, statements like (1_1) concerning moral obligation just do not express nomological laws, soft or otherwise.

Counterfactuals. With the current alternative, (1_1) and (3_1) become the universally quantified counterfactuals (1b) and (3b), respectively:

(1b) For any individual *x*, if it were the case that *x* is a moral agent who is available and able to come to Jones's assistance, it would be the case that *x* ought prima facie to provide Jones with aid.

(3b) For any individual *x*, if it were the case that *x* is a moral agent who is available and able to come to Jones's assistance, it would be the case that *x* ought not prima facie to provide Jones with aid.

A tempting view indeed for anyone who demands an account of the truth-conditions for our target sentences while insisting on an ontology devoid of non-presently existing individuals. Unfortunately, any such theoretical attraction notwithstanding, this counterfactual interpretation is fraught with difficulties, including each of the following prominent ones:

First, as we learned from Stalnaker (1968) and Lewis (1973), transitivity and contraposition both fail for counterfactuals. And yet, (I) and (II) seem harmlessly valid:

(I) For any individual *x*, if *x* is a moral agent who is available and able to come to Jones's assistance, *x* ought prima facie to provide Jones with aid. For any individual *x*, if *x* ought prima facie to provide Jones with aid, *x* will attempt to help Jones. Thus, for any individual *x*, if *x* is a moral agent who is available and able to come to Jones's assistance, *x* will attempt to help Jones.

(II) For any individual *x*, if *x* is a moral agent who is available and able to come to Jones's assistance, *x* ought prima facie to provide Jones with aid. Thus, for any individual *x*, if it is not so that *x* ought prima facie to provide Jones with aid, *x* is not a moral agent who is available and able to aid Jones.

Of course (I) and (II) would not be valid unless the embedded notion of conditionality in each case obeyed transitivity and contraposition.

Second, while there is room for genuine disagreement over the correct rule of truth for a counterfactual $A \mathrel{\square\!\!\rightarrow} B$, to facilitate matters I assume the one provided in Lewis 1973:

> $A \mathrel{\square\!\!\rightarrow} B$ is true (at world *w*) if and only if either (i) there are no possible *A*-worlds (in which case $A \mathrel{\square\!\!\rightarrow} B$ is vacuously true) or (ii) some *A*-world where *B* holds is closer (to *w*) than is any *A*-world where *B* does not hold.

Next, suppose with Kripke (1980) *genetic essentialism*—any presently existing individual necessarily has the origin it in fact has. Return now to Jones and permit me to expand on his background. Rebounding from a failed and childless marriage, Jones, vowing not to contribute to a world of overpopulation, underwent a successful and irreversible vasectomy three years ago. That is, we have the truth of (4_1):

(4_1) No presently existing individual is a biological offspring of Jones.

Now surely any reason for treating (1_1) and (3_1) as (1b) and (3b), respectively, should likewise dictate that (5_1) and (6_1) are to be parsed as (7_1) and (8_1), in turn:

(5_1) For any individual *x*, if *x* is a moral agent who is available and able to assist Jones and *x* is a biological offspring of Jones, *x* ought prima facie to aid Jones.

(6_1) For any individual *x*, if *x* is a moral agent who is available and able to assist Jones and *x* is a biological offspring of Jones, *x* ought not prima facie to aid Jones.

(7_1) For any individual x, if it were to be the case that x is a moral agent who is available and able to assist Jones and x is a biological offspring of Jones, it would be the case that x ought prima facie to aid Jones.

(8_1) For any individual x, if it were to be the case that x is a moral agent who is available and able to assist Jones and x is a biological offspring of Jones, it would be the case that x ought not prima facie to aid Jones.

With all of this, however, presentism comes to grief; owing to the Lewis rule of truth for counterfactuals, genetic essentialism, and the truth of (4_1), presentism demands that (7_1) and (8_1) are both true. But in the scenario involving Jones it seems clear that whereas (5_1) is true, (6_1) is false. By parity of reasoning, the proposal that (1_1) and (3_1) are to be construed as (1b) and (3b), in order, should be rejected.

Conditional Obligations. At this juncture the presentist directs our attention to fairly recent developments in deontic logic. After van Fraassen (1972), Lewis (1974), and others, a statement of *unconditional* obligation is represented as OA, where O is the familiar monadic deontic operator of standard deontic logic. OA is adjudged true (at a world w) if and only if A is true in all of the deontically ideal worlds (relative to w). In sharp contrast, a statement of conditional obligation is represented as $O(A/B)$, with $O(/)$ a newly introduced dyadic deontic operator. $O(A/B)$—the assertion that under conditions satisfying B it is obligatory that A is satisfied—obeys a different (and weaker) rule of truth: some value realized at some B-world where A holds is better than any value realized at any B-world where A does not hold (Lewis 1974: 4). To accompany this axiological interpretation of conditional obligation, Lewis supplies the following axioms and rules of inference (Lewis 1974: 11-12):

R1. All truth-functional tautologies are theorems.
R2. If A and $A \supset B$ are theorems, so is B.
R3. If $A \equiv B$ is a theorem, so is $O(A/C) \equiv O(B/C)$.
R4. If $B \equiv C$ is a theorem, so is $O(A/B) \equiv O(A/C)$.
A1. $P(A/C) \equiv \sim O(\sim A/C)$.
A2. $O(A \,\&\, B/C) \equiv [O(A/C) \,\&\, O(B/C)]$.
A3. $O(A/C) \supset \mathrm{P}(A/C)$.
A4. $O(T/C) \supset O(C/C)$.
A5. $O(T/C) \supset O(T/B \vee C)$.
A6. $[O(A/B) \,\&\, O(A/C)] \supset O(A/B \vee C)$.
A7. $[P(\bot/C) \,\&\, O(A/B \vee C)] \supset O(A/B)$.
A8. $[P(B/B \vee C) \,\&\, O(A/B \vee C)] \supset O(A/B)$,

where $P(/)$ reads 'it is permissible that...given that...,' T stands for tautology, and $\perp$ is the negation of any tautology. In the above logic of conditional obligation, this feature is salient for our purposes here: since $A \supset OB$ does *not* imply $O(B/A)$, the latter (unlike the former) is not automatically true when A is false.

Return to (1_1) and (3_1). The challenge confronting presentism is to provide an account of (1_1) and (3_1) meeting the constraint that (1_1) is true and (3_1) is false with quantification over just presently existing individuals. According to the present suggestion, (1_1) and (3_1) become statements of universally quantified conditional obligation. Thanks to the rule of truth for $O(A/B)$, it is urged, (1_1)—so construed—is indeed true whereas (3_1)—so interpreted—is surely false, and this remains so even when the quantification involved is presentistic.

Against this intriguing proposal, I offer these objections:

First, in the various systems of conditional obligation articulated by van Fraassen (1972), Lewis (1974), and more recently Åqvist (1987) and Feldman (1986), the detachment principle $O(A/C) \supset [C \supset OA]$ is not a theorem. And yet, I submit, (III) is plainly valid:

> (III) For any individual x, if x is a moral agent who is available and able to come to Jones's assistance, x ought prima facie to aid Jones. Scott is a moral agent who is available and able to assist Jones. Thus, Scott ought prima facie to aid Jones.

If so, however, the notion of conditionality embedded in (1_1) cannot be the one embodied in the above systems of conditional obligation.

Second, as I have argued at length elsewhere,[6] each of the systems of conditional obligation in question succumb to one or more versions of the notorious paradoxes of deontic logic. Without rehearsing the details of these paradoxes, the following observations are pertinent here.[7] To generate one of the paradoxes against a particular system of deontic logic, a possible situation is described and a set of natural language sentences is produced where the sentences in question all seem true if the possible situation were to occur. Next, it is documented that under the most judicious representations of the natural language sentences within the deontic system at stake the result is a logically inconsistent set. This is of course powerful evidence that such a deontic system fails to provide a theoretically viable account of the natural language target sentences. Now sentences just like (1_1) and (3_1) loom large in some of the deontic paradoxes, most notably the Paradox of the Knower and the Contrary-to-Duty-Imperative Paradox.[8] And consequently, if the above systems of conditional obligation fall prey to one or more of these paradoxes, this is ample reason to find that (1_1) and (3_1) are not to be

treated as statements of universally quantified conditional obligation.

Conclusion as regards the deontic case. With this negative verdict of six alternative presentistic interpretations of (1_1) and (3_1), I scarcely claim to have exhausted all of the positions in logical space facing the presentist. Still, I do think I have addressed the most promising ones. If so, until and unless some other interpretation is offered that suits presentism, it appears quite appropriate to theorize that non-presently existing individuals are to be invoked for a correct account of deontic sentences like (1_1) and (3_1). Turning to actualism, in my 1996a and forthcoming a, I provide a negative verdict against seven actualistic treatments of (1) and (3). Once again, then, without some other construal of these sentences that fits actualism, it seems proper to quantify over possible but non-actual objects for an account of items like (1) and (3). But let's not end here. For there awaits another but very different problematic case for both actualism and presentism.

3. INTENTIONAL VERBS

Like Chisholm (1986), assume an actualism that includes these key ingredients: an ontology confined to actual individuals and attributes (some exemplified, others not); a relational account of believing; and, a Russellian treatment of definite descriptions. Since a position of this sort requires that no person ever has genuine *de re* beliefs toward non-actual individuals, the question becomes acute as to the proper treatment of items such as

(9) Ponce de Leon searched for the fountain of youth.

After all, if Ponce de Leon may be said to have really searched for the fountain of youth, to have hoped to find it and the like, he presumably can be said to have entertained beliefs of or about the object of his search. To deny that Ponce de Leon had any *de re* beliefs toward the fountain of youth, therefore, dictates that one who adopts the aforementioned ontological position embrace one of these alternatives: (a) deny the truth of (9); or (b) interpret (9) so that its truth does not stand Ponce de Leon in a genuine *de re* relation to the (non-actual) fountain of youth.

Chisholm, quite correctly, rejects option (a). Instead, against the background of his theoretically elegant account of believing as a relation between a believer and an attribute,[9] Chisholm proposes that the intentional verb in (9) be taken as expressing a dyadic *searched for* relation holding between Ponce de Leon and an *attribute*. That is, Chisholm would have us parse (9) as follows (Chisholm 1986: 56-57):

(10) Ponce de Leon endeavored to find the attribute of being a unique site of a unique fountain of youth.

Because there *is* such an attribute even though it fails to be exemplified, (9)—so construed—does not require that Ponce de Leon bear a genuine *de re* relation to the nonexistent fountain of youth.

A similar actualist view is advanced independently by David Kaplan in his classic essay "How to Russell a Frege-Church" (1975). To begin with, Kaplan rightly observes that Russell's own primary-secondary scope distinction for eliminating descriptions within intensional contexts fails in the case of (9), owing to the fact that the intentional verb there takes no sentential complement. In accord with Chisholm, Kaplan suggests, why not model (9) semantically (and ontologically) as the bearing of a dyadic relation between Ponce de Leon and an attribute where the resulting paraphrase of (9) turns out much like (10) above (Kaplan 1975: 729).

To my mind, there are ample reasons for rejecting the Chisholm-Kaplan model for items such as (9). And I have so argued at length, where the critique is successively refined in Tomberlin 1988, 1994, 1996a, and forthcoming a. This negative verdict of the Chisholm-Kaplan account, if correct, prompts another serious test for actualism:

> *Challenge Two*. Tender a credible treatment of (9) meeting this constraint: (9) is true but its truth does not require Ponce de Leon to stand in a *de re* relation to some non-actual individual.

What now of presentism in the case of intentional verbs? Clearly enough, the presentist faces a parallel difficulty: take an instance of (α)—call it (β)—

(α) *x* searched for *y*

where (in β) the singular term replacing '*x*' refers to a presently existing individual, the singular term replacing '*y*' does not, and yet (β) formulates a truth. Like actualism before, (β) generates the following challenge for presentism:

> *Challenge Two*. Supply a credible account of (β) meeting this constraint: (β) is true but its truth does not require a presently existing individual to bear a *de re* relation to a non-presently existing one.

4. HOW NOT TO MEET THE SECOND CHALLENGE

In Fitch 1996 there is a novel and intriguing actualistic treatment of (9), one promising a straightforward response to Challenge Two. His proposal roundly deserves close and careful examination.

By Fitch's lights, when confronted with troublesome (9), the actualist need only "go adverbial". Very roughly, the suggestion is that (9) becomes

(9*) Ponce de Leon searched for a-unique-fountain-of-youthly.

Here 'a-unique-fountain-of-youthly' behaves as an *adverbial modifier* of the now *monadic* predicate 'searched for'. According to this proposal, (9), so construed, does not ascribe a relation between Ponce de Leon and the non-actual fountain of youth. Quite the contrary, parsed as (9*), (9) is seen to ascribe a complex but non-relational property to the single individual Ponce de Leon. As such, we are told, the truth of (9) does not require an actual individual to stand in a *de re* relation to some non-actual object. And consequently the second challenge is supposedly conquered.

Without worrying over the missing semantics for this adverbial treatment of (9), there appear to be decisive objections to any such account. For consider:

(α) *x* searched for *y*.

Now if the singular terms replacing '*x*' and '*y*' should both refer to actual concrete individuals, let us assume, then even under the Fitch proposal the resulting instance of (α) ascribes a dyadic relation between those very individuals. (When Bob *searched for* his missing daughter last night, that is, he really did stand in the searched for relation to his daughter.) Suppose, however, that whereas the singular term replacing '*x*' picks out an actual concrete individual, the one replacing '*y*' does not, as in (9):

(9) Ponce de Leon searched for the fountain of youth.

By the Fitch model, we are to hold in effect that 'the fountain of youth', as it occurs in (9), does not function as a singular term. (After all, (9) is parsed as (9*).) With the present interpretation of the Fitch proposal, therefore, the instances of (α) come in (at least) two sorts: when '*y*' is replaced by a singular term denoting a concrete individual, the instance in question ascribes a dyadic relation between actual concrete objects. And yet, if the term that replaces '*y*' fails to pick out a concrete individual, the instance at issue ascribes a complex but non-relational property to one actual object.

This alleged shift in semantical behavior of the various instances of (α) seems incredulous on two counts: first, outside of a prior commitment to an

ontology devoid of possible but non-actual individuals, the semantical shift at stake appears impossible to independently motivate or support; second, any such view requires intolerably that *logical form* turns on matters of contingent fact—to know what sort of proposition is expressed by an instance of (α) I must already know whether the singular terms involved do or do not refer to concrete but contingent objects.

As formulated so far, our objection to the Fitch model has centered around the pivotal assumption that under this model instances of (α) ascribe a dyadic relation between actual concrete individuals when the singular terms replacing '*x*' and '*y*' both refer to concrete objects. What happens if this assumption is abandoned? Why not, that is, interpret the Fitch model in such a way *that every* instance of (α) receives the sort of treatment accorded to (9), even when the singular term replacing '*y*' refers to a concrete individual? As I see it, there is a fatal objection to any such view. For even if the *searched for* relation never holds between actual concrete individuals, this surely is not true for the *loves* relation—Bob really does bear the latter relation to his missing daughter, Jane. But then the Fitch model, under the current interpretation, cannot do justice to truths like

(11) Bob searched for Jane, his missing daughter he deeply loves.

After all, 'his missing daughter he deeply loves' incontestably ascribes the dyadic *loves* relation between Bob and Jane. It follows, therefore, that (11) won't be true unless 'Jane', as it occurs in 'Bob searched for Jane', refers to Bob's missing daughter. And this is precisely what 'Jane' fails to do according to the present interpretation of the Fitch model.

This is no way to preserve actualism (or presentism).[10,11]

ENDNOTES

[*] It is with great pleasure that I dedicate the present essay to Alvin Plantinga—a friend and teacher.

[1] See Bealer 1993, Bigelow 1996, Chisholm 1990, Hinchliff 1996, and Menzel 1991. For extensive references on presentism see the bibliography in Bigelow 1996.

[2] There are in fact *grades* of actualism. Alvin Plantinga (1985), for example, endorses actualism as the view that there neither are nor could have been objects that do not actually exist. But Nathan Salmon (1987) embraces actualism only by affirming the first half of Plantinga's characterization while explicitly rejecting the second (and modal) half. In addition, there are more technical characterizations of actualism in Menzel 1990 and Fitch 1996. As the reader may verify, the discussion here applies to all of the above. For extensive references on actualism, see the bibliographies of Tomberlin and McGuinness 1994 and Tomberlin 1996a.

[3] As set out here, the present case combines features of Case One and Case Two in Tomberlin and McGuinness 1994.

[4] As opposed to substitutional quantification. For criticisms of the latter, see Tomberlin 1990, 1993, and forthcoming b.

[5] For recent discussions of soft laws, see Antony 1995, Horgan and Tienson 1990, and Schiffer 1991.

[6] In Tomberlin 1981 and 1986, I evaluate the conditional obligation systems of van Fraassen (1972), Mott (1973), and Al-Hibri (1978) negatively against the Contrary-to-Duty Imperative Paradox and the Knower Paradox, respectively. I document that Lewis (1974) falls prey to the Knower Paradox in my 1989a. In Tomberlin 1989b, it is argued that Feldman (1986) fails against both the Knower Paradox and the Contrary-to-Duty Imperative Paradox. And I establish that Åqvist (1987) succumbs to a version of the Contrary-to-Duty Imperative Paradox in Tomberlin 1991b.

[7] For additional discussion, see Feldman 1990 and Tomberlin 1995.

[8] See, for example, Tomberlin 1989b, and 1991b.

[9] There is an extended critique of Chisholm's ingenious version of self-ascription for believing in Tomberlin 1990b and 1991a. The very different formulations of self-ascription in Brand 1983, 1984 and Lewis 1979 and 1986 are critically evaluated in Tomberlin 1987 and Tomberlin 1989c, respectively.

[10] Should the presentist seek to apply the Fitch model to items like (β), the negative critique here of Fitch on actualism carries over *mutatis mutandis*. Linsky and Zalta 1994 contains an original and important version of actualism. I critically examine their view in Tomberlin 1996a. Linsky and Zalta reply in their 1996.

[11] For rewarding correspondence and/or discussion, I am grateful to David Armstrong, George Bealer, John Biro, Roderick Chisholm, David Cowles, Michael Devitt, Kit Fine, Greg Fitch, Gilbert Harman, Terry Horgan, David Kaplan, Bernard Kobes, Bernie Linsky, Kirk Ludwig, Bill Lycan, Chris Menzel, Al Plantinga, Greg Ray, Nathan Salmon, Bob Stalnaker, Ed Zalta, and my colleagues Frank McGuinness, Jeff Sicha, and Takashi Yagisawa. I do not mean to imply, of course, any agreement on their part with what I have argued here.

Note concerning the present paper. It is perhaps fitting that Jim Tomberlin's final paper is to be published in a volume honoring the contributions of Alvin Plantinga who was greatly admired and respected by Jim. It is also fitting that Jim's last paper concerns the debate over actualism, a subject he has written extensively on in recent years. This work meant a great deal to Jim and he continued to work on this paper until the very end when he literally could not work any more. He was disappointed that he was not going to be able to continue to fight against actualism in the coming years, but he hoped that his series of papers on this topic would not be forgotten (and that I and other actualists would some day "see the light"). Philosophy was one of the three great loves of Jim's life (Hilda, his wife, and good food and wine being the other two) and he worked on philosophy until the very end of his life. Jim was a loyal and helpful friend to many in philosophy, he will be missed by all of us.

G.W. FITCH
Arizona State University

REFERENCES

Antony, Louise. 1995. Law and order in psychology. *Philosophical Perspectives* 9: 429-496.

Åqvist, Lennart. 1987. *Introduction to Deontic Logic and the Theory of Normative Systems*. Napoli, Bibliopolis.

Bealer, George. 1993. A solution to Frege's puzzle. *Philosophical Perspectives* 7: 17-60.

Bigelow, John. 1996. Presentism and properties. *Philosophical Perspectives* 10: 35-52.

Brand, Myles. 1983. Intending and believing. In *Agent, Language, and the Structure of the World*, ed. James E. Tomberlin. Indianapolis, Hackett.

Brand, Myles. 1984. *Intending and Acting*. Cambridge, MA: MIT Press.

Castañeda, Hector Neri. 1983. Reply to Alvin Plantinga. In *Agent, Language, and the Structure of the World*, ed. James E. Tomberlin. Indianapolis, Hackett.

Castañeda, Hector Neri. 1986. Replies. In *Hector-Neri Castañeda*, ed. James E. Tomberlin. Dordrecht, D. Reidel.

Castañeda, Hector Neri. 1989. *Thinking, Language, and Experience*. Minneapolis, University of Minnesota Press.

Chisholm, Roderick M. 1981. *The First Person.* Minneapolis: University of Minnesota Press.

Chisholm, Roderick M. 1986. Self-Profile. In *Roderick M. Chisholm*, ed. Radu J. Bogdan. Dordrecht, D. Reidel.

Chisholm, Roderick M. 1990. Referring to things that no longer exist. *Philosophical Perspectives* 4: 545-556.

Feldman, Fred. 1986. *Doing the Best We Can*. Dordrecht, D. Reidel.

Feldman, Fred. 1990. A simpler solution to the paradoxes of deontic logic. *Philosophical Perspectives* 4: 309-342.

Fitch, G.W. 1994. Non-denoting. *Philosophical Perspectives* 7: 461-486.

Fitch, G.W. 1996. In defense of Aristotelian actualism. *Philosophical Perspectives* 10: 53-72.

Hinchliff, Mark. 1996. The puzzle of change. *Philosophical Perspectives* 10: 119-136.

Horgan, Terence and John Tienson. 1990. Soft laws. *Midwest Studies in Philosophy* 15: 256-279.

Kaplan, David. 1975. How to Russell a Frege-Church. *Journal of Philosophy* 72: 716-729.

Kripke, Saul. 1963. Semantical considerations on modal logic. *Acta Philosophica Fennica* 16: 83-94.

Kripke, Saul. 1980. *Naming and Necessity*. Cambridge, MA: Harvard University Press.

Lewis, David. 1973. *Counterfactuals*. Cambridge, MA: Harvard University Press.

Lewis, David. 1974. Semantic analyses for dyadic deontic logic. In *Logical Theory and Semantic Analysis*, ed. S. Stenlund. Dordrecht, D. Reidel.

Lewis, David. 1979. Attitudes de dicto and de se. *Philosophical Review* 88: 513-543.

Lewis, David. 1986. *On the Plurality of Worlds*. Oxford: Blackwell Publishers.

Linsky, Bernard and Edward N. Zalta. 1994. In defense of the simplest quantified modal logic. *Philosophical Perspectives* 8: 431-458.

Linsky, Bernard and Edward N. Zalta. 1996. In defense of the contingently nonconcrete. *Philosophical Studies* 84: 283-294.

Menzel, Christopher. 1990. Actualism and possible worlds. *Synthese* 85: 355-389.

Menzel, Christopher. 1991. Temporal actualism and singular foreknowledge. *Philosophical Perspectives* 5: 475-508.

Mott, Peter L. 1973. On Chisholm's paradox. *Journal of Philosophical Logic* 2: 197-211.

Plantinga, Alvin. 1985a. Self-Profile. In *Alvin Plantinga*, eds. James E. Tomberlin and Peter van Inwagen. Dordrecht, D. Reidel.

Plantinga, Alvin. 1985b. Replies. In *Alvin Plantinga*, eds. James E. Tomberlin and Peter van Inwagen. Dordrecht, D. Reidel.
Plantinga, Alvin. 1987. Two concepts of modality. *Philosophical Perspectives* 1: 189-232.
Salmon, Nathan. 1987. Existence. *Philosophical Perspectives* 1: 49-108.
Schiffer, Stephen. 1991. Ceteris paribus laws. *Mind*: 100: 1-17.
Stalnaker, Robert. 1968. A theory of conditionals. In *Studies in Logical Theory*, ed. N. Rescher. Oxford: Blackwell Publishers.
Tomberlin, James E. 1981. Contrary-to-duty imperatives and conditional obligation. *Noûs* 15: 357-375.
Tomberlin, James E. 1986. Good samaritans and Castañeda's system of deontic logic. In *Hector-Neri Castañeda*. Dordrecht, D. Reidel.
Tomberlin, James E. 1987. Critical review of Myles Brand, Intending and Acting. *Noûs* 21: 45-63.
Tomberlin, James E. 1988. Semantics, psychological attitudes, and conceptual role. *Philosophical Studies* 53: 205-226.
Tomberlin, James E. 1989a. Deontic logic and conditional obligation. *Philosophy and Phenomenological Research* 50: 107-114.
Tomberlin, James E. 1989b. Obligation, conditionals, and conditional obligation. *Philosophical Studies* 55: 81-92.
Tomberlin, James E. 1989c. Critical review of David Lewis, On the Plurality of Worlds. *Noûs* 23: 117-225.
Tomberlin, James E. 1990a. Belief, nominalism, and quantification. *Philosophical Perspectives* 4: 573-579.
Tomberlin, James E. 1990b. Critical review of R. Bogdan, ed., Roderick M. Chisholm. *Noûs* 24: 332-342.
Tomberlin, James E. 1991a. Belief, self-ascription, and ontology. In *Consciousness*, ed. Enrique Villanueva. Atascadero, CA: Ridgeview.
Tomberlin, James E. 1991b. Critical review of Lennart Åqvist, Introduction to Deontic Logic and the Theory of Normative Systems. *Noûs* 25: 109-116.
Tomberlin, James E. 1993. Singular terms, quantification and ontology. In *Naturalism and Normativity*, ed. Enrique Villanueva. Atascadero, CA: Ridgeview.
Tomberlin, James E., and Frank McGuinness. 1994. Troubles with actualism. *Philosophical Perspectives* 8: 459-466.
Tomberlin, James E. 1995. The paradoxes of deontic logic. *Cambridge Dictionary of Philosophy*, ed. in Robert Audi. Cambridge: Cambridge University Press.
Tomberlin, James E. 1996a. Actualism or possibilism? *Philosophical Studies* 84: 263-281.
Tomberlin, James E. 1996b. Perception and possibilia. In *Perception*, ed. Enrique Villanueva. Atascadero, CA: Ridgeview.
Tomberlin, James E. Forthcoming a. Naturalism, actualism, and ontology. *Philosophical Perspectives*.
Tomberlin, James E. Forthcoming b. Quantification: objectual or substitutional? In *Philosophical Issues* 8, ed. Enrique Villanueva. Atascadero, CA: Ridgeview.
Van Fraassen, Bas C. 1972. The logic of conditional obligation. *Journal of Philosophical Logic* 1: 417-438.

Chapter 2

PROPERTIES

Peter van Inwagen
University of Notre Dame

Although this paper makes extensive and essential use of the concept "abstract object," I am not going to try to explain or give any sort of account of this concept. That would be another paper. I will use the name 'platonism' for the thesis that there are abstract objects, and 'nominalism' for the thesis that there are no abstract objects. It has been suggested (I'm thinking of John Burgess and Gideon Rosen and their book *A Subject without an Object: Strategies for the Nominalistic Interpretation of Mathematics* (1997)[1]) that although a lot of philosophical work has been devoted to the question whether real analysis (or some other substantial part of mathematics) can be interpreted or revised or reconstructed in terms acceptable to nominalists, not nearly enough work has been devoted to the question why anyone should care whether something was acceptable to nominalists. It seems to me, however, that it is perfectly evident that nominalism is to be preferred to platonism, and perfectly evident *why* nominalism is to be preferred to platonism. And if nominalism is to be preferred to platonism, it is no great mystery why a philosopher of mathematics should want to have available a nominalistically acceptable reconstruction of all of, or some essential core of, mathematics.

And why do I say that nominalism is to be preferred to platonism? Since that question is not my topic, I will simply gesture vaguely at an answer. Platonists must say that reality, what there is, is divided into two parts: one part *we* belong to, and everything in "our" part is more like us than is anything in the other part. The inhabitants of the other part are radically unlike the things in our part—any given object in the other part is vastly *more* unlike any object *x* in our part than anything in our part is unlike

T. M. Crisp, M. Davidson and D. Vander Laan (eds.), Knowledge and Reality, 15-34.

x—and we can't really say much about what the things in the other part are like. (Compare the task of describing the properties of a pen and the properties of the number four.) It seems to me to be evident that it would be better not to believe in the other part of reality, the other category of things, if we could manage it. But we can't manage it. In the first part of this paper, I shall try to explain why we can't get along without *one* kind of abstract object: properties.

1. WE CAN'T GET ALONG WITHOUT PROPERTIES

How can the dispute between those who affirm and those who deny the existence of properties (platonists and nominalists) be resolved? The ontological method invented, or at least first made explicit, by Quine and Goodman (and illustrated with wonderful ingenuity in David and Stephanie Lewis's "Holes") suggests a way to approach this question.[2] Nominalists and platonists have different beliefs about what there is. Let us therefore ask this: How should one decide what to believe about what there is? According to Quine, the problem of deciding what to believe about what there is is a very straightforward special case of the problem of deciding what to believe. (The problem of deciding what to believe is, to be sure, no trivial problem, but it is a problem everyone is going to have somehow to come to terms with.) If we want to decide whether to believe that there are properties—Quine tells us—we should examine the beliefs we already have, the theses we have already, for whatever reason, decided to believe, and see whether they "commit us" (as Quine says) to the existence of properties. But what does this mean? Let us consider an example. Suppose we find the following proposition among our beliefs:

> Spiders share some of the anatomical features of insects.

This proposition may be expressed in what Quine calls the canonical language of quantification as follows:

> It is true of at least one thing that it is such that it is an anatomical feature and insects have it and spiders also have it.

(The canonical language of quantification does not essentially involve the symbols '∀' and '∃' and it does not essentially involve variables. There is *no* difference in meaning between 'It is true of at least one thing that it is such that it is an anatomical feature and insects have it and spiders also have it' and '∃*x* (*x* is an anatomical feature and insects have *x* and spiders also have *x*)'.)

But, obviously, if it is true of at least one thing that it is such that it is an anatomical feature and insects have it and spiders also have it, then at least one thing is an anatomical feature. And what is an anatomical feature if not a property?

Does this little argument show that anyone who believes that spiders share some of the anatomical features of insects is committed to platonism, and, more specifically, to a belief in the existence of properties? How might a nominalist respond to the argument? Suppose we present the argument to Nora, a convinced nominalist (who believes, as most people do, that spiders share some of the anatomical features of insects). Assuming that Nora is unwilling simply to have inconsistent beliefs, there would seem to be four possible ways for her to respond to it:

(1) She might become a platonist.

(2) She might abandon her allegiance to the thesis that spiders share some of the anatomical features of insects.

(3) She might attempt to show that, despite appearances, it does not follow from this thesis that there are anatomical features.

(4) She might admit that her beliefs (her nominalism and her belief that spiders share some of the anatomical features of insects) are apparently inconsistent, affirm, as an article of her nominalistic faith, that this inconsistency is apparent, not real, and confess that, although she is confident that there is some fault in our alleged demonstration that her belief about spiders and insects commits her to the existence of anatomical features, she is at present unable to discover it.

Possibility (2) is not really very attractive. It is unattractive for at least two reasons. First, it seems to be a simple fact of biology that spiders share some of the anatomical features of insects. Secondly, there are many, many sentences, sentences that seem to express "simple facts," that could have been used in place of 'Spiders share some of the anatomical features of insects' in an essentially identical argument for the conclusion that there are properties. Possibility (4) is always an option, but no philosopher is likely to embrace it except as a last resort. What Nora is likely to do is to try to avail herself of Possibility (3). If she does, she will attempt to find a *paraphrase* of 'Spiders share some of the anatomical features of insects', a sentence that (i) she could use in place of this sentence, and (ii) does not even *seem* to have 'There are anatomical features' as one of its logical consequences. If she can do this, she will be in a position to contend that the commitment to

the existence of anatomical features that is apparently "carried by" her belief about spiders and insects is only apparent. And she will be in a position to contend—no doubt further argument would be required to establish this—that the apparent existence of anatomical features is *mere* appearance (an appearance that is due to certain forms of words we use but needn't use).

Is it possible to find such a paraphrase? (And to find paraphrases of all the other apparently true statements that seem to commit those who make them to the reality of properties?) Well, yes and no. 'Yes' because it is certainly possible to find paraphrases of the spider-insect sentence that involve quantification over some other sort of abstract object than anatomical features—that is, other than properties. One might, for example, eliminate (as the jargon has it) the quantification over properties on display in the spider-insect sentence in favor of quantification over, say, concepts. No doubt any work that could be done by the property "having an exoskeleton" could be done by the concept "thing with an exoskeleton." But—here's the 'No'—a nominalist will be no more receptive to an ontology that contains concepts than to an ontology that contains properties. When I say it is not possible to get along without asserting the existence of properties, therefore, what I mean is that it is not possible to get along without asserting the existence of properties—*or* something that a nominalist is not going to like any better than properties.

Now Quine, the founder of the feast, would very likely want to break in at this point and tell us that we can find paraphrases of the spider-insect sentence that require "quantification over" (as he would say) no abstract objects but *sets*—an "ontic commitment" (as he would say) much to be preferred to an ontic commitment to properties. This is an important thesis, and Quine's arguments in support of his thesis are important arguments. I am afraid that in this paper I can do no more than acknowledge the existence of Quine's thesis and his supporting arguments.

Let us ask this. Is it possible to provide sentences like 'Spiders share some of the anatomical features of insects' with *nominalistically acceptable* paraphrases? My position is that it is not. I cannot hope to present an adequate defense of this position, for an adequate defense of this position would have to take the form of an examination of all possible candidates for nominalistically acceptable paraphrases of such sentences, and I cannot hope to do that. The question of nominalistically acceptable paraphrase will be answered, if at all, only as the outcome of an extended dialectical process, a process involving many philosophers and many years and many gallons of ink. I can do no more than look at one strand of reasoning in this complicated dialectical tapestry. My statement, "We can't get along without properties" must be regarded as a promissory note. But here is the ten-dollar co-payment on the debt I have incurred by issuing this note.

Suppose a nominalist were to say this: "It's easy to find a nominalistically acceptable paraphrase of 'Spiders share some of the anatomical features of insects'. For example: 'Spiders are like insects in some anatomically relevant ways' or 'Spiders and insects are in some respects anatomically similar'." A platonist is likely to respond as follows (at least this is what *I'd* say):

> But these proposed paraphrases seem to be quantifications over "ways a thing can be like a thing" or "respects in which things can be similar." If we translate them into the canonical language of quantification, we have sentences something like these:
>
> > It is true of at least one thing that it is such that it is a way in which a thing can be like a thing and it is anatomical and spiders are like insects in it.
> >
> > It is true of at least one thing that it is a respect in which things can be similar and it is anatomical and spiders and insects are similar in it.
>
> These paraphrases, therefore, can hardly be called nominalistically acceptable. If there are such objects as ways in which a thing can be like a thing or respects in which things can be similar, they must certainly be *abstract* objects.

What might the nominalist say in reply? The most plausible reply open to the nominalist seems to me to be along the following lines.

> My platonist critic is certainly a very literal-minded fellow. I didn't mean the 'some' in the open sentence 'x is like y in some anatomically relevant ways' to be taken as a *quantifier*: I didn't mean this sentence to be read '$\exists z$ (z is a way in which a thing can be like a thing and z is anatomical and x is like y in z)'. That's absurd. One might as well read 'There's more than one way to skin a cat' as '$\exists x\, \exists y$ (x is a way of skinning a cat and y is a way of skinning a cat and $x \neq y$)'. I meant this open sentence to have no internal logical structure, or none beyond that implied by the statement that two variables are free in it. It's just a form of words we learn to use by comparing various pairs of objects in the ordinary business of life.

And here is the rejoinder to this reply:

> If you take that line you confront problems it would be better not to have to confront. Consider the sentence 'x is like y in some physiologically relevant ways'. Surely there is some logical or structural or syntactical

> relation between this sentence and '*x* is like *y* in some anatomically relevant ways'? One way to explain the relation between these two sentences is to read the former as '$\exists z$ (*z* is a way in which a thing can be like a thing and *z* is physiological and *x* is like *y* in *z*)' and the latter as '$\exists z$ (*z* is a way in which a thing can be like a thing and *z* is anatomical and *x* is like *y* in *z*)'. How would *you* explain it? Or how would you explain the relation between the sentences '*x* is like *y* in *some* anatomically relevant ways' (which you say has no logical structure) and '*x* is like *y* in *all* anatomically relevant ways'? If neither of these sentences has a logical structure, how do you account for the obvious validity of the argument
>
> > Either of two female spiders of the same species is like the other in all anatomically relevant ways.
> >
> > *Hence*, an insect that is like a given female spider in some anatomically relevant ways is like any female spider of the same species in some anatomically relevant ways?
>
> If the premise and conclusion of this argument are read as having the logical structure their syntax suggests, the validity of this argument is easily demonstrable in textbook quantifier logic. If one insists that they have no logical structure, one will find it difficult to account for the validity of this argument. That is one of those problems I alluded to, one of those problems it would be better not to have to confront. (One of thousands of such problems.)

I suggest that we can learn a lesson from this little exchange between an imaginary nominalist and an imaginary platonist: that one should accept the following condition of adequacy on philosophical paraphrases.

> Paraphrases must not be such as to leave us without an account of the logical relations between predicates that are obviously logically related. Essentially the same constraint on paraphrase can be put in these words: A paraphrase must not leave us without an account of the validity of any obviously valid argument.

Accepting this constraint has, I believe, a significant consequence. This consequence requires a rather lengthy statement.

> Apparent quantification over properties pervades our discourse. In the end, one can avoid quantifying over properties only by quantifying over other sorts of abstract object—"ways in which a thing can be like a thing," for example. But most philosophers, if forced to chose between quantifying over properties and quantifying over these other objects

> would probably prefer to quantify over properties. The reason for this may be illustrated by the case of "ways in which a thing can be like a thing." If there really are such objects as ways in which a thing can be like a thing, they seem to be at once intimately connected with properties and, so to speak, more *specialized* than properties. What, after all, would a particular "way in which a thing can be like a thing" be but the sharing of a certain property? (To say this is consistent with saying that not just any property is such that sharing it is a way in which a thing can be like a thing; sharing "being green" can plausibly be described as a way in which a thing can be like a thing, but it is much less plausible to describe sharing "being either green or non-round"—if there is such a property—as a way in which a thing can be like a thing.) And if this is so, surely, the best course is to accept the existence of properties and to "analyze away" all apparent quantifications over "ways in which a thing can be like a thing" in terms of quantifications over properties.

It is the content of this lengthy statement that I have abbreviated as "We can't get along without properties."

This argument I have given has some obvious points of contact with the so-called Quine-Putnam indispensability argument for mathematical realism.[3] But there are important differences between the two arguments—I mean besides the obvious fact that my argument is an argument for the existence of properties and not an argument for the existence of specifically mathematical objects. It should be noted that my argument is not that we should believe that properties exist because their existence is an indispensable postulate of science. Nor have I contended that the scientific indispensability of properties is *evidence* for the existence of properties. I have not maintained that, because of the scientific indispensability of properties, any adequate account of the success of science must affirm the existence of properties. For one thing, my argument has nothing in particular to do with science. Science does indeed provide us with plenty of examples of sentences that must in some sense, on some analysis, express truths and also, on the face of it, imply the existence of properties. For example: 'Many of the important properties of water are due to hydrogen bonding'. But our everyday, pre-scientific discourse contains a vast number of such sentences, and these will serve my purposes as well as any sentences provided by the sciences. If our spider-insect sentence is insufficiently non-scientific to support this thesis, there are lots of others ('The royal armorer has succeeded in producing a kind of steel that has some but not all of the desirable characteristics of Damascus steel'). My argument could have been presented in, say, the thirteenth century, and the advent of modern science has done nothing to make it more cogent.

More importantly, I have not supposed that the fact (supposing it to be a fact) that quantification over properties is an indispensable component of our discourse is any sort of *evidence* for the existence of properties. That's as may be; I neither affirm that thesis nor deny it. It is simply not a premise of my argument, which is not an epistemological argument. Nor is my argument any sort of "transcendental" argument or any sort of inference to the best explanation; I have not contended that the success of science, or the success of our everyday, pre-scientific discourse, cannot be accounted for on nominalistic premises. Again, that's as may be. If I have appealed to any general methodological principle, it is only this: If one doesn't believe that things of a certain sort exist, one shouldn't say anything that demonstrably implies that things of that sort do exist. (Or, at any rate, one may say such things only if one is in a position to contend, and plausibly, that saying these things is a mere manner of speaking—that, however convenient it may be, it could, in principle, be dispensed with.) This methodological rule does not, I think, deserve to be controversial. We would all agree, would we not, that if *p* demonstrably implies the existence of God, then atheists who propose to remain atheists shouldn't affirm *p*? Or not, at any rate, unless they can show us how they could in principle dispense with affirming *p* in favor of affirming only propositions without theological implications?[4]

I suppose I ought to add—the point needs to be made somewhere—that if one *could* show how to eliminate quantification over properties in a nominalistically acceptable way, that achievement, by itself, would have no ontological implications. After all, Quine has shown how to eliminate quantification over everything but pure sets, and Church has shown how to eliminate quantification over women.[5] The devices of Quine and Church would be of ontological interest if "containing only pure sets" or "not containing women" were desirable features for an ontology to have. But they're not. If what I said in my brief opening remarks is right, however, "containing no abstract objects" *is* an advantage in an ontology.

I will close this part of the paper with a point about philosophical logic—as opposed to metaphysics. My argument fails if there is such a thing as substitutional quantification; and it fails if there is such a thing as quantification into predicate positions. (Or so I'm willing to concede. If either substitutional quantification or quantification into predicate positions is to be found in the philosopher's tool kit, then defending my thesis—"We can't get away with it"—becomes, at the very least, a much more difficult project.) I say this: substitutional quantification and quantification into non-nominal positions (including predicate positions) are both meaningless. I have argued elsewhere for the meaninglessness of substitutional quantification.[6] I accept Quine's arguments for the meaninglessness of quantification into predicate positions.[7]

2. IF WE AFFIRM THE EXISTENCE OF PROPERTIES, WE OUGHT TO HAVE A THEORY OF PROPERTIES

By a "theory of properties," I mean some sort of specification of, well, the *properties* of properties. If one succeeds in showing that we cannot dispense with quantification over properties, one's achievement does not tell us much about the intrinsic features of these things. In my opening remarks, I said that we didn't know much about the properties of properties. I am now making the point that the sort of argument for the existence of properties I have offered does not tell us much about the nature of properties. The whole of our discourse about things, on the face of it, defines what may be called "the property role," and our argument can be looked on as an attempt to show that something must play this role. (The property role could, in principle, be specified by the Ramsey-style methods that Lewis sets out in "How to Define Theoretical Terms."[8]) But it tells us nothing about the intrinsic properties of the things that play this role that enable them to play this role. In "Holes," Bargle argues that there must be holes, and his argument is in many ways like our argument for the existence of properties. That is, he uses some ordinary discourse about cheese and crackers to define the "hole role," and he attempts to show that one can't avoid the conclusion that something plays this role. Argle, after an initial attempt to evade Bargle's argument, accepts it. He goes on, however, to show how things acceptable to the materialist can play the hole role. In doing this, he spells out the intrinsic properties of the things he calls holes (when they are holes in a piece of cheese, they are connected, singly-perforate bits of cheese that stand in the right sort of contrast to their non-cheesy surroundings), and he, in effect, shows that things with the intrinsic properties he assigns to holes are capable of playing the role that Bargle's argument shows is played by something-we-know-not-what.

We are not in a position to do, with respect to properties, anything like what Argle has done with respect to holes, for, as I have observed, we cannot say anything much about the intrinsic properties of properties. The plain fact is: we platonists *can't* describe those somethings-we-know-not-what we say play the property role in anything like the depth in which Argle describes the things that (*he* says) play the hole role. Argle can describe the things he calls 'holes' as well as he can describe anything; we platonists can describe any concrete object in incomparably greater depth than we can any property.

I wish it weren't so, but it is. Or so *I* say. Some will dissent from my thesis that properties are mysterious. David Lewis is a salient example. If Lewis is right about properties, the property-role is played by certain *sets*, and one can describe at least some of these sets as well as one can describe any set.[9] In my view, however, Lewis is not right about properties. In the

next section, I will explain why I think this. (A qualification. I have said that, according to Lewis, certain sets are suitable to play the property role. In Lewis's view, however, it may be that our discourse defines at least two distinct roles that could equally well be described as "property-roles." Although—Lewis tells us—the sets he calls 'properties', can play *one* of the property roles, they are unsuited for the other (or the others)—if indeed our discourse does define two or more roles that can plausibly be described as "property-roles."[10])

3. LEWIS'S THEORY OF PROPERTIES AS SETS (WITH SOME REMARKS ON MEINONGIAN THEORIES OF PROPERTIES AS SETS)

According to Lewis the property "being a pig" is the set of all pigs, including those pigs that are inhabitants of other possible worlds than ours. But, in saying this, I involve myself in Lewis's notorious modal ontology. Let us, for the moment, avoid the questions raised by Lewis's modal ontology and say that Lewis's theory is one member of a species of theory according to all of which the property "being a pig" is the set of all possible pigs. Members of this species differ in their accounts of what a possible pig is. (That is to say, they differ in their accounts of what a *possibile* or *possible object* is, for we are interested not only in the property "being a pig" but in properties generally. According to all theories of this kind, every property is a set of *possibilia* and every set of *possibilia* is a property.) Lewis's theory will be just the member of this species according to which possible objects are what Lewis says possible objects are, and will be like the other members of the species on all points not touching on the nature of possible objects. The other members of the species are Meinongian theories, or all of them I can think of are.

What is a possible object? A Meinongian, or, rather, a neo-Meinongian like Terry Parsons or Richard Sylvan, has a simple answer to this question.[11] Just as a possible proposition is a proposition that is possibly *true*, and a possible property is a property that is possibly *instantiated*, a possible object is an object that is possibly *existent*. And, the neo-Meinongians maintain, objects are not necessarily and automatically existent. Although any object must *be*, there are objects that could fail to *exist*. In fact, most of the objects that are *do* fail to exist, and many objects that do exist might have been without existing. (Paleo-Meinongians would not agree that any object must be; they contend that many objects, so to speak, don't be.)

What is to be said about neo-Meinongianism? What Lewis says seems to me to be exactly right: the neo-Meinongians have never explained what they

mean by 'exist'.[12] We anti-Meinongians and they mean the same thing by 'be'. We anti-Meinongians say that 'exists' and 'be' mean the same thing; the neo-Meinongians say that this is wrong and 'exists' means something else, something other than 'be'. (And, they say, the meanings of the two verbs are so related that—for example—the powers that exist must form a subset of the powers that be.) Unfortunately, they have never said what this "something else" is. I would add the following remark to Lewis's trenchant critique of neo-Meinongianism. The only attempt at an explanation of the meaning of 'exists' that neo-Meinongians have offered proceeds by laying out supposed examples of things that are but do not exist. But, in my view, the right response to every such example that has ever been offered is either "That does too exist" or "There is no such thing as that." And, of course, if there is no distinction in meaning between 'be' and 'exist', then neo-Meinongianism cannot be stated without contradiction. If 'be' and 'exist' mean the same thing, then the open sentence 'x exists' is equivalent to '$\exists y$ $x = y$'. And, if that is so, 'There are objects that do not exist' is logically equivalent to 'Something is not identical with itself'. Since neo-Meinongians obviously do not mean to embrace a contradiction, their theory depends on the premise that 'exist' means something other than 'be'. But, so far as I can see, there is nothing for 'exists' *to* mean but 'be'. In the absence of further explanation, I am therefore inclined to reject their theory as meaningless.

Let us turn to Lewis's version of the properties-as-sets-of-possible-objects theory. According to Lewis, there are no objects that do not exist. Objects, however, may be divided into those that *actually* exist and those that do not actually exist. The category "possible object" comprises both those things that actually exist and those things that exist but do not actually exist ("*merely* possible objects"). But what do we mean when we say of objects that do not actually exist that they, nevertheless, exist? Isn't a flying pig an excellent example of an object that doesn't actually exist? And isn't it true of any flying pig that it doesn't exist—doesn't exist without qualification? No, says Lewis. Flying pigs are not objects of which we can correctly say that they do not exist "in the philosophy room." Outside the philosophy room, in the ordinary business of life, we can say, and say truly, that flying pigs do not exist, despite the fact that we say truly in the philosophy room that there are flying pigs. When we say, outside the philosophy room, that there are no flying pigs, our use of the quantifier is like that of someone who looks in the fridge and says sadly, "There's no beer." When I say, in the philosophy room, "There are flying pigs, but they're one and all non-actual," I'm saying this: There are [an absolutely unrestricted quantifier; the philosophy room is just that place in which all contextual restrictions on quantification are abrogated] flying pigs, and they're spatio-temporally unrelated to me'.

The problem with Lewis's theory, as I see it, is that there is no reason to think that there is anything spatiotemporal that is spatiotemporally unrelated to me, and, if there *is* anything in this category, I don't see what it has to do with modality.[13] Suppose there *is* a pig that is spatiotemporally unrelated to me—or, less parochially, to us. Why should one call it a "merely possible pig"—or a "non-actual pig"? Why are those good things to call it? This is not the end of the matter, however. Even if a pig spatiotemporally unrelated to us *can't* properly be called a merely possible pig, it doesn't follow immediately that Lewis's theory of properties is wrong. If what Lewis calls the principle of plenitude is true—if, as Lewis maintains, there exists (unrestricted quantifier) a pig having, intuitively speaking, every set of properties consistent with its being a pig—, then there might be something to be said for identifying the set of all pigs (including those spatiotemporally unrelated to us) with the property "being a pig." (If there exist pigs having every possible combination of features, there must be pigs that are spatially or temporally unrelated to us: if every pig was spatially and temporally related to us, there wouldn't be room for all the pigs Lewis says there are.) There might be something to be said for this identification, that is, even if the set of all pigs couldn't properly be called 'the set of all pigs, both actual and merely possible'. But even if there are pigs spatiotemporally unrelated to us, there is, so far as I can see, no good reason to accept the principle of plenitude—even as it applies to pigs, much less in its full generality.

On the face of it, the set of pigs seems to represent far too sparse a selection of the possible combinations of characteristics a pig might have for one to be able plausibly to maintain that this set could play the role "the property of being a pig." According to both the neo-Meinongians and Lewis, the set of pigs has a membership much more diverse than most of us would have expected, a membership whose diversity is restricted only by the requirements of logical consistency (for Lewis) or is not restricted at all (for the neo-Meinongians). If I am right, both Lewis and the Meinongians have failed to provide us with any reason to accept this prima facie very uncompelling thesis.

4. A THEORY OF PROPERTIES

There is only one real objection to Lewis's theory of properties: it isn't true. It is a model of what a good theory should be, insofar as theoretical virtue can be divorced from truth. In this, the final section of this paper, I present a theory of properties that, or so *I* say, does have the virtue of truth. Alas, even if it has that virtue, it has few others. Its principal vice is that it is very nearly vacuous. It can be compared to the theory that taking opium is followed by sleep because opium possesses a sleep-inducing virtue. That

theory about the connection of opium and sleep, as Lewis has pointed out, is not *entirely* vacuous; it is inconsistent with various theses, such as the thesis that taking opium is followed by sleep because a demon casts anyone who takes opium into sleep. The theory of properties I shall present, although it is pretty close to being vacuous, is inconsistent with various theses about properties, and some of these theses have been endorsed by well-known philosophers. (A proper presentation of this theory would display properties as members of a more inclusive class of entities, relations. But I will not attempt to discuss relations within the confines of this paper.)

The theory I shall present could be looked on as a way of specifying the property role, a way independent of and a little more informative than specifying this role via the apparent quantifications over properties that are to be found in our discourse. This theory identifies the property role with the role "thing that can be said of something." This role is a special case of the role "thing that can be said." Some things that can be said are things that can be said *period*, things that can be said *full stop*. For example: that London has a population of over seven million is something that can be said; another thing that can be said is that no orchid has ever filed an income-tax return. But these things—'propositions' is the usual name for them—are not things that can be said *of* anything, not even of London and orchids. One can, however, say *of* London that it has a population of over seven million, and one can also say this, this very same thing, of New York. And, of course, one can say it of Mexico City and of Oxford. (It can be said only falsely of Oxford, of course, but lies and honest mistakes are possible.) I will assume that anything that can be said of anything can be said of anything else. Thus, if there are such things as topological spaces, one can say of any of them that it is a city with a population of over seven million, or that it has never filed an income-tax return. I don't know why anyone would, but one could.

Let us call such things, propositions and things that can be said of things, *assertibles*. The assertibles that are not propositions, the things that can be said *of* things, we may call *unsaturated* assertibles. I will assume that the usual logical operations apply to assertibles, so that, for example, if there are such assertibles as "that it has a population of over seven million" and "that it once filed an income-tax return," there is also, automatically as it were, the assertible "that it either has a population of over seven million or else has never filed an income tax return." (In a moment, I shall qualify this thesis.) It follows that the phrase I used to specify the unsaturated-assertible role—"things that can be said of things"—cannot be taken too literally. For if there are any unsaturated assertibles, and if there are arbitrary conjunctions and disjunctions and negations of such unsaturated assertibles as there are, it will be impossible for a finite being to say most of them of anything. "Things that can be said of things" must therefore be understood in the sense "things that can in principle be said of things" or perhaps "things of a type such that

some of the simpler things of that type can be said of things" or "things that a being without limitations could say of things." All these ways of qualifying 'said of' could do with some clarification, but I cannot discuss the problems they raise here. (One possible solution to the problem raised by human limitations for our role-specification would be to substitute something like 'is either true or false of' for 'can be said of' in our specification of the unsaturated-assertible role. This is, in my view, a promising suggestion, but I do think that 'can be said of' has certain advantages in an initial, intuitive presentation of the theory of properties I shall present.)

It seems to me that there are such things as unsaturated assertibles: there are things that can be said of things. It seems to me that there is an *x* such that *x* can be said of *y* and can also be said of *z*, where *z* is not identical with *y*. One of the things you can say about the Taj Mahal is that it is white, and you can say that about the Lincoln Memorial, too. (I take it that 'about' in this sentence is a mere stylistic variant on 'of'.) If, during the last presidential campaign, you had heard someone say, "All the negative things you've said about Gore are perfectly true, but don't you see that they're equally applicable to Bush?", you wouldn't have regarded this sentence as in any way problematical; not logically or syntactically or lexically problematical, anyway. (And if the speaker had said 'perfectly true *of him*' instead of 'perfectly true' your only objection could have been that this phrasing was wordy or pedantic.) I say it seems to me that there are such things. I certainly see almost no reason to *deny* that there are such things, other than reasons that are reasons for denying that there are abstract objects of any sort. (For assertibles of any sort, if they exist, are certainly abstract objects.) I say 'almost no reason' because there are, I concede, powerful "Russellian" objections to admitting assertibles into our ontology. If there are things that can be said, there are things that can be said of things that can be said. And it seems evident that one of them must be "that it can't be said truly of itself." But that way paradox lies. I will not discuss this problem, for the simple reason that it is a problem that confronts anyone who has a theory of properties—or a theory of sets. (But here is a qualification I promised a moment ago. Perhaps there is such an assertible as "that it can be said truly of itself" but, for the reason I have just alluded to, no such assertible as "that it can't be said truly of itself.")

I propose, therefore, that properties be identified with unsaturated assertibles, with things that can be said of things. It seems unproblematical that unsaturated assertibles can successfully play the property role. And I would ask this: what is the property whiteness but something we, in speaking of things, occasionally predicate of some of them? And what is predicating something of something but *saying* the former *of* the latter? Well, perhaps someone will say that it sounds wrong or queer to say that whiteness is one of the things we can say of the Taj Mahal. I don't think that arguments that

proceed from that sort of premise have much force, but I won't press the point. Anyone who thinks that unsaturated assertibles—from now on I'll say simply 'assertibles'—can't play the property role but is otherwise friendly to my arguments may draw this conclusion from them: there are, strictly speaking, no properties, but assertibles may be pressed into service to do the work that would fall to properties if it were not for the inconvenient fact that there are no properties to do it. If we suppose that there are assertibles, and if we're unwilling to say that assertibles are properties, what advantage should we gain by supposing that there are, in addition, things that we *are* willing to call properties?

Now if properties are assertibles, a wide range of things philosophers have said using the word 'property' are false or unintelligible. For one thing, a property, if it is an assertible, cannot be a part or a constituent of any concrete object. If this pen exists, there are no doubt lots of things that are in some sense its parts or constituents: atoms, small manufactured items; perhaps, indeed, every sub-region of the region of space exactly occupied by the pen at *t* is at *t* exactly occupied by a *part* of the pen. But "that it is a writing instrument," although it can be said truly of the pen—and is thus, in my view, one of the properties of the pen—is not one of the parts of the pen. That it is not is as evident as, say, that the pen is not a cube root of any number. Nor is "that it is a writing instrument" in any sense present in any region of space. It makes no sense, therefore, to say that "that it is a writing instrument" is "wholly present" in the space occupied by the pen. In my view, there is just nothing *there* but the pen and its parts (parts in the "strict and mereological sense"). There are indeed lots of things true of the pen, lots of things that could be said truly about the pen, but those things do not occupy space and cannot be said to be wholly (or partly) present anywhere.

If properties are assertibles, it makes no sense to say that properties are somehow more basic ontologically than the objects whose properties they are. A chair cannot, for example, be a collection or aggregate of the properties ordinary folk say are the properties of a thing that is not a property, for a chair is not a collection or aggregate of all those things one could truly say of it. Nor could the apparent presence of a chair in a region of space "really" be the co-presence in that region of the members of a set of properties, because, if for no other reason, there is no way in which a property can be present in a region of space. (I hope no one is going to say that if I take this position, I must believe in "bare particulars." A bare particular would be a thing of which nothing could be said truly, an obviously incoherent notion.)

Properties, if they are assertibles, are in no way objects of sensation. If colors are properties and properties are assertibles, then the color white is the thing that one says of something when one says of it that it is white. And this assertible is not something that can be seen—just as extracting a cube root is

not something you can do with a forceps. We never see properties, although we see *that* certain things have certain properties. (Looking at the pen, one can see that what one says of a thing when it one says it's cylindrical is a thing that can be said *truly* of the pen.) Consider sky blue—the color of the sky. If it is not true now, it was certainly true ten thousand years ago that nothing was sky blue. Let's suppose, for the sake of the illustration, that it's true now. (If I say that nothing is sky blue, it's not to the point to tell me that the sky is sky blue or that a reflection of the sky in a pool is sky blue, for there is no such thing as the sky and there are no such things as reflections. And don't tell me I perceive a sky-blue quale or visual image or sense-datum, for there are no qualia or visual images or sense-data. I may be sensing sky-bluely when I look at the sky on a fine day, but that shows at most that something has the property "sensing sky-bluely"; it does not show that something has the property "being sky blue.") Now some philosophers who would agree with my thesis that nothing is sky blue infer from this proposition the conclusion that it's possible to *see* the property "being sky blue." After all, this property is in some way involved in the visual experience I have when I look at the sky, and this fact can't be explained by saying that I'm seeing something that has it, for nothing has it. And what is there left to say but that I see the uninstantiated property "being sky blue"? I would answer as follows: since the property "being sky blue" is just one of those things that are available to say about a cup or a sheet of wrapping paper or a shirt (or, for that matter, human blood or the Riemann curvature tensor), we obviously don't *see* it. It's involved in our sensations when we look at the sky only in this Pickwickian sense: we're sensing in the way in which visitors to the airless moon would sense during the lunar day if the moon were surrounded by a shell of sky-blue glass. And why *shouldn't* we on various occasions sense in the way in which we should sense *if* an X were present when there's in fact no X there?

Some philosophers have said that existence is not a property. Are they right or wrong? They are wrong, *I* say, if there is such a thing to be said about something as that it exists. And it would seem that there is. Certainly there is this to be said of a thing: that it might not have existed. And it is hard to say how there could be such an assertible as "that it might not have existed" if there were no such assertible as "that it exists."

Some philosophers have said that there are individual essences or haecceities, "thisnesses" such as "being *that* object" or "being identical with Alvin Plantinga." Are they right or wrong? They are right, *I* say, if one of the things you can say about something is that it is identical with Alvin Plantinga. Is there? Well, it would seem that if Plantinga hadn't existed, it would still have been true that he might have existed. (It would seem so, but it has been denied.) And it is hard to see how there could be such a saturated

assertible as "that Alvin Plantinga might have existed" if there were no such unsaturated assertible as "that it is Alvin Plantinga."

Some philosophers have said that although there are obviously such properties as redness and roundness, it is equally obvious that there is no such property as "being either red or not round." They have said that in their view, the world, or the Platonic heaven, is "sparsely," not "abundantly," populated with properties. Are they right? If properties are assertibles, only one answer to this question seems possible: No. If one of the things you can say about something is that it is red and another thing you can say about something is that it is round, then, surely, one of the things you can say about something is that it is either red or not round. (Mars is either red or not round, and *that*, the very same thing, is also true of the Taj Mahal and the number four—given, of course, that all three objects exist.) It is, of course, our answer to this question—"abundantly"—that eventually leads to our troubles with Russell's Paradox. But, again, the alternative doesn't seem possible.

Some philosophers have denied the existence of uninstantiated properties. Is this a plausible thesis? If properties are assertibles, it is a very implausible thesis indeed, for there are obviously things that can be said of things that can't be said *truly* of anything: that it's a—non-metaphorical—fountain of youth, for example. (No doubt someone, Ponce de León or some confidence trickster, has said this very thing about some spring or pool.) Having answered the question whether there are uninstantiated properties, at least to my own satisfaction, I'll briefly consider a couple of related questions. Are there such things as *necessarily* uninstantiated properties? Yes indeed, for one of the things you can say about Griffin's *Elementary Theory of Numbers* is that it contains a correct proof of the existence of a greatest prime. (You can say it about *Tess of the D'Urbervilles*, too.) And, of course, if one of the things you can say about something is that it is round and another thing you can say about something is that it is square, then (by a principle I've endorsed several times), one of the things you can say about something is that it is both round and square.

Some philosophers have said that properties exist only contingently. This would obviously be true if there could not be uninstantiated properties, but it would be possible to maintain that there are uninstantiated properties and that, nevertheless, some or all properties are contingently existing things. Could this be? Well, it would certainly seem not, at least if the accessibility relation is symmetrical. One of the things you can say about something is that it is white. Are there possible worlds in which there's no such thing to be said of anything? Suppose there is such a world. In that world, unless I'm mistaken, it's not even possibly true that something is white. Imagine, if you don't mind using this intellectual crutch, that God exists in a world in which there's no such thing to be said of a thing—not "said *truly* of a thing": "said

of a thing *simpliciter*"—as that it is white. Then God, who is aware of every possibility, is not aware of the possibility that there be something white. (If God could be aware of or consider the possibility that there be something white, he would have to be aware that one of the things that can be said of something is that it is white.) Therefore, there must be no such possibility in that world as the possibility that there be something white. Therefore, with respect to that possible world, the possible world that is in fact actual is not even possible; that is to say, in that world, the world that is in fact the actual world doesn't exist. But then the accessibility relation is not symmetrical. And I should want to say about the proposition that the accessibility relation is symmetrical what Gödel said of the axioms of set theory: it forces itself upon the mind as true. Admittedly, there are steps in this argument that can be questioned and have been questioned—or at least the corresponding steps in certain very similar arguments have been questioned. (I give one example of an objection, not the most important objection, that could be made to this argument: the argument at best proves that 'that it is white' denotes *an* object in, or with respect to, every possible world; it doesn't follow from this that this phrase denotes the *same* object in every possible world.) But the argument seems convincing to me. At any rate, it is the argument that will have to be got round by anyone who wants to say that properties do not exist necessarily.

There are many other theses and questions about properties than those I have considered. But the theses and questions I have considered are all those, or so it seems to me, to which the theory of properties as assertibles is relevant. The fact that this theory is inconsistent with various theses about properties shows that, although it may be very close to being vacuous, it does not manage to be entirely vacuous.[14]

ENDNOTES

[1] See Burgess and Rosen 1997, Part 1A, "Introduction".

[2] Quine 1961; Quine 1960: Chap. VII; Goodman and Quine 1947; Lewis and Lewis 1983.

[3] See Putnam 1971, reprinted in its entirety in Laurence and Macdonald 1998.

[4] For an important objection to this style of reasoning, see Melia 1995. I intend to discuss Melia's paper elsewhere; to discuss it here would take us too far afield. I wish to thank David Manley for impressing upon me the importance of Melia's paper (and for correspondence about the issues it raises).

[5] In 1958, Alonzo Church delivered a lecture at Harvard, the final seven paragraphs of which have lately been making the e-mail rounds under the title (not Church's), "Ontological Misogyny." In these paragraphs, Church wickedly compares Goodman's attitude toward abstract objects to a misogynist's attitude toward women. ("Now a misogynist is a man who finds women difficult to understand, and who in fact considers them objectionable incongruities in an otherwise matter-of-fact and hard-headed world. Suppose then that in

analogy with nominalism the misogynist is led by his dislike and distrust of women to omit them from his ontology.") Church then shows the misogynist how to eliminate women from his ontology. (In case you are curious: We avail ourselves of the fact that every woman has a unique father. Let us say that men who have female offspring have two modes of presence in the world, primary and secondary. Primary presence is what is usually called presence. In cases in which we should normally say that a woman was present at a certain place, the misogynist who avails himself of Church's proposal will say that a certain man—the man who would ordinarily be described as the woman's father—exhibits secondary presence at that place . . . .) "Ontological Misogyny" came to me by the following route: Tyler Burge, Michael Zeleny (Department of Mathematics, UCLA), James Cargile.

Quine's reduction of everything to pure sets (well, of physics to pure sets, but physics is everything for Quine) can be found in his 1976. I thank Michael Rea for the reference.

[6] Van Inwagen 1981. The arguments presented in this paper are similar to the more general arguments of William G. Lycan's fine paper, "Semantic Competence and Funny Functors" (1979). Van Inwagen 1981 is reprinted in my 2001.

[7] See the section of Quine 1970 entitled "Set Theory in Sheep's Clothing" (pp. 66-68).

[8] Lewis 1983: 78-95.

[9] See Section 1.5, "Modal Realism at Work: Properties," of Lewis 1986, pp. 50-69.

[10] See Lewis 1999a. See especially the section entitled "Universals," pp. 10-24.

[11] See Parsons 1980 and Routley 1980 (Richard Routley = Richard Sylvan).

[12] See Lewis 1999b.

[13] I have gone into this matter in a great deal of detail in van Inwagen 1986.

[14] A longer version of this paper (with the appropriately longer title "A Theory of Properties") will appear in *Oxford Studies in Metaphysics*.

REFERENCES

Burgess, John and Gideon Rosen. 1997. *A Subject without an Object: Strategies for the Nominalistic Interpretation of Mathematics*. Oxford: Oxford University Press.

Goodman, Nelson and W. V. Quine. Steps toward a constructive nominalism. *Journal of Symbolic Logic* 12: 105-122.

Laurence, Stephen and Cynthia Macdonald, eds. 1998. *Contemporary Readings in the Foundations of Metaphysics*. Oxford: Blackwell Publishers.

Lewis, David. 1983. *Philosophical Papers: Volume I*. Oxford: Oxford University Press.

Lewis, David. 1986. *On the Plurality of Worlds*. Oxford: Blackwell Publishers.

Lewis, David. 1999a. New work for a theory of universals. In *Papers on Metaphysics and Epistemology*, 8-55. Cambridge: Cambridge University Press.

Lewis, David. 1999b. Noneism or allism? In *Papers on Metaphysics and Epistemology*, 152-163. Cambridge: Cambridge University Press.

Lewis, David and Stephanie Lewis. 1983. Holes. In David Lewis, *Philosophical Papers: Volume 1*, 3-9. New York: Oxford University Press.

Lycan, William G. 1979. Semantic competence and funny functors. *The Monist* 64: 209-222.

Melia, Joseph. 1995. On what there's not. *Analysis* 55: 223-229.

Parsons, Terence. 1980. *Non-Existent Objects*. New Haven, CT: Yale University Press.

Putnam, Hilary. 1971. *Philosophy of Logic*. New York: Harper & Row.

Quine, W. V. O. 1960. *Word and Object*. Cambridge, MA: the MIT Press.

Quine, W. V. O. 1961. On what there is. In *From a Logical Point of View*, 1-19. Cambridge, MA: Harvard University Press.

Quine, W. V. O. 1970. *Philosophy of Logic*. Englewood Cliffs, NJ: Prentice-Hall.

Quine, W. V. O. 1976. Whither physical objects? In *Essays in Memory of Imre Lakatos*, eds. R. S. Cohen, P. K. Feyerabend and M. W. Wartofsky, 497-504. Dordrecht: D. Reidel.

Routley, Richard. 1980. *Exploring Meinong's Jungle and Beyond: An Investigation of Noneism and the Theory of Items*. Canberra: Departmental Monograph #3, Philosophy Department, Research School of Social Sciences, the Australian National University.

Van Inwagen, Peter. 1981. Why I don't understand substitutional quantification. *Philosophical Studies* 39: 281-285.

Van Inwagen, Peter. 1986. Two concepts of possible worlds. *Midwest Studies in Philosophy* 11: 185-213

Van Inwagen, Peter. 2001. *Ontology, Identity and Modality: Essays in Metaphysics*. Cambridge: Cambridge University Press.

Chapter 3

SO YOU THINK YOU EXIST?

In Defense of Nolipsism

Jenann Ismael and John L. Pollock
University of Arizona

Human beings think of themselves in terms of a privileged non-descriptive designator—a mental "I". Such thoughts are called "*de se*" thoughts. The mind/body problem is the problem of deciding what kind of thing I am, and it can be regarded as arising from the fact that we think of ourselves non-descriptively. Why do we think of ourselves in this way? We investigate the functional role of "I" (and also "here" and "now") in cognition, arguing that the use of such non-descriptive "reflexive" designators is essential for making sophisticated cognition work in a general-purpose cognitive agent. If we were to build a robot capable of similar cognitive tasks as humans, it would have to be equipped with such designators.

Once we understand the functional role of reflexive designators in cognition, we will see that to make cognition work properly, an agent must use a *de se* designator in specific ways in its reasoning. Rather simple arguments based upon how "I" works in reasoning lead to the conclusion that it cannot designate the body or part of the body. If it designates anything, it must be something non-physical. However, for the purpose of making the reasoning work correctly, it makes no difference whether "I" actually designates anything. If we were to build a robot that more or less duplicated human cognition, we would not have to equip it with anything for "I" to designate, and general physicalist inclinations suggest that there would be nothing for "I" to designate in the robot. In particular, it cannot designate the physical contraption. So the robot would believe "I exist", but it would be wrong. Why should we think we are any different?

T. M. Crisp, M. Davidson and D. Vander Laan (eds.), Knowledge and Reality, 35-62.

1. THE MIND/BODY PROBLEM

I look around and see the world, and when I do I see it from a certain perspective. I see the world as a spatial system with myself located in it, and I see it from the perspective of where I am. My perceptual system locates objects with respect to me. For example, my visual system represents objects in a polar coordinate system with myself at the origin—the focal point. On the basis of my perceptions I make judgements about the way the world is, and adopt goals for changing it. Most of my goals are egocentric—I want to change my own situation in the world. I am equipped for this purpose with various causal powers. I have the ability to perform actions that have effects on my surroundings. These causal powers are centered on my location in the world. I have a body, and I act on the world by moving various parts of my body.

This simple self-description of myself and my place in the world seems uncontroversial, but it leads to perplexing philosophical problems. Although I am intimately connected with my body, and can only act on the world via my body, I do not think of myself as simply *being* my body. When I turn my gaze downwards and see my own body, I think of myself as being "up here looking down". This follows from the way my perceptual system represents the objects I see as being in front of me, with I myself being located at the focal point of my visual field. Anything that I can see is in a different physical location than I am. This includes the parts of my body that I can see, and so they can be neither me nor a part of me. The focal point of my visual field is located inside my head, between my eyes, so I think of myself as being "in here". This leaves open the possibility that I am some physical system that is a proper part of my body and located inside my head—perhaps my brain, or my pineal gland. But it also seems to leave open the possibility that I am something entirely different from my body that is simply residing there in my head. Thus is born the mind/body problem—what kind of thing am I, and what is my relationship to my body?

Familiar philosophical jargon puts this by saying that I am a "self", and then asking what kind of thing selves are. Philosophers have traditionally attacked the mind/body problem by observing that they have various kinds of self-knowledge and then spinning out the consequences of that knowledge. It should be noted that this is the approach that generated the problem in the first two paragraphs. Although we will stop short of rejecting this approach, we will call it into question, and we will entertain the radical solution to the mind/body problem that we call "nolipsism"—there are no selves. Literally, we do not exist. It will be argued that there is more to be said for this position than might be supposed, although, of course, if it is true then *we* cannot say it.

2. PRIVILEGED ACCESS

How might one address the mind/body problem? One venerable strategy has been to focus on the fact that I seem to have privileged access to myself. This is manifested in several different ways. One is Descartes' *cogito* argument. Necessarily, if I have a thought then I exist. Thus if I think that I exist, it follows that I do exist. This is something I cannot be wrong about. Does this show something interesting about selves? It suggests that we can at least be confident that we exist and hence that nolipsism is false. But it will be argued below that this reasoning is fallacious.

Another kind of privileged access is my introspective access to my own mental states. I can tell, in a way that no one else can, that I am having certain thoughts, that the apple on the table looks a certain way to me, or that my finger hurts. The states and events that I introspect are "mental". Presumably there are corresponding physical states and events occurring in my body and causally responsible for my being in the mental states or for the occurrence of mental events. It would be parsimonious to identify the mental states and events with their physical counterparts, but there are familiar arguments to the effect that they are distinct. Jackson's "Mary argument" (Jackson 1986) seems to establish that what I know when I know how red things look to me is distinct from any physical facts about the physical structure of the world. It is tempting to conclude that mental states are not physical states, but that is a non sequitur. All that follows immediately is that the objects of knowledge are different, i.e., mental propositions and concepts are different from propositions and concepts about the physical counterparts of mental states and events.

Token physicalism argues that mental events are the same events as the corresponding physical events, in the same sense that a flash of lightning is the same event as the corresponding electrical discharge. This is based upon a general view about the individuation of events, and we find it convincing. This has the consequence of identifying mental events with physical events, but leaves other kinds of mental objects unexplained. For example, having or feeling a pain is identified with a neurological event, but the pain itself is distinct from the having of the pain—it is not an event. As such, this strategy does not identify the pain with anything physical. The same point can be made about perceptual images, qualia, etc. There does not seem to be anything physical that is even a candidate for being identified with these mental objects. For example, an image cannot be identified with neural activity. The latter is an event, and if it can be identified with anything mental, it must be the having of the image rather than the image itself. Similarly, a pain can recur. Each occurrence of it is a separate mental event, but the pain is something different from any of its occurrences. Our mental

lives are densely populated with such mental objects. We have introspective access to them, and they are apparently not physical. It seems this should tell us something about what sort of thing we are, although it is not clear exactly what conclusion we should draw from this.

The connection between I and my thoughts, percepts, and other mental states and occurrences is perplexing. I "have" my thoughts and percepts. It is tempting to say that they occur "in me". Presumably my having them has physical counterparts occurring within my body. (Note, however, that the counterparts might not occur within that part of the body that is a candidate for being me, i.e., that is located at the focal point of my visual perception.) What is it that makes them *my* thoughts and percepts? It is not just that their physical counterparts occur in my body. It is at least possible that two different persons, with distinct mental lives, could share a body or part of a body. Consider split brain cases, multiple personalities, and perhaps even Siamese twins. So what makes a thought or percept mine? It seems to be a nonphysical fact about it. If this is right, perhaps it should be concluded that I am not a physical thing.

3. *DE SE* REPRESENTATIONS

The traditional approach to the mind/body problem is to take at face value our internal view of ourselves, and try to find a theory of the relationship between mind and body that accommodates it. Our self-description is accepted uncritically as data. We are going to call this strategy into question, but preparatory to doing this let us to call attention to an important aspect of our self-representation. It is essentially *de se*.

A *de se* representation is one that is expressed with the first-person pronoun "I". The peculiar logic of *de se* representation was brought to the attention of philosophers by a collection of articles by Casteñeda and Perry.[1] We will adapt an example of Perry's to bring out its most important features. Imagine a man, Rudolph Lingens, who finds himself, emerging from a nap, lost and suffering from amnesia in the Stanford Library. He has no beliefs except those he acquires on the basis of his immediate experience. He has no identifying knowledge of himself or his location. His wallet is gone, and there are no signs in sight. He speaks truly when he says "I don't know who or where I am". Suppose, as he wanders the stacks, picking up and flipping through random volumes, he happens on a biography that contains a complete account of his own history. He reads the entire book without an inkling that it is he who is being described. Nothing in the historical account itself, nothing in the objective third-person facts about Rudolph Lingens, tells him that he, himself is that man. He could have a complete account not

only of his own life, but of the entire history of the world, beginning to end, and it would give him no clue as to his own identity. It would be as useful to him in his ignorance as the map of a city would be to a lost man who is unable to identify his location on the map. He might even read with interest how Lingens once woke in the Stanford Library in an amnesiac fog, and think to himself, "Poor bloke, I know how he felt". Unless he knows that he himself is Rudolph, and he himself is in the Stanford Library, nothing in the objective account of the facts could convey that information. There is nothing that the author could have added, employing only descriptive vocabulary, that would do the trick. Just as the lost man needs for someone to point out his location on the map, Lingens needs a pointer to his identity and location in the world. He is missing a crucial piece of information—information he would express with the exclamation "I am Rudolph Lingens and I am in the Stanford Library". That is not captured in an objective account of the history of the world. It must supplement it.

The crucial observation here is that thoughts formulated using "I" and "here" cannot be reformulated using only descriptions of persons and places. The same thing is true of "now". "I", "here", and "now" are non-descriptive designators. Lingens can know every purely descriptive fact there is to know and still not be able to infer who or where he is or what time it is. We will refer to "I", "here", and "now" as *reflexive* designators.

4. REFLEXIVE DESIGNATORS

Let's list the semantic oddities of the *de se* representation "I":

(i) each person can think of himself using "I" without knowing any identifying fact about himself,
(ii) one can possess a complete, objective description of himself, a list of all of one's intrinsic properties and relations to other objects, intrinsically described, without knowing whether "I" applies to it,
(iii) one cannot refer to someone else using "I", no matter how mistaken his self-conception, no matter, even, if everything he believes about himself is true of someone else.

The first two were illustrated in the example of the previous section. We can adapt it to illustrate the third; imagine that Lingens wakes up, not amnesiac, but deluded. Suppose that he wakes up believing that everything Elvis Presley believes of himself is true of him (i.e., Lingens). So he and Elvis have, property for property, identical descriptive self-conceptions, and yet, undeniably, refer to different people when they utter "I".

How this works is a complicated question that requires some delicacy in setting out; the information expressed by Lingens exclamation "I am Lingens, and I am in the Stanford Library" is analogous to that conveyed by the placement of the red dot on a map. The red dot picks out a physical location (in physical space, not on the map) simply by being there. It also indicates a location on the map, and thus coordinates the map with physical space. Similarly, a person's thought refers to a place as *here* simply by virtue of her being there. Her thought refers to a time as now simply by being at that time. And she thinks of herself as *I* simply by being that person. These representations secure their designata non-descriptively—simply by virtue of the cognizer's having a location in time, physical space, or the space of persons. In this, they are like the red dot on the map, although *now* and *here* are more like moving dots. They are like the pointer on a GPS (global positioning system) that moves across the map displayed as the GPS moves.[2]

An observation that will be important later is that reflexive designators can designate different kinds of things, and it may not be more than conventionally determinate what they designate. E.g., does the pointer on my GPS designate itself, or the GPS, or its location, or what? For our use of the GPS, it makes no difference which we say, and we could conventionally stipulate any of these answers. Functional facts about the GPS do not determine a designatum, and they are all that could determine a designatum "objectively". So it is open to us to adopt whatever conventional stipulation we care to adopt, or to leave the matter undetermined, in which case there is really no fact of the matter about what the pointer represents.

5. THE NEED FOR *DE SE*

The mind/body problem arises from the fact that we think of ourselves in a special non-descriptive way that, by virtue of being non-descriptive, leaves open the question "What am I?" That is, we employ a *de se* designator in our routine cognition. It begins to seem mysterious that we should do this. What is the point of having a *de se* designator at all, particularly if it gets us into such a philosophical muddle? What will be argued is that there are purely computational pressures on the design of a sophisticated cognitive agent that can only be satisfied by providing it with various kinds of reflexive designators, including *de se* designators. Sophisticated cognitive agents literally cannot be made to work in a complex environment unless they are equipped with *de se* designators.

These observations involve an important change of perspective on the mind/body problem. The traditional approach to the mind/body problem takes our internal view of ourselves at face value, and tries to find a theory

of the relationship between mind and body that accommodates it. Our self-description is accepted uncritically as data. We are going to approach things in a different way by looking from the outside, in, assuming nothing about selves but that they are designata of *de se* designators, and seeing what can be learned from an examination of the functional role of the designator. What are the conditions under which a being has a need for *de se* designators, and how do they give rise to the problem of understanding the relationship between minds and bodies? This is to adopt the "design stance".

Suppose we want to build a sophisticated cognitive agent—a robot capable of performing intellectual tasks analogous to those performed by human beings. What would this involve? We will assume without argument that a human-like cognitive agent thinks about things in the world in terms of mental representations of them, and that at least some important parts of human rational thought involve manipulating mental representations. We can think of these mental representations as comprising a system of "mental symbols". Building a cognitive agent involves implementing a system of cognition in an underlying physical structure—a physical (perhaps biological) computer. Our claim will be that the need for reflexive designators in general, and *de se* designators in particular, arises from the demands of practical reasoning in a cognitive agent capable of functioning in a complex and unpredictable environment. We assume that practical reasoning consists of: (1) the adoption of goals as the objects of some kind of conative state that we will noncommittally call "valuing"; (2) epistemic reasoning about how to achieve goals; and (3) the selection and execution of courses of action discovered in (2). Rather simple considerations give rise to the need for a mental *here* and *now*, and increasing complexity gives rise to the need for *de se* designators.

5.1 *Now* in Epistemic Reasoning

Perception is only possible in a changing world because, after all, perception changes the agent. Truth in such a world must accordingly be indexed to times, and a cognitive agent that possesses knowledge of the way the world is at different times needs a way of indexing its beliefs to times. One way to do this—the human way—is to include designators for times in the agent's system of mental representations.

It will be useful to contrast various kinds of cognitive agents with a chess-playing computer. The latter could be implemented as a simple agent that plays chess by reasoning about what to do. (Real chess computers don't work this way, but any real chess program could be re-implemented within OSCAR[3] so that the agent uses the same search algorithms but reasons about what moves to make.) The importance of this example is that, as we will

argue, the chess agent is able to engage in practical reasoning while making only minimal use of reflexive designators. If we are to explain reflexive designators in human cognition as arising from the computational needs of human practical reasoning, we must explain how human practical reasoning differs from that of the chess agent, and how that difference gives rise to the need for reflexive designators.

At first blush it seems that the simplest version of the chess agent does not need a mechanism for temporal indexing, because it does not store beliefs about other times. Its beliefs are only about the current state of the chess board. However, if it is to choose its moves on the basis of practical reasoning, then it must be able to conceive of the board having one state at the present time and a different state at some future time. First, its goal (e.g., black wins) is about the future. That is, the goal is that *there will be* a board position of a certain sort (a winning position for black). To plan for the achievement of those goals, the agent has to have beliefs to the effect that different kinds of moves will have specific effects on the board position, i.e., that if the board is initially in a certain position it will subsequently be in another position. This requires being able to distinguish between board positions occupied at different times. However, it does not require the ability to actually think about the times themselves. It requires no more than a temporal ordering of positions. The agent needs a way of representing what comes before what, but this does not require designators for times.

A natural need for temporal designators does not seem to arise until the agent begins to form beliefs about the physics of its environment. Then it needs a way of representing temporal duration. This seems to require temporal designators, i.e., the ability to think about times rather than just the passage of time.

The need for the reflexive mental designator *now* arises from more sophisticated cognitive or computational pressures. *Now* refers to the *current time*. For the chess computer to reason about how to achieve goals, it must be able to distinguish between its current board position and possible future board positions. This by itself does not require a designator for the current time. The belief of the chess agent could instead use a tensed copula, giving it the form "The position is *B*" (as opposed to "The position is *B* at the present time". The tensed copula relieves the agent of the need for a representation of the current time.[4] Given the tensed copula and temporal reference, we can define the reflexive representation *now* as "the time it is". But if the agent has temporal ordering and the tensed copula, without temporal reference, we cannot define a temporal designator for the current time. A reflexive temporal designator can only be introduced when the agent has temporal representations in general, and the latter only seem to be

necessary for the agent to have rather sophisticated physical knowledge of how the world works.

For practical reasoning, the agent must be able to distinguish between the current state of the world and possible future states. This requires at least the tensed copula. The tensed copula can be defined in terms of *now*, viz., "*P* is true" means "*P* is true at the present time". So if the agent has temporal representations in general, then the tensed copula and *now* are interdefinable. It is worth noticing that neither can be defined "descriptively", as "the time that satisfies description *D*". If that were to work, description *D* would have to be a different description for each instant of time, and so there would be no general description that could do the job. Thus very general cognitive pressures require the agent to have some way of thinking non-descriptively of the present time.

5.2 *Here* in Epistemic Reasoning

Perception provides humans with an egocentric view of the world. Visual, tactual, proprioceptive, and perhaps some other modes of perception have a "perspective", and the human agent has a position in space relative to that perspective. Roughly, we perceive the world from where we are. We can imagine agents that differ from us in this respect. For example, the chess agent has input (from the keyboard) that updates its knowledge of the board positions in the game it is playing. But its knowledge is about "the game", "the board", etc. As it is only aware of one game, board, etc., there doesn't seem to be a need for reflexive designators for spatial location. We can similarly imagine an artificial agent with "distributed sensors" that have fixed positions in the world. For example, the agent might reside in a room with video cameras mounted in each corner of the ceiling. The information derived from perception (via the video cameras) would still give perceived objects spatial locations, and that requires a coordinate system, but that coordinate system might be shared by several similar agents all residing in the same room and having physical implementations in different bodies.

5.2.1 Vision

In contrast, human visual perception is perspectival, providing knowledge of objects relative to an egocentric coordinate system. Roughly, this is a polar coordinate system with the agent at the origin. The beliefs the agent acquires via perception are beliefs about what is going on at particular spatial locations identified with reference to this perceptual coordinate system. The beliefs actually make reference to locations in the coordinate system. This requires a way of representing the locations, and hence of

representing the coordinate system itself. One way of picking out a coordinate system is relative to the locations of some specific objects, however we cannot form beliefs about objects until we perceive some objects, and that involves a prior ability to form beliefs about locations in our perceptual coordinate system. So the perceptual coordinate system cannot be anchored conceptually by reference to objects in the world. We must be able to represent locations in this coordinate system before we can form beliefs about objects in the world. The only way to do this is to have a designator *here* that designates the location of the origin of the coordinate system, and a designator *before* (*in front of here*) indicating a direction from here. We also need designators like *up*, and *right* or *left* indicating orientation.

The designators *here*, *before*, and *up* cannot get their content from descriptions relating them to objects in the world, because we must be able to employ these designators prior to acquiring perceptual knowledge of objects in the world. Given a *de se* designator, we might try defining *here* as "where I am now", *before* as "in front of me", and *up* as "in an upward direction relative to me". The first definition is plausible, but the others are not. "In front of me" and "in an upward direction relative to me" already presuppose the directionality and orientation relative to *here* that is being defined.

If we are building an agent, and it is only intended to function in a narrowly circumscribed environment whose general properties we know, we might give the agent built-in knowledge of that environment (an "a priori world model"), including built-in knowledge of its own body. This would make it possible to have a description that picks out the agent's body uniquely, and then we could design the agent's cognitive architecture in such a way that perception gives it beliefs about states of the world located relative to its body (designated descriptively). In this case, *here*, *before*, and *up* can be descriptive designators constructed in terms of a descriptive designator designating the agent's body. Notice, however, that the descriptive designators we choose must play a privileged role in the agent's epistemic norms. The agent's epistemic norms must automatically locate perceived objects relative to the object (body) described. That is necessary for the agent to be able to acquire knowledge of its surroundings simply on the basis of perception. Thus we cannot require the agent to *discover* where, in its visual field, is the object (body) described. If the agent had to do that before judging where perceived objects are, it would not be able to get started. In this sense, the descriptive designator we choose for picking out the body isn't really functioning descriptively.

The biggest problem with designing an agent in this way is that it is "brittle" in the sense that it will not be able to function in an environment

that differs in any way from its built-in world model. The agent will have to judge that perceived objects are located in proximity to the object described by the privileged designator even when things go wrong and nothing fits the description. If the agent subsequently discovers that nothing fits the description, that will defeat all of its earlier perceptual judgments and all of its putative contingent knowledge of the world will evaporate.

If an agent must be able to function in a wide variety of environments with rather unpredictable properties, the use of a descriptive designator is not an option. A "flexible" agent needs the designators *here*, *before*, and *up* as anchors for the coordinate system used by perception, and these designators cannot be descriptive. They must be primitive elements of the agent's cognitive architecture. Objects in the world are represented as having locations picked out by descriptive designators defined in terms of these reflexive designators rather than the reflexive designators getting their content from some objects in the world. The reflexive designators just act as anchors for relating different perceived objects. Once an agent has a fair amount of knowledge of the world, it can ask where *here* is, and answer this with respect to its body or some other interesting objects, but cognition must begin by employing *here*, *before*, and *up* as primitive designators.

So in "flexible" agents, visual perception provides information about *here*, *before*, *up*, and also *now*. *Here*, *before*, and *up* generate a three-dimensional spatial coordinate system, and if *now* is supplemented with temporal reference we get a four-dimensional coordinate system. (Note that for temporal reasoning we need temporal directionality, i.e., past and future directions of time, just as we need *before* and *up* for spatial reasoning.)

5.2.2 Touch

Touch (haptic perception), like vision, is perspectival, locating objects in a polar coordinate system whose origin is centered on the body. However, at least in humans, the origin of the tactual coordinate system is not in the same place as the origin of the visual coordinate system. Introspectively, the origin of the tactual coordinate system is located somewhere on the body below the head. Furthermore, the *before* and *up* dimensions of the tactual coordinate system are often oriented differently from those for the visual coordinate system. For instance, if I am looking over my shoulder, what is before me visually is behind me tactually. And if I am looking between my legs, what is up visually is down tactually. This indicates that there are distinct visual and tactual *here*'s, *before*'s, and *up*'s.

Although vision and touch provide information about the world via separate coordinate systems, we regard the objects perceived tactually to be in the same physical space (and usually to be the same objects) as those

perceived visually. It has often been noted that it is a contingent fact that vision and touch give us knowledge of the same physical space. Presumably we could build an agent that had to discover this fact by a combination of induction and inference to the best explanation. For most agents this is a completely predictable aspect of their environment and so it makes cognition more efficient to simply build this into the agent's cognitive architecture. However, this is more difficult to achieve than might be supposed. The source of the difficulty is the observation that the visual and tactual coordinate systems are different and not even stably correlated. To get a stable correlation we must at least take account of proprioception, which provides information about how the visual and tactual sensors are oriented with respect to each other. Presumably, with this added information, we can build into the agent's cognitive architecture the expectation that vision and touch provide information about a common space and (generally) common objects. Of course, there are visual objects like shadows, rainbows, and holograms that have no tactual correlates, and there are tactual objects like wind or objects felt in the dark that may lack visual correlates, so all of this must be rather complicated. However, we will not pursue the details here.

It is worth noting that although vision and touch are perspectival, locating objects in a polar coordinate system with the agent at the origin, when we think about the physical world abstractly we think of it in terms of a fixed three-dimensional space and we think of ourselves as moving around in it, rather than thinking about it in terms of one of our perceptual coordinate systems.

5.3 *De Se* Goals

An agent only capable of epistemic cognition about the physical world around it does not seem to have need for a way of thinking of itself. This is particularly obvious if it is just an idle spectator rather than a causal force on its environment. So although agents having moderate epistemic sophistication need the reflexive designators *here* and *now*, they do not need *de se* designators. When do *de se* designators become necessary? Our suggestion will be that the need for *de se* designators arises in part from the goal structure of sophisticated practical reasoners and in part from the need to reason about how to achieve goals.

Human goals tend to be personal (although not exclusively so). Goals derive from what the agent values, and valuing is egocentric in humans. Humans tend to value states of affairs in which they themselves play a particular role. If there were a description (e.g., "the first type 17 robot constructed") that is guaranteed to pick out the agent in any world it is apt to be in, its conative machinery could generate valuings of states of affairs

involving that description rather than a *de se* designator, and the resulting goals would be guaranteed to be "about" the agent itself. If the agent also had knowledge (perhaps built-in) about how to achieve such goals, then it could engage in full-fledged practical reasoning without having *de se* designators. However, for general-purpose agents operating in unpredictable environments, or extremely variable environments, there will be no such description. The only way to formulate personal goals for such agents is by using a non-descriptive designator.

Humans have many different kinds of goals. I have low-level goals such as the alleviation of *my* hunger, but also high-level goals concerning such things as *my* country, *my* as yet unborn children, the books *I* will write over the next twenty years, *my* personal appearance, *my* knowledge of astrophysics, *my* summer vacation, etc. Even a goal like world peace is really egocentric. What I value is peace in *my* world among beings like *me*. These goals can only be formulated using a *de se* designator. Agents capable of having such goals must be constructed so that their conative machinery produces valuings of *de se* states of affairs, i.e., produces *de se* goals.

Our conclusion is that *de se* goals are essential in agents that (1) have wide-ranging personal goals (i.e., goals in which they themselves figure in a privileged way), and (2) their operating conditions are sufficiently unpredictable to make it impossible for either evolution or their designer to seize upon a descriptive designator beforehand and build that into their conative and cognitive machinery.

5.4 Knowing about Actions and their Effects

It does no good to have goals unless the agent can figure out how to achieve them. In order to reason about how to achieve a goal, an agent must make judgments about what actions it can perform and what their likely effects are. These ability-judgments are *de se*—in practical reasoning, what is at issue is what *I* can do.

There is a difference between doing something on purpose (intentionally) and doing it accidentally or having it simply happen to you. For instance, there is a difference between your moving your arm and your arm moving without your willing it. For practical reasoning, we want to know what we can do intentionally and what is apt to happen if we do. A simple agent might have this knowledge built into it, but a more sophisticated agent must be able to acquire new knowledge about what it can do as its skills and physical capabilities change. It seems that the judgment that I will be able to do something in certain circumstances is generally based inductively on the observation that I often have done it in those circumstances. This requires me to have the ability to tell (not necessarily infallibly) that I am doing or

trying to do something intentionally (e.g., moving my arm) rather than its just happening to me without my initiating it. It seems that this is something humans can introspect—we can tell what we are trying to do. It is hard to see what other alternative there could be.[5] It seems that the cognitive architecture of a practical reasoner must contain machinery for introspecting what one is trying to do. The output of such an introspection module will be *de se—I* am trying to do such-and-such. Note that a properly equipped agent may be able to make such judgments without knowing what it is to do something intentionally. It certainly need not have at its disposal any kind of philosophical analysis of intentional action. It might not even have the "in principle" ability to find such an analysis. That would not hamper its ability to engage in practical reasoning. All that practical reasoning requires is that the agent makes such judgments and uses them in deciding what to do.

We first generated the need for *de se* representations by looking at egocentric goals and noting that they must be *de se*. It is hard to imagine how we could have an agent none of whose goals are egocentric. But it is worth noting that even if an agent's goals were not egocentric it would still need *de se* representations to reason about what it can do intentionally. So this constitutes a separate source for the need for *de se* representations.

5.5 Reasoning about How to Achieve *De Se* Goals

Reasoning about how to achieve goals requires more than judgments about what we can do. It also requires judgments about what is apt to happen if we do those things. If the goals are *de se*, this requires the agent to engage in epistemic reasoning about *de se* propositions. How is that possible?

I possess several different kinds of *de se* goals. Some are about my inner states—e.g., the alleviation of my hunger or pain. Others are about my body—I want to get a haircut. Still others are not directly about my body but are about things causally connected to my body—I want my children to get a good education. Most involve a mixture—I want to attend a chamber music festival, I want to read a new novel by my favorite author, I want to dine with friends at a new restaurant.

Consider my goal of alleviating my hunger. To achieve this goal, I must learn that my ingesting certain substances will usually be followed by diminished hunger. Furthermore, I must learn that I can ingest such substances by intentionally moving my body in certain ways under specified circumstances. Both of these facts that I must learn are *de se*. To learn that my ingesting certain substances will usually be followed by diminished hunger, I must be able to tell that it is *I* that is doing the ingesting. To do this I must be able to pick out my own body in the world. Similarly, to learn that

I can move my body in certain ways, I must be able to tell that it is *my* body that is moving.

I do things by moving my body or parts of my body in various ways. So to reason about the effects of my actions, I must be able to identify my own body. It is interesting that that does not seem to require me to be able to locate myself (as opposed to my body) except insofar as I am at the same location as my body. Human beings are aided in locating their bodies by the fact that the point from which they see is located on their body, and the movements of their limbs are generally readily apparent perceptually. This makes it convenient for humans to be built so that they regard themselves as being at the focal point of visual perception, and to think of themselves as having physical extremities projecting outwards from that location and enabling them to act upon the world. However, we can imagine cognitive agents in which these matters are not so nicely organized. Consider a "distributed" agent that is confined to a single room and whose perception is provided by video cameras permanently mounted in the corners of the room. The visual field of such a system need not encode information in a polar coordinate system. It can use a straight-forward three-dimensional coordinate system of the sort that humans use to represent physical space. Suppose the seat of cognition for this agent is a box of electronics mounted on the ceiling, and those electronics remotely control robot hands mounted on little electric carts that run around on the floor. This agent will still have *de se* goals and need *de se* beliefs about what it can do, and for this purpose it will need beliefs about the locations of its hands. As the robot hands are able to move around the room and carry out physical tasks, there will be no way to assign them a fixed location in the agent's visual field (its visual representation of the world). How might this agent acquire the kind of *de se* knowledge about its own actions that is required for achieving *de se* goals? One way to do this would be to let proprioception provide the agent with *de se* knowledge (it certainly does in humans). If the agent can tell proprioceptively when it is moving in certain ways, then it could correlate its movements with the movements of a specific body in its visual field, and then inference to the best explanation might lead it to conclude that the movements of that body are *its* movements. This requires that proprioception provides information about bodily movements in a form that enables the agent to identify them with visually perceived bodily movements. Proprioception must yield more than just knowledge of what the movements feel like. It must yield spatial characterizations of a sort that can be compared with the spatial characterizations generated by vision.

As long as the robot hands can be readily perceived, and the agent can sense its hand movements proprioceptively, it can discover inductively which hands it can control. At least in principle, it can then learn inductively

that it can alleviate its "hunger" by backing its robot hands up to a wall socket where their batteries are recharged. The point of this example is that such an agent can engage in practical reasoning about how to achieve *de se* goals without vision providing it with any *de se* beliefs. Introspection must provide *de se* beliefs to the effect that the agent is hungry, and proprioception must provide *de se* beliefs about what it is doing, but vision need not.

If vision does not produce *de se* beliefs, then it provides no direct basis for the agent to make a judgment about where it is. But the distributed agent has no need for such a judgment. All it must be able to determine is where its robot hands are, and that is something it does inductively by discovering which hands it can control. We might be tempted to insist that although the agent does not judge itself to have a location, it nevertheless does. Its location is the distributed location consisting of the locations of all of its robot hands. But why should we say that? What about the seat of cognition mounted on the ceiling? Should that also be counted as part of the location of the agent? If that is to be counted, how about the city power plant that produces the electricity used by the seat of cognition? There does not seem to be any clear line to be drawn between what counts as part of the agent and what counts as facilities supporting the operation of the agent.

It is not clear that the distributed agent actually has a location. This is because it has no need to reason about its own location. In this respect, it is quite unlike human beings. We take perception to locate perceived objects with respect to ourselves, and so conversely we are located in a certain place relative to the objects we perceive. A partial explanation for why we are so constructed is that, unlike the distributed agent, our sensors move around in the world and so cannot, without further inference, locate perceived objects in a fixed three-dimensional reference frame. Because our sensors move, our view of the world must be perspectival. However, at least in principle such perspectival judgments need only locate objects with respect to *here*, not necessarily with respect to *me*. This suggests that it is only a matter of convenience that human perception locates objects relative to the self.

The general lesson in all of this is that in order for an agent to be able to reason about how to achieve *de se* goals, it must have epistemic norms enabling it to identify its own actions in the coordinate system used by its perceptual system. Human epistemic norms do this in part by locating the self at the origin of the visual coordinate system and locating the body (the locus of actions) in close proximity to the self. But it appears that this is just one way to solve the problem. There could be agents that did not locate themselves in physical space at all, and they would still be able to reason about how to achieve *de se* goals.

5.6 *De Se* Memories

The next thing to observe is that to engage in practical reasoning about the achievement of egocentric goals, a cognitive agent must have beliefs that are about itself at different times. First, egocentric goals are about the agent's future situation. To reason about how to achieve them, the agent must form beliefs about what will be true of it in the future if it does various things. These are *de se* beliefs about the future. Second, the agent must reason inductively about what it can do and what the effects of its actions are likely to be. This requires monitoring what one did in the past and what happened to oneself as a result. These are *de se* beliefs about the past. Presumably, our *de se* beliefs about the future are based inductively on such *de se* beliefs about the past.

An agent's access to the past is ultimately via memory. Any other source of historical knowledge, such as the testimony of others, must be validated by appeal to either memory or previously validated sources of historical knowledge. Memory is fallible, just as is the evidence of our senses, but it must provide us with at least defeasible justification for believing what we seem to remember. Otherwise, we would have no access to the past at all.

De se knowledge of our own past states must derive ultimately from *de se* memories. We can also have purely descriptive memories in which we judge that we were one of the characters described, but the latter identity (that we are the person described) is just further *de se* historical knowledge. As a matter of logic we cannot infer *de se* conclusions from a set of purely descriptive beliefs. So if we are to have *de se* historical knowledge, some of it must come in the form of *de se* memories, and we must treat the latter as giving us defeasible justification for believing their pronouncements.[6]

De se memories make it possible for the agent to reidentify itself over time. It can know that *it* is the one that did so-and-so because it remembers doing it. An agent's *self* is the designatum of its *de se* designator. If we encounter a novel kind of agent, like the distributed agent of section 5.5, and we want to know what its *de se* designator designates, we must take its own pronouncements about its self-identity seriously. We have no other access to what it is thinking about. For example, our initial inclination may be to identify the distributed agent (the robot's self) with the set of its robot hands. But suppose the robot hands are removed from the room each night for maintenance, and new hands left in their place. Suppose the robot tells us that it gets new hands each night, but they all work alike—enabling it to plan ahead for the achievement of long term goals that may require the use of its hands over a period of several consecutive days. We ask how it knows this, and it replies that it remembers this happening every night of its "life", and it remembers formulating such long term goals and pursuing them over a

period of several days. If we grant that the robot's *de se* designator does designate something, and we are trying to understand what that is, we have nothing to go on except its reports of its own persistence. If a hypothesis about the robot's self-identity conflicts with our only access to the persistence of the robot, i.e., with its reports of its own persistence, then the hypothesis cannot possibly be warranted. So we would have to reject the claim that the robot's self is identical with the collection of robot hands.

5.7 Reflexive Designators and Computational Pressures

The general conclusion to be drawn from this section is that purely computational pressures deriving from the requirements of situated cognition in a rational agent give rise to the need for the reflexive designators *now*, *here*, and *I*. Purely epistemic considerations require *now* and *here* in agents operating in complex environments. The need for *de se* goals derives from the requirements of practical reasoning in agents with widely varying personal goals, and the need for *de se* beliefs derives from the need to be able to reason about how to achieve *de se* goals. *De se* beliefs are also needed for reasoning about what the agent can do in attempting to achieve goals, and for reasoning about past and future states of the agent.

To better understand the nature of these conclusions, let us make a tripartite distinction. First, we can distinguish between the mental states involved in thought (propositional attitudes) and their propositional objects. Let us take propositions to be the "logical" objects of thought, and give them however much structure that requires. In particular, they may contain various kinds of designators designating individual objects. Thus we do not think of propositions as being sets of possible worlds. What we can noncommittally call "sentences" in our system of mental representation "express" propositions. Propositional attitudes consist of *believing-true*, *hoping-true*, *fearing-true*, etc., propositions. They do this by employing mental sentences in various ways.

The manipulation of mental representations is implemented in a physical computational system—a physical (perhaps biological) computer. Computers can be described at various levels of abstraction, and at some levels it is appropriate to talk about "virtual machines" manipulating symbols. For example, we might write a LISP program that manipulates lists of numerals. Let us call these computer symbols *c-symbols*. We may be begging some questions here against certain construals of connectionism, but it is our conviction that connectionism is best viewed as a theory about lower levels of implementation and a connectionist architecture that correctly models human cognition must make room for a high level description in terms of c-symbols.

Thus we are led to a tripartite distinction between mental representations, c-symbols, and propositions and their constituents. When we have *de se* thoughts, the propositions we entertain contain logical items we can call *de se designators*, and our mental sentences contain "syntactical" items we can call *de se representations*. There will also be *de se c-representations*, which are just computer symbols used in the implementation of *de se* thought. We don't mean to make any metaphysical hay out of these distinctions. We just want it to be clear whether we are talking about mental items, computational items, or the propositions and propositional constituents they represent.

If we are to build an agent, we do that by implementing the cognitive architecture in a physical computational system. In effect, we program a computer that is connected to the world in various ways. What we have described as the computational pressures giving rise to reflexive designators are really remarks about how to program such a computer to enable it to carry out various tasks. What the computational pressures require most directly is dedicated c-symbols that are treated in special ways during cognitive processing. These c-symbols "correspond to" reflexive mental representations and reflexive designators in the corresponding propositions. However, what is needed to implement epistemic and practical reasoning is the c-symbols. The mental representations and propositional designators are there (if they really are) just because of the c-symbols. And what is needed vis-à-vis the c-symbols is that they play a purely computational role in the implemented cognition. The c-sentences generated by the agent's conative and perceptual systems must contain the reflexive c-symbols and the computational processes must make use of that to mesh the outputs of the systems properly, in effect enabling the agent to form goals and acquire c-beliefs about how to achieve them.

6. WHAT AM I?

Now let us return to the mind/body problem. The problem arises from the fact that we think of ourselves in a non-descriptive (*de se*) way. I am whatever my *de se* designator designates. Because my *de se* designator is non-descriptive, it is not transparent what kind of thing it designates. How then can we find out what we are?

It is plausible to suppose that the referent of any mental term is determined by its functional role in thought together with the way in which the agent's body is situated in the world.[7] The latter allows the agent's causal connections to the world to play a role in determining reference. This is a general remark about the contents of a cognizer's thoughts. Applying it to *de se* representations, it follows that the referent of my *de se* representations

has to be determined by my built-in rules for reasoning with *de se* representations together, perhaps, with facts about how my body is situated in the world. If there is a fact of the matter about what kind of thing I am, it must follow from these computational and causal facts about my cognitive system. This is the *determinate reference principle*.

Suppose we build a sophisticated robot. To enable it to engage in sophisticated practical cognition, we must equip it with a *de se* c-symbol, thus enabling it to think (or at least c-think) of itself in a *de se* way. It then becomes an open question what the robot's *de se* c-symbol represents. Just as for human beings, there is a potential distinction between the robot's body and its self (the object of its *de se* c-thoughts). They may be the same thing, but that remains to be determined. If the robot is sufficiently intelligent, it may become very interested in this question. However, in building the robot, there is no need for us to equip it with the resources for answering the question "What am I?" *directly*. The robot will be able to perform its routine cognitive tasks entirely adequately without knowing the solution to the mind/body problem. If there are facts about the robot's cognition that determine the referent of its *de se* representations, and the robot is sufficiently intelligent, then it can in principle solve the mind/body problem. On the other hand, if there are no facts about the robot's cognitive architecture that determine a solution to the mind/body problem, it follows from the determinate reference principle that there is no fact of the matter about what its *de se* representations represent. If there is nothing such that it is a fact that the robot's *de se* representations represent that thing, then there is nothing that they represent. And again, that need be no obstacle to a robot's performing its routine cognitive tasks or getting around in the world. It is a useful fiction for the robot to c-think that "it" exists, but there is no need for that to be true. If the robot c-thinks "I exist" without there being anything that "I" designates, then there is nothing that actually *thinks* "I exist". The robot just c-thinks there is. Of course, the robot's body exists, but that need not be designated by the robot's *de se* c-symbol.

It is clear that the cognitive architecture of an agent with a *de se* representation need not determine its referent "directly", i.e., there is no need for a simple rule built into the agent's cognitive architecture enabling it to immediately conclude "I am my body" or "I am a non-physical being" or "I am a supervenient object, supervening on my body by virtue of my body's computational organization". But it is too quick to conclude that because the agent is not equipped with a simple rule of this sort, there is no answer to the question "What am I?" that is forthcoming from some more complex argument employing general inference schemes that serve the agent elsewhere. In fact, searching for such arguments is exactly the business the philosopher of mind is in.

In evaluating standard philosophical arguments that purport to answer the question "What am I?", it will be useful to consider how uncompelling they are when applied to our robot. Because we are antecedently convinced that we exist, we find such arguments more compelling when we view them from the inside as applied to ourselves than we do when we apply them to a robot.

What kinds of arguments are there that purport to determine the referent of a *de se* designator? Let us rehearse a few familiar ones. The most obvious is an abductive argument alleging that the simplest explanation for what we know about ourselves is that we are identical with our bodies. Here we take it for granted that we exist and that our mental states are determined by the physical states of our body. Given this data, it is explanatory to hypothesize that I am identical with my body.

A view insisting that there is nothing non-physical in the world and hence that we must be the most convenient physical thing associated with our activities, viz., our body, is certainly a simple view. The trouble is, it really doesn't explain everything that we think we know about ourselves. For example, most people believe either that when they die they cease to exist even if their body continues to exist for a while, or that they can continue to exist even if their body is destroyed. In either case it follows that they are not their body. However, it is a little hard to see how to defend either of the premises on which this argument turns.

Another familiar argument from the philosophical literature involves brain transplants.[8] If my brain is transplanted to another body, it is tempting to suppose that I will go with it, in which case I am not identical to my (whole) body. Note that this argument seems to turn on the presupposition that our *de se* memories will go with our brain. It was remarked above that general computational considerations require that we reidentify ourselves by appeal to those *de se* memories. However, for the same reason, intuitions regarding brain transplants are not robust. For example, we can imagine a professional football player who learns that he has an inoperable brain tumor. It is not out of the question that he would opt for a brain transplant so that he can continue to play football, particularly if he is told we can do a core dump of his memories and personality traits to a computer and upload them to his new brain after the operation. Note that the pull of this example also turns in part on the observation that we reidentify ourselves across time by appealing to *de se* memories. By restoring the football player's memories in his new brain, we ensure that, *as he remembers it*, he is still the same football player.

A variant of the brain transplant argument that seems stronger is a kind of *Ship of Theseus* argument. Suppose that at some time in the future many medical procedures are performed like some car repairs are now performed. Doctors maintain a repository of body parts, and if I injure my arm they

simply remove it and replace it with another arm. They repair my damaged arm at their leisure and put it into cold storage to be used for another patient. This might not work with brains, but presumably it would work with most other organs. We can imagine that over time Jones and I, both of whom are accident prone, end up purely by chance exchanging all of our major body parts. The body I then have has a stronger claim to being the same body as the one Jones used to have than it does to being the same body I used to have, but this does not tempt me to conclude that I am really Jones. So it seems doubtful that I am the same thing as my body.

Perhaps a more plausible view would be that I am some part of my body, perhaps my brain or some still smaller seat of cognition. This seems a bit ad hoc, but if there are neurological parts of my body that could not be destroyed without destroying me, this at least avoids the preceding argument.

However, we began this paper with a different argument to the effect that I am not my body. We are now in a position to construct a variant of that argument that is more compelling that the original version may have seemed. Human beings locate perceived objects spatially with respect to themselves. That has the converse effect of locating them with respect to visually perceived objects. In human beings, the location of the self relative to perceived objects is made possible by locating the self at the focal point of the visual field. It is not computationally necessary to do that. There is nothing obviously wrong with building an agent whose cognitive architecture resulted in its locating itself six inches to the left of that focal point. We find that perverse when we try to imagine it, but that is because our cognitive architecture enforces the identification of our location with the focal point. But the only reason for having an agent locate itself in space is to provide a convenient reference point for use in relating perceived objects to one another. Different kinds of agent architectures could work just as well for this purpose.

Human beings have their eyes embedded in the fronts of their heads, and accordingly they locate themselves somewhere inside their heads. That is where it appears to them visually that they are. But imagine a somewhat different kind of creature whose eyes were mounted on the ends of willowy stalks extending outwards some distance from the head. The focal point of the visual field of such an agent might be three feet in front of its head, and it would be natural to construct the cognitive architecture of such an agent so that it took itself to be located at that focal point. The interesting thing about this example is that there need be nothing physical that is at that location. So if the self is genuinely there, then it isn't anything physical.

We can get the same conclusion by imagining a human being with a malformed head that has a big empty space in the middle of it. Suppose that

just happens to be the location of the focal point of that person's visual field. Then for her too, there isn't anything physical where she thinks she is. So identifying the self with a physical part of the body does not explain important beliefs that the person has about herself.

Could we insist that the agent is just wrong about where she is? The only access we have to the designatum of the agent's *de se* designator is her beliefs about herself. We have noted that, as a human being, it is an essential part of her cognitive architecture that she believes she is where she seems to be vis-à-vis her visual field. The agent's epistemic access to the world is via beliefs like "There is an apple on the table before me". The agent cannot forsake her belief about her location with respect to her visual field without giving up such beliefs as this, and giving up all of these beliefs would undercut all of her contingent knowledge of the world. The agent will then be left without any objective information she could use to try to locate herself somewhere else.

At this point, it is useful to consider the distributed agent again. The distributed agent need not have any beliefs about where it is. This is because its visual perception is non-perspectival. One might say that the distributed agent is wherever its robot hands are, but that cannot be right if the robot changes hands every night. We might instead suppose that the agent is where it center of cognition is, viz., in the box on the ceiling, and its effectors are the radios that send out signals controlling the robot hands. Given that the agent has no beliefs about its own location, it is hard to see what could decide this question. In fact, if we ask it where it is, the robot will say, "I don't know what you mean. Physical location is not applicable to me." It seems to us most reasonable to just deny that the distributed agent has a physical location. It is more like a deity that views the room from outside that coordinate system and directly manipulates events in the room. If the agent has no physical location, then of course it is not anything physical.

These days, most of us are physicalists and believe that there is nothing non-physical in the world. But faced with arguments like the above, some philosophers have been tempted to bite the bullet and conclude that the self is non-physical. They have then been faced with the task of explaining what kind of a non-physical thing they might be. We do not feel that convincing answers to this question have been given. Still, one might be convinced that even without an account of what non-physical selves are like, we are forced to conclude that that is what we are.

We don't think so. Let us return to the design stance. If the argument that we are something non-physical is compelling when applied to human beings, it should be equally compelling when applied to robots with the same cognitive architecture. Suppose we want to build a robot that is capable of cognition of human-like sophistication. So we build a physical computational

system, and provide it with a cognitive architecture by suitable programming. All we put into the robot was a bunch of physical stuff. By virtue of the way we programmed it, we provided it with a *de se* c-symbol, but we didn't provide it with anything for the *de se* c-symbol to represent. There is no need for that in order to get the robot's cognition to work properly. A *de se* c-symbol is required as an anchor-point for tying various aspects of cognition together. It enables c-thoughts about perception, conative states, and intentional action to interact with each other in the ways required for sophisticated practical cognition. But for that purpose, it makes no difference at all whether there is anything that the *de se* c-symbol represents. And in building our robot, we have not built in anything for the *de se* c-symbol to represent, so there seems to be no reason at all to think that somehow a shadowy non-physical self sneaked in. Such a hypothesis has no explanatory power. We can explain everything there is to explain about how the robot works without recourse to its having a non-physical self. Positing non-physical selves seems tantamount to positing a ghost in the machine. There is no more reason for thinking there is a ghost in my robot than there is for thinking there is a ghost in my attic.

The preceding considerations reflect the fact that the abductive argument is completely different when applied in the first-person to ourselves and when applied in the third-person to the robot. In applying it to ourselves we take it for granted that we exist and that our mental states are determined by the physical state of our body. Given that data, it would be explanatory to identify myself with some suitable physical structure. But in the case of the robot, the existence of the self is part of what is at issue. It is not part of the data to be explained. The data we have regarding the robot concern its physical constitution and its behavior. Identifying the robot's self with some physical structure is completely unexplanatory. The robot's physical constitution is what it is because we built the robot that way, and the robot behaves as it does because we programmed it to manipulate c-symbols in the way required for sophisticated cognition. To hypothesize a robot self and real thoughts (as opposed to c-thoughts) is completely gratuitous. It buys us nothing. So it is scientifically disreputable to suppose the robot really has a self.

Suppose we all agree that when we are finished building it there isn't going to be anything there but the robot body—the implemented cognitive system. In particular, we are not going to create some kind of mystical non-physical self. Believing this, we may set ourselves the task of enabling the robot to engage in practical and epistemic c-reasoning without having any false c-beliefs to the effect that it is something other than a complicated lump of plastic, silicon, and titanium. What is interesting is—*that cannot be done*! The arguments of section five show that the only way to build a general-purpose

robot capable of practical c-reasoning is to provide it with a *de se* c-symbol and program it to use that symbol in certain specific ways in c-reasoning. The result will replicate those aspects of the human cognitive architecture that lead us into the mind/body problem, and if the robot is smart enough they will lead it there as well. That is, the robot will c-conclude that "it" is not a physical robot, and its c-reasoning will be unexceptional when viewed from the perspective of the epistemic norms implemented in its cognitive system. Norms that are necessary to make practical c-reasoning work also lead the robot inexorably to the (false) c-conclusion that there is something there other than the robot.

For example, consider the *cogito* argument with respect to the robot. The robot can run that argument just like we do. It can c-think "I think", and then go on to c-infer that "it" exists and that "it" is not identical with its body, without either of those c-beliefs being true. The *cogito* fails because the robot can have the c-thought "I exist" without there being anything that has the thought "I exist".

The point of this is that the arguments that led us to conclude that we are non-physical selves are not plausible when we apply them in the third-person to the robot. They just lead us to the conclusion that the robot is wrong in c-believing that "it" (a self distinct from the physical robot) exists. Shouldn't we think that we are like the robot? The same computational pressures that lead the robot to c-believe that it exists lead us to believe that we exist. If there is no reason to think that there is anything there in the robot to make its c-belief true, shouldn't we be equally dubious about ourselves?

7. IS THIS INTELLIGIBLE?

The preceding arguments are, we feel, strong. But the conclusions are perverse. It would be irrational for me to conclude "I do not exist". My conceptual framework mandates believing various things about myself, such as my location relative to my visual field. It follows immediately from these beliefs that I exist. E.g., if I am at a certain location then I exist. I cannot, rationally, give up the belief that I exist.

On the other hand, something similar is true of the robot, but we are inclined to say that the robot's c-belief is false. It makes a difference whether we are thinking of rational agents from the inside or the outside. Thinking of our robot from the outside, we can insist that it is wrong in c-believing "I exist", but we can simultaneously insist that it would be irrational for the robot to c-believe otherwise. But when I think about myself, I cannot get outside of my own epistemic norms. Judging that the robot is rational in

c-thinking that it exists is tantamount in myself to simply c-thinking that I exist. I cannot, rationally, do otherwise.

This remains perplexing, however. It does not seem reasonable to conclude that the robot has somehow come to embody a non-physical self. In order for that to be the case, it would have to be its computational organization that somehow brings that non-physical self into existence, but how could that be? On the other hand, assuming that *I* do exist, it seems that the only possible explanation for this would be that *I am* brought into existence by my body's computational organization. And if I am willing to say this about myself, why should I be reluctant to say it about the robot?

Our conclusion is that we really don't know what to conclude. We lay these arguments out for your perusal, and you should draw your own conclusions (if you exist).

8. CONCLUSIONS AND COMPARISONS

This paper has two parts—a constructive part and a skeptical part. In the constructive part we investigate the logical role of reflexive designators in rational cognition. There is a rich literature on reflexive designators, going back to Castañeda and Perry. The original interest in reflexive designators was from the perspective of the philosophy of language. Pollock briefly investigated their role in practical cognition in Pollock 1988. Perry takes the issue up in some of his recent work,[9] arriving at similar conclusions to Pollock. We have taken the matter further here, correcting various aspects of earlier proposals and arguing that purely computational pressures deriving from the logical structure of rational cognition dictate the need for reflexive designators in sophisticated agents. Furthermore, we have argued that different aspects of rational cognition give rise to the different (temporal, spatial, and personal) reflexive designators.

The skeptical part of the paper derives from the observation that *de se* designators could serve their requisite functional role in cognition without actually designating anything. If we incorporate such designators into the cognition of a robot, there is reason to be skeptical about their designating anything. But then, why shouldn't we be equally skeptical about our own existence? The difficulty is that we cannot be skeptical about our own existence. Our epistemic norms do not allow that. So we have what seems to be a fairly strong argument for nolipsism, but it is not a view we can endorse.

Dennett is well known for having suggested related views, but upon close inspection it is not clear what his views actually are. In Dennett 1981, he proposes a range of cases in which the brain, the body, and the center of the

perceptual perspective come apart. He doesn't explicitly draw conclusions, but seems to gravitate towards locating himself where his brain is. The general recipe for generating these kinds of cases is clear enough. Those considered by Dennett involve an embarrassment of riches—too many material candidates for selfhood, i.e., too many space-occupying hunks of matter with a claim to being me, and in the end, a case in which there are a pair of selves, located in the same place. We add to the pot a range of cases in which there are no natural material candidates for selfhood, nothing that occupies the center of my perceptual perspective, no enduring hunk of matter whose actions I control, nor even, necessarily, anything like a brain, a spatially localized bit of organic matter where computation goes on, and which serves as the causal basis for my experience. And, unlike Dennett, we give arguments from the functional role of *de se* designators rather than merely telling stories.

It is unclear what Dennett wants to conclude from his stories. In some places, he describes the self as a kind of fiction; a unified agent posited as part of a hermeneutic activity designed to explain our own behavior. He writes:

> We are all virtuoso novelists, who find ourselves engaged in all sorts of behavior, more or less unified, but sometimes disunified, and we always put the best "faces" on it we can. We try to make all of our material cohere into a single good story. And that story is our autobiography. The chief fictional character at the center of that autobiography is one's self (Dennett 1992).

But in other places (e.g., Dennett 1989) he says that the self is perfectly real, but abstract—an organization that tends to distinguish, control and preserve portions of the world.

Nolipsism is perhaps most closely allied with a class of views associated with Lichtenberg, Wittgenstein, Anscombe, and sometimes Schlick, for which Strawson coined the term *no-subject views* (Strawson 1959: 95). But it is important to emphasize that we stop short of endorsing nolipsism. The point of the skeptical part of the paper is simply that the earlier constructive account of the functional role of *de se* designators provides the basis for what seems to be a rather strong argument for nolipsism. One would normally be inclined to endorse such a strong argument were it not that our own computational structure (our epistemic norms) make it impossible for us to accept the conclusion.

ENDNOTES

[1] Castañeda 1976; Castañeda 1968; Perry 1977; and Perry 1979. For more recent work by Perry, see Perry 2001.

[2] Modern GPS's often contain digital maps and display the location of the GPS on a small LCD screen.

[3] OSCAR is the artificial rational agent constructed by John Pollock and described in Pollock 1995.

[4] It is perhaps worth noting that the English word "now" is actually an adverb, not a pronoun. "Now" means "at the current time". This relates it closely to the tensed copula. It is unclear whether this observation is of importance.

[5] G. E. M. Anscombe addressed this question at length in Anscombe 1957. But in the end she did not produce an account of how such self-knowledge is possible. Her conclusion was simply that part of what it is to do something intentionally is to know that you are. She gave no explanation for how you can know that.

[6] For a more extended argument to this effect, see Pollock and Cruz 2000.

[7] For a detailed account, see Chapter 5 of Pollock 1989.

[8] See Shoemaker 1963.

[9] Perry 1990 and Perry 1998.

REFERENCES

Anscombe, G. E. M. 1957. *Intention*. Oxford: Basil Blackwell.

Castañeda, H. N. 1968. On the logic of attributions of self-knowledge to others. *Journal of Philosophy* 65: 439-456.

Castañeda, H. N. 1976. On the logic of self-knowledge. *Noûs* 11: 9-22.

Dennett, Daniel. 1981. Where am I? In *Mind's I*, eds. Douglas Hofstadter and Daniel Dennett. Basic Books.

Dennett, Daniel. 1989. The origins of selves: Do I choose who I am? *Cogito*: 163-173.

Dennett, Daniel. 1992. The self as a center of narrative gravity. In *Self and Consciousness: Multiple Perspectives*, ed. F. Kessel, P. Cole, and D. Johnson. Hillsdale, NJ: Erflbaum.

Jackson, Frank. 1986. What Mary didn't know. *Journal of Philosophy* 83: 291-295.

Perry, John. 1977. Frege's theory of demonstratives. *Philosophical Review* 86: 474-497.

Perry, John. 1979. The problem of the essential indexical. *Noûs* 13: 3-22.

Perry, John. 1990. Self-notions. *Logos*: 17-31.

Perry, John. 1998. Myself and I. In *Philosophie in Synthetischer Sicht*, ed. Marcelo Stamm, 83-103. Stuttgart: Klett-Cotta.

Perry, John. 2001. *Reference and Reflexivity*. Stanford: CSLI Publications.

Pollock, John L. 1988. My brother in the machine. *Noûs* 22: 173-212.

Pollock, John L. 1989. *How to Build a Person*. Cambridge, MA: MIT Press.

Pollock, John L. 1995. *Cognitive Carpentry*. Cambridge, MA: MIT Press.

Pollock, John L. and Joseph Cruz. 2000. *Contemporary Theories of Knowledge*, 2nd edition. Lanham, MD: Rowman and Littlefield.

Shoemaker, Sydney. 1963. *Self-Knowledge and Self-Identity*. Ithaca, NY: Cornell University Press.

Strawson, P. F. *Individuals*. London: Metheun.

Chapter 4

SUBSTANCE AND ARTIFACT IN AQUINAS'S METAPHYSICS

Eleonore Stump
St. Louis University

The concept of a substance is fundamental to ancient and medieval metaphysics, but it is not so easy to understand what this notion comes to. It is sometimes taken to be the concept of a thing which can exist on its own, apart from other things. But that this interpretation of the concept cannot be right can readily be seen by reflecting on the fact that substances are typically distinguished from artifacts. It certainly seems as if an artifact such as an ax can exist on its own apart from other things, at least as much as a flower; and yet a flower is a substance, and an ax is not. In this paper, I explore Aquinas's account of the nature of a substance by looking at it in the context of his general metaphysics and by focusing on the way in which he attempts to distinguish substances from artifacts. I begin by exploring his fundamental understanding of the nature of a material thing, and I conclude by considering what Aquinas's account of substance commits him to as regards the relation of a composite to its components.[1]

1. MATTER AND FORM

Aquinas thinks that any macro-level material thing is matter organized or configured in some way, where the organization or configuration is dynamic rather than static. That is, the organization of the matter includes causal relations among the material components of the thing as well as such static features as shape and spatial location. This dynamic configuration or

T. M. Crisp, M. Davidson and D. Vander Laan (eds.), Knowledge and Reality, 63-79.

organization is what Aquinas calls 'form'.[2] A thing has the properties it has, including its causal powers, in virtue of having the configuration it does; the proper operations and functions of a thing derive from its form.[3]

Like many contemporary philosophers, Aquinas also recognizes levels of organization. What counts as matter for a macro-level object may itself be organized or configured in a certain way; that is, it may be possible to decompose the matter of a thing into material and formal components.[4] For Aquinas, the lowest-level material component which counts as matter organized in a certain way is an element.[5] An element is composed of matter and form; but if we conceptually strip away the form or configuration of an element, the matter that remains is not itself a matter-form composite. All that remains when an element is conceptually stripped of its form is prime matter, that is, matter which cannot itself be decomposed further into matter and form.

Prime matter is thus matter without any organization at all, "materiality" (as it were) apart from configuration. When it is a component in a matter-form composite, prime matter is the component of the configured composite which makes it the case that the configured thing can be extended in three dimensions and can occupy a particular place at a particular time. But by itself, apart from form, prime matter exists just potentially; it exists in actuality only as an ingredient in something configured.[6] So we can remove form from prime matter only in thought; everything which exists in reality is configured in some way. For this reason, Aquinas sometimes says that form is the actuality of anything.[7] Configuration or organization is necessary for the existence of anything at all; without form, nothing is actual. Consequently, although matter is not necessary for the existence of a thing, on Aquinas's view, form is. For Aquinas, to be is to be configured.

2. SUBSTANTIAL AND ACCIDENTAL FORMS

Aquinas takes it that the forms of material objects can be divided into two sorts, substantial forms (that is, the substantial forms of the things that are primary substances) and accidental forms. For present purposes we can understand his distinction between these two sorts of forms roughly in this way. The difference between the substantial and the accidental forms of material objects is a function of three things: (1) what the form organizes or configures; (2) what the configuration effects; and (3) what kind of change is produced by the advent of the configuration.

(1) A substantial form of a material thing configures prime matter. An accidental form, on the other hand, configures something which is an actually existing complete thing, a matter-form composite.[8] Or to put the

same point in a different way, if we conceptually strip away a substantial form from a material thing (and don't immediately replace it with another substantial form of some sort), what is left cannot exist in actuality. Nothing that is actual consists only of prime matter plus accidental properties. But if we strip away any particular accidental form, what is left is still an actually existing complete thing, and it remains the same complete thing it was before the accidental form was stripped away. (On the other hand, it is not possible to strip away all accidental forms from a material thing. It is necessary to a material thing that it have accidental forms, even if it isn't necessary that it have one rather than another accidental form.)

(2) For this reason, configuration by a substantial form brings it about that a thing which wasn't already in existence comes into existence. Since any thing that comes into existence exists as a member of a kind, the substantial form of a thing is thus also responsible for a thing's belonging to a particular primary kind or lowest species. On Aquinas's views, every substance is a member of exactly one lowest species or primary kind. (For Aquinas, the species of substances can also be ordered hierarchically under genera, which can themselves be ordered hierarchically under higher genera till one comes to the highest over-arching genus, which is just substance.) Configuration by an accidental form, on the other hand, brings it about only that an already existing thing comes to have a certain property, without ceasing to be the thing (or the kind of thing) it was.[9] Accidental forms are thus responsible for the non-essential properties of a thing; the addition or removal of an accidental form does not alter the species to which the whole belongs or the identity of the whole.[10]

(3) The change produced by the advent of a substantial form is therefore the generation of a substance. The change produced by the advent of an accidental form, by contrast, is only an alteration of one and the same thing.[11]

It is clear from these claims that any material thing which actually exists has a substantial form. But Aquinas's claims about substantial form also imply that no existing material thing has more than one substantial form.[12] A composite which consists of prime matter configured by a substantial form couldn't itself be one component among others of a larger whole configured by yet another substantial form. That is because a substantial form of a material thing configures prime matter; but if a substantial form were to configure what is already configured by a substantial form, then it would be configuring a matter-form composite, not prime matter. (Of course, the new substantial form might simply replace the previous one, but in that case the composite would still be configured by only one substantial form.)

Furthermore, Aquinas's claims about substantial forms limit the way in which already existing things can be combined into a substance. Barnacles

have a substantial form, and so do starfish. If a barnacle attaches itself very firmly to the back of a starfish, that attaching will not constitute the generation of a substance. If it did, there would be one thing—the barnacle-starfish composite—which had more than one substantial form, the form of the barnacle and the form of the starfish.[13] So what the attachment of the barnacle to the starfish effects, on Aquinas's views, is just that two complete things come to have a property or properties which they did not have before, as, for example, the property of being fastened together. The new configuration of the barnacle attached to the starfish will thus be an accidental one. Any case in which two already existing material substances come together into some kind of composite without ceasing to exist as the things they were before they came together will similarly be a case of alteration rather than generation, and the new composite will be configured with an accidental, rather than a substantial, form.[14]

Aquinas holds that any ordinary artifact is configured only with an accidental form. The production of an artifact, such as an ax with a metal blade attached to a wooden handle, brings together already existing things—a metal thing and a wood thing—which in the new composite still remain the things they were before being conjoined. An artifact is therefore a composite of things configured together into a whole but not by a substantial form.[15] Since only something configured by a substantial form is a substance, no artifact is a substance.

3. THE NATURE OF SUBSTANCES

Elements—earth, air, fire, and water—are substances, and so is a material made of one element.[16] Furthermore, different elements can combine to form a compound which is itself a substance.[17] So, for example, earth and fire can combine to form flesh. But they can do so only in case the substantial form of each combining element is lost in the composite and is replaced by the one substantial form of the whole compound.[18] Furthermore, the substances which are compounds of elements, such as flesh or blood, can combine into one thing, such as an animal, only in case these compounds also are not substances in their own right in the newly composed whole.[19] On Aquinas's view, the components of a whole are actual (rather than potential) things existing in their own right only when the composite of which they are components is decomposed.[20] If this were not so, says Aquinas, then there would be as many substances in one thing such as a human being as there are parts in him, a conclusion Aquinas clearly regards as absurd.[21]

An objector might suppose here that, for example, flesh in an animal is the same as flesh existing on its own. Since Aquinas is willing to grant that

flesh existing on its own is a substance, it seems that flesh must be a substance when it is in an animal as well. Consequently, the objector might maintain, Aquinas's principle that there cannot be more than one substantial form in a thing is violated. In an animal, there will be at least both the substantial form of the flesh and the substantial form of the animal.

But this objection to Aquinas fails to take into proper consideration his understanding of form. On Aquinas's view, flesh existing on its own does not have the same form as flesh in an animal. That is because flesh in an animal can perform the functions proper to that flesh in a way that flesh existing on its own cannot.[22] The proper function of flesh (or any other constituent of the whole) is given by the substantial form of the whole. When it exists on its own, without being configured by the form of the whole animal, no component of an animal functions as it does when it is in the whole. And so flesh in an animal, unlike flesh which exists on its own, is configured only by the one substantial form of the animal and not by the substantial form of flesh as well.[23]

In the context of his philosophical theology, Aquinas gives a helpful summary of his views of composition.[24] A composite can be constituted of two or more components in three ways, he says. (a) It can be composed of complete things which in the composite remain as the complete things they were before being conjoined; their conjoining is thus effected by an accidental form such as order or figure. Artifacts such as heaps and houses are composites of this sort; substances are not. (b) It can be composed of complete things that do not remain complete things in their own right but lose their own substantial forms in the resulting composite. A mixture composed of diverse elements is an example of this sort of whole. Mixtures of this sort are substances if they exist on their own,[25] but not if they are themselves components of a substance. (c) It can be composed of things which are not complete things or substances in their own right but which make one complete substance by their union. The coming together of prime matter and substantial form to constitute a substance is his example here.[26]

One implication of these views of Aquinas's is that no part of a substance counts as a substance in its own right as long as it is a component of a larger whole that is a substance. That is because the substantial form which such a part would have if it existed on its own is lost when it becomes part of a composite substance and is replaced by the one substantial form of the composite.[27]

Aquinas's claims about substance here can perhaps be understood with a contemporary analogy. Hydrogen and oxygen are the components of water, and we can decompose a water molecule into hydrogen and oxygen. But when water exists as a whole, as water, we don't actually have hydrogen or oxygen; we have water. Furthermore, the substantial form of water informs

prime matter and not oxygen and hydrogen. That is, the configuration of a water molecule is a configuration of the materiality of the whole molecule; it isn't a configuration of hydrogen added to a configuration of oxygen. If oxygen and hydrogen each kept precisely the configuration they had in their isolated state and were just somehow pushed together, the resulting composite wouldn't be water. In order to get water, the configuration that oxygen had and the configuration that hydrogen had before oxygen and hydrogen conjoined into a water molecule are replaced by a new configuration that includes, for example, the polar co-valent bond between hydrogen and oxygen. Furthermore, the thing that emerges from the joining of hydrogen and oxygen—the water—has characteristics and causal powers different from either hydrogen or oxygen because of that configuration of the whole. Aquinas explains the idea in this way:

> the nobler a form is the more it dominates corporeal matter and the less it is submerged in it and the more it exceeds it in its operation or power. And so we see that the form of a mixed body has a certain operation which is not caused from the qualities of the elements [of which that body is composed].[28]

We can put the general point at issue here the other way around: if we divide a composite substance into its components, we may turn what was one substance into several substances.[29] For example, according to Aquinas, there are simple living things (such as certain worms) which can be cut in half to form two living things of the same sort.[30] During the time that the parts were parts of the composite substance and did not exist on their own, they were not actual substances themselves.[31] There were not actually two worms in the one worm before it was cut in two.

Furthermore, the whole worm ceases to exist when it is divided and the two new worms come into existence.[32] In the case of fission involving animals which are not human, then, Aquinas's view implies that the career of the whole substance lasts only as long as the whole is not fissioned; each of the fissioned parts is a different substance from the whole substance existing before the fission. (What Aquinas would say about the thought experiments of contemporary philosophy in which a human person is fissioned cannot be inferred from this part of his metaphysics alone because there are special characteristics of the form which is the human soul that complicate the case.)

So a material substance comes into existence when prime matter is configured by a substantial form. The components that existed before being woven together by that configuration (if there are components that were previously existing things) cease to exist as things in their own right,[33] and a new thing is generated. Elements are the most fundamental composites of

matter and form, and all other material substances are composed of them. But when different elements come together to form a compound, their substantial forms are replaced by a new substantial form which configures the newly generated whole. An artifact, on the other hand, comes into existence when things which already exist as things in their own right are rearranged in such a way that each of the rearranged parts remains the thing it was, and the whole composite is united by an accidental, rather than a substantial, form. For this reason, Aquinas thinks that the resulting composite is one thing in some weaker sense than is at issue in the case of a substance.

4. SUBSTANCES AND ARTIFACTS

Artifacts are thus things which can exist on their own but are not substances. For that matter, so are parts of substances.[34] A severed human hand is not a substance; but, at least for a while, it can exist on its own. What keeps parts of substances from counting as substances for Aquinas is that a part of a substance isn't a complete thing in its own right. It can be defined only with some mention of the whole in its definition. A hand, for example, is an appendage of a human being. So, at best, for Aquinas, the ability to exist on its own is a necessary but not a sufficient condition for something's being a substance.

It would be helpful to be able to say here with some precision just what the difference between substances and other sorts of things is, in order to shed light on Aquinas's view of the nature of a substance. But it is difficult to give a non-circular analysis of Aquinas's concept of substance or substantial form, in my view.

When Aquinas himself gives a careful characterization of substance,[35] he tends to describe a substance as a thing which has a nature such that the thing can exist on its own.[36] But, of course, one wants to know why this description cannot apply to an artifact. If the answer is that natures are the sort of thing had only by substances, not by artifacts or parts of substances, then 'nature' is a technical term defined in terms of substance, and so the description of substance is circular.

One might take a clue from the preceding description of Aquinas's view of a part of a substance and try adding a conjunct to Aquinas's characterization of substance, in this way: a substance is something (i) which has a nature such that the thing can exist on its own and (ii) which is a complete thing in its own right. But this also is not sufficient to give a non-circular way of differentiating substances from artifacts. There seems to be no reason why we should not think that an artifact is a complete thing unless we have

some understanding of complete thing such that only substances can be complete things.

Given Aquinas's understanding of artifacts as a collection of substances conjoined by an accidental form, we might try adding yet a third conjunct, (iii), which doesn't include mention of another complete thing (such as a primary substance) in its definition.

But it isn't at all clear that even this formula is adequate. An ax, for example, apparently has a nature such that it can exist on its own, and it does seem to be a complete thing; so if it is excluded by this conjunctive definition of substance, it must be in virtue of the third conjunct. But it is hard to see why an ax cannot be defined without mention of the substances which compose it. Why couldn't an ax be defined in terms of its function, for example, rather than in terms of its material components?

Finally, one might try excluding artifacts from the category of substances on the grounds that there is a close connection between being a complete thing and having a substantial form, such that only composites configured by substantial forms are complete things. In that case, artifacts will in fact be excluded. But now the characterization of substance is circular again, since complete things are defined in terms of substantial forms, which only substances have.[37]

It may be that the best clue for finding a non-circular distinction between substance and artifact lies in Aquinas's insistence that substantial forms configure prime matter, but that the parts of an artifact retain their own substantial forms within the larger whole they compose. There are various notions of emergence in the philosophical literature, and they are usually restricted to properties. But for present purposes we can understand the emergence of a whole W roughly in this way: W is an emergent thing if and only if the properties and causal powers of W are not simply the sum of the properties and causal powers of the constituents of W when those constituents are taken singillatim, outside the configuration of W. On Aquinas's account of substance and with this rough understanding of the notion of an emergent thing, a substance is an emergent thing with respect to its parts, which lose their own substantial form in constituting the whole. By contrast, it is much easier to see an artifact such as an ax just as the sum of its parts, and to see the causal powers and the properties of an ax as the sum of the causal powers and properties of the constituents of the ax. Even philosophers who are willing to countenance the notion of emergent things might balk at considering an ax emergent with regard to its parts.

But the promise of this way of distinguishing substances and artifacts in Aquinas's metaphysics is considerably diminished by considering, say, styrofoam. On the face of it, styrofoam appears to be an artifact insofar as it is the product of human design, but it seems closer to water than to axes as

regards emergence. It may be that if Aquinas had known some of the products of contemporary technology, he would have found the distinction between substance and artifact much harder to make crisp and clear. Alternatively, it may be that he would have thought that not all products of human design count as artifacts. Maybe styrofoam is a substance, but one that human beings help bring into existence, in much the same way that human design goes into the production of new breeds of dogs, without its being the case that a dog is an artifact. If Aquinas were willing to countenance such things as styrofoam as substances, then the notion of an emergent thing could be used as the basis for a distinction between substances and artifacts.

5. COMPOSITES AND THEIR COMPONENTS

With this much clarity about Aquinas's understanding of the differences between substances and artifacts, I want now to turn to his views about composition, in particular the relation of a composite whole to those components that make it up.

Perhaps the first point to note is that, on Aquinas's understanding of substantial forms, there cannot be two material substances in the same place at the same time.[38] I have already explained that any thing can have only one substantial form on Aquinas's views. Since any given matter occupying a particular place at a time can be configured by only one substantial form and since only a thing configured by a substantial form is a substance, it is clear that there cannot be two whole material substances coincident in place and time.

The general point here holds also for artifacts.

Suppose, for example, that at t_1 there is a lump of bronze which a sculptor fashions into a bronze statue that comes into existence at t_2. On Aquinas's view, the lump of bronze is a thing whose matter is bronze and whose form is the configuration that makes the bronze a lump. When the sculptor makes the lump into a statue, the matter which is the bronze is preserved; but the configuration which made that matter a lump is lost and is replaced by a new configuration which makes the matter a statue. If the statue is melted down, then the matter of the bronze is preserved, and it may again acquire the configuration of a lump; but the configuration of the statue will be lost, and so the statue will cease to exist. Thus, although they are composed of the same matter, in virtue of having different forms the lump and the statue are not the same thing.

On the other hand, the lump and the statue cannot exist at the same time and place as separate things, because one and the same matter cannot at one

and the same time have the configuration of a lump and the configuration of a statue. (Aquinas therefore subscribes to a dictum also argued for in contemporary philosophy: one thing cannot be itself and another thing.[39])

One might object that in the space occupied by the statue, or the place occupied by the lump, there is bronze as well as a statue or a lump, so that there are after all two material things in one place, whether the bronze is a lump or a statue. But for Aquinas, the bronze considered in itself, apart from the configuration of the statue or the configuration of the lump (or some other configuration), isn't a thing at all. To be a thing requires having a form of a whole of some sort. If the bronze has the form of a statue, the thing that exists is a statue; if it has the form of a lump, the thing that exists is a lump. Without any form of a whole, the bronze is not a thing.

So, for Aquinas, the bronze and the statue are not identical, and yet they are not separate things either. Instead, the statue is a composite material thing which has the bronze as a material constituent. For Aquinas, constitution is not identity. Lynne Rudder Baker summarizes positions of this sort by saying,

> For a long time, philosophers have distinguished the 'is' of predication... from the 'is' of identity. ... If the constitution view is correct, then there is a third sense of 'is', distinct from the other two. The third sense of 'is' is the 'is' of constitution (as in 'is (constituted by) a piece of marble') (Baker 1999: 151).

Although on Aquinas's view nothing is identical to its constituents, the constitution relation is nonetheless a unity relation. For that reason, even if a property or causal power is conferred on a thing just in virtue of its having one constituent or another, the property or causal power is a property or causal power of the whole thing. For example, on Aquinas's views, the substantial form of a substance is responsible for the fact that that substance has a certain nature and certain causal powers associated with that nature.[40] Socrates has the power to reason in virtue of having a human substantial form. Nonetheless, the thing to which the operation of those powers is attributed is the substance—Socrates, for example—and not his particular substantial form.[41] In consequence, rationality is predicated of Socrates *simpliciter*.[42]

In fact, the constitution relation lets us make a distinction among the properties of a composite. The whole can have a property either in its own right or else derivatively, in virtue of the fact that one of its constituents has that property in its own right; and the same point applies, *mutatis mutandis*, to the parts of a whole.[43] Baker emphasizes the fact that in cases in which a whole borrows a property from its parts, the property in question is nonetheless genuinely to be attributed to the whole. She says,

> Borrowing walks a fine line. On the one hand, if x borrows H from y, then x really has H—piggyback, so to speak. .... [I]f I cut my hand, then I really bleed. ... I borrow the property of bleeding from my body, but I really bleed. But the fact that I am bleeding is none other than the fact that I am constituted by a body that is bleeding. So, not only does x really have H by borrowing it, but also—and this is the other hand—if x borrows H from y, there are not two independent instances of H: if x borrows H, then x's having H is entirely a matter of x's having constitution relations to something that has H non-derivatively.[44]

Although Aquinas does not draw this distinction among properties explicitly, his metaphysical views about constitution provide for it, and he relies on it in one place after another. So, for example, he argues that whatever follows naturally on the accidents or the parts of a substance is predicated of the whole substance on account of the accident or part in question. As he puts it in a discussion of the actions of parts and wholes,

> the action of a part is attributed to the whole, as the action of an eye is attributed to a [whole] human being but never to another part [of a human being], except perhaps *per accidens*, for we do not say that a hand sees because of the fact that the eye sees.[45]

And in another place, he explains that a man is said to be curly on account of his hair or seeing on account of the function of the eye.[46] Similarly, in discussing the powers of the soul, the substantial form of a human being, Aquinas says,

> We can say that the soul understands in the same way that we can say that the eye sees; but it would be more appropriate to say that a human being understands by means of the soul.[47]

Here a property (understanding) of a metaphysical part, the soul, and a property (seeing) of an integral part, the eye, are transferred to the whole, the person, which in effect borrows these properties from its parts.

6. ALTERATION AND REPLACEMENT

For all material things other than human beings, Aquinas thinks that the form of the whole comes into existence with the whole and goes out of existence when the material composite ceases to exist. (Although the substantial form of a human being also comes into existence just with the existence of the whole human being, including the whole human body, in Aquinas's view the soul can exist on its own even after the composite as a

whole has ceased to exist. But human substantial forms are the only exception to the general rule for the forms of material things. There are no disembodied forms of cows or axes.)

Since the forms that conjoin such things as cows and axes are also individuated by matter, however, one might wonder to what extent a change in the matter constituting a material thing is compatible with the persistence of the form of that thing and thus with the persistence of the thing itself, for forms which cannot exist apart from the material composite they configure.

In the case of material composites which are substances, the form that conjoins the whole configures prime matter. So any matter coming into the composite loses its own substantial form and is configured by the one substantial form of the whole. Furthermore, on Aquinas's views, the substantial form of a thing is individuated by matter under indeterminate dimensions, that is, by matter considered as indeterminately extended.[48] Consequently, a change in a particular quantity or quality of matter will not affect the identity of the form. Considerable change in matter is therefore compatible with the persistence of the individual substantial form.

The case of artifacts is different, however. In the case of artifacts, the individual form conjoining the whole is also individuated by its matter; but it is an accidental form, and it configures complete material things, rather than prime matter. The composite whole is thus a collection of substances, and these substances are the matter of the whole. Furthermore, new matter coming into the whole isn't given its substantial form by the form of the whole; it keeps whatever substantial form it had before it became part of the composite. Finally, like the forms of all material composites except human beings, the accidental form uniting an artifact comes into existence with the existence of the collection of substances it configures and goes out of existence when the collection ceases to exist. For these reasons, there is some reason to think that, for Aquinas, if all the substances comprising the artifact were removed and replaced by other substances, the original artifact would cease to exist. But if that is right, then the persistence of the form of an artifact, and so the persistence of the artifact, is not compatible with the replacement of all the material parts of the thing it configures.

How many of the material parts of an artifact can be replaced compatible with the persistence of the particular accidental form configuring the artifact as a whole is much less clear.[49] One would, of course, like to have some principled way of drawing the line between a change of matter which is just an alteration in the artifact, and a change of matter which changes the identity of the form of the whole, so that the original form, and consequently the original artifact, no longer exist. As far as I can see, however, nothing in Aquinas's metaphysics mandates a particular way of drawing this distinction. But perhaps this result is not such a bad one. The heap of stones that is

an Egyptian pyramid can survive the replacement of one old stone with a new one. It can't survive the replacement of all the old stones with a whole set of new ones; in that case, we have a replica of that pyramid, and not the original pyramid itself. But perhaps there is no definite answer to the question when in the process of putting a new stone for an old one we have crossed the line from repairing the old pyramid to constructing a replica of it.[50]

7. CONCLUSION

I have here examined elements in that part of Aquinas's metaphysics which has to do with his theory of material substances, in its context in his appropriation of Aristotle's hylomorphic account of material objects. Aquinas's view of the nature of a material substance as composed of prime matter informed by a substantial form has implications for his general metaphysics. In particular, it commits him to the thesis that there is only one substantial form that informs the whole of a substance, and this in turn has implications for his views of the difference between substances and artifacts. Although it is difficult, as I have shown, to find in Aquinas a principled distinction between substances and artifacts which is not circular, I have suggested one way in which to understand that distinction that works for simple artifacts and that might be able to handle all cases if Aquinas were willing to grant that sometimes the products of human artistry and design count as substances rather than as artifacts. In addition, on Aquinas's views of material objects, there cannot be two things coincident in time and place. For him, the matter of a thing is not a thing in its own right when it is part of the whole of which it is a component. On the other hand, it is also the case on his views that a whole is not identical to the things that constitute it. Aquinas's name can therefore be added to the list of those philosophers who hold that constitution is not identity. Finally, according to Aquinas, the fact that the form of a substance is a substantial form which configures prime matter while the form of an artifact is an accidental form that configures things which are substances implies a difference between substances and artifacts as regards their sensitivity to change of components. As long as a thing retains the same substantial form, its matter can vary without a change in the identity of the thing. But because an artifact is configured only with an accidental form which conjoins components that are themselves substances, its identity is more sensitive to a change in matter. Aquinas's account of substances thus implies that in considering whether a material object remains the same when its material components are replaced, we need to make a distinction between those things that are substances and those that are not.

ENDNOTES

[1] Some contemporary philosophers make a distinction between composition and constitution and between components and constituents. This is not a distinction Aquinas recognizes, however, and so in this paper I will use these terms interchangeably. This paper is taken, with revisions, from my 2003.

[2] There is a very helpful discussion of Aristotle's concept of form in Grene 1972. She argues that Aristotle's concept of form is very like the contemporary biological concepts of organization or information. (I am grateful to Shawn Floyd for calling Grene's article to my attention.) For a helpful attempt to explicate a notion at least closely related to the Aristotelian concept of form which is at issue in this part of Aquinas's metaphysics, see Fine 1999. Fine does an admirable job of discussing this notion in the context of contemporary mereology and showing what the Aristotelian notion can do that cannot be done equally well with mereological schemes. He says, "I should like to suggest that we take the bold step of recognizing a new kind of whole. Given objects *a, b, c, ...* and given a relation *R* that may hold or fail to hold of those objects at any given time, we suppose that there is a new object– what one may call 'the objects *a, b, c, ...* in the relation *R* ' " (1999: 65).

[3] See, for example, *Summa contra gentiles* IV.36 (3740). I am here making a conceptual distinction between the organization of a thing and the properties the thing has in virtue of being organized in that way, and in what follows I will sometimes speak of a form's conferring certain properties on the whole it configures.

[4] See, for example, *De principiis naturae* 2 (346).

[5] *De principiis naturae* 3 (354); see also *Sententia super metaphysicam* V.4.795-798 and VII.2.1284. The role of prime matter in Aquinas's metaphysics is sometimes misunderstood because Aquinas's notion of the elements isn't taken into account. So, for example, Peter van Inwagen thinks that he differs from Aristotle (and consequently others, such as Aquinas, who accept the notion of prime matter) because, unlike the upholders of prime matter, he believes that "matter is ultimately particulate" (van Inwagen 1990: 3, 15). But van Inwagen is clearly concerned with the ultimate constituents into which material objects could be actually decomposed. Aquinas also thinks that the ultimate constituents into which a material object can be decomposed are particulate, in the sense that they are matter-form composites which can't be divided into any smaller matter-form composites. Prime matter is never actual; its existence is only potential and conceptual. Consequently, prime matter could never be one of the actual constituents into which a material thing could be decomposed. The division of form from prime matter can occur only in thought.

[6] *De principiis naturae* 2 (349); see also *Sententia super metaphysicam* VII.2.1289-1292.

[7] *De principiis naturae* 1 (340).

[8] For the claims about what substantial and accidental forms configure, see, for example, *De principiis naturae* 1 (339).

[9] For the claims about what the forms bring into existence, see *De principiis naturae* 1 (339).

[10] Lynne Rudder Baker makes an interesting case for the claim that sometimes the primary kind of a thing is given not simply by characteristics intrinsic to the thing, as Aquinas's Aristotelian analysis here suggests, but rather by external, relational or historical features of the thing. (See Baker 2000: 46-58.) I think that she is right on this score, but that her view completes Aquinas's position rather than undermining anything in it.

[11] See, for example, *Summa contra gentiles* IV.48 (3834-3835).

[12] To avoid confusion, it might also be helpful here to emphasize that Aquinas's point is a point about substances. Statues are not substances but artifacts; for Aquinas there can be more than one substantial form in an artifact.

[13] Someone might wonder why one shouldn't say that the new composite has the substantial form of a barnacle-starfish, so that the new composite would after all count as a substance. It helps in this connection to understand the notion of a form as a dynamic organization and a substantial form as the dynamic organization that confers on the organized thing those properties essential to its being a member of a particular species. On Aquinas's view, a substance cannot have more than one substantial form. So if the composite barnacle-starfish were a substance, it would have only one substantial form. In that case, the substantial form of the barnacle and the substantial form of the starfish would be replaced by one single new substantial form. But when the barnacle attaches to the starfish, it certainly seems as if the species-conferring dynamic organization of the barnacle remains the same, and so does the species-conferring dynamic organization of the starfish. The barnacle does not cease to be a barnacle in virtue of its attaching to the starfish, and the starfish does not cease to be a starfish in virtue of having a barnacle attached to it. Consequently, each of them retains its original substantial form after the barnacle's attaching itself to the starfish, and there are two substantial forms in the composite. For this reason, the resulting composite does not have its own substantial form and does not count as a substance.

[14] See, for example, *Summa theologiae* IIIa.2.1.

[15] *De principiis naturae* 1 (342).

[16] Cf. *De principiis naturae* 3 (354), where Aquinas talks about water being divided into water until it is divided into the smallest bits that are still water, namely, the element *water*.

[17] See, for example, *Compendium theologiae* 211 (410), where Aquinas discusses the case in which the combination of elements constitutes a complete inanimate thing which is a suppositum, that is, an individual in the genus of substance.

[18] *Summa contra gentiles* IV.35 (3732); cf. also *Sententia super metaphysicam* VII.17.1680 and VII.16.1633.

[19] See, for example, *Compendium theologiae* 211 (409-410) where Aquinas discusses the way in which elements combine and uses the examples of flesh and a hand to make this point.

[20] *Sententia super metaphysicam* VII.16.1633.

[21] *Summa contra gentiles* IV.49 (3846).

[22] For an analogous attitude in service of a different metaphysical position, see Olson 1997: 135-140.

[23] See, for example, *Summa theologiae* IIIa.5.3, where Aquinas explains that such flesh which is not informed by the substantial form of a human being is called 'flesh' only equivocally, and *Summa theologiae* IIIa.5.4 where he makes the more general claim that there is no true human flesh which is not completed by a human soul. (Cf. *Sententia libri De anima*, II.1.226 and *Sententia super metaphysicam*, VII.9.1519.) See also *Sententia super metaphysicam* VII.11.1519 and *Summa contra gentiles* IV.36 (3740) where Aquinas explains that the substantial form of a thing confers on that thing operations proper to it.

[24] *Summa theologiae* IIIa.2.1.

[25] It would seem that a mixture existing on its own is just another case of the sort of composite listed under (c): a composite of prime matter and substantial form. It may be that Aquinas lists it as a category by itself because he thinks that in a mixture the elements that came together to form it aren't entirely absorbed into the whole but are still "virtually present", that is, present in their powers although not present as substances. Cf. *Summa contra gentiles* II.56.

[26] See also, for example, *Summa contra gentiles* IV.35 (3730-3734), which has a slightly more detailed taxonomy of composites. Aquinas distinguishes there between the composition of one from many which is accomplished only by order, as in a city composed of many houses, and composition accomplished by order and binding together, as in a house conjoined of various parts.

[27] See, for example, *Compendium theologiae* 211 (410-411), where Aquinas explains this general point in connection with the composition of the incarnate Christ.

[28] *Summa theologiae* Ia.76.1.

[29] See, for example, *Compendium theologiae* 212 (418).

[30] *Sententia super metaphysicam* CII.16.1635-1636.

[31] See, for example, *Sententia super metaphysicam* VII.13.1588 and VII.16.1633.

[32] Cf. *Sententia super metaphysicam* VII.16.1635-1636.

[33] The point of saying that they go out of existence as things in their own right is to preclude the misunderstanding that these things cease to exist *simpliciter*. They continue to exist as components of the whole. Analogously, when an apple is eaten, it ceases to exist as an apple, but all its matter continues to exist and (at least for a time) constitutes some of the components within the eater.

[34] See, for example, *Quaestiones quodlibetales* V.2.1, where Aquinas explains why a part of a substance is not itself a substance; see also *Compendium theologiae* 211 (409).

[35] For technical reasons involving medieval logic, it isn't possible for Aquinas to give a definition, in his sense of 'definition', for substance. That is because a definition for him consists in an analysis of the thing to be defined into genus and differentia. But because substance is a genus which doesn't itself belong to any higher genus, it isn't possible to assign substance to a genus. Consequently, substance can't be defined in the usual medieval way.

[36] See, for example, *Quaestiones quodlibetales* 9.3.1 ad 2. In other places, he gives somewhat different characterizations. So, for example, in *De unione verbi incarnati* 2, he describes substance as that to which it belongs to subsist *per se* and *in se*, whereas it belongs to an accident to be in something else. In this passage, he stresses that substance exists not only *per se* but also *in se*, in order to distinguish a substance from a part of a substance. A part of a substance does not exist *in se*; it exists in the whole substance of which it is a part. See also *De unione verbi incarnati* 2 ad 3.

[37] Sometimes Aquinas delineates a substance in terms of an Aristotelian condition for substances: a substance is what has an intrinsic principle of motion. This condition looks promising when one thinks of animate substances; but Aquinas also recognizes inanimate substances, and there the Aristotelian condition looks much less promising. Water is a substance, and so is a quantity of water. It has an intrinsic principle of motion insofar as it is naturally inclined to fall to the earth. But the same can be said of an ax. It is true that the ax is inclined to fall to the earth only insofar as it is material, and not insofar as it is an ax; but, then, the same thing seems to be true of a quantity of water.

[38] Aquinas would therefore share Peter van Inwagen's intuition that two objects (or substances) cannot be composed of all and only the same proper parts at the same time (van Inwagen 1990: 5).

[39] Van Inwagen 1994. Aquinas makes the point explicitly in *Sententia super metaphysicam* VII.13.1588; see also *Summa theologiae* IIIa.17.1 obj.1, where Aquinas says that where there is one thing and another, there are two things, not one. (Although this line occurs in the objection, it is not disputed in the reply to the objection.)

[40] See, for example, *Summa contra gentiles* IV.36 (3740) where Aquinas claims that there is an operation proper to the nature of anything which derives from the substantial form of that thing. See also *Summa theologiae* IIIa.17.2 where Aquinas says that a suppositum is that which exists and a nature is that whereby it exists.

[41] *Summa theologiae* IIIa.2.3.

[42] *Summa contra gentiles* IV.48 (3835).

[43] Baker speaks in this connection of something's having a property independently, rather than in its own right, and she gives a helpful analysis of what it is for anything to have a property independently. See Baker 1999: 151-160.

[44] Baker 1999: 159-160.

[45] *Summa theologiae* Ia.76.1.

[46] *Summa contra gentiles* IV.48 (3835).

[47] *Summa theologiae* Ia.75.2 ad 2.

[48] Cf. *In Boethii De Trinitate* 2.4.2.

[49] In Fine 1999, Fine tries to provide hylomorphic theories of wholes and parts with a distinction which explains and accounts for the fact that sometimes a change of parts is compatible with the continued existence of the whole of which they are parts and sometimes it is not. As Fine puts it, a material composite has both a rigid embodiment (one which cannot be changed compatible with the continued existence of the composite) and a variable embodiment (which can be changed without the composite's going out of existence). Fine's explanation and use of this distinction is helpful, but, as far as I can see, it does not provide all that is needed for a principled way of drawing the line between a change of matter which is just an alteration and one which constitutes generation of something new.

[50] Whether or not there is vagueness in the real world is, of course, a matter of debate among philosophers.

REFERENCES

Baker, Lynne Rudder. 1999. Unity without identity: a new look at material constitution. *Midwest Studies in Philosophy* 23: 144-165.

Baker, Lynne Rudder. 2000. *Persons and Bodies: A Constitution View*. Cambridge: Cambridge University Press.

Fine, Kit. 1999. Things and their parts. *Midwest Studies in Philosophy* 23: 61-74.

Grene, Marjorie. 1972. Aristotle and modern biology. *Journal of the History of Ideas* 33: 395-424.

Olson, Eric. 1997. *The Human Animal*. Oxford: Oxford University Press.

Stump, Eleonore. 2003. *Aquinas*. New York: Routledge.

Van Inwagen, Peter. 1990. *Material Beings*. Ithaca, NY: Cornell University Press.

Van Inwagen, Peter. 1994. Composition as identity. In *Philosophical Perspectives*, vol. 8, ed. James Tomberlin, 207-219. Atascadero, CA: Ridgeview Publishing Co.

Chapter 5

EPISTEMOLOGY AND METAPHYSICS

William P. Alston
Syracuse University

1.

Lest the reader recoil in horror from the prospect of someone's setting out to present comprehensive systems of epistemology and metaphysics in a short paper, let me reassure said reader that this is far from my intention. Instead I will be exploring ways in which these two areas of philosophy are interrelated, ways in which they impinge on each other, lean or otherwise depend on each other, and the implications this has for what is possible and appropriate in their pursuit.

I begin by looking at a recurrent refrain in recent philosophy—that epistemology does not count as 'first philosophy'. We hear this from such divergent voices as Willard Van Orman Quine and Richard Rorty, and in less strident tones from such advocates of 'naturalized epistemology' as Alvin Goldman, Fred Dretske, and Hilary Kornblith. On a sensible reading of this dictum, the point is that epistemology cannot be pursued without taking for granted, and employing, various pieces of knowledge (or at least beliefs with a strongly positive epistemic status) that we already possess. The Cartesian project of doubting everything one previously supposed one knew, shutting oneself in a room, and settling on the criteria for knowledge (warranted belief) *before* deciding on what items deserve such appellations, is a quixotic procedure. We cannot move a step in settling questions in epistemology or anything else unless we have something to go on, some

T. M. Crisp, M. Davidson and D. Vander Laan (eds.), Knowledge and Reality, 81-109.

resources to employ in distinguishing plausible from implausible answers, and in evaluating those that fall in the former group. Once we try to imagine ourselves carrying out the Cartesian enterprise, we see how thoroughly impossible it is. It would be like trying to make a choice of what to cook for dinner in the absence of any edibles. A careful scrutiny of Descartes' procedure in the early *Meditations* reveals that he by no means restricted himself to the Spartan regime he announced at the outset.

So epistemology cannot be the very first of our intellectual enterprises. We must already have some knowledge before we can reflect on what it is to know something, just as we must already have some obligations before we are in a position to reflect on what it is to be obliged to do something. In this essay I will use "metaphysics" in an outrageously broad sense to range over anything that we can know about what the world, or some portion thereof, is like. Thus "metaphysical knowledge" will be what might more appropriately be called "factual knowledge". I plead guilty to the charge of misusing the term. My only excuse is that by and large the beliefs about, and knowledge of, what the world is like that I will be thinking of as relevant to epistemological inquiry will be of a very general sort, such as the conditions under which a way of forming beliefs is reliable (along with whether particular ways of belief formation meet those conditions) and the design plan of human beings created by God.

With only this much by way of prologue, I will turn to a couple of examples of epistemologists recognizing that conclusions about the epistemic status of certain kinds of beliefs depend on non-epistemological facts about the way things are, depend on "metaphysical" considerations in my inflated use of 'metaphysical'. First, as is appropriate for a volume with the purpose of this one, consider Alvin Plantinga's case for warranted theistic belief and for warranted Christian belief in his *Warranted Christian Belief* (2000) (hereinafter WCB). And first let's note Plantinga's account of what is required for a belief's being "warranted" (his term of choice for the epistemic evaluation of belief), a status he identifies as that "quality or quantity (perhaps it comes in degrees), whatever precisely it may be, enough of which distinguishes knowledge from mere true belief" (2000: 153).

> …a belief has warrant for a person *S* only if that belief is produced in *S* by cognitive faculties functioning properly (subject to no dysfunction) in a cognitive environment that is appropriate for *S*'s kind of cognitive faculties, according to a design plan that is successfully aimed at truth. We must add, furthermore, that when a belief meets these conditions and does enjoy warrant, the *degree* of warrant it enjoys depends on the strength of the belief, the firmness with which *S* holds it (2000: 156).

He distinguishes the *de jure* question about Christian (and theistic) belief, "Is it warranted to accept Christian (or theistic) belief", from the *de facto* question, is Christian (or theistic) belief true? We may take the former as an epistemological question and the latter as a metaphysical question. What Plantinga goes on to argue is that a positive answer to the *de jure* question depends on a positive answer to the *de facto* question.

Let's first see how this works out for the more generic theistic belief. Section III of Chapter 6 is entitled "The *de Jure* Question is Not Independent of the *de Facto* Question". He begins that section as follows.

> And here we see the ontological or metaphysical or ultimately religious roots of the question as to the rationality or warrant or lack thereof for belief in God. What you properly take to be rational, at least in the sense of warranted, depends of what sort of metaphysical and religious stance you adopt. It depends on what kinds of beings you think human beings are, what sorts of beliefs you think their noetic faculties produce when they are functioning properly, and which of their faculties or cognitive mechanisms are aimed at the truth…. And so the dispute as to whether theistic belief is rational (warranted) can't be settled just by attending to epistemological considerations; it is at bottom not merely an epistemological dispute, but an ontological or theological dispute (2000: 190).

This dependence is then more specifically presented in terms of what Plantinga calls an "Aquinas-Calvin model" (A-C model) of theistic belief formation.

> …the basic idea…is that there is a kind of faculty or a cognitive mechanism, what Calvin calls a *sensus divinitatis* or sense of divinity, which in a wide variety of circumstances produces in us beliefs about God. These circumstances, we might say, trigger the disposition to form the beliefs in question; they form the occasion on which those beliefs arise. Under these circumstances we develop or form theistic beliefs—or, rather, these beliefs are formed in us; in the typical case we don't consciously choose to have those beliefs. Instead, we find ourselves with them, just as we find ourselves with perceptual and memory beliefs (2000: 178-179).

Plantinga goes on to say that when belief in God is formed in accordance with the model it has warrant without being based on inference of any sort.

> On this model, our cognitive faculties have been designed and created by God; the design plan, therefore, is a design plan in the literal and paradigmatic sense. It is a blueprint or plan for our ways of functioning, and it has been developed and instituted by a conscious, intelligent agent.

> The purpose of the *sensus divinitatis* is to enable us to have true belies about God; when it functions properly, it ordinarily *does* produce true beliefs about God. These beliefs therefore meet the conditions for warrant; if the beliefs produced are strong enough, then they constitute knowledge (2000: 179).

And so the A-C model constitutes a theological claim that provides support for the epistemological claim of warrant for belief in God when formed in accordance with the model.

> You may think humankind is created by God in the image of God—and created both with a natural tendency to see God's hand in the world about us and with a natural tendency to recognize that we have indeed been created and are beholden to our creator, owing him worship and allegiance. Then, of course, you will not think of belief in God as in the typical case a manifestation of a belief-producing power or mechanism that is not aimed at the truth. It is instead a cognitive mechanism whereby we are put in touch with part of realty—indeed by far the most important part of reality…. On the other hand, you may think we human beings are the product of blind evolutionary forces; you may think there is no God and that we are part of a godless universe. Then you will be inclined to accept the sort of view according to which belief in God is an illusion of some sort, properly traced to wishful thinking or some other cognitive mechanisms not aimed at the truth (Freud) to a sort of disease or dysfunction on the part of the individual or society (Marx) (2000: 190-191).

To sum up in an oversimplified maxim. "Theistic belief is warranted if and only if it is true."

Now for the more specific Christian belief, which, of course, includes theistic belief in Plantinga's sense of that term. Here Plantinga deploys an "extended Aquinas-Calvin model". As that term implies, it includes the "unextended" model, to reflect the way Christian belief includes theistic belief, but adds to it.

> First, it adds that we human beings have fallen into sin, a calamitous condition from which we require salvation—a salvation we are unable to accomplish by our own efforts…. Our fall into sin has had cataclysmic consequences, both affective and cognitive. As to affective consequences, our affections are skewed and our hearts now harbor deep and radical evil: we love ourselves above all, rather than God. There were also ruinous *cognitive* consequences. Our original knowledge of God and of his marvelous beauty, glory, and loveliness has been severely compromised…. In particular, the *sensus divinitatis* has been damaged and deformed…. Still further, sin induces in us a *resistance* to the deliverances

> of the *sensus divinitatis*, muted as they are by the first factor; we don't want to pay attention to its deliverances. We are unable by our own efforts to extricate ourselves from this quagmire; God himself, however, has provided a remedy for sin and its ruinous effects…. This remedy is made available in the life, atoning suffering and death, and resurrection of his divine Son, Jesus Christ. Salvation involves among other things rebirth and regeneration, a process (beginning in the present life and reaching fruition in the next) that involves a restoration and repair of the image of God in us (2000: 205).

Thus far the extension of the A-C model amounts simply to the core of traditional distinctively Christian belief. Now we come to the part that implies that and how Christian belief, when formed in a certain way, enjoys warrant.

> …we now come to a more specifically cognitive side of the model. God needed a way to inform human beings of many times and places of the scheme of salvation he has graciously made available…he chose to do so in the following way. First, there is Scripture, the Bible, a collection of writings by human authors, but specially inspired by God in such a way that he can be said to be its principal author. Second, he has sent the Holy Spirit…. A principal work of the Holy Spirit with respect to us human beings is the gift of *faith*.... By virtue of the internal instigation of the Holy Spirit, we come to see the truth of the central Christian affirmations. Now faith is not just a cognitive affair… it is a repair of the madness of the will that is at the heart of sin. Still, it is *at least* a cognitive matter. In giving us faith, the Holy Spirit enables us to see the truth of the main lines of the Christian gospel as set forth in Scripture…. Still further, according to the model, the beliefs thus produced in us meet the conditions necessary and sufficient for warrant; they are produced by cognitive processes functioning properly (in accord with their design plan) in an appropriate epistemic environment…according to a design plan successfully aimed at truth; it they are held with sufficient firmness, these beliefs qualify as *knowledge*….(2000: 205-206)

Plantinga's argument for the warrant of theistic and Christian belief on the extended A-C model is, of course, much more detailed and sophisticated than this brief review brings out. But at least what I have presented makes clear that it counts as a way of claiming the dependence of epistemology on metaphysics.

Before going further I must make an important distinction between types of epistemological theses. Continuing to speak in terms of warrant, first there is the very basic and general issue as to what is required for a belief to have warrant. That was spelled out in the above quote from WCB (2000: 156).

As I also made explicit, there is an even less specific thesis about warrant that amounts to identifying it by its epistemological role—that it is that quantity enough of which suffices for making a true belief into knowledge. But that doesn't have enough epistemological meat to make it into my taxonomy.

My second type consists of claims as to the more specific conditions under which the very general conditions specified by the first level can be satisfied. Thus in WCB Plantinga claims that the proper function, etc. conditions laid out at the first level are frequently satisfied by, e.g., memory beliefs, perceptual beliefs, belief formed by rational intuition, and, to get to the specific concern of WCB, beliefs concerning "the great truths of the gospel" formed under the illumination of the Holy Spirit. These more specific beliefs as to (some of) the conditions under which true beliefs satisfy the first level conditions for warrant constitute a second, more specific type of substantive epistemological claims. Finally, as a third still more specific type, we have assertions that one or another particular belief with a particular propositional content is warranted by virtue of satisfying one of the set of conditions specified by the theses of the second type. WCB maintains that such beliefs as that Jesus of Nazareth is God incarnate in human form, that Jesus suffered death on the cross to make satisfaction for our sins and reconcile us with God, and that Jesus rose from the dead can all be warranted by virtue of satisfying the last condition mentioned above, that they are formed by encountering the Bible under the illumination of the Holy Spirit.

I make these distinctions in order to point out that the dependence of epistemology on metaphysics is much more obvious for the second and third types than for the first. And, indeed, Plantinga's pointing out this dependence in WCB is primarily for the second type, and, by derivation, the third. The center of his claim for this dependence in WCB is his position that his "Extended A-C model" of the formation of Christian belief, which he takes to be a case of proper functioning of our cognitive faculties, is itself a theological commitment, and that if that theology is rejected, the epistemological claim to warrant for beliefs so formed is no longer acceptable. And so the epistemological thesis that Christian beliefs can receive basic, i.e., not dependent on reasons, warrant in this way leans on a theological position for its acceptability.

But the first type thesis of the basic requirements for warrant also depends on metaphysics for its support—though not so obviously on the face of it. Since my main concern here is with the second and third types, I will confine myself to noting metaphysical presuppositions on which the force of the first type thesis depends, viz., the thesis that a belief's being warranted is a matter of the its being produced by cognitive faculties functioning properly

in an environment that is appropriate for those faculties, according to a design plan that is successfully aimed at truth. This presupposes that there is a distinction between proper and improper function of cognitive faculties, that what is proper is determined by a "design plan", and that the design plan for human cognitive faculties (or some of them) is aimed at truth. These are substantive assumptions as to what some segment of the world is like, and as such go beyond any evaluative epistemological claims.

2.

For a second example of the recognition of a dependence of epistemology on metaphysics (in my extended sense of 'metaphysics') I turn to another Alvin—Alvin Goldman, one of the pioneers in "naturalized epistemology". Goldman's central term of positive epistemic status is "justified". I would prefer not to conduct this discussion in these terms, since in Alston 1993 I gave reasons for denying that 'justified' succeeds in picking out any crucially important, objective epistemic status of beliefs, thereby setting myself against a large proportion of the Anglo-American epistemological establishment, including my former self. But although I could, at some considerable expense of space, translate Goldman's points into the pluralistic "epistemic desiderata" framework I now support, my concerns can just as well be met by speaking Goldmanese.

I can be more crisp in presenting the Goldman example since his way of resting epistemology on metaphysics is much more familiar than Plantinga's.

The basic idea is that the most crucial requirement for a belief's being justified is that it be formed in a reliable way, one that would generally produce a (much) greater proportion of true than of false beliefs, in situations of the sort that we generally encounter. (This is my way of spelling out what reliable belief formation amounts to, but I do not believe that it seriously diverges from Goldman's way of thinking of it.) There are various subsidiary clauses, of a roughly "internalistic" sort in Goldman's formulations of a theory of justified belief, but since the reliability requirement is the central part of what he takes for justification, I can safely concentrate on that portion.[1] With this caveat we can take reliable belief formation as Goldman's first type thesis of the conditions for the justification of belief.

Then for the second type we get various more specific, but still very general, conditions under which one or another kind of belief formation is more or less reliable. If fully developed this would involve specifications of what makes for more or less reliability for perceptual beliefs, memory beliefs, introspective beliefs, beliefs based on various kinds of reasoning, and so on.

And where do we look for such specifications? In so far as we go beyond "folk psychology", and doing so is not infrequently required for maximum precision and accuracy, we must look to cognitive psychology. It is cognitive psychology that plays the chief metaphysical role in Goldman's recognition of the epistemology-metaphysics dependence. This would or could play a like role in Plantinga's second type theses as well, though there would be theological assumptions in the background even here with respect to the design plan and the proper-improper distinction. But since Goldman has no tendency to throw epistemological bouquets at Christian belief, or other religious belief, the field is left to cognitive psychology (and, secondarily, folk psychology) alone. And so Goldman writes a large and important book entitled *Epistemology and Cognition* (1986), the first part of which is an epistemological development of a first level position on justification and knowledge, and the second part of which mines contemporary cognitive psychology for materials with which to construct a second type account of the conditions under which one or another mode of belief formation is, *inter alia*, more or less reliable.

I shouldn't give the impression that I think that all epistemologists agree with the two Alvins in recognizing the dependence of epistemology on metaphysics. Quite the contrary. A suitably chastened version of epistemology as first philosophy is not dead. Roderick Chisholm, the most distinguished American epistemologist of the second half of the 20th century, was adamant throughout his career in taking epistemic assessments of particular beliefs to be intuitively obvious on their own, not beholden to anything else for support. And he was convinced that the deliverances of such intuitions was a sufficient foundation for erecting a system of epistemology. And many others have followed him in this, including Richard Fumerton, Richard Foley, Richard Feldman, and even others not named 'Richard'. Even if some version of this kind of position can be successfully defended, it does not follow that epistemology is able to deal effectively with a thoroughgoing skepticism (as contrasted with simply ignoring it). But it would show that important epistemological results can be established (*pace* complete skepticism) without calling on any support for metaphysics in order to do so. My aim in this paper is not to decide the issue between epistemologists who advocate the dependence of epistemology on metaphysics and those who deny it. What I want to do, rather, is to go along with the former group and see where their position leads us.

3.

So I assume that some meta-epistemology of the Plantinga sort or the Goldman sort is acceptable. Fundamental epistemological principles can be (warrantedly) arrived at only on the basis, *inter alia*, of metaphysical principles. Well and good. Or is it? Suppose some troublesome critic should point out that we are still not home free with the epistemological conclusions unless we are warranted in accepting the metaphysical bases in question. If we are quite unjustified in accepting the Extended A-C model of (some) Christian belief formation, or in accepting relevant results of contemporary cognitive psychology about the degree of reliability of different ways of forming beliefs, then however much these factual assumptions would support the epistemological conclusions, *if they were warranted*, we are still no further forward. And how do we determine whether the metaphysical claims on which we are relying are sufficiently warranted? Why, of course, only by applying some epistemological principles as to the conditions under which claims like that are warranted. For the question of whether certain claims (metaphysical or otherwise) are warranted (justified) is, obviously, an epistemological question and hence can only be answered on the basis of applying epistemological principles. But not just any epistemological principles. What we require are warranted, justified, acceptable epistemological principles. And see where this has brought us. To get epistemological results we have to rely on metaphysics. But such reliance will support those results only if the metaphysics is itself warranted. And to determine whether that is the case we have to rely on epistemology. So epistemology relies on metaphysics, which relies on epistemology, which relies on metaphysics, which…. It seems that we are chasing our tails, or, more elegantly put, going around in circles. Now if the epistemology we begin by seeking to establish were wholly different from the epistemologies that pop up at all subsequent stages, and likewise for the metaphysical stages, the circularity would not be vicious. But there seems to be no chance of that, at least with the Plantinga version (I will consider the situation for Goldman later). To illustrate this, consider Plantinga making a type 2 epistemological claim, e.g., that Christian beliefs formed in accordance with the extended A-C model (E1) are thereby warranted. His metaphysical-theological reason for this is that such beliefs are formed by the internal instigation of the Holy Spirit (M1). And to what epistemological principle does he appeal to support taking M1 to be warranted? Given his position, he will inevitably appeal to E1. Someone else might take the warrant of M1 to be generated by theological reasoning. But since Plantinga has available as a source of warrant for Christian beliefs a formation in accordance with the extended A-C model, and since M1 is a Christian belief that he takes to be so

formed, and since he takes such a formation as far superior as a basis for ascribing warrant to such beliefs to any sort of reasoning, this leaves him no real alternative to circling back to E1. And this in essence is how this position winds up in a very small circle. No doubt the day of reckoning could be postponed by other thinkers by relying on other plausible bases for ascribing warrant, and other metaphysical supports for this. But since the supply of both plausible grounds for ascribing warrant and plausible metaphysical supports for ascribing warrant on a certain basis is fairly severely limited, the only advantage that can be gained by such moves is making the resulting circle larger, without saving it from being vicious.

One could take this as a conclusive reason for abandoning our "epistemology depends on metaphysics" position, and switching to the Chisholmian "epistemology needs no support from metaphysics" position. But the solace afforded by this move would be short lived. To bring out why, I need to recur to some points developed in "Epistemic Circularity", in Alston 1989 and in Alston 1993. In the first of these discussions I consider what is involved in trying to show that a certain mode of belief formation is a reliable one. (Parallel points could be made for showing that the mode, or its results, enjoy warrant or some other positive epistemic status.) An argument for a positive epistemic status suffers from *epistemic circularity* when it essentially relies on premises that themselves do the intended job of adequately supporting the conclusion only if they themselves enjoy the positive epistemic status attributed in the conclusion. Thus I argued in that article, and in Alston 1993, that any otherwise effective argument for the by and large reliability of the practice of forming certain kinds of beliefs on the basis of sense perception itself makes use or premises that are justified (warranted…) only if that sense perceptual practice is reliable. I then suggested a generalization of this conclusion to other basic modes of belief formation—memory, introspection, rational intuition, various kinds of reasoning, etc. If these considerations are on target, then the Chisholmian approach to epistemology, which takes as its starting point the rational intuition of justificatory statuses of beliefs, is faced with a circularity of the same form as that we encountered with the epistemology-metaphysics relationship, except that this one is purely intra-epistemological. For if the reliability of rational intuition cannot be shown except by relying on premises we acquire from rational intuition, we are depending on the reliability of rational intuition to support the claim that rational intuition is reliable. The difficulty is essentially the same.

Moreover, even if not all otherwise effective arguments for the reliability of a basic practice of belief formation depends on premises that we obtain from that very practice, we still run into either circularity or an infinite regress in trying to establish a positive epistemic status for the outputs of a

doxastic (belief forming) practice. For suppose that we can establish the reliability of rational intuition by a purely empirical argument that contains no premises from rational intuition. Then we will be faced with questions about the credentials of the doxastic practices that yield the premises of this argument, e.g., about sense perceptual practice. And if that practice in turn can be shown to be reliable by an argument the premises of which are yielded by still another practice, say inductive inference, we are faced with the same question about that practice. Clearly, if we continue with this questioning, we will eventually either loop back onto reliance on a practice that appeared earlier in the process, or the process will regress infinitely. Neither of these alternatives is palatable. And since we are far from commanding an infinity of basic doxastic practices, it is the circularity alternative that is our situation. Hence even if we can evaluate epistemological statuses while remaining within the bounds of epistemological inquiry, avoiding metaphysics altogether, we will still be in the position of one or more parts of epistemology depending to certain other parts for their credentials, while the latter in turn depend on the former parts for their credentials.

4.

After this detour into Chisholmian territory we can return to the form of bothersome circularity that we seem to get into by taking epistemology to depend on metaphysics. The next step is to give a more discriminating formulation of just what the difficulty is, and of the view for which it is a difficulty. There are several reasons for denying that "epistemology depends on metaphysics" is an adequate formulation of what gives rise to the circularity. One that was already taken care of in the above has to do with what kind of dependence is involved. It is an epistemological, rather than a conceptual or exegetical dependence. It consists in requiring support from metaphysical propositions if the epistemological claims in question are to be warranted, or enjoy some other positive epistemic status. And another firming up we have already done concerns what epistemological claims are "in question". The distinction between three levels of epistemological theses was designed to handle that indeterminacy. It will be recalled that we decided to focus on the second and third levels, particularly the second (what it takes for a beliefs of a certain sort to be warranted (justified, ……)).

But another loose end emerges when we note that in WCB Plantinga contents himself with the claim that Christian belief, at least Christian belief formed in accordance with the extended A-C model, is warranted *provided it is true*. For the A-C model, or something like, is a crucial component of

Christian belief. And so if Christian belief is true, then it is true that Christian belief can be acquired in accordance with the A-C model, and that provides for an adequate warrant for beliefs so acquired. So the warrant (or rationality) of Christian belief cannot be attacked while leaving open the question of its truth. That is by no means a trivial claim; it goes against a widespread view that whether Christianity is true or not, we can be secure in the view that it lacks adequate warrant. But note that it falls short of providing any sufficient reason for holding that Christian belief *is* true. *If* it is true, it is warranted, but until we have sufficient reason (or warrant) for regarding it as true, we are still left dangling with respect to the question of whether it is warranted. This is not a criticism of WCB. I underline the point in order to introduce a distinction between two readings of "epistemology epistemically depends on metaphysics". That could be read as "epistemological claims (of the sort being considered) are warranted only if certain metaphysical claims would give them adequate support if they are warranted". Call this the *conditional* reading. Or it could be read more strongly by cancelling the 'if', i.e., by taking the condition so introduced as being satisfied. On this stronger *categorical* reading it is "epistemological claims (of the sort being considered) are warranted because they are adequately supported by warranted metaphysical claims".

I make this distinction because on the conditional reading the dependence of epistemology on metaphysics does not give rise to any bothersome circularity. If with Plantinga we stop at saying that if certain metaphysical claims are true (or warranted), then certain epistemological claims are warranted, that doesn't give rise to the circularity adumbrated in section 3. We don't get the "metaphysics depends on epistemology" direction of dependence because nothing is being claimed about the epistemological status of metaphysical claims. It is only the categorical thesis, the one that commits itself to the positive epistemic status of the metaphysical claims in question, that gives rise to the epistemic dependence of metaphysics on epistemology, thereby completing the circle of mutual dependence.

We are still not finished with the task of precising the claim that the epistemic dependence of epistemology on metaphysics gives rise to an apparently unacceptable circularity, for we have still not sufficiently identified just what it is that creates the circularity. To do that we must make still another distinction. To get into this recall the way in which I initially sought in section 3 to render plausible the idea that a bothersome circularity is involved.

> Fundamental epistemological principles can be (warrantedly) *arrived at* only on the basis, *inter alia*, of metaphysical principles. Well and good. Or is it? Suppose some troublesome critic should point out that we are

> still not home free with the epistemological conclusions unless we are warranted in accepting the metaphysical bases in question. If we are quite unjustified in accepting the Aquinas-Calvin model of (some) Christian belief formation, or in accepting certain results of contemporary cognitive psychology about the relative reliability of different ways of forming beliefs, then however much these factual assumptions would support the epistemological conclusions, *if they were warranted*, we are still no further forward. And *how do we determine* whether the metaphysical claims on which we are relying are sufficiently warranted? Why, of course, only by *applying* some epistemological principles as to the conditions under which claims like that are warranted. For the question of whether certain claims (metaphysical or otherwise) are warranted (justified) is, obviously, an epistemological question and hence can only be answered on the basis of applying epistemological principles. But not just any epistemological principles, of course. What we require are warranted, justified, acceptable epistemological principles. And see where this has brought us. To *get epistemological results* we have to rely on metaphysics. But such reliance will do anything to support those results only if the metaphysics is itself warranted. And to *determine* whether that is the case we have to rely on epistemology to tell us. So epistemology relies on metaphysics, which relies on epistemology, which relies on metaphysics, which…. (Italics added in this reformulation.)

The italicized terms and phrases make it explicit that the circularity is thought to arise because we are speaking here of an *inquiry*. What is envisaged is a process of *seeking* true (or warranted) answers to epistemological questions. And when we derive answers to those questions from metaphysical claims, we are not satisfied until we have *determined* whether those latter claims are themselves warranted. And in order to do that we must *apply* epistemological principles to the issue at hand. And hence we are enmeshed in the circle. So, to put it in a nutshell, the alleged circle arises because we are concerned not only to get true (warranted) answers to questions, but also to assure ourselves *that* the answers are true (warranted). In other words, we don't consider the inquiry completed until we not only have warranted beliefs on the first level, but also warranted answers to second level questions about the epistemic status of the beliefs on the first level. It is this demand for warranted second beliefs about the epistemic status of our first level (epistemological and metaphysical) beliefs that lent plausibility to the allegation of a bothersome circularity.

To see that this is not the only way to think of epistemic dependence, go back to another phrase in the section 3 passage. "…we are still not home free with the epistemological conclusions unless we are warranted in the

metaphysical bases in question." And that account of what it takes to be "home free" with an epistemological conclusion simply requires that "we *are warranted* in accepting the metaphysical basis in question". But to *be* warranted in accepting that metaphysical basis (and to accept the initial epistemological thesis *on that basis*) it is not necessary that we have *shown* the metaphysical basis in question to be warranted, whether by "applying epistemological principles" or in any other way. *Being* warranted in a belief (and the same goes for being 'justified', 'rational' and enjoying other positive epistemic statuses of beliefs) is one thing, and *showing* that we are warranted, etc., is quite another. Just as being healthy is one thing and showing that one is healthy is quite another. This is a fundamental and crucial distinction for epistemology, albeit more honored in the breach than in the observance. So the distinction I have been working up to is that between the following:

1. *Being* warranted in believing that *p* depends on *being* warranted in believing that *q* (and basing the former belief on the latter).

2. *Showing* that one is warranted in believing that *p* depends on *showing* (or, more modestly, being able to show) that one is warranted in believing that *q*.

If it is only in sense 2 that the epistemic dependence of epistemology on metaphysics carries with it a like dependence of metaphysics on epistemology, then it is only when we are sufficiently reflective to require dependence in sense 2 that we fall into circularity. And why should not dependence in sense 1 be sufficient to make the general point that epistemology depends on metaphysics?

Uncovering this hornet's nest raises two questions. A. If we are satisfied with sense 1, lower level dependence, are we free from any troublesome circularity? B. Should we, as reflective philosophers be satisfied with a claim to a sense 1, lower level dependence, or must we aim at a claim to a sense 2, higher level dependence? I shall take these questions in order.

5.

On the first level the dependence of epistemology on metaphysics takes the following form: an epistemological thesis (*E*) is warranted only if one or more metaphysical theses (*M*) are warranted. (Call this dependence claim E-M.) What would it take for E-M to give rise to a circularity of the sort about which we are worried? It would have to be the case that E-M could be true only if there is a converse dependence of just this sort on the first level.

That is, only if the very same *M* is warranted only if *E* is warranted. But it does not seem that any such necessary condition must hold. One problem with giving a thoroughly convincing argument for this is that it is not clear in general what it takes for a metaphysical thesis to be warranted. But I suggest that we can marshal a fairly strong argument by considering a certain possibility for what warrants metaphysical theses and considering whether being warranted in one or more epistemological theses must be a part of that. It may be that in doing so we can see a general reason why the answer to this question may be negative in each case.

Consider the possibility that a metaphysical thesis may be warranted because it is self-evident. This status has been claimed for many metaphysical theses, e.g., *properties presuppose a subject in which they inhere*, *nothing exists without a sufficient reason for its existence*, and *no cause is subsequent to its effect* (call this last thesis, 'C'). I am not concerned here to defend the claim that these or any other metaphysical claims are self-evident, only to consider whether this possible source of warrant must contain some warranted epistemological thesis. If the self-evidence of a proposition does confer warrant on it, this is only because a certain epistemological proposition is *true*, viz., that self-evidence is sufficient for warrant. But this does *not* imply that for a subject, *S*, to be warranted in believing that C, and warranted because of the self-evidence of C, *S* must be warranted in believing the epistemological proposition that self-evidence is sufficient for warrant. *S* may just recognize the self-evidence of C and thereby straightaway believe it with no doubt or hesitation, thereby being warranted in so believing. If self-evidence really is sufficient for warrant, and C is self-evident to *S*, then *S* is warranted in believing that C, whatever else, whether of an epistemological character or otherwise, *S* is warranted in believing. To suppose otherwise is to commit a fallacy analogous to that exposed by Lewis Carroll in his famous article, "What Achilles Said to the Tortoise". Carroll's point was that if we start by supposing that *p* logically implies *q* and then think that this requires that the logical relation between *p* and *q* (or the general pattern of inference exemplified by that relation) is part of what is required for the implication to hold, then we will never succeed in giving a complete specification of what logically implies what in any case, unless we arbitrarily cut off the ascent to higher level requirements at some point. The analogue to that here is that if we begin by supposing that the (recognized) self-evidence of *p* is sufficient for *S* to be warranted in believing that *p*, and then go on to suppose that a warranted belief by *S* in the sufficiency of self-evidence for warrant (an epistemological claim) is also part of what is required for *S*'s being warranted in believing that C, we will never succeed in giving a complete specification of what is sufficient for *S*'s being warranted in believing that C, unless, again, we arbitrarily cut off the

multiplication of higher level conditions at some point. The general point here is that whatever it is that is sufficient for warranted belief that *p*, it cannot be the case that a warranted belief in the higher level epistemological proposition that *that* is sufficient is also necessary for warranted belief that *p*. If the latter were necessary, then we were mistaken in supposing that what we first identified as sufficient for warrant was indeed sufficient.

Put the point in another way. If what we originally identified as sufficient for a warranted belief that *p*, viz., *W*, could be sufficient only if accompanied by a warranted belief that *W* is sufficient for warrant, then it is impossible for anything to be sufficient for warrant. The search for a sufficient condition would give rise to a contradiction or to a vicious infinite regress. If we begin by taking *W* as sufficient for a warranted belief that *p*, and are then forced to take *W* as sufficient only if supplemented by *X* (a warranted belief that *W* is sufficient for a warranted belief that *p*), we have the first horn of this dilemma. For this would imply that *W* is sufficient to warrant a belief that *p* only if another condition is required for warrant, and hence is not sufficient. And where *p* is true only if -*p*, it (logically) cannot be true, i.e., is self-contradictory. If, on the other hand, we make the higher level epistemological addition not that what we supposed gave *S* the warrant for believing that *p* (*W*) is sufficient for warrant, but rather a warranted belief that *S* is warranted in believing that *p*, we then escape the fate of self-contradiction involved in the first horn. For now the additional condition does not involve the claim that *W* is not sufficient for warrant, and so we don't fall into supposing that *W* is sufficient only if it is not sufficient. But we run straight into the scarcely less palatable alternative of a vicious infinite regress. For having begun with the idea that *W* (self-evidence) is sufficient for warrant, we find ourselves forced to modify the claim by adding to it (to make it sufficient) a warranted belief that *S* is warranted in believing that *p*. This is to abandon the claim that *W* is sufficient for warrant and to accept something more inclusive as sufficient for warrant, viz., *W* plus a warranted belief that *S* is warranted in believing that *p*. But then, by the same principles, that larger allegedly sufficient condition (call it *X*) must be analogously enlarged in order to be sufficient by the still higher level warranted belief that *S* is warranted in believing that *S* is warranted by *X* in believing that *p*. And that additional enlargement in turn will be sufficient for warrant only if further enlarged by the still higher level belief that…. And so on ad infinitum.

These considerations show that we cannot establish the *necessity* of an epistemological component in what warrants metaphysical beliefs by relying simply on general considerations concerning what is required for the warrant of any belief, at least considerations that involve level ascents of the sorts just exemplified. The reason this is not sufficient to prove that no metaphysical theses can depend for their warrant on warranted belief in epistemological

theses is that it remains conceivable that such warrant might obtain in a way that is quite independent of self-defeating level ascents of the sorts just discussed. And not just barely conceivable. There are some plausible examples. Plantinga himself has argued that if our cognitive faculties are generally reliable (epistemological thesis), then certain metaphysical these are thereby eliminated, specifically a combination of naturalism and the most usual forms of contemporary evolutionary theory. (Naturalism is metaphysical in the strict sense, and evolutionary theory counts as such in my relaxed sense.) To be sure, these are denials of metaphysical theses rather than the genuine article, but this (alleged) epistemic dependence is in the same ball park. And Plantinga has also suggested an epistemological argument for the existence of God, based on the claim (a companion to the argument against evolutionary naturalism just mentioned) that a designer God would provide the best explanation for the general reliability of our cognitive faculties. Even if this is so, he would presumably not claim that it renders theism *warranted,* so that, if true, a belief in theism would *thereby* count as knowledge; but if the argument is cogent, it would at least bestow some kind of positive epistemic status on theism, and that is a kind of dependence of theism on a warranted epistemological thesis for its warrant. Moreover realism about the physical universe (metaphysical thesis) could with some reason be claimed to be a necessary condition of the supposition that we can obtain knowledge of the physical universe in the ways we generally take to be successful at that (epistemological thesis). So it is a live possibility that some metaphysical beliefs could derive positive epistemic status from epistemological beliefs that have positive epistemic status.

But even if that derivation holds in some cases, that is a long way from establishing a circular epistemic dependence of epistemology and metaphysics. What is required for that is not just particular cases of each direction of epistemic dependence. The circular dependence that looks to be troublesome is one in which we have an epistemological thesis, *E*, that depends on for its warrant on the warrant of a metaphysical thesis, *M*, and *M* itself in turn depends for its warrant on the warrant of *E*. And furthermore it must be the case that *E* could not be warranted without *M*'s being warranted, and *M* could not be warranted without *E*'s being warranted. We need a *necessary reciprocal* epistemic dependence between two specific propositions. And the above considerations fail to establish this for first level dependence. Even if in some cases warrant or justification or whatever for a metaphysical thesis is derived from warrant or whatever that attaches to epistemological theses, that is a long way from showing that we get necessary reciprocal epistemic dependence between a pair (or many pairs) of propositions, one epistemological and the other metaphysical.

So if we are confronted with a problem about circularity. it obtains only when we ascend to a second level, when we are not content with its simply being the case that we are warranted in a given case but demand that it be shown (rendered plausible, confirmed, adequately supported…) that we are so warranted. Please note that this demand for establishing that we are warranted in certain beliefs does not itself involve the self-defeating level ascent that I was exposing above. One need not fall into any sort of confusion or contradictions or regresses by simply seeking to establish particular second level epistemological conclusions, as well as first level non-epistemological conclusions. The self-defeating enterprise rejected above arises only when we are *forced* to a higher level showing at *each* level we attain.

This brings us to the second question distinguished at the end of the last section. Are there legitimate motives for attempting to answer higher levels questions about the epistemic status of our beliefs on a lower level? The answer can be short and sweet. Insofar as we are reflective beings, we will sometimes seek answers to such questions, and the more reflective we are the more often we will do so. This is in no way in conflict with what I said earlier is a fundamental though oft neglected point about epistemology—that one can be warranted (justified…) in a belief without being warranted (justified…) in supposing oneself to be. Granting that, one can reasonably and legitimately be interested in whether one is warranted in a belief as well as being interested in, e.g., whether the belief is true or in whether the substance one is looking at is cadmium if that is the content of the belief in question. And with respect to the particular problem under scrutiny here, one can reasonably and legitimately be interested in the epistemic status of a metaphysical thesis used to support an epistemological principle, even if that judgment of that epistemic status itself requires the support of just the same metaphysical thesis just mentioned. And so even if a bothersome reciprocal, circular dependence of epistemology and metaphysics arises only on higher level reflection and questioning, it remains a problem that is well worth exploring.

My next task is to nail down the way in which what I have just presented as giving rise to an apparently bothersome circularity applies to both of the Alvins. Plantinga presents us with an ideally clear and simple case in which the very metaphysics that is invoked to support the epistemological claims itself depends for its support on those same epistemological claims. The claim that Christian belief, when acquired in the extended A-C model way, is warranted (CB) is supported by the theological ("metaphysical") claim that this is a way in which Christian belief can be, and is, acquired (CM).

And CM in turn is supported by the claim that it is warranted. Indeed, since CM is a part of the Christian beliefs the warrant of which set off the whole circular process, we may as well say that CM is supported by warranted CB, as the whole supports the part. And so we have a simple, straightforward case of a reciprocal epistemic dependence of an epistemological thesis, CB, and a metaphysical thesis, CM.

The Goldman case is not quite so simple. The closest analogue to the above would be one in which the results of cognitive psychology, or the epistemically relevant sub-set thereof (leaving folk psychology out to simplify the presentation), plays the same role as Christian belief does for Plantinga. We begin with the epistemological claim that the results of cognitive psychology are warranted (CP). That is supported by the "metaphysical" claim that the procedures of cognitive psychology (to idealize considerably) constitute a reliable way of producing beliefs that are, in general, true (PR). And this, in turn is supported by the claim that those beliefs are warranted (the original epistemological thesis, CP), and hence are likely to be true. This last support relation is not as strong as the whole-part relation in the Plantinga case; PR is not a part of CP. But we still get a reciprocal relation of epistemic dependence between an epistemological claim (CP) and a metaphysical claim.

But reciprocal mutual dependence is not restricted to the maximally simple one step circular relationship. It can be more indirect, involving a lengthier process. Here is a Goldmanian illustration of that sort. Take our initial epistemological claim to be that perceptual beliefs when acquired in normal conditions (suppose we can spell out what makes conditions "normal") are prima facie justified (PB). This is supported by an account of the nature of perception and perceptual belief formation that implies that in normal conditions it is a reliable way of forming beliefs (RPB). That account in turn is supported by, *inter alia*, a wealth of empirical evidence acquired by sense perception (EV). And the beliefs involved in accepting that evidence are justified, in turn, by PB. And so PB is supported by RPB, which is supported by EV, which is supported by PB. These two schemata are only a sample of what can be extracted from Goldmanian reliabilist epistemology. We could construct many more with different epistemological claims as starting points. And so our model of reciprocal epistemic dependence of epistemology and metaphysics is not without realization in contemporary philosophy. Hence the bothersome circularity that seems to be involved threatens the cognitive enterprise, and it is well worth seeking to defuse that threat.

6.

So let us proceed to do so. I will begin by recalling my attempt to deal with a more general circularity problem. This problem arises from the fact that any attempt to show that our basic doxastic (belief forming) practices (DP's) are by and large reliable will involve using premises drawn from the very practice under scrutiny, or from another practice that in turn, or…. If we persist long enough along this line, we will be forced into a circularity most simply illustrated by the fact that eventually the practice with which we began is assumed to be reliable by virtue of accepting premises drawn from that practice.

In reacting to this situation we are confronted with a version of the level distinction that popped up with respect to the epistemology-metaphysics relationship. There we saw that as long as we are content with its being the case that the metaphysical theses on which basic epistemological principles depend for their warrant are themselves true and warranted, we can avoid any reciprocal epistemic dependence of metaphysics on epistemology. But as soon as we seek higher level knowledge or warranted belief as to the epistemic status of those metaphysical theses, or as soon as we demand that it be shown that they are true and/or warranted, we do get into an apparently unacceptable mutual epistemic dependence of basic epistemological theses and basic metaphysical theses. And so it is with the more general problem. One can show that, e.g., sense perceptual doxastic practice (SP) is reliable by deriving that conclusion by a cogent argument from premises (even premises arrived at by SP) that are warranted, so long as we do not require that those premises be shown to be warranted, or require that we be warranted in supposing them to be warranted. But as soon as we try to satisfy any such higher level requirements, we are forced into an apparently unacceptable circularity. For, in seeking to validate the epistemic credentials of the premises in question, which are drawn from SP, we are forced to presuppose the reliability of the very doxastic practice the reliability of which the argument was invoked to support.

In Chapter 4 of *Perceiving God* (1991), I sought to deal with this dilemma by invoking what I called there a "practical rationality" argument. Since any otherwise cogent argument for the reliability of a basic doxastic practice is infected with epistemic circularity, we must find some other way of dealing with the problem if we are not to relapse into a thoroughgoing scepticism.[2] The other way I accepted there was an argument that it is "practically rational" to engage in, and take as generally reliable, any doxastic practice to which we find ourselves firmly committed. Briefly stated, the argument runs as follows. Given that a complete abstention from belief formation is not a viable alternative, we must form beliefs in some

way or other. But even if we were able to replace some or all of our present DP's (and presumably we are not able to do this with the most firmly established ones such as SP), we would no more be able to establish their reliability without epistemic circularity than we are able to do this with their predecessors. And the cost of such massive changes in our cognitive lives, even if possible, would be enormous. Hence the only practically rational course is to stick with what we have until and unless one or another of our DP's turns out to be riddled with persistent and massive internal discrepancies or persistent and massive contradiction of its outputs with the outputs of other equally or better established DP's. Though this falls short of a "theoretical" establishing of the reliability of our basic DP's, it at least shows that it is practically rational to use them and to take them as reliable. And hence it is the best we can do, given our human condition.

Unfortunately, or, following Socrates fortunately, this strategy turned out to be ill-advised. For one thing, it was subjected to devastating criticism by Al Plantinga in WCB, Ch. 4 section II (it's amazing how this name keeps popping up in this paper). Al constructed a complex and ingenious argument that involved attributing to me a conception like Rawls's notion of an "original position" from which basic commitments can be evaluated. Clearly, if I were to construct an argument from such a position, I had to have something to go on to use as premises. And that gives rise to the question of just what it is legitimate to assume for this purpose in the original position, and what, if anything, makes it legitimate rather than arbitrary. Al argued with customary skill that any choice of such assumptions would, just because of the "originality" of the original position, turn out to be arbitrary.[3]

I accept that Al's argument disposes of my practical rationality defense as I presented it in PG. But I don't think anything as complicated as his argument is required to do the job. By bringing out what is behind the force of his argument, a much simpler way of refuting my contention becomes apparent. What is behind his argument is the point that that my practical rationality argument itself suffers from epistemic circularity. It was designed to be an argument for the practical rationality of our basic DP's generally. And, like any argument, it requires a practical assumption of reliability for some DP's or other in order to obtain premises with a positive epistemic status. Even if the argument is solely based on rational reflection on our cognitive endeavors, it presupposes the general reliability of rational intuition, memory, and certain forms of reasoning. And in fact it depends on much more than that. If it has force not just for the case of the arguer but for human beings generally, as I was assuming, it presupposes that the substitution of different ways of forming belief would be extremely costly for human beings generally. And our assurance that this is so comes from

what we have observed of our fellows, as well as on our own experience. And so it presupposes the general reliability of SP as well. But if we are going to be subject to epistemic circularity anyway, we may as well argue straightforwardly for the reliability of SP and other basic DP's, rather than contenting ourselves with the much weaker conclusion that it is practically rational to take them to be reliable.

But having seen the ill-advisedness of reliance on the practical rationality argument, we can make use of some of the intuitions behind it to devise a simpler way out of the circularity dilemma. The basic point is that if we are engage in rational reflection about anything, there is no real alternative to making use of whatever resources we find ourselves in possession of when we do so. There is no place to start except from where we are at the moment. One can hardly get more truistic than that. And where we are at any moment after we have reached the stage of rational reflection is a place that contains firmly entrenched DP's, as well as all the beliefs (and, we hope, knowledge) about ourselves, our fellows, our physical environment, the way things go with all this, etc., that we have acquired by engaging in the aforementioned DP's. What alternative is there to taking all this for granted for the time being, and until and unless we see sufficient reason for abandoning some of it? I can see only one alternative. Even granting that we have to take something for granted if we are to move one step in rational reflection on issues, there is the "Cartesian" alternative of attempting to reduce what we take for granted to some bare minimum by "bracketing" a large part of it and trying to make do with what remains. Various choices of an austere starting point have been attempted. We might restrict ourselves to self-evident propositions and deductive logic. This is one interpretation of the Cartesian enterprise. Another possibility is to begin with bits of immediate experience and deductive logic, plus, if we become more daring, inductive inference. An even more daring project would add particular perceptual beliefs about the physical environment. But none of these projects have provided what was expected of them. The first one only makes contact with the world of particular things by cheating at various points. The second, involving a phenomenalist construal of physical objects in terms of contingencies in sense experience, has been almost universally abandoned as hopeless. Even if we add SP and its outputs to the initial starting point, we never get to general natural laws, much less high level scientific theories. In addition to the resounding failure of all such projects they are subject to Thomas Reid's "undue partiality" argument, which lays on them the charge of arbitrariness in the choice of starting point, a charge that has a high degree of initial plausibility given the variety of starting points chosen by different philosophers. (Since they are *starting points*, the nature of

enterprise prevents the proponent of each from arguing for the superiority of one choice over the others.) Here is Reid's vivid presentation of this perspective.

> The skeptic asks me, Why do you believe the existence of the external object which you perceive? This belief, sir, is none of my manufacture; it came from the mint of Nature; it bears her image and superscription; and, if it is not right, the fault is not mine: I even took it upon trust, and without suspicion. Reason, says the skeptic is the only judge of truth, and you ought to throw off every opinion and every belief that is not grounded on reason. Why, sir, should I believe the faculty of reason more than that of perception?—they came both out of the same shop, and were made by the same artist; and if he puts one piece of false ware into my hands, what should hinder him from putting another (1970: 207)?

Hence there is no fully reasonable alternative to using *everything* we find ourselves confidently accepting by way of DP's, beliefs, principles, etc. in tackling any intellectual problem with which we are concerned.

The application of this approach to the specific concern of this paper—the reciprocal epistemic dependence of epistemology and metaphysics—should be clear. In deciding epistemological questions of my second and third types (what it takes for a particular positive epistemic status and which beliefs satisfy those requirements) we are free to make use of whatever we take ourselves to know or warrantedly believe about whatever is relevant to the matter at hand—whether this concerns theology, science, common lore, particular matters of observation, or metaphysics in the strict sense. And when questions arise about any of these matters, then, again, we are free to make use of whatever we take ourselves to know or warrantedly believe about whatever is relevant to a considered decision, including what we accept confidently about the conditions under which one or another decision on the matter satisfies appropriate standards of justification, warrant, rationality, or whatever. Since we have no reasonable alternative to using whatever we take ourselves to have in settling whatever intellectual questions concern us, there is no bar to using metaphysical considerations to settle epistemological questions, and using epistemological considerations, *inter alia*, to settle metaphysical questions. On the contrary, in the light of the principle of using what we have to work with, this procedure recommends itself as just what should be expected.

Note that this way out of the dilemma, like its cousin in the more general epistemic circularity problem, resolutely turns its back on any pretension to possess a uniquely privileged position from which to give a guaranteed final word on any question. And, in particular, there is no claim to a uniquely privileged position from which to settle questions of epistemic evaluation.

Any such evaluation, like any position on metaphysical or any other issues, is in principle subject to further review, examination, assessment, and criticism. The position being recommended here fully grasps the fallible character of the human situation.

7.

To be sure, not everyone will rest content with my proposal. I can hear cries of outrage in the offing. I will consider two such protests.

1. "This suggestion is tantamount to giving *carte blanche* to every irrational, unfounded, cockeyed way of settling questions of which human beings are capable, and they are legion. You are proposing to approve of anyone's using whatever way of forming beliefs they are habituated to and feel comfortable with, however outrageous they may be. And you think it is perfectly all right for anyone to use as premises in their reasoning anything they firmly believe, no matter how absurd. In other words, you are abandoning any pretense at holding thinkers subject to epistemological norms, and making a distinction between sound and unsound reasoning, between warranted and unwarranted conclusions. Under the cover of "There is no place to begin except where we are at the moment", you make no distinction between more and less epistemically acceptable "places" from which to set out to settle questions. How is this anything other than the most blatant irrationalism, the most unfettered permissiveness in our cognitive lives?"

My reply to this outburst is in several stages.

A. The first is a straightforward plea of not guilty to the charges. There is nothing in the position sketched in section 6 that carries with it the dire consequences just alleged. The fact that, *faute de mieux*, everyone works (cognitively) with the DP's and the beliefs that one has at one's disposal at the moment has no tendency to place all such resources on a level epistemologically, no tendency to wipe out distinctions between them with respect to justification, warrant, rationality, or other epistemological desiderata. To be sure, anyone who draws such distinctions and considers some thinkers to be epistemologically superior to others in their cognitive endeavors does so by the use of the DP's and beliefs that are firmly entrenched in his/her psyche. And perhaps this would be taken by my imaginary protester as one more example of what he is protesting. But what is the alternative to this other than someone's operating from a position that is itself immune from the chances and changes of human life, a position that has not been built up by socialization that shaped the emergence of the ability to carry on reflective and critical thinking—in short Thomas Nagel's "view from nowhere". But that is not possible for human beings. Possible for God, yes, but not for the likes of

us. Something like this would be possible for us if God were to arrange for us to be born with a impeccably reliable set of DP's, and perhaps an impeccably warranted set of initial beliefs to go with them. But God, for whatever reason, has not chosen to set things up that way. Instead he has so arranged things that each of us undergoes cognitive development that is strongly influenced by the social and cultural matrix in which we live. And even after we reach the age of reflective thinking, we continue to be further shaped by influences from other fallible human beings with whom we are in contact. This being the case, none of us enjoys a view from nowhere, but rather a view that bears the marks of a "somewhere" in which we live and move and have our being. And, the contingencies of human social life being what it is, it is not surprising that there will be differences between the individual perspectives that result. And not just factual differences but evaluative, normative differences as well; some of us turn out to have better DP's at our disposal and more epistemically respectable outputs of our DP's to work with. And when such comparisons are made not everyone will make them in the same way, for whoever makes them is operating from what s/he has to go on. And, again, that is the human condition. But, again, that does *not* imply that every starting point for inquiry is equally good on any relevant dimension of epistemic evaluation. It does imply that any basis for such evaluations is itself subject to epistemic evaluation, but that is the human condition. And if you don't like that, don't blame me, to paraphrase Reid.

B. The second part of the answer is just a corollary to A. There is a sense in which everyone, no matter how irrational or misguided by my, your, or commonly accepted principles, is doing the same in reflective thinking, viz., proceeding on the basis of what seems rational in the way of DP's and what seems warranted, true or likely to be true, in the way of beliefs. It is in the nature of our situation as human beings that this should be so. But just as the fact that everyone competing in a race is, if prepared to make the necessary effort, doing the best they can with the limbs, etc. at their disposal does not imply that they are all equally good runners, so it is here. And it certainly does not imply that they all reach the finish line at the same moment.

2. The second objection I will consider is one that is more familiar to epistemologists and also less emotionally involved than the first. It runs as follows. "Haven't you made the time honored (or dishonored) move in the face of circularity problems, viz., deciding to live with the circularity by embracing a coherentist epistemology according to which all epistemic support is ultimately circular, since it consists in the mutually supportive relations between beliefs in some total, or very large, system of beliefs? True, you did not present the position in these terms. But your references to carrying on reflective thought by the use of what DP's and beliefs you have

available to you at the time show that you are thinking of taking the curse from circular support by placing it in a total system that turns out to be the ultimate object of epistemic evaluation, individual beliefs being assessed in terms of the overall structure of the system, along with how the individual belief fits into it. And isn't this position subject to all the devastating objections frequently urged against coherentist epistemology? Can't there be many equally coherent and mutually incompatible systems of belief? Doesn't resting everything on the internal coherence of systems of beliefs shut us off from "input from the world" that is needed for us to have the chance of having beliefs that are true to the way their objects are? And so on."

This is a natural as well as a powerful objection. And the response to it will be complex. There are several complications to be made explicit. First, we must be clear about the restricted role given to coherence in my position. Remember that we were faced with what seemed to be a troublesome circularity only where we are dealing with higher level questions about the epistemic status of beliefs, dealing with the determination of the right answers to questions about the epistemic status of beliefs on a lower level. Where no such epistemic ascent is involved, nothing I have said suggests that beliefs owe whatever positive epistemic status they enjoy to the coherence of a system into which they coherently fit. There is no pressure from the considerations of this essay to reject a foundationalist account at the lowest level, at least a modest foundationalism that allows some subsidiary role for reciprocal epistemic support. Let me take a moment to spell out how that might work for the epistemology-metaphysics relationship.

To take the most extreme possibility first, the above discussion of that relationship leaves open the possibility that both the metaphysical and epistemological beliefs under discussion have an immediate positive epistemic status, one that in no way depends on support from other beliefs, whether in a coherentist way or otherwise. To keep this from being completely bizarre, this immediate positive epistemic status would have to be prima facie, subject to being overridden by conflict with other beliefs with positive epistemic status, as well as only *partial* (not sufficient by itself to make the belief, if true, a case of knowledge). But even with that qualification, the idea that epistemological, and especially metaphysical, theses could enjoy a significant positive epistemic status that is independent of support from other theses seems hard to credit. So to get a more plausible form of the idea, let's suppose that there are various kinds of beliefs that enjoy prima facie (and perhaps only partial) positive epistemic status, apart from any support they have from other beliefs. These kinds might include (some) perceptual beliefs, introspective beliefs, and beliefs formed by rational intuition. Then by acceptable forms of inference

other beliefs could acquire positive epistemic status from those starting points, though the epistemic status of most of those other beliefs would derive from more or less extended chains stretching back to the ultimate foundations rather than directly from the latter. And in addition the non-foundational beliefs would typically derive from the foundations a positive epistemic status that itself is only prima facie and, perhaps, only partial. And here we are within sight of a plausible non-coherentist model for the epistemology-metaphysics relationship. Let's say that the epistemological and metaphysical beliefs of the sorts considered at the outset of this paper each receive a prima facie, and perhaps partial, epistemic support from chains of support reaching back to the foundations, thought of, let's say, as receiving a prima facie, and perhaps partial, positive epistemic support from experience, apart from any support from other beliefs. Since this "initial" support is only partial, it is in need of further support of a different kind, one that does not count as even indirect support by the foundations. And here is where there is an opening for the reciprocal support that appeared to be troublesome. Because of the reciprocity, this kind of support does not trace back to the foundations, at least not wholly. Consider Plantinga's claim that a warranted status for basic Christian beliefs can be supported by the A-C model of Christian belief formation, and vice versa. What saves this reciprocal support thesis from collapsing into a pure coherence theory is the fact that each side of the reciprocal support relation has some significant degree of positive epistemic status apart from the support by its partner in the reciprocal relationship. Their symbiotic relationship strengthens in each case the positive epistemic status each derives from other sources. This feature of the epistemological structure being suggested is often called a "coherence" feature. The excuse for that term is the reciprocity involved. But the reciprocal relations in question can be purely local. There need be no dependence of these local reciprocities, much less all epistemic statuses of particular beliefs, on the coherence of the total system in which they figure. Hence the standard criticisms of coherentism mentioned above have no application. The position is in no way committed to taking the internal coherence of a system of beliefs to be either sufficient or necessary for the positive epistemic status of each, or even any, particular belief. No appeal to the total system need come into the picture at all. Hence the fact that there can be equally coherent and incompatible systems is no problem for it. And as for input from the world, the fact that an important source of positive epistemic status on this account is *experience* nicely takes care of that desideratum. Moreover, it is crucial to realize that even if an appeal to the coherence of a total system were one element in the epistemology, the system involved is a *dynamic* one, not a *static* one.[4] And so the constant updating from experience that goes on is a crucial part of the picture. Hence

the epistemological structure of human belief and knowledge is not constrained to take a coherentist form in any way that gives rise to crippling objections.

But when we turn to the ultimate issues engendered by pushing reflection on epistemological issues to the point of wondering how we can be warranted (justified) in believing anything, given the pervasiveness of epistemic circularity, a different picture emerges. The answer to these worries suggested above did turn on taking the only alternative to be taking what we find ourselves with in the way of DP's and of beliefs, and working with that pro tem, leaving open the possibility that any piece of that package might have to be rejected or revised depending on how things go when working with whatever other parts we are using at some subsequent time. It does sound as if this could be put in terms of a coherentism as follows. We are entitled to accept any DP or belief we find ourselves with until and unless the demands of coherence require us to sacrifice or modify it to maintain as coherent a total system of DP's and their belief products as possible. And doesn't this fall victim to (some of) the standard objections to coherentism?

NO. The main point here is the one elaborated just above. The correct (best, most adequate, most warranted…) epistemology for the higher level epistemological beliefs that are involved in higher level investigations as to the epistemic status of lower level beliefs is no more constrained to coherentism, or any other particular epistemology, by the "we have to work pro tem with what we have available" position than is the epistemology of lower level beliefs. To understand this we need to be more explicit about that position, what its purpose is, what constraints there are on it, more generally, what sort of position it is. First, what gives rise to the position is the attempt to live with pervasive epistemic circularity. The problem is—how can we be rational (warranted…) in using any DP's and accepting any beliefs when any attempt to show that any DP is reliable or that any belief is true, warranted, justified… runs into epistemic circularity if pushed far enough? And the answer is that it is part of the human condition to be in this situation. And we have to work with what we have available to us just because that is the only thing we can do. So the answer to the question "How can it be rational to accept beliefs under these conditions?" is that there is no alternative open to us. There is no rational alternative because there is no alternative. So the position is one as to how it is acceptable to proceed in inquiry (in the only way possible), and as to what attitude we should take toward this, realizing that this is the only possibility open to us. But just because it is that kind of position, it is *not* a position as to what the best (correct…) epistemology is for higher level epistemological beliefs. That is left completely open by that position and by the considerations that lead to it. So far as it goes, the correct epistemology might be coherentism (or would be if it were not faced with

devastating objections, as I believe it to be), some modest form of foundationalism, some mixed view, or whatever. Those options are all left open for the epistemology of higher level beliefs just as much as for the epistemology of lower level beliefs. Or, more precisely, the only requirement it makes on an epistemology is that it make room for some reciprocal support of beliefs. But that leaves many options open. And so, though the "we have to work with what we have" position can be put in a way that makes it sound like a coherence theory, that is, at most, a way we can think of what we are doing when we engage in inquiry under the influence of that position.

ENDNOTES

[1] For a spread of add-ons to Goldman's basically reliabilist account of epistemic justification see the following: Goldman 1979, Goldman 1986, Ch. 5, and Goldman 1992.

[2] Actually, the situation there was depicted as somewhat less stark than this. I did point out that there is such a thing as "significant self support" for a doxastic practice, exemplified for SP by the way in which the use of SP leads us to accept a view of things that imply a general reliability of SP. Since we do not get this result from any DP, including discreditable ones like crystal ball gazing, the result is by no means trivial, even though it does involve epistemic circularity. But since I do not claim that significant self support suffices by itself for conclusive support, I will leave it to one side in this discussion.

[3] There was a good deal more to Al's argument, for example an exploration of what conception of rationality is being used. But sufficient unto the day….

[4] As BonJour insisted in his coherentist days. See his 1985: 144-145.

REFERENCES

Alston, William P. 1989. *Epistemic Justification*. Ithaca, NY: Cornell University Press.

Alston, William P. 1991. *Perceiving God*. Ithaca, NY: Cornell University Press.

Alston, William P.1993. *The Reliability of Sense Perception*. Ithaca, NY: Cornell University Press.

BonJour, Laurence. 1985. *The Structure of Empirical Knowledge*. Cambridge, MA: Harvard University Press.

Goldman, Alvin. 1979. What is justified belief? In *Justification and Knowledge*, ed. George Pappas. Dordrecht: D. Reidel.

Goldman, Alvin. 1986. *Epistemology and Cognition*. Cambridge, MA: Harvard University Press.

Goldman, Alvin. 1992. Epistemic folkways and scientific epistemology. In *Liasons: Philosophy Meets the Cognitive and Social Sciences*. Cambridge, MA: MIT Press.

Plantinga, Alvin. 2000. *Warranted Christian Belief*. New York: Oxford University Press.

Reid, Thomas. 1970 (originally published 1764). *An Inquiry into the Human Mind*. Ed. Timothy Duggan. Chicago, IL: University of Chicago Press.

Chapter 6

HISTORICIZING THE BELIEF-FORMING SELF

Nicholas Wolterstorff
Yale University

1.

There's a picture of the belief-forming self that hovers over all of Plantinga's discussion in his three volumes on warrant. In the first volume one catches only glimpses of the picture; in the second and third volumes it is fully revealed. Our belief-forming selves consist of belief-forming faculties for whose operation there is a design plan, this design plan determining the difference between the proper functioning and the malfunctioning of these faculties. For some if not most of these faculties, the purpose governing the design is that the faculty, when functioning in the sort of environment for which it is designed, will produce true beliefs. The picture, as Plantinga acknowledges, is a very Reidian picture, both in this general contour and in the description he offers of the workings of various individual faculties.

The reason this picture hovers over Plantinga's discussion is that this is the picture of the belief-forming self that dominates his account of warrant. Suppose that the faculty which produced some belief is one of those that was successfully designed to produce true beliefs. Then whether or not the belief is warranted is determined by whether or not the faculty was functioning properly in a congenial environment. That is to say, whether or not the belief is warranted is determined by whether or not it was produced by a faculty functioning as designed in the sort of environment for which it was

T. M. Crisp, M. Davidson and D. Vander Laan (eds.), Knowledge and Reality, 111-135.

designed—assuming that the faculty was successfully designed to produce true beliefs.

In this paper I shall argue that this picture of the belief-forming self is either mistaken or misleading—depending on what one means by a "faculty." Most of one's belief-forming dispositions are not to be found in our human design plan; or not to be found there in the form in which they actually exist in one. The belief-forming dispositions that each of us actually possess at any particular time have been produced in us across the course of our lives. Though their production has occurred *in accord with* our design plan, they are not themselves *to be found within* our design plan. Furthermore, it is not one's designed belief-forming self that is relevant to the determination of warrant, but the belief-forming self produced in accord with the design.

To use language favored by those in the continental tradition, we must *historicize* our understanding of the belief-forming self. The language, though accurate, is nonetheless treacherous, carrying misleading connotations. What shapes one's belief-forming self is not just one's induction into one and another historical tradition—and especially not just one's induction into some such monumental tradition as the Western tradition, the American tradition, and the like. What shapes one's belief-forming self is one's own personal history, including but by no means confined to what has been handed on to one from others.

In Plantinga's discussion one comes across indications, sprinkled around in the text, that what I have just said expresses his own settled view on these matters: the belief-forming dispositions that constitute our belief-forming selves at any point in our personal history, though produced in accord with our design plan, are nonetheless for the most part not to be found within our design plan; and it is our personal-historically shaped belief-forming selves that are relevant to the determination of warrant, rather than our selves as designed. He says, for example, that "there is also learning, which also, in a way modifies the design plan. More exactly the design plan specifies how learning new facts and skills will lead to changes in cognitive function" (Plantinga 1993: 43).[1] Thus the picture that hovers, so I claim, over his discussion, is in fact misleading as to his true intent.

The reason for the discrepancy is not difficult to discern. Plantinga's emphasis on the design plan distracts us—and to some extent distracts him as well—from the fact that our actual belief-forming selves, though produced *in accord with* the design, are not *as* designed. Something similar happens for Reid. Reid's passion for digging down to what he called the "original principles" of our belief-forming selves easily distracts his readers from what he calls the "acquired principles," to which he in fact devotes a considerable amount of attention. The picture that hovers over his discussion does not accurately represent his actual thought.

In short, it is easy to fall into thinking of Reid, and of those who think along Reidian lines, as failing to take seriously the fact that we, as belief-forming selves, are in good measure creatures of history; it is easy to think of Reid and Reidians as ignoring history in their passion for talking about a shared and stable human nature. The truth, so I contend, is otherwise. Whatever be the impression given by their emphases, careful reading makes it indubitable that it is the explicitly held view of Reid and Reidians that our belief-forming selves are (in good measure) the creatures of history—not just of the broad sweeping currents of social history, but of our personal histories, these incorporating, each in its own way, those broad currents. What the Reidian emphatically adds, however, and emphasizes, is that this personal formation occurs in accord with our nature. How else could it occur? Reidian *doxastic anthropology*—as I shall call it—represents a blending of the natural and the historical.

What follows is an essay in doxastic anthropology in which I identify some of the principal ways in which our belief-forming selves are the product of our personal histories coupled with our design plan. By intent there will be no originality in what I have to say, since one of the points I want to make is the historical point that a recognition of the historicized character of the belief-forming self was a prominent feature of the thought of Hume and Reid—and in a rather curious way, a component of Locke's thought as well. At the end of my discussion I will consider in what way, if any, Plantinga's account of warrant has to be modified when we recognize with full clarity that the belief-forming self is a historically-shaped self.

2.

A response that I am likely to receive from those learned in the continental tradition of contemporary philosophy, to my insistence that we must historicize our understanding of the belief-forming self, is that this has already been done, more than adequately, in our recent past by Hans-Georg Gadamer in his theory of interpretation. Why re-invent the wheel? Be it granted that text-interpretation has been up front center in the contemporary continental tradition whereas perception was the focus of attention in the epistemology of the Enlightenment and continues to be so in contemporary analytic philosophy; nonetheless these two traditions can and should be brought into play with each other by way of analytic philosophers appropriating and absorbing Gadamer's thoughts on the belief-forming self—and by Gadamerians conceding that there are adumbrations of Gadamer's thought in Hume and Reid.

So let me begin by arguing that nothing of the sort is to be found in Gadamer, appearances to the contrary. Gadamer, no doubt against his own will, remained a captive of his own stereotype of Enlightenment thought.[2]

Gadamer framed his account of interpretation in opposition to what he understood to have been Schleiermacher's account, Schleiermacher being, in his judgment, the greatest of his predecessors. I judge Gadamer's understanding of Schleiermacher to be mistaken on central points; but that will make no difference for my concerns here.[3]

In his reading of the history of theory of interpretation (hermeneutics) from the magisterial Reformers up to Schleiermacher, Gadamer follows in the footsteps of Dilthey, differing mainly in his appraisal of some of the claims made in that stretch of theory and of some of the changes in interpretation which occurred.[4] Here is Gadamer's summary of Dilthey's narrative:

> Dilthey's studies on the origin of hermeneutics manifest a convincing logical coherence, given the modern concept of science. Hermeneutics had to rid itself one day of all its dogmatic limitations and become free to be itself, so that it could rise to the significance of a universal historical organon. This took place in the eighteenth century, when men like Semler and Ernesti realized that to understand Scripture properly it was necessary to recognize that it had various authors—i.e., to abandon the idea of the dogmatic unity of the canon. With this "liberation of interpretation from dogma" (Dilthey), the collection of the sacred Christian writings came to be seen as a collection of historical sources...(1989: 176).

Though the history Dilthey traced was the history of Scripture hermeneutics among Protestants from the Reformers to Schleiermacher, it's clear from this passage (assuming the correctness of Gadamer's interpretation of Dilthey) that Dilthey saw its significance as going well beyond its being a development within Protestantism. In that stretch of interpretation theory, so Dilthey claimed, interpretation became "free to be itself." As Gadamer puts Dilthey's view, "hermeneutics comes into its own only when it ceases serving a dogmatic purpose—which, for the Christian theologian, is the right proclamation of the gospel—and begins functioning as a historical organon" (1989: 177).

What is it for interpretation to be liberated from all "dogmatic bias" so as to "come into its own"? Is it to allow assumptions of a non-dogmatic character to guide the enterprise of interpretation, in place of Christian dogmatic assumptions? Definitely not. It is to be liberated from all "historical conditions," to use Gadamer's phrase. Scripture is to be interpreted in its own terms, the parts in the light of the whole and the whole in the light of the parts. What replaces dogmatic interpretation is not some

other form of ideological interpretation but the hermeneutical circle. The same is true for other texts. And the goal of interpretation for Schleiermacher, so Gadamer claims, following Dilthey, is to understand the author's production of the text by reversing the process of composition—starting now from text and moving to the train of thought which gave rise to the text whereas the author moved from that train of thought to his production of the text. One cannot fully understand that authorial train of thought, however, without going beyond that train as such to the author himself. Hence it is that, in Gadamer's words, "what is to be understood is…not only the exact words and their objective meaning, but also the individuality of the speaker or author" (1989: 186).

Such understanding is possible, on Gadamer's interpretation of Schleiermacher, because however different we may be from the author, we nonetheless share with him our human nature. Corresponding to "the production of genius" is an act of "divination" on the part of the interpreter, this act of divination presupposing "a kind of con-geniality" (1989: 189) between interpreter and author. "The ultimate ground of all understanding must always be a divinatory act of con-geniality, the possibility of which depends on a pre-existing bond between all individuals" (1989: 189). Though an author "is always an alien individuality that must be judged according to its own concepts and criteria of value," nevertheless that alien individuality can "be understood because I and Thou are of the same life" (1989: 199).[5]

Gadamer's claim, in short, is that "romantic hermeneutics," of which Schleiermacher is the paradigmatic example, took "homogeneous human nature as the unhistorical substratum of its theory of understanding," and proposed, in the practice of interpretation, the freeing of "the con-genial interpreter from all historical conditions" (1989: 290).[6]

3.

Gadamer's objections to "romantic hermeneutics" and to nineteenth century historicism, which, so he compellingly argues, shares with romantic hermeneutics its fundamental assumptions, are essentially two-fold: the presupposed philosophical anthropology is mistaken, and the description of how interpretation ought to proceed is accordingly untenable.

Gadamer acknowledges borrowing heavily from early Heidegger in the anthropology he proposes as replacement for that presupposed by "romantic hermeneutics." Let me highlight two elements of this anthropology. First, it's in the nature of the human being always to be projecting itself ahead of itself in such modes as expecting, anticipating, planning, committing, and so forth. Not only do we exist within time in the way which the traditional

philosophers, Kant above all, emphasized; we live within time in this much more specific way of "throwing" ourselves ahead of ourselves.

And second, all of us are so deeply shaped by the traditions into which we have been inducted that there is no possibility of rendering that shaping inoperative, even for a time, so as to employ only our "homogeneous human nature" in the performance of some activity. The "disregarding of ourselves" (1989: 299) urged on us by romantic hermeneutics is an impossibility. We are, in that way, inextricably historical beings. "Belonging to traditions belongs just as originally and essentially to the historical finitude of Dasein as does its projectedness toward future possibilities of itself. Heidegger was right to insist that what he called "thrownness" belongs together with projection" (1989: 262).[7]

Our projecting ourselves ahead of ourselves and our being thrown into traditions are not dimensions of the self that merely co-exist, side by side; the *projections* we make are themselves formed in us by our induction into tradition. Though it's in the nature of the human being to project itself, one's particular projections—for the most part anyway—are not the product of one's generic human nature but of finding oneself thrown into tradition. Our belonging shapes our projections. Thus the project of overcoming all prejudices, all pre-judgments, so as to become an example of The Human Being Itself, "this global demand of the Enlightenment," itself proves "to be a prejudice, and removing it opens the way to an appropriate understanding of the finitude which dominates not only our humanity but also our historical consciousness" (1989: 276).

Interpretation, like all other activities, must be understood in the light of this philosophical anthropology; or to put it the other way round, our analysis of the structure of interpretation must reveal "the existential futurality" (1989: 261) of the human being. We have to think through "how hermeneutics, once freed from the ontological obstructions of the scientific concept of objectivity, can do justice to the historicity of understanding" (1989: 265). Let me highlight the general point Gadamer makes here, and then his specific application thereof.

The person who interprets an historical text is herself functioning as an historical being; historicity is not only a dimension of the text interpreted but of the agent interpreting. Given the Enlightenment project of freeing ourselves from tradition, it's tempting to view this unavoidable historicity of the interpreter as constituting an obstacle to understanding. Quite the contrary; what enables one to understand a text from the past is not "con-geniality" with the author resulting from a common "homogeneous" human nature, but belonging to a shared tradition. Not "homogeneity" but "belonging" is what makes interpretation possible (1989: 262). "Belonging to a tradition is a condition of hermeneutics" (1989: 291). "The historicity of human Dasein in its expectancy and its forgetting is the condition of our being able to

represent the past. What first seemed simply a barrier, according to the traditional concept of science and method, or a subjective condition of access to historical knowledge, now becomes the center of a fundamental inquiry" (1989: 262). Understanding a text or event from history does not require liberation from one's own historicity but is and must be itself "a historically effected event" (1989: 300).

Gadamer's application to interpretation of this general point takes the form of an analysis of how interpretation proceeds. One approaches any act of interpreting a text with pre-judgments (praejudicia, prejudices) concerning the text in hand—pre-judgments concerning both the meanings of the words used and the things said. Some of these anticipations concerning meaning and content prove not to be borne out by the text; they are refuted, disconfirmed. When that happens, one forms new anticipations. Some of those will in turn not be borne out by the text; at those points one forms yet new anticipations concerning meaning and content; and so on, until finally one is no longer pulled up short in one's anticipations by the text. This cyclical process is what Gadamer proposes in place of Schleiermacher's circle. Let's have it in his own words. He is speaking of meaning in this particular passage; but on the following page he makes clear that he is proposing the same analysis for content:

> Interpretation begins with fore-conceptions that are replaced by more suitable ones. This constant process of new projection constitutes the movement of understanding and interpretation. A person who is trying to understand is exposed to distraction from fore-meanings that are not borne out by the things themselves. Working out appropriate projections, anticipatory in nature, to be confirmed "by the things" themselves, is the constant task of understanding. The only "objectivity" here is the confirmation of a fore-meaning in its being worked out. Indeed, what characterizes the arbitrariness of inappropriate fore-meanings if not that they come to nothing in being worked out? (1989: 267)

And how do we come by suitable anticipations? "The anticipation of meaning [and content] that governs our understanding of a text is not an act of subjectivity, but proceeds from the commonality that binds us to the tradition" (1989: 293).

It follows, says Gadamer, that "temporal distance is not something that must be overcome" when we interpret. This was "the naïve assumption" of romantic hermeneutics, and of its blood cousin, historicism; namely, that we must, upon removing all our own pre-judgments, "transpose ourselves into the spirit of the age, think with its ideas and its thought, not with our own, and thus advance toward historical objectivity. In fact the important thing is to recognize temporal distance as a positive and productive condition enabling understanding. It is not a yawning abyss but is filled with the

continuity of custom and tradition, in the light of which everything handed down presents itself to us" (1989: 297).

4.

There is much to be said about this model of interpretation, both pro and con; on this occasion I shall confine myself to just one point, namely, that close scrutiny of what Gadamer says concerning the confirmation and disconfirmation of anticipations concerning meaning and content reveals that, in spite of all his protestations, Gadamer has not liberated himself from the a-historicism that he himself attributes to the Enlightenment.

Let's have before us a sample of the passages relevant to this claim.

(1) "All correct interpretation must be on guard against arbitrary fancies and the limitations imposed by imperceptible habits of thought, and it must direct its gaze 'on the things themselves' (which, in the case of the literary critic, are meaningful texts…). For the interpreter to let himself be guided by the things themselves is obviously not a matter of a single, 'conscientious' decision, but is 'the first, last, and constant task.' For it is necessary to keep one's gaze fixed on the thing throughout all the constant distractions that originate in the interpreter himself….Working out [the] fore-projection, which is constantly revised in terms of what emerges as he penetrates into the meaning, is understanding what is there" (1989: 266-7).

(2) "A person who is trying to understand is exposed to distraction from fore-meanings that are not borne out by the things themselves. Working out appropriate projections, anticipatory in nature, to be confirmed 'by the things' themselves, is the constant task of understanding" (1989: 267).

(3) "How do we discover that there is a difference between our own customary usage and that of the text? I think we must say that generally we do so in the experience of being pulled up short by the text. Either it does not yield any meaning at all or its meaning is not compatible with what we had expected" (1989: 268).

(4) "We cannot stick blindly to our own fore-meaning about the thing if we want to understand the meaning of another. Of course this does not mean that when we listen to someone or read a book we must forget all our fore-meanings concerning the content and all our own ideas. All that is asked is that we remain open to the meaning of the other person or text…. [I]f a person fails to hear what the other person is really saying, he will not be able to fit what he has misunderstood into the range of his own various expectations of meaning" (1989: 268).

(5) "The important thing is to be aware of one's own bias, so that the text can present itself in all its otherness and thus assert its own truth against one's own fore-meanings" (1989: 269).

(6) "Methodologically conscious understanding will be concerned not merely to form anticipatory ideas, but to make them conscious, so as to check them and thus acquire right understanding from the things themselves. This is what Heidegger means when he talks about making our scientific theme 'secure' by deriving our fore-having, fore-sight and fore-conception from the things themselves" (1989: 269).

Notice the language: anticipations are not "borne out" by "the things themselves," we are "pulled up short" by the text, we must "hear what the other person is really saying," the text must be allowed to "assert its own truth" against our anticipations, and so forth. Though he never discusses the matter on its own, Gadamer's analysis presupposes that there is in the interpreter some process, some operation, whereby, upon reading the text, beliefs are more or less reliably formed in the interpreter concerning what the text does mean and say. The emphasis in Gadamer's discussion is all on the anticipations that we bring with us to the activity of interpreting texts, anticipations formed in us by our induction into tradition. But the analysis presupposes a distinct belief-forming dimension of the self which accounts for beliefs getting formed concerning the meaning and content of the text by way of actually reading the text. Tradition forms *anticipations* of meaning and content in us; reading the text forms *beliefs* concerning actual meaning and content. Unless our engagement with the text produced beliefs in us concerning the text, we could neither confirm nor disconfirm our anticipations; we would have nothing to measure them against. But about this distinct, belief-forming, dimension of the self, Gadamer says nothing in *Truth and Method*; his analysis presupposes it, but he does not discuss it.

And now for the point: there is nothing in Gadamer's allusions to this belief-forming dimension of the self which suggests that it too is touched by our historicity. Quite to the contrary. When Gadamer speaks of our inextricable historicity, it is always the projectional dimension of the self that he has in view, his thesis being that the deliverances of that dimension of the self are determined in good measure by our "thrownness" into tradition; he never has in view that dimension of the self whereby we form beliefs about the meaning and content of the text upon reading it. Indeed, the language he uses when he talks about anticipations being confirmed or disconfirmed by the text itself carry the unmistakable suggestion that here tradition has no influence; one's engagement with the text yields beliefs about the meaning and the content of the text against which one tests the anticipatory beliefs one received from tradition. So presumably what he assumed to be at work at this point in the process is our "homogeneous human nature." What else? What he emphasizes is that we may be so attached to our particular pre-judgments that we never allow the text to speak to us; what he suggests is that *in* allowing it to speak to us, there is

then operative in us a dynamic or process which is not itself in turn formed and shaped by tradition.

An implication is that even the point of historicity that Gadamer emphasizes is deprived of the high significance he assigns it, deprived of it by that aspect of the belief-forming self in whose functioning Gadamer fails to acknowledge any historicity. If careful reading of the text is capable of producing in me true beliefs about its meaning and content, then why are those projectional *praejudicia* necessary? That we approach texts with *praejudicia* concerning meaning and content proves to be a purely contingent matter so far as the activity of interpretation is concerned. It may be helpful to have some anticipations; but it is by reading the text that one comes to know what it means and says, thus confirming or disconfirming one's anticipations. The most fundamental aspect of the belief-forming self operative in Gadamer's account of hermeneutics is a non-historicized aspect.

And even the point where Gadamer does see history at work reveals, on close scrutiny, the assumption of the existence of a homogeneous human nature. Consider those tradition-shaped anticipations that we bring with us to the activity of interpretation. Does there not have to be, and was not Gadamer assuming that there is, something in all of us which accounts for the fact that we believe things on the say-so of others? The point I made above is that Gadamer assumes that there is in all of us a faculty or complex of dispositions that produces beliefs in a person upon her reading a text. What he also takes for granted is that there is in all of us a faculty or set of dispositions which leads us to believe what people tell us. His central claim is that, in interpretation, we bring deliverances of the latter faculty or set of dispositions along with us to the enterprise of interpretation so as to hold them up for comparison against deliverances of the former faculty or set of dispositions, accepting or rejecting deliverances of the latter depending on whether they coincide with or contradict deliverances of the former. Gadamer's analysis of interpretation assumes a "homogeneous" belief-forming nature with these two aspects. To say it again: Gadamer, in spite of himself, has not broken free from the assumptions that he himself attributes to the Enlightenment.

5.

Hume's analysis of inductively-formed beliefs is as well-known as anything in modern philosophy. Why is it, Hume asked, that upon experiencing the occurrence of one event, I expect the occurrence of an event of another particular type? For example, why is it that, upon seeing the car in front of me hit a patch of ice, I expect it to swerve?

It's not because I have an innate disposition to believe, upon seeing a car before me which is proceeding in a certain way hit a patch of ice, that it will swerve. In particular, it's not because my innate capacity of reason tells me that if a car proceeding as this one is proceeding hits a patch of ice, it will swerve. For there is no logical connection here, and so reason cannot tell me any such thing. The explanation instead goes as follows. A regularity in my experience has been that when a car which is proceeding in this way hits a patch of ice, it swerves. This regularity in experience has produced in me a new belief-forming disposition corresponding to the experienced regularity, specifically, this disposition: the disposition to believe, upon seeing a car which is proceeding in this way hit a patch of ice, to believe that that car will swerve. Of course, if I am to form this disposition I must have a *disposition* to form this disposition, and others like it; and that disposition to form particular inductive dispositions has to be innate in me. It's not a point Hume emphasizes; but if there were not in us an innate disposition to form inductive dispositions upon experiencing regularities of a certain sort, how could experience of such regularities ever have that result?

In short, what's to be learned from Hume is that one component of the historicized belief-forming self consists of all those belief-forming dispositions that we acquire by undergoing experiences of a sort apt for the formation of those particular dispositions—inductive dispositions being prime examples of such, though perhaps there are others as well.

Then again, maybe this is not something to be learned from Hume's analysis, since Reid offers an alternative analysis of the same phenomena which would seem to be just as plausible. The author of our nature, Reid says, has

> implanted in human minds an original principle by which we believe and expect the continuance of the course of nature, and the continuance of those connections which we have observed in time past. It is by this general principle of our nature, that when two things have been found connected in time past, the appearance of the one produces the belief of the other.[8]

I take Reid's thought to be this: by virtue of an innate belief-forming disposition, upon experiencing a regular correlation between events of two types, I form the general belief that if and when an event of the one type occurs, an event of the other type will also occur: my experience of a regular correlation between cars proceeding in a certain way hitting a patch of ice and those cars swerving produces in me the general belief that if and when a car proceeding thus hits a patch of ice, it will swerve. Now I see the car before me, proceeding thus, hit a patch of ice; that perception produces in me the belief that the car before me is proceeding in that way and is hitting a patch of ice; and by virtue of an elementary exercise of my capacity for

reasoning I infer, from that general belief which I already had, plus this particular belief which I have just now acquired, that the car before me is about to swerve.

I have not on this occasion appealed to the text of Hume to establish that the analysis I attributed to him is in fact the analysis he offered; but assumed that it is. Which analysis is correct, the Humean or the Reidian? I'm not sure; it's my hunch that making a well-grounded choice between the two would require a lot of preliminary work. There is no disposition within my design plan, upon seeing a car proceeding in a certain way hit a patch of ice, to believe that it will swerve. Yet I am now so disposed. My belief-forming self is, in this regard, a product of my personal history. On this, both analyses agree, as they do on the thesis that it was my prior perception of regularities which made me so disposed. The issue in dispute is only whether this new disposition, not to be found in my design plan, is reducible to the combination of a disposition which is to be found in my design plan plus some acquired beliefs, or whether it is not so reducible. Suppose we call a belief-forming disposition which is contained within one's design plan, a *faculty*. The issue in dispute is whether my historically acquired disposition is reducible to a combination of a faculty plus some acquired beliefs, or whether it is not so reducible.

6.

Reid develops a detailed and extremely interesting analysis of "acquired perceptions," as he calls them—for example, my perception that this wine has a fresh peach aroma. A properly historicized understanding of the belief-forming self will give to the dispositions which yield such beliefs a very important place.[9] But since Reid's analysis of acquired perceptions proceeds along pretty much the same lines as his analysis of what accounts for the formation of inductive beliefs, for our purposes here we can move on to consider that aspect of the belief-forming self which accounts for that to which Gadamer devotes all his attention, namely, that aspect which accounts for the fact that we are all inducted into traditions of one sort and another.[10]

Being inducted into tradition is a multi-facetted phenomenon; it includes, for example, the acquisition of certain skills by way of their having been handed on to one. What Gadamer emphasizes, and what I shall likewise emphasize on this occasion, is that it includes the acquisition of certain beliefs by way of their having been handed on to one. As I observed earlier, for this to happen there has to be in one a disposition to believe what people tell one; if one had no such disposition, one's predecessors could talk as much as they wished but nothing would be "handed on." Reid called this disposition, the principle of credulity. It is by virtue of the fact that there is in

each of us a principle of credulity—an innate disposition to "confide in the veracity of others, and to believe what they tell us"[11]—that there is in our human existence the phenomenon of being inducted into the beliefs of a tradition.

Reid spends a good deal of time, in his discussion of the principle of credulity, arguing that our disposition to believe what people tell us, rather than being the product of experience and reasoning, is innate in us—in his terminology, that it is an original principle, in Plantinga's terminology, that it is part of our design plan. "The wise and beneficent Author of nature," says Reid, "who intended that we should be social creatures, and that we should receive the greatest and most important part of our knowledge by the information of others, hath for these purposes, implanted in our natures two principles that tally with each other,"[12] namely, the principle of veracity and the principle of credulity.

Reid's arguments for there being an innate principle of credulity are, to my mind, both intriguing and compelling. On this present occasion my interests lie elsewhere, however; namely, in his claim that our innate credulity disposition gets modified in the course of experience. Let me quote what he says in a central passage. Having uncharacteristically personified Reason, Reid now refers to it as "she" and proceeds to attribute to Reason what he would ordinarily attribute to the person.

> When brought to maturity by proper culture, she learns to suspect testimony in some cases, and to disbelieve it in others; and sets bounds to that authority to which she was a first entirely subject.... And as in many instances, reason, even in her maturity, borrows aid from testimony; so in others she mutually gives aid to it, and strengthens its authority. For as we find good reason to reject testimony in some cases, so in others we find good reason to rely upon it with perfect security, in our most important concerns. The character, the number, and the disinterestedness of witnesses, the impossibility of collusion, and the incredibility of their concurring in their testimony without collusion, may give an irresistible strength to testimony, compared to which, its nature and intrinsic authority is very considerable.[13]

The point is clear: as a consequence of our experience, our innate credulity disposition gets modified in such a way that eventually we believe what certain sorts of people say on certain sorts of topics in certain sorts of situations more firmly than we would have originally, whereas we believe less firmly or not at all what those same people tell us on different sorts of topics or in different sorts of situations, or what other sorts of people tell us. Never, though, do we find ourselves in the position of no longer believing anything that anybody tells us. Though the modification of our innate credulity disposition occurs in accord with our design plan, the exceedingly

complex credulity disposition of an adult at a particular moment in his or her life is not to be found *within* our human design plan; furthermore, it varies considerably from one adult to another depending on their experience of truth-telling and falsehood-telling.

Plantinga agrees, In his chapter on "Other Persons and Testimony" he remarks that "as Reid says, we learn to modify, qualify, modulate our native tendency to believe what others tell us" (1993: 79). He offers, as an example, that "I believe you when you tell me about your summer vacation but not when you tout on television the marvelous virtues of the deodorant you have been hired to sell" (1993: 79).

Reid understands the process of modification, in our tendency to believe what people tell us, as resulting from the interplay of our credulity principle with our other belief-forming faculties and dispositions. Though he is not fully explicit on the matter, he presumably regarded the interplay as occurring in the following way. First what happens is that now and then, after believing what someone says, we subsequently learn that it was false—or at least, come to believe that it was false. The occurrence of such learning presupposes that the proposition we concluded to be false was believed by us with less firmness than some other proposition that we took to be in conflict with it—and also with less firmness than the proposition that that other was indeed in conflict with it. The least firmly held belief gives way. What makes it possible to learn that something one believed on say-so is false is that one's believing it on their say-so is done with less than maximal firmness.

What happens, second, is that, beyond coming to believe that a particular piece of testimony was true and another piece false, we learn to sort what has proved true and what has proved false into types of testimony that regularly prove true and types of testimony that regularly prove false, and to pick up cues to these types. Thus now, when presented with a case of testimony, we believe that it belongs to an unreliable type and we don't believe what we were told, or we believe that it belongs to a reliable type and we believe it.

7.

John Locke's main topic, in Book Four of his *Essay*, was how we ought to regulate the formation of our beliefs. The answer he arrived at was that, when it is a matter of maximal "concernment" to us whether or not some proposition is true, and it is not immediately certain for us that it is, we are to collect a body of sufficiently ample and representative evidence for the truth and/or falsehood of the proposition, determine the probability of the proposition on that evidence, and then believe or disbelieve the proposition with a firmness proportioned to its probability on the evidence.

Then, in the penultimate chapter of the *Essay*, Locke looks back over his discussion in Book Four, observes that people every now and then make mistakes in their judgments of probability, and asks why that is. He delineates several different sources of such error. My interest here is exclusively in the first. "Propositions that are not in themselves certain and evident, but doubtful and false," he says, may be inculcated in us from youth up as (self-evident) *principles*. And

> these have so great an influence upon our opinions, that 'tis usually by them we judge of truth, and measure probability, to that degree that what is inconsistent with our *principles*, is so far from passing for probable with us, that it will not be allowed possible. The reverence [that] is born to these principles is so great, and their authority so paramount to all others, that the testimony not only of other men, but the evidence of our own senses are often rejected, when they offer to vouch any thing contrary to these established rules….For he hath a strong bias put into his understanding which will unavoidably misguide his assent, who hath imbibed wrong principles, and has blindly given himself up to the authority of any opinion in itself not evidently true (IV, xx, 8).

As an example of such a principle, Locke cites the inculcation from youth up in the mind of a "Romanist" of the doctrine of transubstantiation.

I take the following to be Locke's thought. The child who has been reared in Catholicism is now entertaining some proposition which is in fact necessarily true. Not only does he entertain it; he grasps it. If everything were working in an ideal way, his comprehending entertaining of that proposition would evoke in him the belief that that proposition is true. But in this case it does not. The reason it does not is that, having entered the situation believing firmly the doctrine of transubstantiation, he now not only entertains the proposition which is in fact necessarily true but also acquires the belief that this proposition is incompatible with the doctrine of transubstantiation to which he is committed. That pair of beliefs—his belief in the doctrine and his belief that the doctrine is incompatible with the proposition to which he is now attending—inhibits the formation of the belief that the proposition he is entertaining is true, let alone, necessarily true. He might well say to himself that it *seems* to be true; but he takes his experience of its seeming that way to be nothing more than a *mimic* of the experience he has when he entertains a necessary truth. Rather than merely suspending belief, he might come to believe that the proposition is false.

Notice that the reason his acquaintance with the necessary truth fails to evoke the corresponding belief is not that, at this point, his belief-forming system is not working properly. It is working properly—working in accord with its design plan. It's not because of a fluke in the working of his belief-forming system that he acquired his belief in transubstantiation, nor is it

because of a fluke in the working of his system that he has retained that belief; neither is it on account of a fluke in the working of his system that that belief now functions to inhibit the belief that it's a necessary truth he has before his mind's eye. All of these developments are in accord with the design plan of his constitution. The system *was* working just fine and *is* working just fine. But there is, as it were, a glitch in his programming, viz., a false belief, or to speak more precisely, a belief that Locke takes to be a false belief, namely, the belief in transubstantiation. And that glitch is what accounts for the fact that his belief-forming system is not producing what it would produce were it working ideally.

Already in Aristotle one finds the concept of a self-evident proposition. A proposition is self-evident *per se* just in case it is impossible that anybody should grasp and entertain it without "seeing" it to be true and believing it. And a proposition which is self-evident *per se* is self-evident to a particular person if that person grasps and entertains it.[14] Mathematical truths and logical truths—the simpler ones among them—were traditionally cited as examples. But suppose that I do grasp some proposition which the tradition would cite as an instance of self-evidence, and suppose further that I am now entertaining it—staring right at it, as it were. What we have just seen is that whether or not the corresponding belief gets formed in me may well depend on other things I believe at the time and how firmly I believe them. If I firmly believe that the proposition is false—perhaps in my e-mail this morning there was a message from a mathematician friend of mine whom I trust implicitly on matters mathematical announcing that an unexpected result of what he's been working on for eight years is a proof that the proposition is false—if I firmly believe that it is false, then I won't believe the proposition even if I do grasp it and even if it is presently right before my mind's eye. The history of mathematics and logic in the twentieth century is full of exactly such surprising results. One says to oneself that the proposition certainly *seemed* true—maybe it still does—but that in this case, appearance is deceiving.

Locke's example and mine are from the realm of thought; memory and perception also provide us with examples. Suppose that I am perceiving a chair but that the chair is some distance off and that I believe I am looking at a wall covered with some extraordinarily skillful fool-the-eye painting. Given that belief, it may well be the case that my perception of the chair does not evoke in me, as it would if I did not have that belief, the beliefs that I am seeing a chair and that there's a chair there.

A few paragraphs above I spoke of programming, and of a glitch in the programming. The programming of a computer does seem to me a helpful model for thinking of the situation—though let me declare at once that I do not for a moment believe that the human mind *is in fact* a computer. My suggestion is not that the mind is a computer, but that it proves illuminating

to use a certain aspect of the functioning of computers as a *metaphoric model* for thinking about the human understanding.

We human beings are all hard-wired for belief; we all have an innate dispositional constitution which, when activated by one event or another, yields belief.[15] That's the beginning of the matter; and it is this beginning of the matter that Plantinga emphasizes. It cannot be the end of the matter, however, for reasons that have emerged. New dispositions emerge as the result of experiences of certain sorts, prime examples of such acquired dispositions being inductive dispositions. Whether inductive dispositions be analyzed along Reidian or along Humean lines, either way, they are acquired dispositions. And original dispositions get modified, a prime example of such modified dispositions being the credulity disposition as one actually finds it in a particular person at a particular time in his or her life. Thirdly, the beliefs we already have function as programming for the formation of new beliefs. To fully account for the beliefs that actually get formed in a person on a given occasion one has to know not only the belief-forming faculties of that person, but how that person has been programmed. That programming will include what the tradition in its own way characteristically took note of, namely, the concepts possessed by the person. But it will also include the particular contour of belief and non-belief of the person at the time. Beliefs formed by the operation of the system become components of the program with which the system subsequently operates. The beliefs we have already formed are not stored inertly in memory but function as components of our present belief-forming self.

That makes it sound more individual than it is. It's not just me and the world and you and the world. Much of our programming is social, with the consequence that tradition becomes part of our belief-forming self. What we learn from our fellows is not just stored in memory, to be brought out as the occasion demands, but becomes a component of our programming. Thus Gadamer was right, in a way even more profound than he had in mind, when he said, in a passage quoted earlier, that "we are always situated within traditions, and this is no objectifying process—i.e., we do not conceive of what tradition says as something other, something alien. It is always part of us…."

And let me add, though I cannot develop the point, that our doxastic programming includes more yet than the concepts and beliefs we have already acquired. Less obviously perhaps, but no less importantly, it includes one's likes and dislikes, one's sympathies and antipathies, one's loves and hatreds. That's a point characteristic of the Augustinian tradition; to understand God in particular, but reality more generally, one's heart must be cured.

The implication Locke's example highlights is that, in addition to whatever flaws there may happen to be in one's hard-wiring and operating

system, the programming of each of us includes a good many defective components, those defects in programming yielding defects in output. In particular, the falsehood of many of the belief-components in our programming constitute glitches in our programming, the result being that false beliefs get formed when, were it not for the glitch, true ones would have been formed—or no beliefs get formed when, were it not for the glitch, true ones would have been formed.

8.

Several times over in the course of my discussion I have taken note of questions that my discussion raises or suggests but which I do not have space, on this occasion, to deal with. The point just made, about the role of doxastic programming in the formation of new beliefs, raises another of these. If this is indeed how the mind works, how can we ever know whether or not our beliefs correspond to the facts? How can we ever know if and when they are veracious? Reality recedes from our grasp.

Ever since Descartes, we in the modern world have been haunted by the question whether our indigenous hard-wiring yields knowledge of the facts of the world. Our situation now looks even more precarious. Beliefs are not the outcome simply of reality's impact on our hard-wired dispositions for belief-formation; they are the outcome of reality's impact on our hard-wiring *and our programming, with all its glitches*. Always between me and reality there's my particular historical belief-forming self. And that particular historical self is defective—defectively programmed. To use religious language: our fallenness invades not only our ethical selves but our doxastic selves as well. Fallenness is an epistemological topic—as is sin, because for some of the glitches we're culpable. So back to the anxiety: how am I to know where I with my defective programming leave off and where my presumptively reliable indigenous human faculties for belief-formation take over? Best to stick with one's beliefs and be under no illusion that one can reliably compare them with reality. So says the skeptic.

I must confine myself to two observations in response. First, the skeptical conclusion formulated above rests on incoherence. If we have no way of telling where our programmed self leaves off and where our presumptively reliable indigenous self takes over, then of course we cannot know that there are glitches in our doxastic programmings.

And second, we must not allow our examples to lead us into concluding or assuming that our doxastic programming only functions obstructively; much of it also functions *accessively*. Certain belief components in our programming make certain aspects of reality *more reliably* accessible to us than otherwise they would be; and some belief components in our programming make certain

aspects of reality accessible to us which otherwise would not be accessible to us at all. It seems to me that this is how we should think of the learning that results from scientific experiments. If the experimenter did not already hold a vast body of scientific conviction, he could not learn from his experiments what he does learn from them.

So contrary to what our sociologists of knowledge tell us, our programming, though it sometimes constitutes a barrier between us and reality, is often if not usually a means of access. Particularity of programming, in many of its forms, is not prejudice which obstructs, but education which enables, access to reality.

9.

The impression my discussion will have given is that I see myself as having identified three distinct ways in which one's belief-forming self is a product of one's personal history: one acquires new belief-forming dispositions, one's original belief-forming dispositions (faculties) get modified, and one becomes doxastically programmed in such a way that the triggering by events of one's belief-forming faculties is altered.

But closer scrutiny reveals that there may be less difference here than appears. Start with the last. I have analyzed the examples offered as consisting of beliefs one already has either obstructing the workings of a disposition or enabling its working. An event that would trigger a belief-disposition in the absence of a certain belief no longer does so in the presence of that belief; an event that would not trigger a certain belief in the absence of various other beliefs does so when those beliefs are present. But it would also be correct to say that, by virtue of acquiring beliefs that function in this way, the person acquires a new array of belief-dispositions: some prior dispositions get eliminated and some new ones emerge. My mathematician friend telling me that he has proved the falsehood of that theorem eliminates from me the disposition to believe it.

And now two points. First, perhaps we should analyze what happens to one's disposition to believe what people tell one along these lines as well. Yes, that disposition gets modified by one's experience of people telling truths and falsehoods. But perhaps the right way to understand that modification is that one's innate credulity faculty, like one's innate faculty for rational intuition and one's innate faculties of perception, is designed to operate in conjunction with one's doxastic programming. My belief that the testimony I am presented with belongs to a type that is generally false inhibits the triggering of the faculty, whereas my belief that the testimony I am presented with belongs to a type that is generally true not only allows the faculty to be triggered but brings it about that the resultant belief is held

more firmly than would otherwise be the case. This is how the credulity disposition was designed to work.

And second, there is a close similarity—though not indeed identity—between this analysis of the workings of these faculties and Reid's analysis of what accounts for inductive beliefs. Reid differed from Hume in that, whereas Hume thought that inductive beliefs were formed immediately, Reid thought they were the product of inference. What's innate in us, on Hume's view, is the disposition to form inductive-belief dispositions upon perceiving regularities; what's innate in us, on Reid's analysis, is the disposition, upon perceiving those same regularities, to form generalized conditionals about the course of nature—along with the disposition to infer instantiations of these generalizations when confronted with instances of the antecedents of the generalized conditionals.

What's similar between the Reidian analysis of inductive belief formation and our own programming analysis of how perception, rational intuition, and so forth function, is that the acquired dispositions to believe are analyzed entirely in terms of innate faculties operating, as they were designed to operate, in conjunction with already formed beliefs. What's different is the way in which they operate in conjunction with those already formed beliefs. Our faculties of perception, of rational intuition, and perhaps of credulity, are designed to produce beliefs *immediately*, whereas, on Reid's analysis, inductive beliefs are non-immediately produced by inference. Inductive generalizations are produced immediately, however. And though Reid does not make a point of it, surely our innate faculty for the formation of inductive generalizations is designed to work in conjunction with already formed beliefs in the same way that our faculties of perception, of rational intuition, and perhaps of credulity are so designed.

10.

And now, in conclusion, how does Plantinga's theory of warrant fare when we keep clearly in mind the fact that our belief-forming selves are the product of our personal histories, the dispositions constituting these selves being formed (for the most part) in accord with our design plan though not themselves to be found within our design plan?

If the Reidian analysis of what accounts for the formation of inductive beliefs is correct, the formation of inductive beliefs poses no particular puzzle for Plantinga's account. Our innate faculty for the formation of beliefs whose content is inductive generalizations produces those beliefs immediately upon the input of perceived and remembered regularities, whereupon they are stored in memory; our innate faculties of perception immediately yield the belief that an instance of the antecedent of one of those regularities is occurring; and our

innate faculty of inference produces mediately the belief that an instance of the consequence of that generalization is about to occur. No particular puzzles here for Plantinga's account. Indeed, if this is how it goes, the only acquired dispositions are the dispositions to make those inferences.

So let us focus our attention on the other phenomenon that has emerged from our discussion: the fact that, for many if not most of our faculties, to understand why the faculty produced the output it did on a certain occasion, or why it failed to produce a certain output, one has to look not just at the input, the triggering event, but at the beliefs the person held at the time. Call it, for convenience sake, the *programming phenomenon*.

In one way or another, the belief has been formed in me that I am seeing a skillfully painted fool-the-eye scene. I am. But there is also a table, a real table, placed up against the wall. I look at the table. But I do not form the belief that I am seeing a table, nor the belief that there's a table there. Instead I form the belief it's a fool-the-eye table that I am looking at and that it is painted more skillfully than any of the rest of the illusory scene.

I assume that this belief is not only not true but not warranted. Does Plantinga's account explain why it is not warranted?

Well, my faculty for the production of perceptual beliefs has not malfunctioned; it is working exactly as it was designed to work. And that faculty, so Plantinga argues, and I agree, was successfully designed to produce true beliefs. Is it then, perhaps, not functioning in the sort of environment for which it was designed? Is it perhaps functioning in an environment that's not what Plantinga calls "congenial"? Well, it's functioning in a very ordinary environment, nothing at all like those exotic Alpha Centaura environments that Plantinga is so fluent at imagining.

What accounts for the fact that the belief is not warranted, so I suggest, is that there is *a glitch* in the person's *doxastic programming*. Our human faculty for the production of visual perceptual beliefs is designed to operate in conjunction with certain beliefs that one already has. In this case, one of those beliefs is false; that's the glitch. If there's a glitch in the programming, then one must expect something deficient in some of the output. And it really doesn't matter whether the glitch was planted in one by some spiteful little Cartesian demon, or whether it was itself produced by prior functioning of one's innate capacities. Furthermore, that false belief, which in this case functions as a glitch in the programming, may itself have been the output of one's faculties functioning properly; in principle it may itself be a warranted belief.

But maybe I was too hasty in dismissing the suggestion that there was something wrong with the environment.[16] In the more recent presentations of his theory, Plantinga distinguishes between what he calls the cognitive "maxi-environment" and the cognitive "mini-environment."[17] The maxi-environment

is "our cognitive environment as the one we enjoy right here on earth, the one for which we were designed by God or evolution. This environment would include such features as the presence and properties of light and air, the presence of visible objects, of other objects detectable by our kind of cognitive system, or some objects not so detectable, of the regularities of nature, the existence of other people, and so on."[18] By contrast, the mini-environment is specific and detailed. The mini-environment, for a given case of belief-formation, is "all the relevant epistemic circumstances obtaining when that belief is formed."[19] For example, a feature of the mini-environment, for the case we are imagining, is that, when looking at the painted scene from the distance and perspective from which the viewer is looking at it, it is impossible with ordinary unaided eyesight to tell the difference between a real table and a skillfully painted fool-the-eye table.

So which analysis is correct of what went wrong? Is the problem that human eyesight is simply not good enough to tell the difference, from this distance and perspective, between a real table and a fool-the-eye table, or is the problem that the viewer has a glitch in his doxastic programming, that glitch consisting of the belief he already has that he is looking at a fool-the-eye scene? The latter, in my view—though I concede that it may also be true that human eyesight is indeed not good enough to discern the difference.

It is open to Plantinga to seize on this concession and to insist that, whatever may be true about doxastic programming and glitches therein, a satisfactory theory of warrant need take account only of the defective mini-environment, not of those glitches. One way of responding, in turn, would be to flesh out the example a bit, so that, though the human eye is capable of discerning the difference between a real table and a fool-the-eye table from this distance and perspective, nonetheless that capability is over-ridden by the firm prior belief that one is not seeing a real table. But let us instead adopt the more radical recourse of taking note of cases for which there is no relevant cognitive environment at all—that is, purely internal cases. Cases of the Locke-type will do: because of some false belief that one already has, one fails to believe that the necessary truth or falsehood now right before one's mind's eye is in fact a necessary truth or falsehood. To make double sure that there is no relevant cognitive environment, add that the false prior belief was not acquired from testimony but from one's own work in logic or mathematics.

I can imagine a Plantinga-defender now proposing that to cope with such cases we should take a more expansive view of what belongs to one's cognitive environment than anything that Plantinga himself proposes. Plantinga clearly thinks of environment as something external to the self; all his examples point in that direction. The defender proposes that not only should

one's body be regarded as belonging to one's environment, but one's interior life as well, including one's current beliefs.

This strikes me as a non-starter. I fail to see that anything coherent is being proposed, when it is proposed that we regard one's environment as including one's beliefs and feelings. If they are included, what's not included? When a distinction is made between the self and its environment, then surely one's beliefs go with the self rather than with the environment. They belong to the self which finds itself within an environment rather than to the environment within which the self finds itself. There is no concept of a person's environment such that the person's beliefs are included within the person's environment.

My own suggestion is that, rather than trying to stretch his extant account in any such fashion, Plantinga's account be amplified by adding the concept of *doxastic programming* to those other concepts with which he works, namely, the concepts of innate faculty, design plan, proper functioning, being successfully aimed at truth, and congenial environment—understanding programming, let me remind the reader, as a metaphoric model. Our belief-forming faculties are designed in such a way that much of their output functions as programming for subsequent operations of the faculty, the output of those subsequent operations functioning as programming for yet later operations, and so forth, on and on as long as we live. But glitches turn up in the programming in the form of false beliefs. And if that part of the programming which is operative in a certain case contains such a glitch, the resultant belief is not warranted—even if it should just so happen to be true.

That's my suggestion—undeveloped, inchoate. Of course Plantinga has thought about his account of warrant much longer and much more intensively than I have; so perhaps he has a better idea as to how his account of warrant can best deal with the fact that our belief-forming selves are created as we go along—though not *ex nihilo*.

ENDNOTES

[1] See also his discussion of learned perception (1993: 99-101).

On page 22 of Plantinga 1993, Plantinga introduces the concept of what he calls a "snapshot design plan." But what he has in mind is not what I am calling attention to above, but the fact that in human beings there is, as it were, a master design plan for the emergence of a succession of design plans as the individual matures. Maturation is of course not to be identified with learning. This is what he says: "Here we are thinking of the design plan as specifying how the thing works *now*, or at a given time: call this a *snapshot* design plan. But the design plan, at least in the case of organisms, may also specify how the thing will change over time. There is such a thing as maturation; and it can be thought of as involving a master design plan, which specifies a succession of snapshot design plans."

[2] The crucial text is of course Gadamer's *Wahrheit und Methode*, translated as *Truth and Method*. I will be quoting from the Second, Revised Edition, translated by Joel Weinsheimer and Donald G. Marshall (Gadamer 1989).

[3] My critique of Gadamer's interpretation of Schleiermacher can be found in Wolterstorff 2003.

[4] *Cf.* "If…the ideal of the historical enlightenment that Dilthey pursued should prove to be an illusion, then the prehistory of hermeneutics that he outlined will also acquire a quite different significance. Its evolution to historical consciousness would not then be its liberation from the chains of dogma but a transformation of its nature" (Gadamer 1989: 177).

[5] *Cf.* "Schleiermacher's model of hermeneutics is the congenial understanding that can be achieved in the relation between I and Thou. Texts are just as susceptible of being fully understood as is the Thou. The author's meaning can be divined directly from his text. The interpreter is absolutely contemporaneous with his author. This is the triumph of philological method, understanding the mind of the past as present, the strange as familiar" (Gadamer 1989: 240).

[6] The same image, of *homogeneous human nature*, occurs on p. 232: "Here Dilthey is following the old theory that understanding is possible because of the homogeneity of human nature" (Gadamer 1989: 232).

[7] *Cf.* "We are always situated within traditions, and this is no objectifying process–i.e., we do not conceive of what tradition says as something other, something alien. It is always part of us" (Gadamer 1989: 282).

[8] Reid 1997: 197. (Also found on p. 198a-b of Reid 1858.)

[9] Plantinga briefly discusses acquired perception in the last section of his chapter on perception (1993: 99-101).

[10] I discuss Reid's analysis of acquired perceptions in Chapter V of Wolterstorff 2001.

[11] Reid 1858: VI, xxiv, 196b; also Reid 1997: 193-194.

[12] Reid 1858: VI, xxiv, 196a; also Reid 1997: 193.

[13] Reid 1858: VI, xxiv, 197a-b; also Reid 1997: 195.

[14] For most of my own philosophical career I have understood and explained self-evidence along these lines.

[15] If we distinguish between hard-wiring, operating systems, and programming, the contrast I want for the model I am proposing is no doubt programming plus operating system, rather than programming with hard-wiring. But I will continue to blur the distinction between hard-wiring and operating system, and speak of the contrast to our programming as being our hard-wiring.

[16] I thank Thomas Crisp for making this point to me.

[17] See Plantinga 1996: 313ff and Plantinga 2000: 158ff.

[18] Plantinga 1996: 313.

[19] Plantinga 1996: 314.

REFERENCES

Gadamer, Hans-Georg. 1989. *Truth and Method*. 2nd rev. ed. Translated by Joel Weinsheimer and Donald G. Marshall. New York: Crossroads.

Plantinga, Alvin. 1993. *Warrant and Proper Function.* Oxford: Oxford University Press.

Plantinga, Alvin. 1996. Respondeo. In *Warrant in Contemporary Epistemology: Essays in Honor of Plantinga's Theory of Knowledge,* ed. Jonathan L. Kvanvig. Lanham, NY: Rowman & Littlefield.

Plantinga, Alvin. 2000. *Warranted Christian Belief.* New York: Oxford University Press.

Reid, Thomas. 1858. *The Works of Thomas Reid.* Edited by William Hamilton. 5th ed. Edinburgh: Maclachlan and Stewart; London: Longman, Brown and Green.

Reid, Thomas. 1997. *An Inquiry into the Human Mind on the Principles of Common Sense*. Edited by Derek R. Brookes. Edinburgh: Edinburgh University Press.

Wolterstorff, Nicholas. 2001. *Thomas Reid and the Story of Epistemology*. Cambridge: Cambridge University Press.

Wolterstorff, Nicholas. 2003. Resurrecting the author. In *Midwest Studies in Philosophy*, Vol. XXVII. Oxford: Blackwell Publishers.

Chapter 7

A DILEMMA FOR INTERNALISM*

Michael Bergmann
Purdue University

Some objections to internalism in epistemology target only certain versions of it. For example, some objections focus on versions of internalism that are wedded to a deontological conception of justification; others focus on versions of internalism according to which a necessary condition of justified belief is a further justified belief (or the potential for a further justified belief).[1] But there are internalists who are very careful to sidestep these objections by defending positions to which they don't apply.[2] In this paper I develop an objection that applies to *all* species of internalism.

As the title of this paper suggests, the objection is in the form of a dilemma:

(I) An essential feature of internalism is that it makes a subject's actual or potential *awareness* of something a necessary condition for the justification of any belief held by that subject.
(II) This required awareness is either *conceptual* awareness (of a particular kind to be described later) or it is not.
(III) If it *is* conceptual awareness (of the relevant kind), then internalism falls victim to regress problems.
(IV) If it is *not*, then internalism is subject to a prominent objection to externalism.
(V) If internalism is subject either to the regress problems mentioned in (III) or to the prominent objection to externalism mentioned in (IV), then we should not endorse internalism.
(VI) Therefore, we should not endorse internalism.

T. M. Crisp, M. Davidson and D. Vander Laan (eds.), Knowledge and Reality, 137-177.

A familiar feature of this argument is its division of all awarenesses into two types and its conclusion that, for either type, some problem arises for an important epistemological thesis. Wilfrid Sellars' objection to what he called 'the myth of the given' employs a dilemma like this.[3] So does Laurence BonJour's objection to internalist foundationalism (1985: Chap. 4) and Paul Moser's objection to internalist coherentism (1989: 173-76).[4] But what hasn't been widely observed is that an argument employing this sort of dilemma can be marshaled against a broader target, namely, internalism itself.[5] The fact that it can be is striking. It suggests that for many years, lurking in the writings of internalists (such as BonJour and Moser) have been the seeds of internalism's own demise.

The paper will proceed as follows. I will defend premise (I) in section 1, premises (III) and (IV) in section 2, and premise (V) in section 3. (Premise (II) needs no defense.) I believe that this argument, along with the defense I'll give of its premises, provides a formidable objection to internalism. However, it's been my experience that, upon hearing this argument, internalists tend to think that problems with it will become evident if we focus on specific beliefs (to see how the dilemma I pose is supposed to arise in a fleshed out example) or if we focus on specific attempts by internalist epistemologists to avoid this sort of dilemma. For this reason, I will, in section 4, focus on two concrete examples of beliefs that can be used as test cases for my argument. In addition to showing how these examples lend support to my conclusions, I will also consider whether the internalist positions defended by Paul Moser, Richard Fumerton, and Laurence BonJour can handle these examples in a way that enables them to resist my argument. The reason I focus on these three philosophers is that they are among the most able defenders of internalism and they are particularly sensitive to the troublesome issues that lead to the sort of dilemma I propose in this paper. Their failure to avoid being impaled on the horns of this dilemma provides further testimony to the strength of this objection to internalism.

1. REQUIRING AWARENESS

According to premise (I), an essential ingredient of internalism is the requirement (for justification) that there be some sort of actual or potential awareness of something on the part of the subject. Two claims implicit in this premise are: (i) that, according to internalism, actual or potential awareness of something is *required* and (ii) that, according to internalism, it is such awareness on the part of *the subject* that is required. In this section,

I'll defend these claims. In order to do so, I'll need to explain an alternative position they contradict.

The alternative position I have in mind is one that is *suggested* by statements that are fairly common in the epistemological literature. These common statements are probably not intended to stand up to rigorous criticism; they are more like rough and ready ways of characterizing certain aspects of the internalism/externalism controversy. However, if we carelessly take them to be the truth *strictly speaking*, they lead to trouble. In what follows, I'll identify these common statements as well as the alternative position (on what internalism is) that they suggest. But I want to emphasize that I am not so much criticizing anyone for endorsing them as highlighting a confusion we need to be wary of.

The first common statement is that internalism has something to do with the conditions of justification being internal. Another is that a condition is internal only if its satisfaction is in some fairly direct way (e.g., on reflection alone) cognitively accessible. But accessible to whom? Clearly it is possible for a condition's satisfaction to be accessible to one cognizer and inaccessible to another. It doesn't make sense to say of a condition that it is internal *simpliciter*. So how should we understand the fairly common practice of saying, without any explicit relativization, that certain conditions of justification are internal? My recommendation is that we take such sayings as involving an implicit relativization. Claims in the epistemological literature of the form 'C is an internal condition' should be understood as saying that C is internal to normal adult humans. This seems to be the sort of thing that those who speak of internal conditions (without mentioning any relativization) have in mind.

Now, returning to the first statement, what exactly does internalism have to do with the conditions of justification being internal? Well, some externalists say that a reliability condition (e.g., one satisfied by a belief just in case it is produced by a reliable belief-forming process) is sufficient for justification. Internalists complain that this condition is not an *internal* condition—its satisfaction isn't cognitively accessible on reflection alone to normal humans—and so it isn't sufficient for the justification of our beliefs. This sort of interchange suggests that the disagreement between internalists and externalists has to do with whether the conditions necessary for justification are internal conditions: internalists say the conditions necessary for justification *are* internal; externalists say they *aren't*. This way of understanding the internalism/externalism controversy is the alternative position that I said was suggested by some common statements in the epistemological literature.

One way of expressing this position (or at least a part of it) is to say that a sufficient condition for being an internalist is endorsement of:

> I_1: At least one of the necessary conditions of justification is internal (i.e., internal to normal adult humans).

A little reflection shows that this position (i.e., that endorsement of I_1 is sufficient for being an internalist) is contradicted by claims (i) and (ii) implicit in premise (I) of my dilemma. Thus, by explaining what is wrong with this position I will be able to defend premise (I).

Suppose that endorsement of I_1 were sufficient for being an internalist. Then an internalist could consistently endorse the following:

> Julie believes that p. That belief satisfies each of the conditions that are together necessary and sufficient for justification. Furthermore, all of these conditions are internal conditions insofar as they are internal to normal adult humans. But Julie isn't normal. As a result of her abnormality, none of the necessary conditions of justification is internal to *her*. In fact, Julie's belief that p is justified even though she isn't aware (or potentially aware) of anything that contributes to the justification of her belief that p. What's important for justification is that her belief satisfies each of the necessary conditions of justification—not that any of these conditions or anything contributing to the justification of her belief is internal to her (though it's true that each of these conditions happens to be internal to normal adult humans).

But no internalist would say this. No internalist would allow that Julie's belief is justified despite the fact that Julie isn't aware of anything that contributes to the justification of her belief.

In response to the Julie example, one repair of I_1 that naturally comes to mind is:

> I_2: At least one of the necessary conditions of justification is internal *to the subject* (i.e., to the person holding the belief whose justification is at issue).

Unfortunately, endorsement of I_2 is also insufficient for being an internalist. For a supporter of I_2 could accept the following:

> Tanner believes that p. That belief satisfies each of the necessary conditions of justification. Furthermore, each of these conditions is (contingently) internal *to Tanner*. However, there is a possible world in which Tanner believes p and his belief that p satisfies each of the necessary conditions of justification although none of those conditions is internal to him. In that world, Tanner's belief that p is justified even though he isn't aware (or potentially aware) of anything that contributes to the justification of his belief that p. What matters for justification is that the belief in question satisfies each of the conditions necessary for justification—not

> that any of those conditions (or anything contributing to the belief's justifycation) is internal to the person holding the belief. It just happens to be the case that, in the actual world, each of the necessary conditions of justification is internal to Tanner.

But no internalist would say this either. An internalist would insist that it *couldn't* be the case that Tanner's belief is justified when he isn't aware (or potentially aware) of anything contributing to the justification of his belief.

The Julie case supports claim (ii) from premise (I) (i.e., that it is *the subject's* awareness that is at issue) and the Tanner case supports claim (i) (i.e., that this awareness on the part of the subject is *required*). Together, these two cases make it clear that it isn't sufficient for being an internalist to say that some necessary condition of justification is internal to normal humans. What's required for being an internalist is to say that some necessary condition of justification or something contributing to the belief's justification *must be* internal to the subject.[6] Must be for what? The obvious answer is that the condition or justification contributor must be internal if the belief in question is to be justified. So if we construe cognitive accessibility in terms of being actually or potentially aware we can conclude that a *necessary* (though perhaps not a sufficient) condition for being an internalist is endorsement of:

> I_3: S's belief B is justified only if (i) there is something, X, that contributes to the justification of B—e.g., evidence for B or a truth-indicator for B or the satisfaction of some necessary condition of B's justification—and (ii) S is aware (or potentially aware) of X.

And if a necessary condition of being an internalist is endorsement of I_3, then premise (I) of my objection to internalism is true.[7]

Before moving on, I'd like to note briefly one implication of premise (I), namely, that some defenses of internalism use the term 'internalism' in an inappropriate way. For example, Feldman and Conee claim (2001: 2) that it is sufficient for being an internalist that one endorses what they call 'mentalism', the view that justification is determined by the subject's mental states *whether or not the subject is aware (or potentially aware) of those mental states*. John Pollock (1986: 133-34) makes a similar claim when he says that a belief's justification is a function of those states of the believer that are accessible to her automatic processors, *whether or not those states are epistemically accessible (or potentially epistemically accessible) to the believer*.[8] If this were the right way to think about internalism, then someone endorsing the following would correctly be identified as an internalist:

> The justification of our beliefs is determined by those of our mental states that are of kind K. It is highly uncommon for a person to be *aware* (or

> even potentially aware) of mental states of kind K. But it isn't at all uncommon for a person to have mental states of kind K that justify her beliefs. As a result, most of our justified beliefs are justified in virtue of our being in mental states we aren't aware of (or even potentially aware of). Thus, most of our justified beliefs are justified despite the fact that we aren't aware of *anything at all* contributing to their justification.[9]

But no one who held such a view is plausibly construed as an internalist.[10] So the fact that premise (I) conflicts with these accounts of internalism counts, if anything, *in favor* of premise (I) rather than against it.

2. THREE KINDS OF AWARENESS

Having defended premise (I), I now turn to premises (III) and (IV)—the two horns of my dilemma—since, as I noted earlier, premise (II) needs no defense. In laying out these three premises, I alluded to a particular kind of conceptual awareness without saying what kind it is. In order to say what kind it is, I'll need to first say something about the distinction between conceptual and nonconceptual awareness.

Conceptual awareness of X is awareness of X that involves the application of a concept to X. Nonconceptual awareness, by contrast, doesn't involve the application of any concepts. Cows and dogs presumably experience pain of some sort. And presumably these animals are *aware* of such experiences. Yet although they are aware of these experiences, it seems likely that they do not apply any concepts to them. Humans too can be nonconceptually aware of experiences they undergo. The difference is that we are also able to be *conceptually* aware of those experiences (by applying concepts to them) whereas dogs and cows presumably aren't able to be conceptually aware of their experiences.[11]

In light of this understanding of conceptual awareness, we can distinguish two species of it corresponding to two kinds of conceptual awareness requirements for a belief's justification. One kind of conceptual awareness requirement is satisfied by S's belief B only if S conceives of the relevant object of awareness as *contributing to B's justification* (or as *indicating B's truth* or as *being relevant in some way to the appropriateness of holding B*). I will call this sort of requirement a 'conceptual$_1$ awareness requirement' and the awareness it involves 'conceptual$_1$ awareness'. This is the sort of conceptual awareness that is the focus of premise (III). All other conceptual awareness is conceptual$_2$ awareness. It too involves the application of a concept but it doesn't involve application of the sort of concept associated with conceptual$_1$ awareness. For the purposes of my argument, therefore,

there are three kinds of awareness: conceptual$_1$ awareness, conceptual$_2$ awareness and nonconceptual awareness. I will discuss each in turn and, in doing so, defend premises (III) and (IV).

2.1 Conceptual$_1$ Awareness

Suppose the awareness mentioned in I_3 is actual conceptual$_1$ awareness. One way to guarantee this is to interpret I_3 as follows:

> I_4: S's belief B is justified only if (i) there is something, X, that contributes to the justification of B and (ii) S is aware of X in such a way that S justifiedly believes that X is in some way relevant to the appropriateness of holding B.[12]

Now consider the following familiar problem that arises in connection with I_4.[13] In order for S's belief B to be justified, I_4 says that S must have the further justified belief (with respect to something, X_1, that contributes to the justification of S's belief B) that:

> P_1: X_1 is in some way relevant to the appropriateness of holding B.

And according to I_4, in order for her belief that P_1 to be justified S must have the further justified belief (with respect to something, X_2, that contributes to the justification of S's belief that P_1) that:

> P_2: X_2 is in some way relevant to the appropriateness of believing that ⌜X_1 is in some way relevant to the appropriateness of holding B⌝.

And in order for her belief that P_2 to be justified, S must have the further justified belief (with respect to something, X_3, that contributes to the justification of S's belief that P_2) that:

> P_3: X_3 is in some way relevant to the appropriateness of believing that ⌜X_2 is in some way relevant to the appropriateness of believing that ⌜X_1 is in some way relevant to the appropriateness of holding B⌝⌝.

And so on. On this *actual conceptual$_1$ awareness* construal of I_3, therefore, one has a justified belief only if one actually has an infinite number of justified beliefs of ever-increasing complexity. But most of us find it exceedingly difficult even to grasp a proposition like P_5 or P_6 in such a series, much less believe it with justification. Consequently, it's very difficult to see how a supporter of I_4 could resist the conclusion that none of our beliefs is justified. The very ease with which this skeptical conclusion follows from I_4 gives us a reason to reject it.[14]

Now, as I said, the above sort of regress problem—like I_4, the version of conceptual$_1$ awareness internalism it afflicts—is familiar. But perhaps other,

less familiar forms of conceptual$_1$ awareness internalism can escape this sort of difficulty. Consider the suggestion that what internalists require for justification is merely the *potential* for conceptual$_1$ awareness of the relevant fact (rather than actual awareness). A person has potential conceptual$_1$ awareness of a fact if she *could have* a justified belief that the fact obtained. What exactly is involved in this 'could have'? The suggestion certainly isn't that it is merely logically or metaphysically possible that the subject has a justified belief that the fact obtains. Rather, the modality in question has to do with the subject's *abilities*. Thus, 'S could do A' should be understood as something like 'on reflection alone S is able to do A'. This gives us:

> I_5: S's belief B is justified only if (i) there is something, X, that contributes to the justification of B and (ii) S is *able on reflection alone* to be aware of X in such a way that S justifiedly believes that X is in some way relevant to the appropriateness of holding B.

Although I_5 manages to avoid requiring for justification the actual possession of an infinite number of increasingly complicated beliefs, it still leads to trouble. For in order to have the justified belief B, S must be able on reflection alone to justifiedly believe that P_1. And to justifiedly believe that P_1, S must be able on reflection alone to justifiedly believe that P_2. Thus, to justifiedly hold B, S must be able on reflection alone to be able on reflection alone to justifiedly believe that P_2. Given the plausible assumption that *being able on reflection alone to be able on reflection alone* reduces to *being able on reflection alone*, we may conclude that for every P_n in the series, S is justified in her belief B only if she is able on reflection alone to justifiedly believe that P_n. But, as was noted above, one needn't go very far in the series before one comes to a proposition that no human is able to grasp let alone justifiedly believe. So although the *potential awareness* option avoids requiring the actual possession of an infinite number of justified beliefs, it is stuck with requiring the ability to justifiedly hold beliefs of ever-increasing complexity. Like I_4, therefore, I_5 too has skeptical implications that give us a reason to reject it.

Notice that there are two sorts of regresses associated with the above *doxastic* versions of the conceptual$_1$ awareness requirement. First, there is what we might call 'the mental state regress'. In the case of both *actual belief* and *potential belief* understandings of the conceptual$_1$ awareness requirement, what is needed is either an infinite number of beliefs or the potential for that many beliefs. But in addition to the mental state regress (either actual or potential), there is also a complexity regress. And it is the latter that I've relied on in my objections to the above doxastic versions of the conceptual$_1$ awareness requirement. It is because the doxastic versions of the conceptual$_1$ awareness requirement imply that a belief is justified only if

one is able to hold justified beliefs of *ever-increasing complexity* that they are so implausible.

One could try to save the internalist from these problems by dropping one or more of the assumptions I've made along the way. For example, one might suggest that mere belief, not *justified* belief, is required for conceptual$_1$ awareness. But if the internalist has the intuition that merely having a justification contributor isn't enough—that the subject must also have some sort of conceptual$_1$ awareness of that contributor (which includes believing *that* it is a justification contributor)—it seems highly doubtful that the internalist will be impressed by the *mere belief* (no matter how unjustified or insane) that the thing of which she is aware is a justification contributor.

One might also suggest that conceptual$_1$ awareness could involve conceptualization without involving belief. Paul Moser (1989: 186-87) distinguishes doxastic or propositional awareness of X (which involves predicating something of X) from conceptual awareness of X (which involves categorizing X according to some classificatory scheme). Perhaps I_3 should be understood so that it requires merely that the subject be in some sort of cognitive contact with the justification contributor and that she categorize it in the appropriate way. Then there is no requirement that she actually form any judgment. In other words, perhaps the awareness required is conceptual$_1$ but not doxastic.[15]

But concept application can be correct or incorrect; and it can be justified or unjustified in much the same way that believing can be justified or unjustified. So even at the level of concepts, making conceptual$_1$ awareness a requirement for justification gives rise to regress problems (involving ever-increasing complexity and, therefore, skepticism).[16] To see this, consider the following nondoxastic conceptual$_1$ awareness version of I_3, which applies to justification for both concept application and belief:

> I_6: S's belief or concept application, Y, is justified only if (i) there is something, X, that contributes to the justification of Y and (ii) S is aware of X in such a way that S justifiedly applies to X the concept of *being in some way relevant to the appropriateness of Y*.[17]

According to I_6, S's belief B is justified only if

> A_1: S's application to X_1 (a contributor to the justification of B) of the concept *being in some way relevant to the appropriateness of B*

occurs and is justified. And according to I_6, A_1 is justified only if

> A_2: S's application to X_2 (a contributor to the justification of A_1) of the concept *being in some way relevant to the appropriateness of* A_1

occurs and is justified. Likewise, I_6 says that A_2 is justified only if

> A_3: S's application to X_3 (a contributor to the justification of A_2) of the concept *being in some way relevant to the appropriateness of* A_2

occurs and is justified. And so on.

Now consider the concept that is applied in A_3. Spelled out more fully it is:

> being in some way relevant to the appropriateness of S's application to X_2 of the concept ⌜being in some way relevant to the appropriateness of A_1⌝

which, spelled out even more fully, is:

> being in some way relevant to the appropriateness of S's application to X_2 of the concept ⌜being in some way relevant to the appropriateness of S's application to X_1 of the concept ⌜being in some way relevant to the appropriateness of B⌝⌝.

Not an easily digestible concept. And of course it is only the tip of the iceberg when it comes to the complexity we are facing with this regress. Thus, according to I_6, justification for a belief B doesn't depend on the subject having an infinite number of beliefs in propositions of ever-increasing complexity. But it does depend on the application of an infinite number of concepts of ever-increasing complexity.[18] And, for reasons similar to those discussed above in connection with the *potential belief* version of the conceptual$_1$ awareness requirement, it won't help to require merely the potential for applying an infinite number of concepts of ever-increasing complexity. That way too lies skepticism.

Thus, the familiar sort of regress problem that is associated with requiring actual awareness in the form of a further justified belief can't be avoided if one insists that the awareness required for justified belief is conceptual$_1$ awareness. It won't help to require only *potential* conceptual$_1$ awareness. Nor will it help to require only *nondoxastic* conceptual$_1$ awareness. The problem is that the conceptual$_1$ awareness that is being required for the justification of a concept application or a belief involves the application A of a concept whose content is at least slightly more complex than the content of the concept application or belief for whose justification A is required.[19] Thus, since the internalist is likely to deny that unjustified concept application can play a role in the justification of our beliefs (or concept applications) and since she requires conceptual$_1$ awareness, she is forced into a skepticism-inducing complexity regress. We may conclude, therefore, that requiring conceptual$_1$ awareness—whether potential or actual, whether doxastic or nondoxastic—leads to regress

problems. This establishes premise (III) of my argument (since conceptual$_1$ awareness is the sort of conceptual awareness that (III) is speaking of).

2.2 Nonconceptual Awareness

Some internalists are familiar with the regress problems associated with requiring conceptual$_1$ awareness. Moser and Fumerton, for example, are internalists who recognize these problems and, because of this recognition, opt for the second horn of my dilemma by saying that the awareness required for justification is nonconceptual.[20] In taking this route, such internalists face an objection from anti-foundationalists who say that what is nonconceptual—and therefore not capable of being justified—cannot confer justification on a belief based on it (cf. BonJour 1985: ch. 4). The nonconceptual awareness internalist's response to this objection is just to deny an assumption on which the anti-foundationalist objection relies—the assumption that only what is conceptual can confer justification (cf. Moser 1989: 193-94; Fumerton 1995: 74-75).

Externalist foundationalists will have no complaint with this response. For they agree that justification can be conferred by what is nonconceptual. But an externalist foundationalist can point out that in endorsing that response, the internalist opens herself up to a prominent objection to externalism. The objection I have in mind goes something like this:

> *The "Subject's Perspective" Objection to Externalism*: The externalist proposes an analysis according to which S1's belief that p is justified so long as there exists something X contributing to that belief's justification (where X is something like a reliable belief-forming process leading to the formation of the belief in question). The idea is that S1's belief is justified in virtue of X's existence even if S1 isn't aware of X—even if S1 doesn't conceive of X as something relevant to the appropriateness of her belief that p. Now I'll grant [says the proponent of this objection] that if someone else, say S2, were aware of X and conceived of it *as* a contributor to the justification of S1's belief that p, then S2 would have a reason for thinking that p is true. But this doesn't at all suggest that S1's belief that p is justified. True, X is relevant in some way to the appropriateness of S1's belief. But this could be so even if S1 didn't conceive of X as being in any way relevant to the appropriateness of her belief that p. In such a case, it would—from S1's subjective perspective—be an accident that her belief is true.

This objection, or something very much like it, is widely endorsed by internalists (cf. BonJour 1985: 43; Fumerton 1995: 116; Lehrer 1990: 162; Moser 1985: 129). Indeed, I would go so far as to say that a putative

internalist's commitment to internalism would be in doubt if she said that there was nothing to the objection. My claim is that the internalist who rejects the first horn of the dilemma I've proposed (and imposes only a *non*conceptual awareness requirement on justification) is subject to this very same objection to externalism. This should make her very unsympathetic to that objection. But that lack of sympathy puts her allegiance to internalism in question. And it leaves her without much of a complaint against externalism.

Obviously, I need to defend my claim that the internalist who opts for a nonconceptual awareness interpretation of I_3 is subject to the above objection to externalism. Consider an externalist view according to which it is necessary for the justification of S's belief B that B is produced by a reliable belief-forming process. And suppose that S's belief B is produced by a belief-forming process token of a relevant process type that is, in fact, reliable (call this process token 'RP').[21] Now imagine that a proponent of this reliabilist view is presented with the Subject's Perspective objection to externalism and is impressed by it. She decides to add to her account of justification the requirement that S is nonconceptually aware of the reliable process token in question—in this case RP. Will that pacify those who endorse the Subject's Perspective objection to externalism?

It shouldn't. For since the awareness required is nonconceptual, a person can have the required awareness of RP without conceiving of RP in any way—without categorizing it according to any classificatory scheme. But then a person can be nonconceptually aware of RP without conceiving of RP as relevant at all to the appropriateness of her belief. According to the Subject's Perspective objection to externalism given above, if S does not conceive of RP as something relevant to the appropriateness of her belief then, even if RP is relevant to the appropriateness of her belief, it is an accident from S's perspective that her belief is true. Clearly this supposed problem is not solved by adding the requirement that S is *non*conceptually aware of RP.

2.3 Conceptual$_2$ Awareness

Would it help if we added instead the requirement that S has a conceptual$_2$ awareness of RP? No. For S could satisfy this sort of requirement simply by being aware of RP and applying some concept or other to it. And that means that S can have a conceptual$_2$ awareness of RP without conceiving of RP as relevant in any way at all to the appropriateness of her belief B. But then, according to the Subject's Perspective objection to externalism, even if this added requirement were satisfied, it would still be an accident from S's subjective perspective that B is true. For although S applies a concept to RP, she doesn't apply the *right sort* of concept to it. She doesn't apply a concept that involves her conceiving of RP as contributing in

some way to B's justification (or as indicating that B is likely to be true or some such thing). The only way to guarantee that she *does* apply such a concept to RP is to have B satisfy a conceptual$_1$ awareness requirement. Thus, we are forced to concede that by imposing only a conceptual$_2$ awareness requirement, the internalist is vulnerable to the Subject's Perspective objection.

I argued above (in section 2.2) that if the awareness required is nonconceptual, then internalism falls prey to the Subject's Perspective objection to externalism. As we can now see, however, we get the same result if the awareness required is conceptual$_2$ awareness. Since, by definition, nonconceptual awareness and conceptual$_2$ awareness are the only kinds of awareness other than conceptual$_1$ awareness, this establishes premise (IV) of the main argument of this paper—namely, that if the awareness required for justification is not conceptual$_1$ awareness, then internalism is subject to a prominent objection to externalism.

To sum up, we have seen that internalism is motivated largely by the intuitions captured in the Subject's Perspective objection to externalism. But those intuitions, if taken seriously, imply that no belief is justified unless the person holding it is able to justifiedly believe infinitely many other propositions of ever-increasing complexity (or justifiedly apply infinitely many concepts of ever-increasing complexity). The only way to avoid this implication is to make the required awareness something other than conceptual$_1$ awareness. But in doing so, one violates the very intuitions that motivated internalism in the first place.

3. AGAINST INTERNALISM

I have now defended premises (I), (III) and (IV) of my objection to internalism. Those three premises (along with the uncontroversial second premise) entail that internalism is subject either to regress problems or to a prominent objection to externalism. In this section I will complete my argument by defending premise (V) according to which this disjunction of problems implies that we shouldn't endorse internalism.

3.1 Regress Problems

Why not simply admit that justification requires conceptual$_1$ awareness and insist that internalism is true despite the fact that it leads to the sort of regress problems I mentioned in section 2.1? My claim was that, if justification requires conceptual$_1$ awareness, internalism leads to regress problems that suggest that we don't have any justified beliefs. I took these

implications to be a problem for internalism. But some internalists are known for looking askance at the suggestion that skeptical implications are sufficient to discredit a view (cf. BonJour 1985: 12-13; Fumerton 1995: 42-43). So why can't the internalist just confess that her internalism leads to the noted regress problems and then use modus ponens where I use modus tollens, accepting the skeptical consequences rather than rejecting her internalism?

It's worth noting that internalists such as Fumerton (1995: 80-81) resist this sort of response even though they think skeptical implications shouldn't automatically discredit a position. And I think there is a good reason for this resistance. It is *one thing* to think that justification clearly requires something such as a good reason and then to acknowledge that most of our beliefs lack such a reason. An internalist (open to skepticism) who did this would insist that if we have no good reason for any of our beliefs, so much the worse for our beliefs; she wouldn't be inclined to think we don't need a good reason for them after all. But it's *another thing* to think that, in order for any belief to be justified, the person holding it must have (or have the ability to form) an infinite number of beliefs of ever-increasing complexity. For in that sort of case, the problem isn't merely that the requirement has skeptical implications. It's also that the requirement just seems excessive, especially when stated so bluntly. I think this is why the internalist, in response to my discussion of conceptual$_1$ awareness, is not (and shouldn't be) inclined to use modus ponens where I use modus tollens.[22]

It may be helpful to think of these matters in terms of Chisholm's distinction between particularism and methodism (cf. Chisholm 1982: ch. 5). Particularism is the view that we can use, as a starting point in our attempt to identify the criteria for justification, our knowledge of which beliefs are justified and which aren't. Methodism is the view that we should rely on our knowledge of what the criteria for justification are to determine which beliefs are justified and which aren't. The objection to premise (V) that I'm considering in this subsection is a methodist objection. It says that we can just tell that the internalist's proposed criterion for justification (namely, that the subject must have a conceptual$_1$ awareness of something contributing to the justification of her belief) is correct and that if that criterion leads to the conclusion that none of our beliefs are justified, we should just accept that skeptical conclusion.

Now perhaps the methodist approach is, on occasion, a good one. But there are times when the extreme absurdity of the consequences implied by the criterion we started with demands that we use modus tollens with the particularist rather than use modus ponens with the methodist. And this is just such a time. For the skepticism implied by the conceptual$_1$ awareness requirement is global. Every belief of every believer is unjustified. This is

because every believer (with a finite mind) lacks the ability to grasp an infinite number of propositions or concepts of ever-increasing complexity. It turns out, therefore, that if the conceptual$_1$ awareness requirement on justification is correct it is literally impossible for a finite mind to have justified beliefs. Surely this implication is a good reason to reject the proposed criterion. It is doubtful *in excelsis* that the sort of justification epistemologists are trying to analyze is *necessarily* unexemplifiable by the beliefs of finite cognitive subjects such as ourselves.

Furthermore, if the internalist were to accept the consequence that none of our beliefs is justified, her position would be self-defeating in an important way. Her extreme confidence in the correctness of the conceptual$_1$ awareness requirement on justification has led her to the conclusion that that very confidence was misplaced—misplaced in the sense that it is attached to a belief she is not justified in holding. But once she concedes to her particularist opponent that that confidence is misplaced, it is difficult to see why anyone should approve of her persistence in endorsing the conceptual$_1$ awareness requirement. We can make sense of this sort of persistence in the face of absurd consequences only if the persister has extreme confidence in the correctness of the requirement she endorses. But it isn't easy to be understanding when we know that the person with this extreme confidence realizes that that confidence is misplaced.

3.2 The Subject's Perspective Objection to Externalism

Given the conclusions reached so far, the internalist who wants to resist my conclusion is forced to say something like this: "I agree that my position is vulnerable to the Subject's Perspective objection to externalism. But that doesn't mean I shouldn't endorse internalism." Is this an acceptable response? No. For, as I mentioned briefly in section 2.2, this sort of response leads to two kinds of trouble: it puts one's internalism in question and, more importantly, it leaves one without a good reason to prefer internalism to externalism. Let's look at both of these problems in a little more detail.

In response to the suggestion that one's internalism is in doubt if one's view is subject to the Subject's Perspective objection to externalism, an internalist might say the following: "My view is that noninferential justification supervenes on some sort of direct nonconceptual awareness. Clearly, that isn't an externalist view. So my internalism can't be in doubt."[23] The problem with this response is its assumption that, in denying that a position is an internalist one, I'm committed to saying that it is an externalist one. But that assumption is false. Let me explain.

In section 1, I identified one necessary condition for being an internalist, namely, endorsement of I_3. Now consider:

> *I_7*: Each of the necessary conditions of justification is internal (i.e., internal to normal adult humans).

Endorsement of I_7 is insufficient for being an internalist for the same reason that endorsement of I_1 is insufficient for being an internalist—namely, endorsement of either or both is compatible with rejecting I_3[24] But endorsement of I_7 seems to count very strongly against one's being an externalist. So it looks like it is possible for one to be neither an externalist nor an internalist: one need only endorse I_7 and reject I_3. And this shows that the inference from "position X is not an externalist position" to "it is a mistake to think that position X isn't an internalist one" is illegitimate.

One might try to resist my suggestion that a view can be neither internalist nor externalist. After all, the view that all nonexternalist epistemologists are internalists has been a fairly standard one in the epistemological literature.[25] But it is also a fairly standard assumption in the philosophical literature (or at least it follows from entrenched views on what counts as an internalist and what counts as an externalist) that endorsing I_7 is sufficient for *not* being an externalist and that rejecting I_3 is sufficient for not being an internalist. So something has to give. And, although we are speaking of a technical term of art, I still think that, over the past twenty-five years, epistemologists have come to have *some* sense of what is intended by the terms 'internalism' and 'externalism'. This sense lends stronger support to the two views I just mentioned—i.e., that endorsing I_7 is sufficient for not being an externalist and that rejecting I_3 is sufficient for not being an internalist—than it does to the view that one is automatically an internalist if one is not an externalist. I therefore stand by my claim that one's commitment to internalism is at least *questionable* if one's view is vulnerable to the Subject's Perspective objection to externalism.

The other, more significant, problem I mentioned is that one is left without a motivation for preferring internalism to noninternalist views. Conceptual$_1$ awareness internalism is out of the question. That leaves us with either externalism or some form of what I will call 'nonconceptual$_1$ awareness internalism' where by 'nonconceptual$_1$ awareness' I mean awareness that isn't conceptual$_1$ awareness—i.e., awareness that is either nonconceptual awareness or conceptual$_2$ awareness. But why prefer nonconceptual$_1$ awareness internalism to externalism? As I read the literature, there are three main reasons for preferring internalism (of some kind or other) to externalism. The first is that externalism is vulnerable to the Subject's Perspective objection. But of course that reason won't help the nonconceptual$_1$ awareness internalist we are now considering. For she quite sensibly rejects the Subject's Perspective objection since it applies to her own

position and since endorsing it commits one to the regress problems discussed in section 2.1.

The second main reason for preferring internalism to externalism is the view that epistemic justification should be understood *deontologically*. There is much to be said in connection with this supposed reason for internalism. But for now, the only thing we need to mention is that if this view lends any support at all to internalism, it is only to a conceptual$_1$ awareness version of internalism. For it is typically *subjective* deontological justification–i.e., deontological justification thought of as epistemic blamelessness—that is used in support of internalism.[26] But I can have *nonconceptual*$_1$ awareness of whatever you please—good reasons for my belief, strong evidence for it, the reliable process by which it is formed or its satisfying some other necessary condition of justification–without conceiving of the object of such awareness as being even remotely relevant to the appropriateness of my belief. This makes it hard to see how having some sort of nonconceptual$_1$ awareness necessarily moves one in the direction of being more epistemically blameless. However, we *can* see how having conceptual$_1$ awareness seems to help in this regard. Conceptual$_1$ awareness affects our subjective perspective on the appropriateness of our beliefs and, consequently, appears to be relevant to determining the degree to which we are blameless. So we can see how a deontological conception of justification might motivate a *conceptual*$_1$ awareness requirement even though it doesn't seem to provide any motivation for a nonconceptual$_1$ awareness requirement. I don't have the space to develop this point more fully here (it probably deserves a paper in itself), but the basic idea should be clear.

The third of the three main reasons for preferring internalism to externalism is that externalism is, allegedly, subject to the following objection:

> *The "Philosophical Seriousness" Objection to Externalism*: Externalists have not correctly analyzed philosophically interesting epistemic properties. They have either changed the subject to focus on properties that haven't traditionally been the focus of epistemological inquiry or they have failed to fully appreciate and understand the philosophical depth and implications of the properties that have been the focus of traditional epistemology. That externalists have missed the boat in this way is most evident when one looks at their flippant attitudes toward skepticism.

I have devoted an entire paper (Bergmann 2000b) to explaining why this objection fails. The basic idea of that paper is that the Philosophical Seriousness objection fails because the sort of reasoning on which it relies is, if successful at all, too powerful. For if the internalist is right in saying that

the epistemic properties on which externalists focus are philosophically uninteresting, then no epistemic properties at all are philosophically interesting—in which case we've been given no reason to prefer internalism to externalism.

Without these three reasons for preferring internalism to externalism, it is very difficult to see why someone would do so.[27] One wants to ask, "What good does having nonconceptual$_1$ awareness do? How does it do more to contribute to justification than the satisfaction of external conditions does?". Again, the answer can't be that it prevents it from being an accident from the subject's perspective that her belief is true (or likely to be true or justified). But then why think awareness of something is so important if it doesn't have the result that the subject conceives of the thing of which she is aware as relevant in some way to the appropriateness of her belief?

One answer is suggested by Paul Moser. Requiring that the subject have nonconceptual awareness of a justification contributor enables one to avoid an objection to externalism that happens to sound very much like the Subject's Perspective objection. That similar sounding objection is this: *externalism is problematic insofar as it says that a belief can be justified even though, relative to all the subject is nonconceptually aware of, the belief isn't likely to be true* (Moser 1989: 76) The reason Moser's view avoids *that* objection is that another requirement he imposes on justification (in addition to his nonconceptual awareness requirement) is, roughly, that P, the content of the subject's belief, must explain E (the object of the subject's required nonconceptual awareness) better than any other competing proposition S is aware of (1989: 136-37). If P is the best explanation of E and if S is nonconceptually aware of E, then it is *not* the case that *relative to all the subject is nonconceptually aware of* her belief that P isn't likely to be true.

But the fact that Moser's view avoids this *other* objection to externalism shouldn't appease those with internalist sympathies. For, as Moser himself acknowledges (1989: 164), S's belief can satisfy Moser's requirements for justification even if S has no idea that P is the best explanation of E. In fact, given that S needs only *nonconceptual* awareness of E, S needn't even conceive of E as something that might be explained by P or as something that is relevant in some way to P. So, in satisfying Moser's requirements for justification, the most that S's belief that P is guaranteed to have going for it is that some object of awareness to which S may have applied no concepts at all is, unbeknownst to S, best explained by P. That certainly needn't make any relevant epistemic difference *from S's own perspective* to the credentials of her belief that P—no more than the fact (supposing it were a fact) that, unbeknownst to S, the process by which S's belief that P is formed is a

reliable one. So we are left without any motivation to prefer nonconceptual$_1$ awareness internalism to externalism.

This completes my defense of premise (V), the claim that a person shouldn't endorse internalism if her position is vulnerable to the Subject's Perspective objection to externalism. She shouldn't because it puts her internalism in doubt and because it leaves her without a reason to prefer internalism to externalism. "But," one might ask, "what advantages does externalism have over a nonconceptual$_1$ awareness version of internalism?" The main advantage is this: the nonconceptual$_1$ awareness version of internalism is committed to imposing an *unmotivated* requirement—i.e., that the subject have some sort of nonconceptual$_1$ awareness—that externalism isn't committed to imposing.

4. TWO EXAMPLES AND THREE INTERNALISTS

I have now defended the premises of the anti-internalist argument given in the introduction. As I noted earlier, a common reaction from the internalists to whom I've shown this argument is the suspicion that its weaknesses will become manifest either when we apply the dilemma it proposes to particular beliefs or when we consider versions of internalism designed specifically to avoid a Sellarsian type of dilemma. Thus, in order to make the force of my argument more evident to those inclined to resist its conclusion, it will be helpful to focus on some specific examples (I'll focus on a physical object belief and a first-person mental state belief) and to consider the implications of the dilemma I've proposed in connection with them. While doing so it will also be useful to evaluate the responses of three well-known internalist philosophers—Paul Moser, Richard Fumerton and Laurence BonJour—to the challenge of avoiding regress problems while imposing a plausible awareness requirement on justification, since seeing how their responses fail will lend further support to my argument.

4.1 A Physical Object Belief and Moser

Suppose that Jack forms the belief that there is before him a large spherically shaped object (call this belief 'B1'). And suppose that he forms this belief as follows. He walks into a well-lit room that is empty except for a large white ball. As a result of fixing his eyes on the ball in the room, he has a visual experience that is of the same type (color-experience-wise) as the experience that a normal human would have in such circumstances (call this visual experience of a large white ball 'WB'). And Jack's having WB causes the formation of his belief B1. To alleviate any concerns an

externalist might have, let's add that B1 is a reliably formed belief. Now, is Jack's belief justified? In order to answer that question, there is another important question that will naturally come to the mind of an internalist: does Jack conceive of WB as relevant in any way to the appropriateness of B1?

The internalist faces a dilemma when she considers this last question in connection with a case like Jack's. For, since she is an internalist, she wants to insist that it does Jack no epistemic good that B1 is reliably formed unless B1 satisfies some sort of awareness requirement. But the awareness requirement must be either a conceptual$_1$ awareness requirement or not. If it is a conceptual$_1$ awareness requirement, then Jack must be aware of some contributor to the justification of B1 and justifiedly apply to it a concept such as *being relevant in some way to the appropriateness of B1*. But that way lies the complexity regress and skepticism. So the internalist is forced instead to require that Jack's belief satisfy some nonconceptual$_1$ awareness requirement (requiring either nonconceptual awareness or conceptual$_2$ awareness).

Suppose the internalist decides to insist on merely nonconceptual awareness of a contributor to the justification of B1. This seems to be the approach adopted by Paul Moser.[28] According to Moser, S's belief that P is justified only if (a) S is nonconceptually aware of (or presented with) evidence E[29] and (b) S has a *de re* nonconceptual awareness of E's supporting P.[30] In the case of Jack, we can think of this as requiring that he is nonconceptually aware of WB (the visual experience of a large white ball) and that he has a *de re* nonconceptual awareness of WB's supporting B1 (the belief that there is a large spherical object in front of him). But clearly B1 could satisfy those two requirements even if, according to the Subject's Perspective objection, it's the case that *from Jack's subjective perspective* it is an accident that B1 is true. For since Jack is required only to be *non*conceptually aware of WB, he needn't conceive of WB as evidence for anything at all. The same comment applies to Jack's nonconceptual awareness of WB's supporting the proposition that there is a large spherical object in front of him. It is a little difficult to make sense of such a nonconceptual awareness. I have some idea of what a nonconceptual awareness of a pain would be like. But I'm not sure I have much of an idea of what a nonconceptual awareness of *WB's supporting a proposition* would be like. However, we know this much: one can have such an awareness without conceiving of WB's supporting a proposition as WB's supporting a proposition. But if Jack doesn't conceive of WB as evidence for anything and doesn't conceive of WB's supporting the content of B1 as WB's supporting that content, it is hard to see how Jack's subjective perspective on B1 has improved in any way.

Moser is sensitive to this sort of problem and tries to solve it by requiring something more than nonconceptual awareness without requiring conceptual awareness of any sort. The something more he requires is the satisfaction of the following requirement for the justification of S's belief that P:

> as a nondeviant result of this awareness [i.e., the *de re* nonconceptual awareness of E's supporting P], S is in a dispositional state whereby if he were to focus his attention only on his evidence for P (while all else remained the same), he would focus his attention on E (1989: 141).

How shall we understand this requirement? Because he wants to avoid regress problems, Moser makes it very clear (1989: 142) that he wants it to be interpreted in such a way that no conceptual awareness is required. But what exactly is it to be disposed to *focus one's attention on E if one were to focus one's attention only on one's evidence for P*? On one reading, it is just to be disposed to focus on E if one were to focus on E. But that can't be what Moser is saying. The most plausible interpretation I can think of is this: to be in the state described is to be disposed to focus on E if one were to focus on what one takes to be one's evidence for P. Or, to put it another way, it is to be disposed to focus on E if one were to focus on what one conceives of as *being one's evidence for P*.

But that is just to be disposed to apply the concept of *being evidence for P* to E (if one were to apply the concept to anything at all). And surely, it wouldn't be sufficient to be disposed to apply that concept in an unjustified way. So it must be a disposition to apply that concept in a justified way. But that will involve a disposition to satisfy further awareness requirements and to have further dispositions to justifiedly apply concepts. There is here a regress of dispositional concept application of the sort Moser is eager to avoid. Thus, requiring merely nonconceptual awareness, even of WB's supporting the content of B1, doesn't enable B1 to escape the disapprobation of the supporters of the Subject's Perspective objection. And there doesn't seem to be any way of requiring more (enough to satisfy these objectors) without imposing some sort of conceptual$_1$ awareness requirement.

Conceptual$_2$ awareness won't help Jack either. For Jack's conceptual awareness of WB will do him no good (according to the supporters of the Subject's Perspective objection) unless he applies the *right sort* of concept to it—the sort of concept that will make his conceptual awareness a conceptual$_1$ awareness. So long as Jack applies only concepts other than ones like *being indicative of B1's truth* or *being a contributor to B1's justification* or *being in some way relevant to the appropriateness of holding B1*, B1's truth will, according to the Subject's Perspective objection, be an accident from Jack's perspective.

In short, because it is possible for Jack to have a nonconceptual$_1$ awareness of an experience like WB without applying (or being able to apply) concepts like *being indicative of B1's truth* to it, it is clear that such awareness—even if it is conceptual—doesn't enable B1 to avoid the censure of the supporters of the Subject's Perspective objection. Therefore, with respect to physical object beliefs like Jack's, the internalist's only choice is to impose a conceptual$_1$ awareness requirement leading to a complexity regress (and skepticism) or to impose a nonconceptual$_1$ awareness requirement (of either the nonconceptual or conceptual$_2$ variety) which leaves one vulnerable to the Subject's Perspective objection to externalism.

4.2 A First-Person Mental State Belief

As we have just seen, the problem with requiring merely nonconceptual$_1$ awareness for the justification of physical object beliefs is this: one could be aware (even conceptually aware) of an experience like WB without conceiving of it as *being relevant in some way to the appropriateness of holding B1*. The only way to avoid this problem is to impose a conceptual$_1$ awareness requirement, which leads to a complexity regress and skepticism. Does this same sort of problem arise in connection with beliefs about first-person mental states? Yes. But it isn't easy to see that it does because it is difficult to think clearly about the type of example needed to *demonstrate* that it does. The remainder of this paper will be devoted to a careful consideration of this very important (and difficult to think about) type of example and its implications for internalists.

Suppose that Lucy is being appeared to redly and that she believes that she is being appeared to redly (call this belief 'B2'). And suppose that there is nothing of which Lucy is aware to which she applies a concept like *being in some way relevant to the appropriateness of holding B2*. It's true that she is aware of her being appeared to redly. And perhaps she even conceives of that experience in some way (e.g., as a rather uninteresting experience). But, due to her severe cognitive malfunction, she doesn't (and isn't able to) conceive of that experience as something relevant to the appropriateness of holding B2.

This may seem a little hard to swallow. We can see how, in the case of the perceptual belief discussed earlier, Jack might be conceptually aware of WB, the visual experience of a large white ball, without applying to it the concept of *being in some way relevant to the appropriateness of holding B1* (where B1 is his belief that there is a large spherically shaped object in front of him). But is it even possible for Lucy to be conceptually aware of her being appeared to redly without applying to it the concept of *being in some way relevant to the appropriateness of holding B2* (where B2 is her belief

that she is being appeared to redly)? I don't see why not. It is one thing to have a nonconceptual$_1$ awareness of one's being appeared to redly. It's another to apply to one's being appeared to redly a concept like *being in some way relevant to the appropriateness of holding B2* (or *being indicative of the truth of B2* or *contributing to the justification of B2*). Granted, the malfunction involved in having such an awareness without applying (or being able to apply) such a concept would be extreme. But that doesn't make it impossible. Perhaps, in addition to suffering from some sort of cognitive defect, Lucy is also in the grip of a philosophical theory which strongly supports her disinclination to apply to her experience of being appeared to redly a concept like *being in some way relevant to the appropriateness of holding B2*. The philosophical theory (which she accepts on the authority of a philosophically astute but malicious prankster) gets Lucy interested in denying what seems to us to be utterly obvious. And the cognitive malfunction makes it possible for her to do what we are incapable of doing, namely, genuinely refraining from a concept application that is so obviously correct.[31]

This example provides us with a case of someone who is appeared to redly, who believes that she is appeared to redly, and who is conceptually aware of her being appeared to redly even though, according to proponents of the Subject's Perspective Objection, it is an accident from her perspective that her belief that she is being appeared to redly is true. For although Lucy is aware of her being appeared to redly and although she applies a concept to the object of that awareness, she doesn't (and isn't able to) apply to it a concept like *being in some way relevant to the appropriateness of holding B2*. It's true that the malfunction involved in this example is extreme. But the point is that it is possible and the proponent of the Subject's Perspective Objection is committed to requiring (for the justification of B2) that such malfunction not be actual.

So even in the case of first-person mental state beliefs, the internalist must choose between the extreme skepticism induced by conceptual$_1$ awareness requirements and the vulnerability to the Subject's Perspective objection that results from imposing only nonconceptual$_1$ awareness requirements. Let's briefly consider what Fumerton and BonJour would say about this example of Lucy's first-person mental state belief.

4.2.1 Fumerton

Because of his concern to avoid regress problems, Fumerton tries to steer clear of conceptual awareness requirements of any kind. He says (1995: 75) that a person's belief that p is (noninferentially) justified only if she is directly acquainted with the fact that p, her thought that p, and the relation of

correspondence holding between her thought that p and the fact that p. Concerning this direct acquaintance he says the following:

> Acquaintance is *not* another intentional state to be construed as a non-relational property of the mind. Acquaintance is a *sui generis relation* that holds between a self and a thing, property or fact. To be acquainted with a fact is not by itself to have any kind of propositional knowledge or justified belief ... One can be acquainted with a property or fact without even possessing the conceptual resources to *represent* that fact in thought ... Acquaintance is a relation that other animals probably bear to properties and even facts... (1995: 74-75).

Clearly, he has in mind some sort of nonconceptual awareness.

In giving the Lucy example, I've already stipulated that she is nonconceptually aware of the fact that p (i.e., the fact that she is being appeared to redly). We can add that she is also nonconceptually aware of her thought that she is being appeared to redly. Finally, we can add that she is nonconceptually aware of the relation of correspondence holding between her thought that she is being appeared to redly and the fact that she is being appeared to redly. As I noted earlier, it is a little difficult to make sense of nonconceptual awareness of things like this relation holding (vs. things like being in pain). But for our purposes, what's important is that we know that this awareness or direct acquaintance is nonconceptual (i.e., it needn't involve the application of any concepts).

Now what happens to Lucy's subjective perspective on B2 when we add to our example that Fumerton's conditions on justification are satisfied by B2? Because the direct acquaintance required is nonconceptual, Lucy can be directly acquainted with the fact that she is being appeared to redly *without conceiving of the object of this awareness as being in any way relevant to the justification or truth of B2* (this is because nonconceptual awareness is the sort of thing that can occur without the application of any concepts to the object of awareness). Furthermore, she can be directly acquainted with the relation of correspondence holding between her thought that she is being appeared to redly and the fact that she is being appeared to redly *even if she has no idea* that the relation of correspondence holds between these two items (again, this is because nonconceptual awareness is the sort of thing that can occur without the application of *any* concepts). Thus, Lucy's belief that B2 can satisfy Fumerton's requirements even if she conceives of her being appeared to redly as no more relevant to B2 than is the mild pain in her left knee. It is, therefore, exceedingly difficult to see how these direct acquaintances improve things from *Lucy's subjective perspective*. If things were epistemically bleak from her perspective *before* she had these nonconceptual acquaintances Fumerton requires, there is no reason to think

they will be less bleak afterwards. The only way for things to improve from Lucy's perspective is for her to apply to the objects of these direct acquaintances a concept such as *being relevant in some way to the appropriateness of B2*. But to require that is to impose a conceptual$_1$ awareness requirement on justification, something Fumerton refuses to do, knowing full-well that it leads to regress problems. I conclude, therefore, that Fumerton's account of the justification of first-person mental state beliefs is vulnerable to the Subject's Perspective objection.

4.2.2 BonJour

According to BonJour (2001: 30-31), S's belief that she is being appeared to redly is justified if S is being appeared to redly. For to have an experience of being appeared to redly is to have a *conscious* experience of being appeared to redly. And to have a conscious experience of being appeared to redly is to have the experience and to be *aware* of its nonconceptual sensory content.[32] Thus, if Lucy is being appeared to redly, then she is aware of her being appeared to redly. Furthermore, according to BonJour, an awareness of one's being appeared to redly is an awareness of an excellent reason for believing that one is being appeared to redly. The nonconceptual sensory content of that experience is an excellent reason for such a belief because it *matches the description* that is included in the propositional content of that belief.

Let's grant to BonJour that to be appeared to redly entails that one is aware of being appeared to redly and that to be aware of one's being appeared to redly is to be aware of an excellent reason for believing that one is being appeared to redly. That isn't enough for BonJour to avoid the Subject's Perspective objection to externalism. The fact that Lucy's awareness of the sensory content of her experience is an awareness of what *in fact* counts as an excellent reason for holding B2, doesn't imply that Lucy conceives of the sensory content of her experience as providing any indication of B2's truth. The supporter of the Subject's Perspective objection will insist that it is only if Lucy applies to her experience of being appeared to redly a concept like *being indicative of the truth of B2* that B2's truth isn't an accident from Lucy's perspective. And in the example given above, Lucy *doesn't* apply that sort of concept to her experience of being appeared to redly.

According to the proponent of the Subject's Perspective objection to externalism, even if Lucy is in possession of a good reason for B2 she is, at best, in the position of Dr. Watson after he has received all the evidence required for proving that the butler did it. Watson is in possession of a good reason for thinking the butler did it but, unlike Sherlock Holmes, he can't see

that it is a good reason. Now suppose Watson forms the belief that the butler did it. From Watson's subjective perspective, the truth of his belief that the butler did it is, according to the Subject's Perspective objection, an accident.[33] In the same way, although Lucy is aware of something that BonJour thinks is an excellent reason for B2, in my example she doesn't conceive of her being appeared to redly as being any reason at all for B2. So from Lucy's perspective it is (according to the Subject's Perspective objection) an accident that B2 is true; it is an accident even though that belief satisfies the conditions BonJour says are sufficient for its justification. The only way for BonJour to avoid this result is to require that Lucy's awareness of her being appeared to redly be accompanied by an application to that experience of a concept like *being indicative of B2's truth*. But this is for BonJour to insist that justification requires conceptual$_1$ awareness. And that leads to the complexity regress and to skepticism. I conclude, therefore, that because BonJour's awareness requirement isn't a conceptual$_1$ awareness requirement his view too falls victim to the Subject's Perspective objection to externalism.

5. CONCLUSION

Why don't internalists recognize that by avoiding regress problems they make their positions vulnerable to the Subject's Perspective objection? I think it's because they aren't recognizing the possibility of the kind of severe malfunction that occurs in my Lucy example. To see the connection between their failure to recognize this possibility and their failure to foresee the problems I identify, let's consider again what Fumerton does. He wants to avoid imposing conceptual$_1$ awareness requirements on justification. But he doesn't want (I assume) to allow for the possibility that it is an accident from the subject's perspective that her beliefs are true (or at least not if those beliefs are to count as justified). In order to guarantee that his view isn't thus vulnerable to the Subject's Perspective objection (and to do so while avoiding regress problems), he says that noninferential justification depends on *nonconceptual awareness* (or direct acquaintance). The reason this guarantee seems to work is that it is tempting to think that it is impossible for it to be an accident from a person's perspective that her belief that, say, *she is being appeared to redly* is true if she is (nonconceptually) aware of her being appeared to redly. But as the Lucy example, with the severe malfunction it involves, shows us, such a scenario isn't impossible. So the supposed guarantee doesn't work. To avoid the Subject's Perspective objection, something more (i.e., a conceptual$_1$ awareness requirement) is needed. And the reason the failure of that supposed guarantee isn't noticed is

that, unless one thinks about the matter carefully, it can seem almost unimaginable for one's introspective belief-forming practices to be as severely damaged as Lucy's are. In short, although internalists such as BonJour, Fumerton and Moser refrain from actually imposing conceptual$_1$ awareness requirements, they seem to have relied on the (false) assumption that our direct acquaintance beliefs cannot fail to satisfy such requirements.

Internalists have two options. They can say that conceptual$_1$ awareness is required for justification; or they can say that nonconceptual$_1$ awareness is required (i.e., awareness that isn't conceptual$_1$ awareness). If they require the former, they are forced into a complexity regress according to which no belief is justified unless the person holding it is able to grasp concepts that are more complex than the human mind can grasp. But if they require *non*conceptual$_1$ awareness instead, their position is subject to one of the most prominent and influential internalist objections to externalism—in which case they should reject that objection and give up their resistance to externalism. Either way, they should not endorse internalism. And the fact that even such able defenders of internalism as Moser, Fumerton and BonJour (who are particularly sensitive to the issues on which I've focused) cannot escape my dilemma provides further confirmation of the strength of the objection I've proposed.

6. APPENDIX: HETHERINGTON'S DILEMMA

I'm grateful to Stephen Hetherington for bringing to my attention two papers of his (see his 1990 and his 1991) that I hadn't noticed when this paper was first subjected to public scrutiny.[34] In them he proposes a dilemma for internalism that in certain ways resembles the one I've presented here. I think it's fair to say that his argument and mine are two different versions of the same kind of objection to internalism. However, as I shall argue in this appendix, internalists have good reasons not to be troubled by the dilemma Hetherington proposes (though none of those reasons count against the dilemma I've proposed above).

6.1 Hetherington's Dilemma Explained

The basic structure of the dilemma Hetherington proposes in the two papers mentioned above can be captured (in a way that is parallel to my own dilemma) as follows:

(I*) An essential feature of internalism is its endorsement of the Transparency Thesis (TT):

> TT: If some [reason or justification-contributor] W is epistemically internal to S, then its being epistemically internal to S is also epistemically internal to S.[35]

(II*) The internalist either gives up her commitment to TT or she doesn't.

(III*) If she doesn't give up her commitment to TT, then the internalist falls prey to regress problems that imply that "epistemic internalism is an empty concept".[36]

(IV*) If the internalist does give up her commitment to TT, then, given (I*), she gives up her internalism and becomes an externalist.

(V*) If an internalist must either admit that epistemic internalism is an empty concept or give up her internalism, then internalism should be rejected.

(VI*) Therefore, internalism should be rejected.[37]

That Hetherington is proposing this dilemma in his 1991 is obvious.[38] It is also obvious that in his 1990, he is proposing a dilemma of this same *form*. For in that paper, not only does he explicitly formulate his objection as a dilemma, he also clearly says that: (i) internalists are committed to saying things that lead to regress problems, (ii) these regress problems imply that epistemic internalism is an empty concept and (iii) the only way to avoid the regress in question is to give up on internalism.[39] But what isn't so obvious is whether his 1990 says that the principle to which internalists are committed and from which a damaging regress can be generated is *TT*. In order to see that both papers are focused on TT, we will need to take a careful look at what each paper says about the regress that Hetherington thinks is generated by internalism. Once I've established that the above dilemma is the one Hetherington is in fact advancing in both papers, we can then turn to an evaluation of it.

In his 1990, Hetherington describes the regress as follows:

> The regress … is this:
>
> According to epistemic internalism about a condition A of your having a justified belief,
>
> (1) A contributes to your justification
>
> only if
>
> (2) You appreciate that A contributes to your justification
>
> only if
>
> (3) You appreciate that your-appreciating-that-A-contributes-to-your-justification (i.e., the appreciating which is (2)) contributes to your justification

only if
(4) You appreciate that the appreciating which is (3) contributes to your justification
only if
(5) You appreciate that the appreciating which is (4) contributes to your justification
only if... And so on (1990: 246).

It is difficult to make sense of the sentence that appears between the colon and the second ellipsis. However, I take it that what Hetherington had in mind was something like the following:

According to epistemic internalism about a (necessary) condition A of your having a justified belief B,
(1) A contributes to your justification
only if
(2) You appreciate that A contributes to your justification.
This means that (2), like A, is a necessary condition of B's justification. But (2)'s truth contributes to your justification for B only if
(3) You appreciate that your-appreciating-that-A-contributes-to-your-justification (i.e., the appreciating which is (2)) contributes to your justification.
This means that (3), like (2), is a necessary condition of B's justification. But (3)'s truth contributes to your justification for B only if
(4) You appreciate that the appreciating which is (3) contributes to your justification.
This means that (4), like (3), is a necessary condition of B's justification. But (4)'s truth contributes to your justification for B only if …

Now suppose Hetherington is right in saying that the internalist—at least one who endorses "epistemic internalism about a condition A of your having a justified belief" B—is committed to thinking that A contributes to your belief B's justification only if you appreciate that it does. Then it seems right to say that, according to internalists, (2) is also a necessary condition of B's justification. But why conclude from *that* that internalists must say that (2)'s truth contributes to your justification for B only if you appreciate that it does? Why can't internalists say instead that A contributes to B's justification only if you appreciate that it does but (2)'s truth contributes even if you don't appreciate that it does?[40]

In the following passage, Hetherington gives his answer to that question:

> Your appreciating-that-A-is-contributing-to-your-being-justified [i.e., (2)] is contributing to your being justified; need you in turn appreciate that it is doing so? What the epistemic internalist should say is that you must. Otherwise, …your appreciating-that-A-is-contributing-to-your-being-justified [i.e., (2)] is itself epistemically external to your being justified. Yet a necessary condition of A's being epistemically internal to you was your appreciating that it is contributing to your being justified. And if the latter aspect of your situation [i.e., your appreciating-that-A-is-contributing-to-your-being-justified] is epistemically external to your being justified, then no doubt the former [i.e., A] is too (1990: 246).

As I understand this passage, he is saying that if you don't appreciate that (2)'s truth contributes to B's justification (i.e., if (2) is epistemically external to you), then you don't appreciate that A is contributing to B's justification (i.e., then A is epistemically external to you).[41] But this entails that:

> P: If A is appreciated by you as contributing to B's justification, then your appreciating that A is contributing to B's justification is also appreciated by you as contributing to B's justification.

And P is basically identical to TT from premise (I*). This becomes clear once we recognize that, according to Hetherington, for something to be epistemically internal to a subject S "is for it to be appreciated by S *as* being his or her reason" (1991: 858). Given this account of what it is to be epistemically internal to S, we can restate TT as follows:

> TT*: If W is appreciated by S as being her reason, then *S's appreciating that W is her reason* is also appreciated by S as being her reason.

Now clearly P and TT* (which is equivalent to TT) are saying pretty much the same thing. And as we just noted above, Hetherington appeals to P in order to generate the regress he tries to identify in his 1990. So even if it isn't obvious from a casual reading, we may conclude that, according to his 1990, the principle to which internalists are committed and from which the regress is generated is TT.

Let's turn now to Hetherington's 1991 to see how the regress identified there (which is explicitly derived from TT) compares with the regress identified in his 1990. If the regresses are the same, this will lend further support to my claim that both papers are proposing the same dilemma.

In his 1991 Hetherington argues (860-61) that according to TT, a potential justification-contributor (or J-contributor), W1, won't be epistemically internal to S unless each of the following is too:

> R: the appreciating-of-W1-as-epistemically-internal-to-S,
> the-appreciating-of-that-first-appreciating-as-epistemically-internal-to-S
> the-appreciating-of-that-second-appreciating-as-epistemically-internal-to-S, ... (860).

But that isn't quite right. To see why, it's important to remember that, as I noted above, TT is equivalent to TT*. By keeping TT* in mind, we can see more clearly which regress Hetherington seems to have had in mind in his 1991. For according to TT*, a potential J-contributor, W1, won't be epistemically internal to S unless each of the following is also epistemically internal to S (i.e., appreciated by S as being her reason):

> R*: (2*) W1's-being-appreciated-by-S-as-being-her-reason,
> (3*) (2*)'s-being-appreciated-by-S-as-being-her-reason,
> (4*) (3*)'s-being-appreciated-by-S-as-being-her-reason,
> (5*) (4*)'s-being-appreciated-by-S-as-being-her-reason, ...[42]

Furthermore, Hetherington thinks that internalists are committed not only to TT (and, therefore, to TT*) but also to:

> Q: At least one of a belief's potential J-contributors must be epistemically internal if the belief is to be justified.[43]

And if the internalist is committed to both TT and Q she must say that each of the appreciations in a regress like R* must be epistemically internal to S if some belief of hers is to be justified.

Notice that if (2*) must be epistemically internal to S in order for some belief B of hers to be justified, it follows that (3*) is necessary for B's justification (since (3*) just is (2*)'s being epistemically internal to S). And if (3*) must be epistemically internal to S in order for B to be justified, it follows that (4*) is necessary for B's justification (since (4*) just is (3*)'s being epistemically internal to S). And so on. But to say that (3*), (4*), (5*), etc. are necessary for B's justification is very much like saying that (3), (4), (5), etc. from Hetherington's 1990 regress are necessary for B's justification. In fact, the two series of propositions are so similar that I think it's fair to say they constitute the very same regress. Thus, it looks like the regress Hetherington was trying to identify in his 1991 is the same regress he was

trying to identify in his 1990. This confirms my suggestion that both papers are proposing the dilemma identified above.

6.2 Hetherington's Dilemma Criticized

Now that we have before us the dilemma Hetherington proposes we can consider why it is that internalists needn't be troubled by it. One weakness of this dilemma is that Hetherington's standard for who counts as an internalist is implausibly strict. Consequently, many philosophers typically viewed as internalists (such as Fumerton, Moser, and BonJour in his more recent work) don't count as internalists according to his standard. Thus, anyone who thinks that such philosophers do count as internalists will think that Hetherington's objections pose no threat to internalism generally but only to versions of it that are extreme enough to endorse TT.

The problem isn't just that some who claim to be internalists aren't internalists by Hetherington's standards. Indeed, when I argue, at the end of section 1, that endorsing mentalism is not sufficient for being an internalist, I too say that not all who claim to be internalists are internalists. But my standard for being an internalist (i.e., that one require, for justification, some sort of awareness of some J-contributor) is much less strict than Hetherington's TT-requirement and much more in accord with accepted views about who counts as an internalist. This makes my standard easier to defend and it prevents internalists from sidestepping my dilemma in the way they can easily sidestep Hetherington's dilemma.

Hetherington's characterization of internalism in premise (I*) is not only too narrow, it is also inadequately supported. In his 1990, he appeals to P—which, as I noted earlier, is basically the same thing as TT—when he explains how the regress he identifies there is generated by internalism. So it is clear that, in his 1990, he is assuming that internalists are committed to TT. However, he doesn't really defend that claim until his 1991 where he is more explicit about the role of TT in generating the regress that he thinks follows from internalism.

Hetherington begins his 1991 defense of (I*)—the claim that internalists are committed to TT—like this:

> For the Transparency Thesis [TT] to be false is for W1, say, to be epistemically within S without S appreciating it as epistemically within (1991: 859).[44]

But this is a mistake. For, contrary to what Hetherington says, the falsity of TT does not imply that some potential J-contributor, such as W1, is epistemically within S without:

(a) S appreciating W1 as epistemically within.

Instead, the falsity of TT implies that W1 is epistemically within S without:

(b) S appreciating W1's being epistemically internal to S as her reason.

Now if the nonoccurrence of (b) entailed the nonoccurrence of (a), then the Hetherington passage just quoted would be correct. But for (b) not to occur is for it to be the case that:

(c) S doesn't appreciate as her reason W1's being appreciated by her as being her reason.

And for (a) to occur is for it to be the case that:

(d) S does appreciate W1 as being appreciated by her as her reason.

Thus, if (c) and (d) can both be true at once, then the nonoccurrence of (b) doesn't entail the nonoccurrence of (a), in which case the passage quoted above from Hetherington is mistaken. And it seems that (c) and (d) *can* both be true at once. A person can appreciate X as being F without appreciating *X's being F* as her reason. So a person can appreciate W1 as being appreciated-by-her-as-her-reason without appreciating *W1's being appreciated-by-her-as-her-reason* as her reason.[45] Hetherington seems to be conflating talk of "S's failure to appreciate an appreciation as a reason" with talk of "S's failure to appreciate something (which needn't be an appreciation) as *being appreciated as a reason*".

In the end, this sort of conflation results in the failure of his 1991 defense of (I*). To see this, consider the following claim Hetherington offers in support of (I*):

> If the Transparency Thesis is [taken by the internalist to be] false then, for the epistemic internalist, justification can depend solely on facts about S which are in no way epistemically internalised by S. That is, if the Transparency Thesis is [taken by the internalist to be] false, then the epistemic internalist becomes an epistemic externalist (1991: 861).

This claim is false. A person can insist that justification depends on facts which are epistemically internal to S (i.e., facts which are appreciated by S as her reasons) while at the same time denying that those appreciated facts are facts about *something's being appreciated by S as her reason*. Consider, for example, Hetherington's case of a child's belief that there is a cat in front of her. (Call that belief 'B' and call whatever it is that B is based on 'W'.) The internalist can insist that B's justification depends not just on W but also on the child's appreciation of W as her reason (for B). And the internalist can insist on this while denying that B's justification depends on the child's

appreciating as her reason *W's being appreciated by her as her reason*. In this way, the internalist can do what Hetherington, in the quotation above, says she must do (i.e., deny that "justification can depend solely on facts about S which are in no way epistemically internalised") and yet still reject the Transparency Thesis.[46]

I've argued that, in both his 1990 and his 1991, Hetherington proposes the dilemma mentioned in the second paragraph of this appendix—a dilemma in which the Transparency Thesis plays a key role. And I've argued that internalists needn't be troubled by that dilemma, first, because they won't accept his characterization of what is essential to internalism and, second, because he gives them no good reason to think they should accept that characterization. None of these remarks applies to my dilemma. Thus, the fact that internalists can escape Hetherington's objection shouldn't make them optimistic about being able to respond adequately to mine.

ENDNOTES

* I'm delighted to include this paper in a volume honoring Alvin Plantinga, largely because he's the person to whom I consider myself most indebted philosophically. My thanks to Jeffrey Brower, Jan Cover, Richard Fumerton, Stephen Hetherington, Timothy McGrew, Kevin Meeker, Trenton Merricks, Alvin Plantinga, Joel Pust, Michael Rea and Matthias Steup for comments on earlier drafts. Thanks also to my commentator, Paul Moser, and to participating audience members for their helpful discussion when I presented this paper at the Central Division Meeting of the APA in Chicago in April 2000. Finally, I would like to thank the Pew Evangelical Scholars Program for providing support while I worked on this paper.

[1] For an example of the first sort of objection, see Alston 1986 and Plantinga 1993a, Chapter 1. For an example of the second, see Alston 1986 and Goldman 1999 (actually, Goldman attacks versions of internalism according to which a necessary condition for justified belief is a further bit of *knowledge* or, at least, the potential for this further bit of knowledge).

[2] See for example Moser 1989 and Fumerton 1995.

[3] It is, admittedly, difficult to find a very clear statement of this sort of dilemma in Sellars' work. But in his 1963: 131-32 he considers two ways to think of episodes of sensing what he calls 'sense contents': as cognitive episodes or as noncognitive episodes. And one of the main burdens of his paper is to show that either way, the proponent of what he calls 'the myth of the given' runs into trouble. See also his 1975.

[4] Let me emphasize that, in presenting this argument, I am not objecting to foundationalism *per se* or to the doctrine of the given. It's also worth highlighting the fact that my argument differs from both BonJour's and Moser's in that my dilemma has to do with being either conceptual or not whereas their dilemmas have to do with being either cognitive (where that means doxastic or propositional) or noncognitive.

[5] I say it hasn't been *widely* observed. But something like it has been noticed by Stephen Hetherington (see his 1990 and his 1991). However, as I argue in the appendix to this paper, Hetherington's papers needn't worry internalists (even though my argument should).

[6] Thus, BonJour says (2001: 21) that internalism imposes the "requirement ... that the justification for a belief must be cognitively accessible to the believer". Notice that he here makes both of the two claims that I emphasized from premise (I).

[7] This corrects what I said in Bergmann 1997. There I failed to distinguish between:

(A) At least one necessary condition of justification other than the no-defeater condition (NDC) *is* internal to normal humans

and

(B) At least one necessary condition of justification other than NDC *must be* internal to the subject (i.e., must be if her belief is to be justified).

(NDC is satisfied by S's belief that p just in case S doesn't take her belief that p to be defeated.) Suppose, as internalists commonly do, that justification is necessary for warrant (that which makes the difference between knowledge and mere true belief). Then my 1997 account entails that endorsement of (A) is sufficient for being an internalist (contrary to what I just asserted in the text) and that rejection of (A) is necessary for being an externalist. I would *now* say the following three things:

(1) endorsement of (B) is necessary for being an internalist
(2) endorsement of (A) is necessary for being a *nonskeptical* internalist
(3) rejection of *both* (A) and (B)–or at least of their analogues with respect to warrant–is necessary for being an externalist.

(Notice, by the way, that requiring NDC for justification is not to endorse I_3 since NDC doesn't require awareness of anything; at most it requires the *absence* of awareness of a certain thing.)

[8] See Pollock and Cruz 1999: 133 and 136 for elaboration that makes it clear that Pollock thinks awareness on the part of the subject is *not* required for a state to be accessible to one's automatic processors–i.e., to the "cognitive mechanisms that direct our epistemic cognition".

[9] This counterexample is designed so that the view being described satisfies the conditions Feldman and Conee say are sufficient for being an internalist position. To make it apply to Pollock's account, replace the term 'mental states' with 'internal states' where those are defined as states accessible (in a nonepistemic sense) to the subject's automatic processors, not to the subject.

[10] It is irrelevant to point out that the view just described is an implausible one. What matters is that it is clearly not an internalist one. As for whether that entails that it is an externalist view, see my discussion below near the beginning of section 3.2.

[11]The point here isn't to make dogmatic pronouncements on the inner life of dogs and cows. I'm just working with common assumptions in order to give the reader some idea of what nonconceptual awareness is.

[12] Or that X is evidence for B or that X is a truth-indicator for B or that X contributes in some way to B's justification. I will often suppress these sorts of alternatives but they should be assumed.

[13] See Moser 1989: 173-176, Fumerton 1995: 64, and Alston 1986: 211 for discussions of this sort of regress problem.

[14] Even internalists shy away from imposing such exceedingly high standards for justified belief. See Fumerton 1995: 64 for this sort of reaction.

[15] One response to this sort of move is to deny that there is any nondoxastic conceptual awareness–to insist that all conceptual awareness is doxastic involving some sort of belief. For the purposes of this paper, I'll simply waive this concern since taking it seriously only makes things *worse* for the internalist.

[16] Moser (1989: 80-81) makes a similar claim though the regress problem he is speaking of is not a complexity regress problem.

[17] In stating I_6 so that it applies to the justification of concept applications as well as beliefs, I'm assuming that the internalist we have in mind here–the proponent of nondoxastic conceptual$_1$ awareness requirement–thinks that the justification of concept application, like the justification of belief, requires conceptual$_1$ awareness. There seems to be no good reason to think justification of *belief* requires conceptual$_1$ awareness if one thinks that justification of *concept application* does not. (Why would an externalist account of the justification of concept application be satisfactory if an externalist account of belief justification isn't?) I'm also assuming that the internalist we have in mind thinks that the conceptual$_1$ awareness required for the justification of *concept application*, like the conceptual$_1$ awareness required for *belief* justification, must itself be *justified* (i.e., that it must involve *justified* concept application). Again, there seems to be no good reason to demand this in the case of belief but not in the case of concept application. (If you think insane or irrational concept application is sufficient, why think concept application is even necessary?)

[18] Notice that even if for every justification contributor X_n, $X_n = X_1$, the series still includes an ever-increasing complexity of concepts applied to X_1.

[19] For example, if the belief for whose justification the concept application is required is the belief that p, then the content of the required concept application will be something like *being in some way relevant to the appropriateness of that belief that p*.

[20] For detailed discussion of their views, see section 4 below.

[21] Every process token is an instance of *some* reliable process type (it is due, in part, to this fact that reliabilism faces the generality problem–see Feldman and Conee 1998). But reliabilism is committed to the view that for each process token, there is a *relevant* type of which it is an instance–a type whose reliability (or unreliability) determines the justification (or lack thereof) of the belief produced by the belief-forming process token in question.

[22] The internalist might protest that she didn't originally start with the intuition that justification requires an infinite number of beliefs of ever-increasing complexity–she started

with something that sounded more plausible. But my claim is that once she sees that her view leads to that bizarre consequence, she should reject her view. The consequence reveals the excessiveness implicit in the original intuition.

[23] This is exactly what Fumerton does say (1988: 449) in response to a point that is similar to the one I'm pressing. To be honest, I'm not at all sure that a view like Fumerton's–where justification supervenes on direct nonconceptual awareness of a certain kind (more details of his view are given in section 4)–*is* a clear case of a nonexternalist view (see Moser 1985, 147 for similar remarks). But I won't press the point here.

[24] My discussion of the Julie case establishes the compatibility of endorsing I_7 and rejecting I_3 in the same way that it establishes the compatibility of endorsing I_1 and rejecting I_3.

[25] One I myself endorsed at one point. See my 1997.

[26] Objective deontological justification of the sort on which Feldman focuses (see his 1988a, 1988b and 2000) is typically *not* employed in support of internalism (certainly not by Feldman who, although he is an internalist and a deontologist, denies that deontology supports internalism). In section 4.2 of my 2000a I discuss some strategies for trying to show that objective deontological justification supports internalism and explain why they fail.

[27] The three objections to externalism mentioned above say that external conditions aren't *sufficient* for a philosophically interesting sort of justification. Assuming, as internalists typically do, that justification is necessary for knowledge, these objections also say that external conditions aren't sufficient for warrant (that which makes the difference between knowledge and mere true belief). Two other objections that deserve mention are the generality problem for reliabilism (see Conee and Feldman 1998) and the charge that, since the beliefs of a Cartesian demon victim can be justified, a reliability condition isn't *necessary* for justification (see Cohen 1984: 280-282; Foley 1985, section I; Moser 1985: 240-241; and Lehrer 1990: 166). However, unlike the objections considered in the text, these latter two objections are *not* directed at externalism generally. Instead, each draws attention to problems with reliabilism in particular. Thus, even if successful, these objections don't threaten *all* forms of externalism in the way in which my dilemma for internalism threatens all forms of internalism. (See Plantinga 1993b: 29 for a defense of the claim that an externalist position–i.e., his own–avoids the generality problem. And see my 2004 where I argue for an externalist account of justification that avoids the evil demon objection to reliabilist accounts.)

Furthermore, the reliabilist can respond to the complaint about a reliability condition not being necessary for justification by remaining silent about justification and saying only that the reliability condition she proposes is necessary for *warrant*. If the externalist does this, then she isn't requiring too much for the internalist's liking: she isn't imposing *any* conditions on justification and, in imposing external constraints on warrant, she isn't thereby contradicting the internalist's own views (since everyone allows for external constraints on warrant). This suggests that the principal point of disagreement between internalists and externalists does *not* have to do with whether externalists are requiring *too much for justification*; rather, it has to do with whether externalists are requiring *enough for warrant.* This latter issue is one on which the three objections to externalism mentioned in the text are focused. It's true that internalists tend to speak about justification, not warrant (just as externalists tend to speak

about warrant and not justification). But given that internalists think justification is necessary for warrant, their claim that awareness is required for justification has implications that contradict the externalist's views on warrant.

[28] Moser (1989: 80) writes:

> The awareness relevant to [internalism] must be essentially nonconceptual to enable those views to avoid the defects of the views criticized [earlier in the book]. Nonconceptual awareness is just awareness that does not essentially involve the application or the consideration of a concept.

[29] On p. 136 of his 1989 he says that being presented with E is necessary for E's being a maximal unconditional probability-maker for P for S. And on p. 141 he says that E's being a maximal unconditional probability-maker for P for S is necessary for the justification of S's belief that P.

On p. 82 Moser says that the notion of *presentation* is conceptually basic in his system. However, he likens it to what Bertrand Russell calls 'acquaintance' (Fumerton too points us to this Russellian notion for help in understanding *his* notion of direct acquaintance–see Fumerton 1995: 75). And he (Moser) says that this presentation is a direct and nonconceptual sort of noticing. On pp. 81-82 of his 1989 he defines awareness in terms of direct attention attraction and direct attention attraction in terms of presentation. So in his system these three notions–i.e., awareness, direct attention attraction and presentation–are intimately linked.

[30] He mentions this requirement on p. 141 of his 1989. I say the required awareness is nonconceptual because on p. 142 he says "the *de re* awareness of such a relation [i.e., of E's supporting P] can be understood via the notion of direct attention attraction introduced [earlier] to clarify the notion of presentation". See the previous note where I point out that Moser thinks of presentation (and direct attention attraction) as a nonconceptual sort of direct awareness.

[31] What's important here is to distinguish an awareness of being appeared to redly from both (a) the application of a concept to the object of that awareness and from (b) the belief–call it 'B2'–that one is being so appeared to. Once these three are distinguished, we need only consider some malfunction (due to brain damage or an evil demon) that prevents the subject from applying (and from being able to apply) to the object of the awareness in question a concept such as *being in some way relevant to the appropriateness of holding B2*. Given that the awareness, the concept application and the belief are three distinct things, I don't see why such malfunction is impossible (at the very least, the suggestion that it is impossible would require some defense). And given the possibility of such malfunction, there is no reason to think the following aren't compossible: (i) S's holding the true belief B2 (that she is being appeared to redly), (ii) S's being aware of her being appeared to redly, and (iii) S's being unable to apply to the object of that awareness a concept like *being in some way relevant to the appropriateness of holding B2.*

[32] See his 2001: 29 and note 14. He recognizes (note 14) that the term 'content' seems a little strange when applied to something nonconceptual and nonpropositional but he says that he uses it because he knows of "no better term for what one is conscious of in having sensory or

phenomenal states of consciousness". But he warns us that "the two sorts of content [propositional and sensory] are importantly different and should not be conflated".

[33] We have to suppose that Watson doesn't have as evidence for the conclusion that the butler did it the belief that Holmes endorses that conclusion (for Watson *can* see that *that* is a good reason for the conclusion that the butler did it).

[34] A version of this paper was first presented at a meeting of the Central Division of the American Philosophical Association in Chicago in April 2000.

[35] This statement of TT is quoted from Hetherington's 1991: 858. As I mention below in the text, Hetherington explains that for something to be epistemically internal to S "is for it to be appreciated by S *as* being his or her reason" (1991: 858).

[36] The claim that the regress problems in question imply that epistemic internalism is an empty concept is made by Hetherington in both his 1990: 247 and his 1991: 862.

[37] One strange thing about this sort of objection to internalism is that there seems to be no need to formulate it as a dilemma. Why not simply drop premises (II*) and (IV*) and change (V*) so that it says "If internalists must admit that epistemic internalism is an empty concept, then internalism should be rejected"?

[38] There he says (p. 859) that "epistemic internalism must accept the Transparency Thesis" and (p. 861) that:

> Epistemic internalism, therefore, faces a dilemma. Either it ceases to require that S epistemically internalise facts about his or her epistemic internalisings, or it does not. (That is, either the Transparency Thesis is [taken by the internalist to be] false or it is not.) ... If the Transparency Thesis is [taken by the internalist to be] false then the epistemic internalist becomes an epistemic *ex*ternalist.
>
> So the Transparency Thesis will be retained ... But then there is the regress.

[39] Hetherington 1990: 246-47.

[40] You might think that Hetherington would answer this question by pointing out that, according to internalists, *nothing* can contribute to a belief's justification unless the subject appreciates it as a justification-contributor. But in his 1990: 245 he says that:

> At least part of what [the internalist] thinks is that *some* aspect *A* of your circumstances is epistemically internal to you and is at least *part* of what makes you justified.

And he adds the following in a note attached to that sentence:

> I say "at least *part*" because, for generality's sake, I am happy to allow that an epistemically internalist condition could be necessary, sufficient, or even something looser still.

This makes it sound like an internalist could allow that there are *other parts* of what makes you justified that are *not* epistemically internal to you. But then, according to internalists, something *can* contribute to the justification of your belief even if you don't appreciate it as a justification-contributor.

[41] I don't quite know what to make of the locution "epistemically external to your being justified" which appears in the passage just quoted and at other places in his 1990 (along with "epistemically *in*ternal to your being justified"). I assume he intended to say "epistemically external to *you*" (and "epistemically internal to *you*").

[42] By comparing R* to R, we can see why R isn't quite right. Consider, for example, the first appreciation mentioned in R. Given Hetherington's account of what it is to be epistemically internal, that first appreciation can be restated as:

the appreciating-of-W1-*as*-being-appreciated-by-S-as-being-her-reason.

Notice that this appreciation differs from each appreciation mentioned in R*. For in this appreciation something is being appreciated, not as *being S's reason*, but as *being appreciated by S as being her reason*.

[43] See the first passage quoted in note 40.

[44] Hetherington uses 'epistemically within S' and 'epistemically internal to S' synonymously.

[45] Hetherington might want to argue that if TT is true, then this *isn't* possible. But even if that were true, how would it save his defense of (I*)? Does he want his argument that the internalist is committed to TT (a principle Hetherington thinks leads to regress problems), to rely on the truth of TT ?

[46] For another example of how this sort of conflation spoils Hetherington's defense of (I*), see his 1991: 859. There he defends the conclusion that internalists are committed to thinking that a potential J-contributor, W1, contributes to the justification of S's belief B only if S appreciates W1 as a J-contributor. And he takes that to support the claim that internalists are committed to TT. But TT is a thesis about appreciations entailing higher-level appreciations of appreciations. Once again, Hetherington seems to be treating a discussion of appreciating facts about *W1* as if it's a discussion of appreciating facts about *appreciating facts about W1*.

REFERENCES

Alston, William. 1986. Internalism and externalism in epistemology. *Philosophical Topics* 14: 179-221. Reprinted in Alston 1989: 185-226. Page references are to reprint.

Alston, William. 1989. *Epistemic Justification: Essays in the Theory of Knowledge*. Ithaca, NY: Cornell University Press.

Bergmann, Michael. 1997. Internalism, externalism and the no-defeater condition. *Synthese* 110: 399-417.

Bergmann, Michael. 2000a. Deontology and defeat. *Philosophy and Phenomenological Research* 60: 87-102.

Bergmann, Michael. 2000b. Externalism and skepticism. *The Philosophical Review* 109: 159-94.

Bergmann, Michael. 2004. Externalist justification without reliability. *Philosophical Issues* 14: 35-60.

BonJour, Laurence. 1985. *The Structure of Empirical Knowledge*. Cambridge, MA: Harvard University Press.

BonJour, Laurence. 2001. Towards a defense of empirical foundationalism. In *Resurrecting Old-Fashioned Foundationalism*, ed. Michael DePaul, 21-38. Lanham, MD: Rowman and Littlefield.

Chisholm, Roderick. 1982. *The Foundations of Knowing*. Minneapolis, MN: University of Minnesota Press.

Cohen, Stewart. 1984. Justification and truth. *Philosophical Studies* 46: 279-95.

Conee, Earl and Richard Feldman. 1998. The generality problem for reliabilism. *Philosophical Studies* 89: 1-29.

Conee, Earl and Richard Feldman. 2001. Internalism defended. *American Philosophical Quarterly* 38: 1-18.

Feldman, Richard. 1988a. Epistemic Obligations. In *Philosophical Perspectives 2: Epistemology*, ed. James Tomberlin, 235-56. Atascadero, CA: Ridgeview.

Feldman, Richard. 1988b. "Subjective and Objective Justification in Ethics and Epistemology." *The Monist* 71: 405-19.

Feldman, Richard. 2000. The ethics of belief. *Philosophy and Phenomenological Research* 60: 667-95.

Foley, Richard. 1985. What's wrong with reliabilism? *The Monist* 68: 188-202.

Fumerton, Richard. 1988. The internalism/externalism controversy. In *Philosophical Perspectives 2: Epistemology,* ed. James Tomberlin, 443-459. Atascadero, CA: Ridgeview.

Fumerton, Richard. 1995. *Metaepistemology and Skepticism*. Lanham, MD: Rowman and Littlefield.

Goldman, Alvin. 1999. Internalism exposed. *The Journal of Philosophy* 96: 271-93.

Hetherington, Stephen. 1990. Epistemic internalism's dilemma. *American Philosophical Quarterly* 27: 245-51.

Hetherington, Stephen. 1991. On being epistemically internal. *Philosophy and Phenomenological Research* 51: 855-71.

Lehrer, Keith. 1990. *Theory of Knowledge*. Boulder, CO: Westview Press.

Moser, Paul. 1985. *Empirical Justification*. Boston, MA: D. Reidel.

Moser, Paul. 1989. *Knowledge and Evidence*. New York: Cambridge University Press.

Plantinga, Alvin. 1993a. *Warrant: The Current Debate*. New York: Oxford University Press.

Plantinga, Alvin. 1993b. *Warrant and Proper Function*. New York: Oxford University Press.

Pollock, John. 1986. *Contemporary Theories of Knowledge*. Savage, MD: Rowman and Littlefield.

Pollock, John and Joseph Cruz. 1999. *Contemporary Theories of Knowledge, 2nd Edition*. Lanham, MD: Rowman and Littlefield.

Sellars, Wilfrid. 1963. Empiricism and the philosophy of mind. In *Science, Perception and Reality*. New York: The Humanities Press.

Sellars, Wilfrid. 1975. The structure of knowledge. In *Action, Knowledge, and Reality: Critical Studies in Honor of Wilfrid Sellars*, ed. Hector-Neri Castañeda. Indianapolis, IN: Bobbs-Merrill.

Chapter 8

EPISTEMIC INTERNALISM, PHILOSOPHICAL ASSURANCE AND THE SKEPTICAL PREDICAMENT

Richard Fumerton
University of Iowa

1. INTRODUCTION

It is a particular pleasure to contribute this paper to a volume honoring Al Plantinga. I have always viewed his work as a model of how to do philosophy and I have learned a great deal from him over the years. It is as a result of philosophical conversation with both Plantinga and his former student Michael Bergmann that I have come to the (always painful) conclusion that I need to revise some of the things that I have said in print. This paper is an attempt to do just that.

In *Metaepistemology and Skepticism* (1996), I implied that the fact that externalists, to be consistent, should allow "track record" arguments in support of their belief that they have first-level justification[1] is a kind of *reductio* of their position. I said the following:

> You cannot use perception to justify the reliability of perception! You cannot use memory to justify the reliability of memory! You cannot use induction to justify the reliability of induction! Such attempts to respond to the skeptic's concerns involve blatant, indeed pathetic, circularity (1996: 177).

The above still seems right to me and, I hope, will strike you as plausible.[2] If one embraces some version of externalism such as reliabilism, if one embraces the view that the reliability of a belief-producing process is

T. M. Crisp, M. Davidson and D. Vander Laan (eds.), Knowledge and Reality, 179-191.

sufficient to generate justified output beliefs (provided that the input beliefs, if any, are justified), then one will need to embrace what I take to be the absurd view that one can use reliable methods of forming belief to justify belief that they are reliable (and thus, that their output beliefs are justified). As I noted in the book, even many externalists seem to get cold feet when it comes to bootstrapping their way up to justified metabeliefs about reliable processes using those very processes. Alston (1993), for example, specifically rejects the legitimacy of track record arguments. Plantinga is harder to read on this issue, but it is interesting to note that in his 2000 defense of Christianity he is certainly hesitant about claiming to have inspired warranted belief that he has inspired warranted belief.

If I had stopped with the remarks quoted above, I wouldn't have anything to retract. I went on, however, to generalize (always a dangerous move in philosophy):

> The fundamental objection to externalism can easily be summarized. If we understand epistemic concepts as the externalist suggests we do, then there would be no objection in principle to using perception to justify reliance on perception, memory to justify reliance on memory, and induction to justify reliance on induction. *But there is no philosophically interesting concept of justification or knowledge that would allow us to use a kind of reasoning to justify the legitimacy of using that reasoning* (1996: 180, emphasis added).

As Michael Bergmann (2000) has pointed out there seems to be (at the very least) some tension between what I said above and what I also said about direct acquaintance, a relation that is crucial to my own internalist account of what constitutes noninferential justification of a sort that satisfies philosophical curiosity. On the view I defended, one has noninferential justification for believing P when one has the thought that P while one is directly acquainted with the fact that P and the correspondence between the thought and the fact.[3] When answering the hypothetical question of how I would justify my belief that there is such a relation of acquaintance, I answered as follows:

> If I am asked what reason I have for thinking that there is such a relation as acquaintance, I will, of course, give the unhelpful answer that I am acquainted with such a relation. The answer is question-begging if it is designed to convince someone that there is such a relation, *but if the view is true it would be unreasonable to expect its proponent to give any other answer* (1996: 77, emphasis added).

Well, what is good for the goose is good for the gander. Why shouldn't a reliabilist, who has, of course, offered a quite different account of

noninferential (and inferential) justification, respond in precisely the same way? If the reliabilist is asked what reason he has to believe that there are belief-independent, unconditionally reliable processes, why shouldn't that reliabilist respond by claiming to have a reliably formed belief that they exist, a belief whose reliability involves relying on a successful track record of using those very processes. The answer is question-begging if it is designed to convince someone that there are such reliable belief-forming processes, but if the account of justification is correct, what other answer would one expect to get from a proponent of the view?

In what follows, I want to explore what I take to be a more felicitous way of presenting the alleged *reductio* against externalism. Our discussion will seek to get clear about precisely what internalists want and why it is that they believe that externalists do not succeed in analyzing a philosophically satisfying concept of knowledge or justification.

2. WHAT DO INTERNALISTS WANT?

In a number of important books and papers, Barry Stroud has tried to make clear what a philosophically satisfying account of knowledge must produce.[4] He usually puts the goal of the philosopher in terms of understanding. Our task as philosophers is to develop a philosophically satisfactory understanding of knowledge in general, or knowledge within a certain specified domain. While he concedes that one can, of course, scientifically study human cognition in the same way that one can study any other natural phenomenon, he argues that scientific investigation could never yield results that satisfy the philosopher. That is because there are certain constraints on how one can legitimately study or investigate knowledge in general. But what are those constraints according to Stroud? As far as I can tell, understanding human knowledge involves coming to know what we know and how we know it. But to be philosophically satisfying (particularly from a first person perspective) our investigation must meet the following conditions:

1) In trying to understand whether and how we know various propositions in a given field of knowledge, we cannot presuppose that we know or even reasonably believe any propositions alleged knowledge of which we are investigating. As a result, we are forbidden from employing as premises any proposition knowledge of which we are trying to understand. So if we are trying to figure out how, if at all, we know propositions about the past, for example, we couldn't use as a premise in reaching our conclusion any truth about the past.

2) In trying to understand how we know various propositions in a given field of knowledge, we cannot presuppose the legitimacy of any of the methods we employ in coming to believe propositions of the sort in question, and therefore cannot use any of those methods in studying the knowledge in question. So, for example, if we are trying to figure out how we know propositions about the external world through perception, we cannot use perception to facilitate our understanding.

There is something very seductive about the above constraints. Historically, philosophers who have taken the problem of skepticism seriously seem to have just taken for granted 2). It was viewed as a the worst sort of question-begging to attempt an inductive justification of induction, or a perceptual justification of the veridicality of perception. Although the problem of memory was not discussed nearly as often, it would presumably be equally illegitimate to employ memory in the attempt to certify the legitimacy of relying on memory.

It is not hard to see why Stroud finds skepticism so difficult to avoid given the above constraints. If the epistemologist's ultimate goal is to understand all knowledge, knowledge in general, and to do so within the constraints posed by 1) and 2), it doesn't take a pessimist to see clouds on the horizon. To understand knowledge in general we would need to satisfy ourselves that all of our methods of arriving at conclusions are legitimate and we would need to do so without using any of those methods! Even if we were to arrive at a purely a priori knowledge of the legitimacy of epistemic principles, we would have left philosophically mysterious a priori knowledge–there would still be one source of knowledge that we haven't been able to study philosophically. In a striking comment that really led me to think about some of these matters afresh, Plantinga (2001: 390) suggested that if the internalist insists on something like 2) in the quest for epistemic security, then even if there were a God, that God would be unable to have knowledge of the sort the internalist wants. That seems right to me.

3. MORE MODEST INTERNALIST GOALS

I have never been comfortable with the emphasis many internalists place on the importance of access to knowledge and justification. As I said, Stroud often seems to locate the epistemologist's target as second-level knowledge (or understanding). Other internalists seem to think that having first-level knowledge or justification is inseparable from having second-level knowledge or meta-justification. But as I have argued elsewhere global access internalism seems to raise immediately the specter of vicious regress.

We should make at least the following distinctions among various sorts of access requirements for knowledge and justification. (In what follows I'll focus on justification, but what I say will apply *mutatis mutandis* to knowledge.)

3.1 Global Actual Access Internalism

The global actual access internalist claims that in order for S to be justified in believing P, S must have access to the fact that he has that justification. The most natural interpretation of "access" here is knowledge or justified belief. Internalists typically also want the meta-justification in question to be noninferential or introspective. So the claim is that in order for some set of conditions J to constitute S's having justification for believing P, S must have an introspectively justified belief that J exists. The modal status of this claim is crucial if it is to be even intelligible. It seems hopeless to argue that this principle is an analytic truth. It doesn't even make sense to suppose that J, by itself, constitutes S's having justification for believing P only if one adds to J S's access to the fact that J obtains. That's tantamount to claiming that J constitutes S's justification for believing P only if doesn't really constitute S's justification for believing P! It is really only J + access to J (call that A) that constitutes S's justification for believing P. But of course (J + A) doesn't really constitute S's justification for believing P either. Global actual access internalism implies that one must have access to (J + A) call that access A*. But (J + A + A*) won't constitute S's justification either for one must have access to that condition..., and so on ad infinitum.

To escape this problem the global access internalists must claim that their principle is some sort of synthetic necessary truth. Given what genuine justification is, there is a necessary connection between possessing it and realizing (knowing or justifiably believing) that one possesses it. Just as P's being true implies that it is true that P is true, even though its being true that P is true is not, presumably, constitutive of P's being true, so the access internalist we are discussing might hold that S's having justification implies that S is aware of that justification without that second-level awareness being constitutive of the first-level awareness. While this sort of strong access requirement might not lead to conceptual regress, it might still seem to lead to vicious regress. Every justified belief requires the having of an infinite number of ever more complex meta-beliefs.

3.2 Global Potential Access Internalism

It might seem that the specter of regress is less ominous if we shift from a requirement that having justification requires actual access to that justification, to the claim that having justification requires only potential access to it. Again, most internalists want the potential access to be

introspective (or noninferential). There are also a number of different interpretations of the critically important concept of potentiality upon which the view relies. Just as with actual access internalism, it is still important that the potential access internalist not regard second-level potential access requirements as constitutive of first-level justification. Rather, to avoid conceptual regress the claim must be that having justification is by its nature tied to something else, the possibility of accessing (introspectively or noninferentially) that justification. Here one can see the natural connection between what one might call internal state internalism, and potential access internalism. The internal state internalist takes a person's having justification for a belief to be an internal feature of the person.[5] Internal states might in turn be viewed as nonrelational properties of mind. Traditional foundationalists have often taken it to be a mark of nonrelational properties of mind that they be introspectively accessible. Thus we can see how an internal state internalist might also end up being a potential access internalist.

While the threat of regress on potential access internalism might seem less severe, it is nevertheless present. To be sure, having justification for believing P doesn't require having an infinite number of ever more complex higher-level beliefs. But it still seems to require the possibility of forming infinitely many ever more complex higher-level beliefs. God might be up to that task, but it is not clear that you and I are.

4. PHILOSOPHICAL ASSURANCE

I think it would be a mistake to dismiss access internalism solely on the grounds that it is fertile ground for skepticism. The philosophical enterprise is by its nature odd. Philosophers ask questions about that which is simply taken for granted by non-philosophers. Those of us who are parents remember fondly the also sometimes frustrating days when our young children would ask a seemingly endless number of "Why" questions. "Why is the sky blue? "You give some sort of answer. It's X and X things appear blue? Your kid then wants to know why X things look blue? The looming regress of "Why?" questions inevitably ends with an impatient parent responding "That's just the way it is," a response that, no doubt, did little to satisfy the child's curiosity. The epistemologist, Stroud argues, wants to know why we can legitimately conclude that a certain way of forming belief is legitimate, and the epistemologist's philosophical curiosity isn't going to be satisfied by being told at any stage of the game that it just is. It's always possible then that the epistemologist is led by philosophical curiosity on a quest that can only end in failure. But if access requirements so obviously

lead to skepticism, one might want very strong reason to believe that they are independently plausible.

4.1 Access Requirements and Defeaters

In a forthcoming paper, Michael Bergmann tries to explain why higher-level requirements for having justification can seem so plausible despite the fact that the externalist's rejection of them is in the end correct. His suggestion, in short, is that when one considers the question of whether or not one's belief is reliably produced (or produced in a trustworthy way), then the fact that one disbelieves in that reliability, or even withholds belief with respect to reliability, does render unjustified the lower-level belief in question. He argues, however, that disbelief and even withholding of belief requires entertaining the proposition in question. With respect to any given proposition, one can believe, disbelieve, withhold belief, or do none of the above. According to Bergmann, if one simply fails to consider the question of whether one's lower-level beliefs are reliably produced, their reliable source may render them justified even in the absence of any doxastic attitudes concerning their source.

In response to Bergmann, I questioned whether the mere subjective attitude (even if wildly irrational) a person takes to the legitimacy of a belief-forming process should affect the epistemic status of the belief formed by that process. But one might plausibly argue for a somewhat different view about potential defeaters. Let's make the traditional distinction between there being justification for a belief and a belief's being justified. There can be justification for S to believe P even if S doesn't believe P, or believes P, but not as a result of possessing justification (not by basing the belief on the justification possessed). By contrast, to have a justified belief that P, there must not only be justification for the belief, but the belief must be based on that justification. Now whether or not an aging person believes that his beliefs based on memory are now unreliable, and whether or not the person in question even considers the question of whether his beliefs based on memory are now unreliable, shouldn't we conclude that if there is good reason for that person to believe that the beliefs are unreliably produced, that defeats whatever justification he might otherwise have had for believing propositions about the past based on memory?

Of course, if we succeed this way in securing the relevance of metalevel justification concerning the unreliability of a belief-forming process to lower-level justification of the belief produced through that process, you know what is coming next. As BonJour argued in *The Structure of Empirical Knowledge* (1987), it would seem plausible to suppose that justification for withholding belief concerning the justificatory status of a lower-level belief

(or the legitimacy of the process that produced it) is equally relevant. With or without my consideration of the question of whether my belief that p is justified, if it would be rational to withhold belief with respect to my having justification for believing p, i.e. if the rational thing to conclude is that it is no more likely than not that my belief that p is justified, surely that fact defeats whatever justification I might otherwise have had for believing P? But haven't we now got all the way to metalevel requirements for justified belief? These metalevel requirements involve justification at the higher levels–they are not particularly concerned with what the person actually believes at the higher levels.

Bergmann argues that there are four possibilities with respect to the higher-level doxastic attitudes one takes towards the appropriateness of a lower-level belief. One might 1) believe that the belief is appropriately formed, 2) disbelieve that the belief is appropriately formed, 3) withhold belief with respect to whether or not the belief is appropriately formed, or 4) have no attitude whatsoever with respect to whether or not the belief is appropriately formed (because one hasn't even considered the matter). But it seems to me that there are only three possibilities with respect to the epistemic justification there is for one to adopt higher-level attitudes towards the justification of lower beliefs. There might be 1) justification for believing the proposition that one's belief that p is justified, 2) justification for disbelieving the proposition that one's belief that p is justified, or 3) justification for withholding belief with respect to whether or not one is justified in believing p. It is very tempting for the internalist to argue that either 2) or 3) defeats one's justification for believing p. Therefore unless there is justification to suppose that one's belief that p is justified, one is unjustified in believing p.

The above argument is very attractive and does provide at least prima facie plausibility for a justificatory access requirement even if that requirement leads to skepticism. But the there is still the worry that we are being led too quickly to a requirement for justification that simply ensures at the outset a victory for skepticism. There is still the concern that as we move up levels we get to propositions so complex that it will be impossible for creatures like us to even entertain them. If we can't entertain a proposition, then it is not clear in what sense there can be justification for us to believe the proposition. Borrowing again from Bergmann's idea, however, we might be able to explain why it is so tempting for traditional foundationalists to suppose at least that whenever they have a first-level noninferentially justified belief, there is also noninferential justification for them to believe that they have that first-level justification. Whenever they succeed in raising the question of whether they have lower-level justification for believing a certain proposition, they will find noninferential justification for believing that they have noninferential

justification for believing that proposition. That is because their ability to raise the question presupposes that they are capable of entertaining at least that metalevel proposition describing justification for a belief about justification. The second-level acquaintance with the first-level acquaintance that is partially constitutive of first-level justification is typically available when one has the conceptual sophistication to entertain the proposition made true by the fact that one has second-level acquaintance with first-level acquaintance. When one is directly acquainted with one's pain in a way that yields noninferential justification for believing that one is in pain, and one has the ability to formulate the question of whether one is justified in believing that one is in pain (formulation of which involves entertaining the relevant proposition), one will find oneself directly aware of the fact that one is directly aware of the pain.

Of course, one shouldn't infer from the fact that whenever one looks for something one finds that it has a certain characteristic, that it necessarily has that characteristic. One shouldn't infer from the fact that whenever one looks for justification for believing that one has a noninferentially justified belief (when one does) one finds the relevant higher-level justification, that there is a necessary connection between having the lower-level justification and having justification for believing that one has it. Consider, by analogy, a pathetic argument for an extreme sort of anti-realism–the view that there is no reality that does not necessarily involve a representation of it. "Go ahead," the anti-realist argues, "give me an example of an unrepresented fact." You won't be able to do so, of course. As soon as you choose some fact as an example, you will have thereby represented it. It should be obvious to everyone, however, that from the tautology that all represented facts are represented, it does not follow that it is necessary that all facts are represented. Similarly, even if it were necessarily the case that whenever we consider the question of whether we have noninferential justification, we find that we are justified in believing that we have it, it doesn't follow that there is any necessary connection between having a noninferentially justified belief and having justification for believing that we have that justification.

There is another obvious fact that complicates the issue. In trying to discover from the first person perspective what the connection is between having a certain kind of justification for believing P and having justification for believing that we have that justification, we need to start with uncontroversial examples of justified beliefs. But what makes a belief an uncontroversial example of a justified belief is presumably the fact that we have strong justification for believing that it is justified. So again it will be a trivial truth that whenever we have an uncontroversial example of a justified belief we will find that we have justification for believing that we have a justified belief.[6] That analytic truth, however, will not secure a necessary

connection between having justification and having access to that justification.

Seeking philosophical assurance by moving up levels seems destined to result in disappointment. We will either get to a point at which we can no longer formulate the relevant question because it has become so complex, or we will simply get frustrated or bored and abandon the project. Again, one might conclude that epistemologists are simply doomed by the nature of their philosophical quest to a life of philosophical disappointment. But is there any other way for the internalists to succeed in their search for philosophical assurance?

5. THE SOURCE OF PHILOSOPHICAL ASSURANCE

Why do I think that one can't use memory to justify the legitimacy of using memory and perception to justify the legitimacy of using perception, but I do think that one can use acquaintance with acquaintance to justify the existence of acquaintance? The answer, is in one sense, simple. On the view I accept, facts about what we are acquainted with are by themselves sufficient for having philosophically relevant justification; facts about what we are caused to believe (reliably or not) by memory and perception are not. Autobiographical reports are all well and good, the externalist will reply, but what's that got to do with philosophical argument? The externalists (who are usually also foundationalists) are perfectly happy with their own accounts of (external) conditions that they claim are sufficient for justification.

What I want to suggest is that one should test the plausibility of a claim about what is genuinely sufficient for having justification by exploring the implications of that claim when moving up levels. Specifically, as I suggested in the quotes with which I began this paper, it seems to me that reliabilists, for example, ought to have no qualms about using a way of forming a belief to justify one's belief that that way of forming beliefs is legitimate. Either the reliability of the belief-forming process is enough, by itself, to yield justified output beliefs or it is not. If it is, then it is no matter what level of belief one is interested in justifying. So if memory and induction are reliable, then through memory and induction I can justify my belief that memory is reliable. I remember seeming to remember doing certain things and I also remember doing them. If induction is a reliable way of forming beliefs about generalizations, I can conclude on that basis that my beliefs about the past based on memory are reliably produced and thus justified. As I said before, it is striking that even many proponents of reliabilism can't quite bring themselves to argue that this is a legitimate way

to justify belief that memory is reliable. To be sure, they might argue that if memory is reliable then we can form justified beliefs about the reliability of memory this way, but they feel uncomfortable simply asserting that they have justified belief about the reliability of memory formed in this way. Why? Because at some level they realize that in asserting the critical antecedent of the conditional claim they go beyond what they are in a position to assert qua philosophers trying to satisfy philosophical curiosity.

The matter is, I think, quite different with what I call acquaintance. I stub my toe and I believe that I am in excruciating pain. What justification do I have for thinking that I'm in pain? How do I know that I'm in pain? My answer is that I am directly aware of the pain itself–the very truth maker for my belief. The pain is "there" transparently before my mind. The thought that is about the pain and the pain that is its object are both constituents of the conscious mental state that I call acquaintance. When all this is so, we are in state that is all that it could be by way of satisfying philosophical curiosity. What more could one want as an assurance of truth than the truth-maker there before one's mind? When one is directly acquainted with pain as one entertains the proposition that one is in pain, there seems to me to be no need, no point, in moving up a level and asking about the justification one has for believing that one is in this state. It is not that one can't ask the question. The question is well-formed and there is, of course, a readily available answer. Just as acquaintance with pain was a completely satisfying way of assuring oneself that one is in pain, so acquaintance with this acquaintance with pain is a completely satisfying way of assuring oneself that one is acquainted with pain.

But again, I would emphasize that it doesn't strike us as even relevant to explore the second-level question as a way of getting a better sort of assurance that one is in pain. Why would it? If I'm right, what is relevant to getting the assurance one wants as a philosopher is getting the pain itself before one's consciousness.[7] In the second-level act of acquaintance the pain is present before consciousness again as a constituent of a more complex state of affairs, but having it before consciousness in that way is no better, so to speak, that having it there as an object of first-level awareness.

The matter is quite different, I think, with belief-forming processes that may or may not be reliable (or that may or may not be functioning properly, or "tracking" facts). Am I noninferentially justified in thinking that I am in pain when I stub my toe? The reliabilist, for example, says that I am provided that my belief is caused by a process that is unconditionally reliable. The philosopher can't resist, at this point, asking the obvious next question. But is my belief caused in the right way? The question is irresistible not because one in general needs second-level justification in order to have first-level justification. The question is irresistible because

having a belief caused in a certain way when we don't know whether or not it is caused in that way is clearly not something that would give us assurance of truth. Strangely enough, some externalists seem to become aware of that fact themselves when they try to apply their analyses at the next level. They realize that track-record arguments aren't really getting us anywhere when it comes to giving us the assurance we seek. The appropriate moral to draw, however, is it that if you can't live with a track-record argument given your claims about what is genuinely sufficient for having justification, then you should abandon those claims. Furthermore, if you feel the need to move up a level to satisfy philosophical curiosity, that too is an indication that you should reconsider your view about what is sufficient for (philosophically relevant) justification at the first-level.

ENDNOTES

[1] Throughout this paper I'll be talking about justification. I'm convinced that what I call justification is the same thing that Plantinga calls warrant. I certainly don't attach to the concept of justification normativity of a sort that Plantinga successfully argues is irrelevant to knowledge. In any event, those happier with talk about warrant may translate my remarks into that terminology if they choose.

[2] For an excellent discussion of some of these issues, one which helped me understand my own views better, see Cohen 2002.

[3] I also suggested that it might be possible to have a noninferentially justified belief based on acquaintance with a fact very similar (but not identical with) the fact that P.

[4] These include Stroud 1984, and a number of papers contained in Stroud 2000.

[5] We must be careful to recognize here the distinction between having justification for a belief and a belief's being justified (a distinction I will return to later in the paper). There can be justification for me to believe some proposition P, but unless I base my belief that P on that justification, most philosophers will deny that the resulting belief is justified. The basing relation is often construed as causal, and most philosophers will reject the idea that the relation of causing is in any sense purely internal to the person whose beliefs are caused. So to be consistent, internalists shouldn't claim that a belief's being justified is a purely internal matter, though they could consistently claim that there being justification to believe a proposition is a purely internal matter.

[6] This is a problem from the first-person perspective. One can still try to generate examples of someone's having justification without having the relevant meta-level justification by looking at second- and third-person ascriptions of justified belief.

[7] I'm focusing here on philosophically satisfying noninferential justification. The story is much more complicated for inferential justification. For inferential justification, what I think the internalist wants is direct acquaintance with probabilistic connections between evidence and conclusion. Wanting this and getting this are two quite different matters.

REFERENCES

Alston, William. 1993. *The Reliability of Sense Perception*. Ithaca, NY: Cornell University Press.

Bergmann, Michael. 2000. Externalism and skepticism. *The Philosophical Review* 109: 113-28.

Bergmann, Michael. Forthcoming. Defeaters and higher-level requirements. *The Philosophical Quarterly*.

BonJour, Laurence. 1987. *The Structure of Empirical Knowledge*. Boston, MA: Harvard University Press.

Cohen, Stewart. 2002. Basic knowledge and the problem of easy knowledge. *Philosophy and Phenomenological Research* 65: 309-329.

Fumerton, Richard. 1996. *Metaepistemology and Skepticism*. Boston, MA: Rowman and Littlefield.

Plantinga, Alvin. 2000. *Warranted Christian Belief*. Oxford: Oxford University Press.

Plantinga, Alvin. 2001. Internalism, externalism, defeaters and arguments for Christian belief. *Philosophia Christi* 3: 379-400.

Stroud, Barry. 1984. *The Significance of Philosophical Skepticism*. Oxford: Clarendon Press.

Stroud, Barry. 2000. *Understanding Human Knowledge*. Oxford: Oxford University Press.

Chapter 9

SCIENTIFIC NATURALISM AND THE VALUE OF KNOWLEDGE

Jonathan Kvanvig
University of Missouri

Philosophical naturalism is, arguably, the dominant philosophical tradition in contemporary western philosophy. Naturalistic theories abound in nearly every area of philosophical investigation, and epistemology is no exception. Just what counts as a naturalistic theory in epistemology is not completely obvious, but the call for and interest in such is unquestionable.

When we begin to ask which theories count as naturalistic ones and which do not, things get more complicated, for it is far from obvious why any extant epistemological theory is anti-naturalistic. In the first section below, I will argue for a particular understanding of naturalism, situating the contemporary interest in it in a foundational attitude of respect for science. This motivation for the view constrains the kind of epistemological theory one can adopt, and my goal is to show how these constraints push inexorably toward a kind of attitudinalism or non-cognitivism in epistemology, an attitudinalism modeled on non-cognitivist approaches in ethics. The force doing the push is the need to account for the value of knowledge, which I will provisionally assume. My ultimate goal is to show how difficult it is for naturalistic views to account for the value of knowledge.

1. NATURALISM AND RESPECT FOR SCIENCE

Richard Feldman has argued recently that everyone, or nearly everyone, in epistemology can legitimately claim to be offering a version of naturalized epistemology (Feldman 2001b). The difficulty of classifying any view as non-naturalistic is, at bottom, the cost of allowing non-reductive versions of naturalism. Consider, for example, what Scott Sturgeon says about mountains:

T. M. Crisp, M. Davidson and D. Vander Laan (eds.), Knowledge and Reality, 193-214.

> I'm a realist about [mountains]. But I'm no reductionist. I do not think a tidy condition can be stated, in non-mountain terms, which coincides with being a mountain. So I deny mountainhood is identical with such a condition. Does that make for trouble? Does realism about mountains interact with non-reductionism to force spooky mountain metaphysics?
>
> Surely not. Mountains are nothing over and above non-mountains. I do *not* say that, though, because I can reduce mountains to non-mountains. I say it because the best picture of the world says it; and I adopt that picture. I say it because the perspective which strikes the best balance of simplicity, strength and the need to resolve Sellars' Problem [concerning the relationship between the manifest and scientific images] *contains* the claim that there are mountains which are irreducible yet not fundamental. The best picture of the world entails non-reductive naturalism about mountains. It says mountains are non-reductively made of non-mountains. So that is what I accept (Sturgeon 2002).

The difficulty for finding any non-naturalist theory of justification follows immediately:

> And so it could be with justification. I'm a realist about it too. But I take seriously the possibility that justification does not align with a tidy descriptive condition. And if it doesn't not, justification isn't identical to any such condition. Would that make for trouble? Would realism about justification plus non-reductionism *force* spooky normative metaphysics? I don't see why. Perhaps the best picture of the world contains the claim that beliefs are justified in a metaphysically derivative but irreducible way. Perhaps the perspective which strikes the best balance of simplicity, strength and the need to resolve Sellars' Problem *endorses* non-reductive naturalism about reason. In the event, we should too. We should accept reason as metaphysically derivative yet irreducible. We should accept non-reductive naturalism about reason (Sturgeon 2002).

The most popular way to articulate Sturgeon's naturalistic vision for reason and justification is in the language of supervenience, allowing us to say that normativity depends on natural facts without being reducible to them. Once we recognize such a naturalistic possibility, however, it is exceedingly hard to find any epistemological theory that cannot legitimately claim to be a naturalistic one. Even theories formulated exclusively in normative terms can endorse non-reductive naturalism by appealing to supervenience relations.

Feldman argues an even stronger point. Since it is far from clear what counts as a naturalistic concept at all, it is far from clear why an appeal to the language of normativity itself is somehow incompatible with naturalism. Chisholm, for example, thinks that there are synthetic a priori truths linking experience with the rationality of certain beliefs (Chisholm 1977), and

Feldman argues that even such an appeal need not be incompatible with naturalism:

> Chisholm apparently thought that, in addition to deductive and probabilistic connections, there was another species of connection between propositions (or between experiences and propositions). His view was that these relations are part of the real, or natural, world. Some may deny that there are any such relations. This seems, once again, to be a dispute about what there is, not a dispute about whether there is something beyond what is natural. In other words, if Chisholm is right, it is quite unclear why terms such as "supports" and other epistemic terms do not belong on the list of naturalistically acceptable terms in the first place. If they do, then even Chisholm can plausibly maintain that epistemic support facts are natural facts. If so, then almost all epistemologists are substantive naturalists (Feldman 2001a).

I think Sturgeon and Feldman are onto something very important here, and it is that no one has a very good idea of what substantive naturalism is committed to, a point argued persuasively by Bas van Fraassen (1996). He argues that if we specify substantive conditions for being a naturalist, we need only wait around for a few decades and the developments of science will show that our view is mistaken. What happens as a result is not a rejection of substantive naturalism, but rather a change in the account of what the view involves. This practice van Fraassen labels "false consciousness" since it shows that there is some underlying viewpoint or attitude responsible for the practice of revising the account. Whether or not we agree with van Fraassen on this point, it is important to distinguish between substantive versions of naturalism and methodological ones, where a methodological version of naturalism arises from respect for science and its methods of investigation. Since science is an investigation of the natural world, any viewpoint that gives honor to the methods of science in the search for truth will incline toward naturalism. Full endorsement of naturalism results when one maintains that no truth can be beyond the epistemic reach of the methods of science, that everything that exists can in principle be investigated by these methods. Some will also add the claim that the only respectable methods for investigating what exists are the methods of science, but naturalism does not require this stronger claim. It is enough that the long arm of science has the entirety of truth within its reach.[1]

Given this construal of naturalism, the attempt to provide naturalistic theories in epistemology is more constrained than we were able to derive above. If one's goal is to provide a naturalistic epistemology and we have only methodological naturalism as our guide, then one can be sure of the naturalistic status of one's theory by employing only those concepts that are already needed in a scientific description of the world. So, for example, causal

theories of knowledge are paradigm instances of naturalistic theories in epistemology, since causality has a presumptive claim to being needed to give an adequate account of the natural world.[2] So too are most versions of reliabilism, according to which beliefs are assessed as candidates for knowledge in terms of nature of the processes or methods that underlie the belief, specifically, whether those processes or methods generally lead to true beliefs. Here the concepts necessary to express the view involve concepts surely at home in the sciences: concepts such as processes and probabilities.

Most versions of virtue epistemology have the same grounds for claiming to be naturalistic theories, inasmuch as they identify the virtues in terms of stable characteristics of a cognizer that yield true beliefs most of the time. Similarly, Plantinga's proper function theory of warrant appears at first glance to have equal claim to being a naturalistic theory, in virtue of the necessity of employing the concept of a proper function in the biological sciences.

This last example reveals a caveat I ought to note about the entitlement to substantively naturalistic credentials on the basis of employing conceptual apparatus central to contemporary science. Even though biology is rife with appeals to proper function, it may be that the concept is implicitly non-natural nonetheless, and Plantinga in particular argues that it is. He argues, that is, that the concept of proper function implies that of a design plan and of a designer (beyond that of Mother Nature) (Plantinga 1993: Chap. 11). Thus, even though his theory is formulated in terms of concepts central to contemporary science, that fact is no guarantee that the theory will be substantively naturalistic.

Even so, constructing a theory using scientifically respectable concepts is the best path to follow for methodological naturalists, since following this path warrants the naturalistic credentials of one's theory. This approach to what counts as naturalistic has the virtue of leaving a large number of theories lacking naturalistic credentials, theories such as coherentism, foundationalism, evidentialism, and some versions of contextualism. What unifies this variety is that each such theory can be expressed employing the concept of evidence (without explaining that concept itself in naturalistic terms): coherentists typically have a holistic conception of evidence, foundationalists typically require the existence of basic evidence, as do contextualists of a certain variety, though they allow the class of basic evidence to vary by circumstance (and evidentialists, most obviously of all, must employ the concept of evidence to express their view).

2. THE PROBLEM OF NATURALISTIC PURITY

Once naturalism is understood in this way, it is easy to see why naturalistic epistemologists gravitate toward various versions of reliabilism. Such approaches employ conceptual apparati clearly at home with the language of science, and they can do so in a way that requires no mysterious supervenience claims about the dependence of the normative on the non-normative.

Even so, endorsing such an approach is no guarantee of producing a naturalistically acceptable theory of knowledge. To see the problem, recall that theories which appeal to evidentialist language are naturalistically suspect. The problem is that naturalistic theories have grave difficulty avoiding such an appeal; that is, they have a difficult time maintaining what I will term "naturalistic purity."

Some examples may help illustrate how a theory might fail the test of purity. According to evidentialism, justification is required for knowledge, and the nature of justification is to be understood in terms of that which is supported by one's evidence. Of course, not every belief for which one has evidence is justified, for one can have evidence for a claim and yet have grounds for doubting it that are sufficient to warrant withholding. A pure version of evidentialism will attempt to explain this grounds-for-doubt qualifier in terms of the concept of evidence itself, and it is not hard to see how to do so. A ground for doubt, according to evidentialism, is a piece of information possessed by a cognizer which is such that in conjunction with any evidence a person has for a certain claim, fails to provide adequate evidence for that claim. So evidentialism seems to have little difficulty passing the purity test: it employs the same conceptual resources to explain the grounds-for-doubt clause as it does to clarify initial, or prima facie, justification.

Not so for reliabilism, however. One might initially think that a belief is candidate for knowledge, according to reliabilism, when it is produced or sustained by a reliable process or method. One difficulty with such a proposal is that one may have grounds for doubting the claim that are sufficient to undermine it as a potential candidate for knowledge. So candidacy for knowledge requires the absence of certain types of grounds for doubt, and a pure version of reliabilism will try to clarify this condition appealing to the language of reliability.

Alvin Goldman's 1979 discussion of this issue is instructive. He attempts initially to deal with such cases in terms of alternative, available reliable processes that would not have resulted in the belief in question. Goldman immediately recognizes that this approach won't quite do as it stands. He points out that requiring an additional process to be used can't be quite

right—if another process is used, the first one won't have been used at all. He also worries about the concept of availability, wondering exactly what it takes for a process to be available (was the scientific method, for example, available to Plato?).

In response to these problems, Goldman gives up the task of constructing a pure version of reliabilism He says, "What I think we should have in mind here are such additional processes as calling previously acquired evidence to mind, assessing the implications of that evidence, etc." (1979: 20). I think Goldman is issuing a promissory note regarding some analysis he is not yet in a position to construct. I think, that is, that he wants to construct a pure version of reliabilism, one which does not borrow language from alien theories of knowledge. Since he doesn't see how to do so, he characterizes the processes in terms the concept of evidence, leaving grounds for complaint from naturalists. For an impure version of reliabilism is not obviously acceptable to naturalists, and one doesn't get a pure version of reliabilism simply by appending the phrase "the process operative when" to the key explicatory concepts of an alien theory.

Plantinga's proper function theory provides another example of the same impurity. The key concept of his theory is that of proper function, and he expresses enjoyment at being able to provide such a naturalistically-acceptable theory of knowledge.[3] In response to the need to rule out possible grounds for doubt, however, Plantinga introduces the concept of a defeater and that of a defeater system, and makes no attempt to explain these concepts in terms of that of proper function.[4] Moreover, the concept of a defeater is pretty clearly an evidential concept, no matter what kind of defeater is involved.[5] So Plantinga provides only an impure version of a proper function theory of knowledge.

Plantinga, of course, holds no stake in a defense of naturalism, for he is convinced that naturalism is false and that the notion of proper function is not, at bottom, compatible with naturalism anyway. For those who thought they had found an ally of naturalism in an epistemology of proper function, however, the point is instructive that Plantinga's version, impure as it is, offers no aid and comfort to his enemies.

Virtue epistemologies of a reliabilist sort face the same problem. The need for a grounds for doubt clause is pressing here as well. Suppose, for example, that one's perceptual faculties count as intellectual virtues or excellences. It is still possible for one to employ such faculties when one's background information argues against using them. In such a case, the temptation Goldman succumbed to presents itself: add a ground for doubt clause formulated in evidentialist language, and thereby insure a failure of naturalistic purity.

Perhaps, though, one might try to avoid the problem as follows. One might think of the problem as a special version of the problem of generality,[6] and the solution to it one of specifying a field of operation in which the use of the faculty always implies positive epistemic status. Consider Sosa's proposal, for example. Sosa relativizes the notion of a virtue to conditions, fields, and environments. He says,

> One has an intellectual virtue or faculty relative to an environment E if and only if one has an inner nature I in virtue of which one would mostly attain the truth and avoid error in a certain field of propositions F, when in certain conditions C. The distinction between E and C is not sharp or important and amounts to a distinction between relatively stable background conditions and relatively episodic conditions (Sosa 1991: 284).

When Sosa gives his final account of a virtue, he supplements the above rough account with a further condition that no broadening of C, F, and E are such that the person would be likely to hold a belief with respect to p and not be likely to be correct regarding p (Sosa 1991: 287).

It is worth noting in passing that this account is not quite what Sosa needs, for it succumbs to an evil demon problem. We can define evil demon counterparts for C, F, and E, and then specify a broadening of each in terms of the disjunction of each with their respective evil demon counterparts. Then for cases in which a person is inclined to form a belief and would still be so inclined in a demon world (perceptual beliefs, for example), then there is a broadening of C, F, and E in which the person is likely to form a belief but is not likely to be correct.

Let us put aside this difficulty, however, for we are not here concerned with the problems raised in epistemology by the possibility of demon worlds, but rather with the question of whether some version of virtue epistemology can pass the test for naturalistic purity. What is needed in Sosa's account for that to be so is for the reference to E, C, and F and the broadening of each to be sufficient to rule out cases where, e.g., a perceptual belief is likely to be correct, even though the believer has background information that makes forming perceptual beliefs in those circumstances suspect. So suppose a person is in such circumstances, i.e., suppose there are values for E, C, and F such that S believes that p in <E, C, F> and is likely to be correct in such circumstances. Suppose further, however, that S has background information casting doubt on the wisdom of forming any such belief in those circumstances.

Sosa's account implies that the belief nonetheless has positive epistemic status, and his response to such cases would appeal, I think, to his distinction between animal and reflective knowledge and his related distinction between aptness and justification (Sosa 1991: 289ff). Justification is a matter of

standard coherence relations on bodies of information, whereas aptness is a matter of having been produced by a virtue, allowing the distinction between animal knowledge, which requires only aptness of belief, and reflective knowledge, which requires justification. We need not devote much attention to the details here, however, for the only way this approach can pass the test of naturalistic purity is to focus exclusively on the concept of animal knowledge. Once the coherence relations involved in Sosa's account of justification are brought into the picture, which includes logical, evidential, and explanatory connections within a body of information, it is clear that the theory has failed the test of naturalistic purity. Furthermore, even if we focus exclusively on the concept of animal knowledge, granting that there is a kind of knowledge that animals possess that need not involve the kind of justification Sosa delineates, it is far from clear that the same concept is applicable to human beings with sophisticated understandings of their epistemic situation, aware of grounds for being suspicious of the reliability of a type of belief-forming mechanism and nonetheless ignoring those grounds for doubt. That is, even if we wish to posit some truncated concept of knowledge appropriate for small children and animals, who lack the sophisticated perspective on themselves that is distinctive of reflective knowledge, we should not extend the application of that concept to those having such a reflective perspective and who choose to ignore its implications.

If we alter the account to accommodate this point, Sosa's theory has no resources to avoid an intrusion by evidentialist concepts into the basic concept of animal knowledge, for that account will need to include a rider to the effect that if one has a reflective perspective on one's situation, then one will need to lack information casting doubt on the wisdom of using one's belief-forming mechanisms in that situation.

Here John Greco's version of virtue epistemology is helpful and illuminating, for Greco attempts to provide a thorough-going virtue account of a problem akin to the one we are addressing. Greco grants the need for both objective and subjective justification in an account of knowledge, and offers a relatively straightforward virtue account of objective justification. What is unique in his view is that he also attempts to provide a virtue account of subjective justification, and if he were successful in this project, we would have a way of developing a virtue account that passes the test for naturalistic purity.

According to Greco, objective justification amounts to a belief being the result of dispositions that make a person reliable regarding that belief in the conditions in question, and subjective justification involves a belief being "the result of dispositions that S manifests when S is thinking conscientiously" (Greco 2000: 218). Greco identifies conscientious thinking

with the "default mode" of being motivated to get to the truth (2000: 191), thus contrasting honest thinking with think that is motivated by non-alethic factors such as greed, prestige, comfort, and the like.

Second, Greco does not identify subjective justification with such proper epistemic motivation; in fact, his definition does not even require proper motivation. For his definition only requires the activity of the dispositions that are present when one is properly motivated, and those dispositions might be active both when one is properly motivated and when one is not. I think this may be too weak a connection between subjective justification and proper epistemic motivation. For example, if the same disposition can accompany both well-motivated and ill-motivated belief, it will be possible to display that disposition ill-motivatedly while rationally believing that an ill-motivated display on this occasion is highly unlikely to get one to the truth. Such a situation strikes me as paradigmatic for lack of subjective justification, rather than one in which subjective justification is present, as Greco's theory implies.

Moreover, Greco's appeal to subjective justification does not explain away the standard counterexamples to simple reliabilism. In BonJour's clairvoyant case,[7] Greco needs to charge the clairvoyant with not manifesting the dispositions that he manifests when trying for the truth. I don't see why the clairvoyant has to be guilty of this charge. The clairvoyant knows better than to trust clairvoyance, but that doesn't imply that he is manifesting dispositions different from the ones manifested for him when aiming for truth. The clairvoyant's failure is that he doesn't take into account possessed defeating information, but one can fail to be do this, and even fail to be disposed to do this, and nonetheless manifest the dispositions one normally does when proceeding honestly.

Most of us do not behave this way cognitively, but that is a contingent fact about us. As we improve cognitively, we learn to monitor for defeating information and we learn to withhold belief when we learn of the presence of defeating information. Even so, in the process of so improving, we often think honestly and display the dispositions that we ordinarily display when honestly trying for the truth without monitoring for defeating information and without withholding belief in the presence of known defeaters. Greco's response to the clairvoyance case requires that the only explanation for retaining a belief in the presence of known defeaters is that we are being moved by dispositions other than those operative when thinking honestly, but there are other options. Habits are often overly general, displaying themselves even where not especially useful or desirable. Transparently honest people sometimes hurt others' feelings by unthinkingly displaying such honesty, and belief formation can exemplify this same feature. The motivations present when one is honestly trying for the truth might be

unthinkingly displayed when attention to the presence of known defeaters would have prevented the display of these motivations.

This point is not merely a trifling failure of detail, but rather a particular instantiation of a more general weakness of reliabilism. On the face of it, knowledge and justification are functions of the information or evidence we possess, but reliabilism wishes to talk in terms of belief-forming mechanisms and so must try to mimic evidential relations with these mechanisms (say, by individuating mechanisms or character traits in such a way that the reliability of these mechanisms coincides with the intuitive idea of information or evidence possessed). I have argued repeatedly that the prospects for successful mimicry are hopeless,[8] and my point here is that the above difficulty is simply another example of the same problem. The concept of defeating information is, intuitively, one concerning the epistemic relationships between semantic contents, and the hope of any version of reliabilism is to mimic these relationships by appeal to the right kinds of mechanisms or character traits. Greco's attempt on this point is not entirely successful, I think, for there are no grounds for thinking that in order to display dispositions involved in trying for the truth, one must be disposed always to avoid belief when defeating information is present. So there is no reason to think that the clairvoyant case is explained away by an appeal to such dispositions.

The point of the discussion to this point has only been a kind of sabotage of the naturalistic program, arguing as I have that it is not obvious what counts as naturalized epistemology and it is far from clear that naturalists have taken seriously enough the requirement to propose a fully naturalistic theory. I want to attack the position more deeply, however, for there is a further problem that shares equal billing among the ignored problems for naturalized epistemology: the problem of accounting for the value of knowledge. I will argue that if naturalists wish to retain the assumption that knowledge claims have truth value, they will be pushed in the direction of virtue epistemology in order to account for the value of knowledge. Since we have already seen that it is not obvious how to be a virtue epistemologist and a naturalist at the same time, the temptation will arise to jettison the assumption. That is, when faced with the difficulty of accounting for the value of knowledge, one's commitment to naturalism may lead one to the extreme position of denying that knowledge claims have truth value, much as a similar naturalistic outlook has led to a similar, non-cognitivist approach in ethics. After goading the naturalist into this position, I will argue that such non-cognitivism in epistemology is untenable.

3. NATURALISTIC EPISTEMOLOGY AND THE VALUE OF KNOWLEDGE

The problem of the value of knowledge is largely ignored and under-appreciated in the history of epistemology. The problem is first introduced by Meno in Plato's dialogue by the same name, where Meno tells Socrates that knowledge is valuable because it is the appropriate guide to action. Socrates provides a counterexample by way of response: if you want to get to Larissa, a guide with true opinions about how to get there is every bit as good as a guide with knowledge. Meno at first balks at the counterexample, but then upon seeing its force, replies, "In that case, I wonder why knowledge should be so much more prized than right opinion, and indeed how there is any difference between them."[9]

Meno's response is instructive. Meno's original conviction was that knowledge is valuable and on pragmatic grounds, but on seeing Socrates' counterexample, he wonders whether our common view that knowledge is valuable is mistaken. His confusion here leads to perplexity about the nature of knowledge—perhaps, he imagines, knowledge is nothing more than true opinion. In Meno's mind, there is an interplay between the question of the value of knowledge and the question of the nature of knowledge, so that failed assumptions about the value of knowledge generate doubts about one's assumptions regarding its nature.

I believe Meno is onto something. Though the question of the nature of knowledge is a prominent and important part of the history of epistemology, the enterprise itself presupposes the value of that about which we theorize. Aristotle reports that all people by nature desire to know, and in theorizing about the object of this desire, we should our presumption that it is a desire for something important. Thus, even though the history of epistemology has not devoted much attention to improving on Meno's failed pragmatic account of the value of knowledge, there is good reason to view the question of the value of knowledge as central to the discipline. We may put the point this way: an adequate epistemology is presumptively subject to twin desiderata, one concerning the nature of knowledge and the other concerning the value of knowledge. First, no epistemology that misconstrues the nature of knowledge can be adequate, and second, an epistemology that undercuts any attempt to explain the value of knowledge is defective.

I said above that an adequate epistemology is "presumptively" subject to these desiderata. This qualifier applies primarily to the requirement concerning the value of knowledge. I do not wish to claim that no epistemology can be adequate without containing an account of the value of knowledge. Nor do I claim that an epistemology is inadequate simply because it is incompatible with the value of knowledge. I do claim, however,

that such incompatibility is a strong reason against the adequacy of such an epistemological theory. Such an implication regarding the question of the value of knowledge would have to be accompanied by an explanation of why knowledge is not valuable.

As I have said, the history of epistemology contains little discussion of the question of the value of knowledge. Given the explicit recognition of the significance of the question in Plato, this failure is somewhat surprising, but I think an explanation for it is not hard to find. The concepts by which knowledge is distinguished from true opinion—concepts such as justification, certainty, infallibility, adequate evidence, accompaniment by an account or reason, having the right to be sure, etc.—are all obviously evaluative concepts, and positive ones at that. As such, it would seem a bit pedantic to spend much time arguing that they add something of value beyond that which true belief provides. So the issue of the value of knowledge remains at the level of a presupposition, in no need of much explicit discussion. It is secure in that position in virtue of the selection of approaches to the question of the nature of knowledge that cite evaluatively positive properties.

What is interesting to note in the present context is the special difficulty faced by naturalistic theories regarding the value of knowledge. The problem that such versions of naturalism face I will term "The Swamping Problem." It arises at the most abstract level because of the following possibility. It is possible for property P to be valuable and for property Q to be valuable, and yet for the property of being both P&Q to have no more value than the value of one of its components. In such a case, the value of one of the components swamps the value of the other component.

The swamping problem would not be relevant to the question of the value of knowledge if that question were merely the question of whether knowledge has value. It follows that knowledge has value if it is composed of items some of which are valuable and not of which are not. The difficulty is that the question of the value of knowledge, as raised by Socrates, is not merely the question of whether knowledge has value. It also involves the issue of whether it is more valuable than related cognitive states such as mere true belief. Once we raise the question of whether knowledge is more valuable than mere true belief, the possibility of encountering the swamping problem is raised since the value of true belief may swamp whatever value there might be in other features of knowledge.

Such a problem is most obvious when the additional item is likelihood of truth. For since true belief is valuable, it is clear that the value of likelihood of truth is parasitic on the value of truth itself. The property of being likely to be beautiful is a valuable feature of a painting, but only because being beautiful is itself a valuable property of a painting. Moreover, if a painting is

known to be beautiful, no additional value is judged to be present by noting that it is also likely to be beautiful. In a phrase, the value of the latter is swamped by the value of the former.

One more example to drive home this point. Suppose I want chocolate, and also want to expend little effort in finding it. I run a search on the web and find a list of places within walking distance of my campus that sell chocolate. I find another webpage that gives a list of places within walking distance that are likely to sell chocolate (the methodology for constructing this list is not given, but perhaps if a place sells any kind of candy it is likely to sell chocolate, and so it is a list of places that sell candy, though not presented as such). The first list is of more interest to me than the second, but that is not my point here. Instead, my concern is with a third webpage, one which is the intersection of the first two: a list of places that both sell chocolate and are likely to sell chocolate. This third list is of no more interest to me than the second, given that my goals are only to find chocolate within walking distance. Moreover, the third list of no more interest to me than the first, even though I have an interest in places that sell chocolate over those that do not and I have an interest in places that are likely to sell chocolate over those that are not likely to sell chocolate. That is, even though both properties are valuable from my perspective, the value of one swamps the value of the other so that the combination of both properties does not yield any enhanced value.

What happens to the swamping problem if one eschews naturalism in epistemology? Put simply, the problem is not severe, for there are several options available for addressing it. First, one might hold that there is a special constituent of knowledge that is intrinsically valuable on its own, independently of any relationship it has to the truth.[10] Another option is to turn more subjective, and identify justification with doing the best by one's own lights in getting to the truth and avoiding error. Such an account of justification explicates its value in terms of its relationship to truth, but doing one's best can have value independently of increasing the likelihood of truth in any way at all. Hence, the value of doing one's best in trying for the truth is not swamped by the presence of truth.

None of this is meant to imply that every theory of knowledge that does not self-consciously aim at being a naturalistic theory of knowledge will be immune from the swamping problem. What I am arguing, instead, is that naturalistic theories of the reliabilist sort face a special burden, which leads to the question of whether there is a special variety of reliabilism that avoids this problem. Virtue epistemologists think there is, for they think that the analogy between virtuous belief and virtuous action is sufficient to undergird the claim that one deserves credit for virtuous belief just as one deserves credit for virtuous action.[11]

Such a strategy must face the issue raised by the role autonomony plays in deserved credit for virtuous action. Actions are voluntary, whereas beliefs are not, and it is quite natural to think that this distinctive feature of action undergirds the practice of praising and giving credit for certain actions.

There are deep and important issues here that I am going to pass over in order to pursue the remainder of the discussion proposed on the topic of naturalism. So instead of pursuing the details, let me gloss quickly what I think a virtue epistemologist needs to say. The first point to insist on is that credit due for action does not require a libertarian conception of freedom to undergird the voluntariness central to any case of credit due for a good action. Instead, credit can be due when it is produced by the right kind of internal cause. In the case of credit due for action, it is important that the internal causes are such that voluntariness occurs. In the case of credit due for belief, it is also important that the belief is the result of internal causes, and even though the right kind of internal causes need not produce beliefs which count as voluntary ones, credit due nonetheless depends on the presence of these causes. In this way, one can deserve credit for certain types of belief just as one can deserve credit for certain types of actions. What is central to the analogy is the internal nature of the causes, not the concept of voluntariness.

There is much more to be said on this issue, but I will leave this discussion with the following provisional conclusion. The issue of the value of knowledge is an important one, and it is worth noting how restrictive a desideratum it is, once one embraces the project of naturalizing epistemology. If knowledge is truly valuable, the only direction I can see for the naturalist to go while still embracing cognitivism is to virtue epistemology.

The problem is that virtue epistemology is, for the naturalist, a house built on sand. When the storms due to questions about naturalistic purity arise, the building is likely to collapse. For no virtue approach appears capable of refraining from appeal to concepts that the naturalist cannot tolerate. In light of this fact, what is a naturalist to do? The temptation is, I think, to turn non-cognitivist about epistemology just as logical positivists turned non-cognitivist about ethics. When one's basic philosophical orientation cannot accommodate a certain domain of truth, one of the two has to go. Thus arises the temptation toward non-cognitivism in epistemology, and it is instructive and not surprising to find in recent discussion the first steps toward such a picture in epistemology. After all, if one of the key problems in epistemology is accounting for the value of knowledge, such theories seem well-suited to explaining that value immediately and directly in terms of the attitudes expressed in using such language. Moreover, the difficulties faced by the

cognitivist in preserving naturalistic purity while attempting to account for the value of knowledge are so difficult that it might seem the path of least resistance to treat at least some of the subject matter along the attitudinal or expressivist lines that are so popular in ethics and value theory more generally. So we can think of the discussion to this point as aimed at pushing the naturalist toward such an attitudinalist or expressivist position.[12] In the remainder of this essay, I want to explore briefly the prospects for this approach.

4. ATTITUDINALISM AND THE VALUE OF KNOWLEDGE

On the face of it, attitudinalism is perfectly adapted to handling the problems for cognitive versions of naturalism discussed above. The technical details of how to handle a grounds for doubt clause disappear, since questions of cognitive content disappear in favor of attitudes expressed. And the question of the value of knowledge seems to be addressed directly, since the attitude expressed in attributions of knowledge is a positive one. There is therefore reason for optimism that attitudinalism provides a solution to the problems facing the naturalist in epistemology.

Attitudinalism can be developed in differing degrees of scope. The widest scope would be an attitudinalism that covered the entirety of knowledge, claiming that the concept was entirely non-cognitive in nature. A weakness of this version of the view is that one the factive character of knowledge, that one cannot know and be mistaken. This point prompts an interest in attitudinalisms of lesser scope, ones that attitudinalize the normative dimensions of knowledge or those that attitudinalize the entirety of the difference between knowledge and true belief.

Concerns about naturalistic purity arise again if one attitudinalizes only part of the difference between knowledge and true belief. For example, suppose one adopts attitudinalism about justification, as does Hartry Field (1996; 1998), while granting that knowledge is more than justified true belief. One will then need a naturalistic account of the fourth condition for knowledge and it is difficult to retain naturalistic purity for such a condition. I do not have the space here to give this issue its due, but let me briefly illustrate the problem. The approaches that do not employ a theory of evidence—approaches such as the relevant alternatives approach and counterfactual approaches in terms of sensitivity and safety—have difficulty avoiding an appeal to the evidentially-laden concept of sameness of total epistemic situation. Relevant alternatives theories normally have to appeal to the concept of close worlds to clarify what makes an alternative relevant, making such theories equivalent to ones which emphasize safety or

sensitivity (where these are clarified respectively by the claims if the person were to hold the belief, it would be true and if the claim were false, the person wouldn't believe it). Counterfactual theories of safety and sensitivity appear to need some notion of sameness of total epistemic situation, however. Possible angels might arrange things so that the sensitivity and safety conditions are satisfied only because new information is given to the cognizer in close worlds, or this difficulty might arise through unusual features due to Mother Nature. In such cases, it is tempting to buttress ones theory by restricting the counterfactuals to require sameness of total epistemic situation. Such a theory, however, has only a thinly-veiled appeal to the concept of evidence: what it is to "have" information and "ignore" it is to fail to respond to its evidential force.

There is further problem for versions of attitudinalism that attempt a cognitivist account of the fourth condition, besides the above one of naturalistic purity, one concerning the difficulty of accounting for the value of knowledge over that of its subparts. As Timothy Williamson puts the point, "[The importance of knowledge] would be hard to understand if the concept knows were the more or less ad hoc spral that analyses have had to become; why should we care so much about that?" (Williamson 2000: 31). It is instructive that the two problems that lead to an interest in attitudinalism re-emerge if one's attitudinalism does not cover the entirety of the gap between knowledge and true belief.

So suppose the naturalist agrees and adopts attitudinalism about the entirety of what must be added to true belief to yield knowledge. Simple versions of attitudinalism have a difficulty we might label "the Spock problem." Purely cognitive beings are possible, even if their survival as a species or subspecies is in jeopardy because if it. Such beings can both have knowledge and ascribe it to others, contrary to the claim of simple attitudinalism that such ascriptions amount to cheering for a particular state of mind.

Still, simple attitudinalism is not the only variety of that viewpoint, and it is not possible here to consider all possible varieties of the position in order to assess its prospects for inclusion in the naturalistic framework. Instead, I want to consider the version adopted by Hartry Field, inasmuch as it depends on what I and many others take to be the most plausible contemporary version of attitudinalism in ethics, that of Allan Gibbard (1990). Gibbard's account is supposed to apply equally to moral and epistemic norms, and he argues that, given an account of natural, Darwinian representation, there is no need to posit normative facts to explain what we are doing when we express normative judgements. Instead, our behavior can be explained solely in terms of the evolutionary value of coordination between minds regarding

which norms to accept and which not to accept. This approach undergirds Field's attitudinalism about justification in epistemology.

The crucial issue for this extension of expressivism outside the domain of morality and action and into the realm of epistemology is this: any defense of the view must appeal to arguments and the adequacy of various proposed explanations, and these arguments and explanations rely on epistemic norms. For example, if I infer that evolutionary theory is better than some alternative because it posits fewer theoretical entities, I rely on the norm that in the search for truth Occam's Razor should be followed. For another example, if I deduce some conclusion by the canons of first-order theory, I rely on the norm that these canons are appropriate in extending our knowledge. And if I reason in accord with Bayes' theorem on some probabilistic matter, I rely on the principle that this theorem is a suitable guide in determining what to believe. The problem these points raise is that, on the face of it, if the norms are not true, the arguments and explanations fail to provide sufficient epistemic grounds for endorsing their conclusion.

Gibbard remains unmoved, however:

> My explanations were of course guided by norms—epistemic norms. Why, say, did no basic tendency toward perfection figure in the explanations [of the beliefs of physicists regarding electrons] I gave? No such thing should be posited, I assumed, when observed patterns can be explained just as well without it. This is a normative judgment, and it and others like it guided me. . . The norms that guide explanation, though, are not themselves parts of the explanation. I did not suggest that we developed our normative capacities because basic tendencies to perfection should not be posted gratuitously. Epistemic norms tell us what constitutes a good explanation, but that does not make them part of that explanation (1990: 122).

Of course, Gibbard is right that the norms that guide reasoning are not themselves part of that very reasoning. Norms are not constituents of explanations, but that is not the appropriate level of concern here. Explanations have presuppositions as well as constituents, and the adequacy of Gibbard's explanation depends on the truth of Occam's Razor in the way distinctive of at least some of these.

What is a presupposition? I offer no general theory of them, but we can say this much: the class of presuppositions is a subset of the class of implications of a statement or set of statements. One quite clear example is the following: the validity of modus ponens presupposes the truth of the corresponding conditional to modus ponens. In the quote above, Gibbard admits that he assumed the truth of Occam's Razor and that it guided his reasoning. Those claims are true, but more is true as well. It is also true that Gibbard's argument would be inadequate if Occam's Razor (properly

formulated) were not true. That is, the quality of his arguments depends crucially on Occam's Razor, in just the way that the validity of modus ponens depends on the alethic status of its corresponding conditional. In the language I am employing, his arguments presuppose the truth of Occam's Razor. If Occam's Razor is false, Gibbard had better look around for a better defense of his attitudinalist views, for the present argument supports his position only if that methodological principle is true.

One can get distracted here by the metaphysically tendentious language of facts, wondering about the distinction between facts and values, but that is not the issue at all. The question is not whether Occam's Razor is a fact, but rather whether it is a presupposition of certain explanations. It's being such requires that it be semantically evaluable, and whatever else it demands in terms of the existence of facts and the like depends on results in the theory of truth. The simple point is that arguments and explanations presuppose the truth of epistemic norms, and if the norms themselves are given non-alethic, attitudinal interpretations, then the explanations and arguments are simply defective in virtue of the fact that their presuppositions are not true.

One might think Gibbard has something of an answer here, in the following way. It is a well-known problem for expressivism how to account for reasoning involving normative conditionals (known as the Geach problem), and Gibbard employs the notion of a normative-factual world (an NF-world) to handle this problem, and one might think the same apparatus can help here.[13]

Here's how Gibbard approaches the Geach problem. We first define an NF-world: it is the opinion-set of a fully opinionated person, one who has a complete and consistent set of opinions containing for every factual claim p, either p or ~p and for every normative claim r, either r or ~r. The appeal to consistency here cannot be understood in usual alethic terms, for the expressivist doesn't think that normative claims are strictly true or false. Instead, consistency for normative judgements is to be understood something like the following: having a set of opinions that is not dilemma-inducing, i.e., there is no option such that one's opinions both prohibit and permit that option.[14] We will see that this need for a non-alethic conception of consistency causes problems, but before we can appreciate the problems, we need to explain the relationship between the attitudes of ordinary cognizers and this technical notion of an NF-world. We define the opinions of less-than-fully-opinionated in terms of those of fully opinionated ones: such opinions are identified with the disjunction of all the NF-worlds that might turn a partially opinionated person into a fully opinionated person. Once we have this apparatus of NF-worlds in place, we can define validity in the usual way, except that we use NF-worlds instead of possible worlds and

thereby avoid any need to appeal to the truth or falsity of normative elements in embedded contexts.

The hope of the attitudinalist is that this apparatus of NF-worlds be used to clarify the way in which an argument might presuppose a norm. Take, for example, an argument that presupposes Occam's Razor. The expressivist approach to such a case requires explaining this notion of entailment in terms of the non-logical notion of consistency outlined above. Instead of claiming that the argument can't be a good one without the norm being true, one will have to argue that anyone who accepts the argument must also accept Occam's Razor, that an inconsistency in attitude is generated in the opinions of a person by accepting the argument and rejecting Occam's Razor. That claim, however, is simply false. As Quine has taught us, nearly any combination of beliefs can be maintained if one is willing to make enough adjustments elsewhere in the system of beliefs. For example, one might accept an argument that presupposes Occam's Razor while at the same time denying that norm, claiming that one's principle is to accept arguments like this one that originally occur to one on Monday morning (we assume that the time the argument first occurred to one is Monday morning). Suitable adjustments elsewhere in the belief-system can be made to avoid any inconsistency in attitude. Gibbard will be hard-pressed to guarantee that enough questioning will unearth some hidden inconsistency, and even if some fully opinionated people would be found inconsistent, others surely would not. Little familiarity is needed with mental illness to realize that obvious entailments can be consistently denied if one is willing to change what is ordinarily accepted in a radical enough fashion.

The proper conclusion to draw is that the presuppositional relationship between norms and arguments is a different issue from the Geach problem, and one that cannot be solved by the resources Gibbard uses to address the Geach problem. What makes the problems different is that the norms in the Geach problem are explicit in reasoning, whereas presupposed norms are not. They can thus be rejected by the reasoner without inducing any inconsistency in attitude, even though they are logically inconsistent with the adequacy of the argument.

Those inclined toward attitudinalism might take the above discussion as a reason to take back any acceptance of the idea that arguments presuppose norms. They might think that the Quinean point that nearly anything can be embedded in some consistent system or other is a reason for thinking that the idea that an argument can presuppose a norm is simply mistaken. Such a response is extreme. What can be accepted, even rationally accepted, is simply not an infallible guide to entailment relations, here or elsewhere.

There are a number of illuminating points of view on this attitudinalist mistake. From a Fregean point of view, the mistake is that psychologizing

some logical relationships is essential to the attitudinalist program, and there is no more reason to think that such limited psychologism is any better than global psychologism. Another perspective on the mistake is broadly Chisholmian. A large part of the history of the last century of epistemology involves the search for true epistemic principles, and some of these principles link non-normative features of the world (e.g., sensory appearances, coherence, etc.) with certain levels of epistemic acceptability. The above problem shows that expressivism cannot countenance this search, for they cannot explain how any of them could be true. In a way, this result may not be that surprising: such principles are fodder for supervenience claims about how the truth of normative claims depends on non-normative features of the world. Perhaps the lesson is that expressivism is far from an alternative explanation of the same data that non-expressivists try to explain: it is, instead, a position forced to deny the data itself.

There is one other illuminating point of view from which to view the above objection to expressivism. If we put this argument in Quinean form, it has especially devastating consequences for the naturalists' respect for science. In Quinean form, what I am arguing for is that epistemic norms have to be alethically evaluable since our best scientific theories require it—indeed, we cannot explain the concept of a best theory without presupposing the truth of some such norms. To turn expressivist about these norms lands one in the same turnip field as the sociobiologists who assert that we have no need for the concept of truth in an understanding of the development of science, failing to take into account that no such viewpoint can be defended by appeal to best explanation or adequate evidence of any sort. Good arguments presuppose logical and epistemic norms, and the concept of presupposition itself involves the concept of semantic evaluation.

5. CONCLUSION

Naturalism is therefore in serious jeopardy of being unable to account for or explain the value of knowledge. Non-cognitivism about epistemic norms undermines every attempt to defend any position, and cognitivist theories have very few good answers to the swamping problem. To the extent that they are successful in avoiding this problem, they do not maintain naturalistic purity, and thus fail to present an acceptable naturalistic account of the value of knowledge. The conclusion to draw, then, it is it is very hard to see how naturalism is compatible with the value of knowledge.

ENDNOTES

[1] One might wish, at this point, to distinguish between naturalism so construed and physicalism, the view that everything in the universe is either physical or physically realized. Naturalism as construed here is logically distinct from physicalism. There is no a priori reason to assume that the methods of science will discover only what is physical, and there is no a priori reason that everything physical is within epistemic reach of the methods of science. But the distinction between the two is not center stage here, so I ignore the distinction in the text.

[2] But cf. Russell 1912: "The law of causality…is a relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm."

[3] "So the view I propose is a radical naturalism: striking the naturalistic pose is all the rage these days, and it's a great pleasure to be able to join the fun" (Plantinga 1993: 46).

[4] Plantinga 1993; see especially his views on the concept of a function.

[5] For more on different kinds of defeaters, see Pollock 1986.

[6] The generality problem was first presented in Feldman and Conee 1985. The problem, as they present it, is that if the processes are individuated too coarsely, then, e.g., all perceptual beliefs are justified or none of them are (because they are all produced by the same process). If the processes are individuated too finely, then only the belief in question will be produced by that process, yielding the result that the process is reliable, and hence the belief justified, if and only if the belief is true. The problem thus issues a challenge to reliabilists to find the right level of generality for their theory.

[7] See BonJour 1985: 38-45. I'm thinking here especially of the case of Norman, p. 41.

[8] See, e.g., Kvanvig and Menzel 1990; Kvanvig 1996; Kvanvig 2000; and Kvanvig 1992: Chap. 5.

[9] Plato, *Meno*, 97c-d.

[10] See, e.g., Chisholm 1989.

[11] See, e.g., Riggs 2003, Greco 2003, and Sosa 2003.

[12] I will not attempt in the text any clarification or taxonomy concerning the varieties of non-cognitivism, and will vary my terminology regularly between terms that ought to be distinguished. Attitudianalism ought to be a view that identifies the non-cognitive element in terms of the *attitudes* of people holding such beliefs, whereas expressivism ought to be a theory about what is involved in certain types of speech acts. These issues are important, but not in the present context, so I will ignore them in the text.

[13] I thank Matt McGrath and Jeremy Fantl for this suggestion, and for other very helpful comments on an earlier draft.

[14] Thanks to Robert Johnson for this idea.

REFERENCES

BonJour, Laurence. 1985. *The Structure of Empirical Knowledge*. Cambridge, MA: Harvard University Press.

Chisholm, Roderick M. 1977. *Theory of Knowledge*, 2nd edition. Englewood Cliffs, NJ: Prentice-Hall.

Chisholm, Roderick M. 1989. Theory of Knowledge, 3rd edition. Englewood Cliffs, NJ: Prentice-Hall.

Feldman, Richard. 2001a. Naturalized epistemology. In the *Stanford Encyclopedia of Philosophy*, http://plato.Stanford.edu/entries/epistemology-naturalized/.

Feldman, Richard. 2001b. We are all naturalists now. Minneapolis: American Philosophical Association Meeting.

Feldman, Richard and Earl Conee. 1985. Evidentialism. *Philosophical Studies* 48: 15-34.

Field, Hartry. 1996. The a prioricity of logic. *Proceedings of the Aristotelian Society* 96: 359-379.

Field, Hartry. 1998. Epistemological nonfactualism and the a prioricity of logic. *Philosophical Studies* 92: 1-24.

Gibbard, Allan. 1990. *Wise Choices, Apt Feelings*. Cambridge, MA: Harvard University Press.

Goldman, Alvin. 1979. What is justified belief? In *Justification and Knowledge: New Studies in Epistemology*, ed. George Pappas, 1-25. Dordrecht, Holland: D. Reidel.

Greco, John. 2000. *Putting Skeptics in Their Place*. Cambridge: Cambridge University Press.

Greco, John. 2003. Knowledge as credit for true belief. In *Intellectual Virtue: Perspectives from Ethics and Epistemology*, eds. Michael R. DePaul and Linda Trinkaus Zagzebski. Oxford: Oxford University Press.

Kvanvig, Jonathan L. 1992. *The Intellectual Virtues and the Life of the Mind: On the Place of the Virtues in Contemporary Epistemology*. Savage, MD: Rowman & Littlefield.

Kvanvig, Jonathan L. 1996. Plantinga's proper function theory of warrant. In *Warrant in Contemporary Epistemology: Essays in Honor of Plantinga's Theory of Knowledge*, ed. Jonathan Kvanvig, 281-306. Savage, MD: Rowman & Littlefield.

Kvanvig, Jonathan L. 2000. Zagzebski on justification. *Philosophy and Phenomenological Research* 60: 191-196.

Kvanvig, Jonathan L. and Christopher Menzel. 1990. The basic notion of justification. *Philosophical Studies* 59: 235-261.

Plantinga, Alvin. 1993. *Warrant and Proper Function*. Oxford: Oxford University Press.

Pollock, John L. 1986. *Contemporary Theories of Knowledge*. Totoway, NJ: Rowman and Littlefield.

Riggs, Wayne. 2003. Understanding virtue and the virtue of understanding. In *Intellectual Virtue: Perspectives from Ethics and Epistemology*, eds. Michael R. DePaul and Linda Trinkaus Zagzebski. Oxford: Oxford University Press.

Russell, Bertrand. 1912. On the notion of cause. *Proceedings of the Aristotelian Society*, xii: 1-26.

Sosa, Ernest. 1991. Intellectual virtue in perspective. In *Knowledge in Perspective: Selected Essays in Epistemology*. Cambridge: Cambridge University Press.

Sosa, Ernest. 2003. The place of truth in epistemology. In *Intellectual Virtue: Perspectives from Ethics and Epistemology*, eds. Michael R. Depaul and Linda Trinkaus Zagzebski. Oxford: Oxford University Press.

Sturgeon, Scott. 2002. Comments. Rutgers, NJ: Rutgers Epistemology Conference.

Van Fraassen, Bas. 1996. Science, materialism, and false consciousness. In *Warrant in Contemporary Epistemology: Essays in Honor of Alvin Plantinga's Theory of Knowledge*, ed. Jonathan Kvanvig, 149-181. Lanham, MD: Rowman & Littlefield Publishers.

Williamson, Timothy. 2000. *Knowledge and its Limits*. Oxford: Oxford University Press.

Chapter 10

NATURALISM AND MORAL REALISM*

Michael C. Rea
University of Notre Dame

My goal in this paper is to show that naturalists cannot reasonably endorse moral realism. In defending this conclusion, I mean to contribute to a broader anti-naturalistic project. Elsewhere (Rea 1998, 2002), I have argued that naturalists must give up realism about material objects, materialism, and perhaps even realism about other minds. Materialism aside, I take realism about material objects and realism about other minds to be important parts of our commonsense metaphysics. Likewise, I take moral realism to be an important part of commonsense morality. Insofar as it conflicts with these important parts of our commonsense view of the world, naturalism is unattractive. Of course, one might doubt that unattractiveness counts as *evidence* against a philosophical position; but, as I'll explain below, I think that naturalism is not a philosophical position, but a research program. Moreover, I have argued elsewhere (Rea 2002) that naturalism, like any other research program, must be adopted or rejected solely on the basis of its pragmatic appeal (or lack thereof). It is for this reason that highlighting unattractive features of naturalism is an important way of attacking it.

Moral realism is the view that there are objective moral facts.[1] There are objective moral facts only if the following two conditions are met: (i) there are moral properties—e.g., properties like being a right action, being a wrong action, being praiseworthy, being depraved, and so on—at least some of which are exemplified by actual objects or events, and (ii) the exemplification of a moral property p does not entail that anyone has beliefs about what exemplifies p, about whether p is exemplified at all, or about the conditions under which p is exemplified. Condition (ii) is meant to express

T. M. Crisp, M. Davidson and D. Vander Laan (eds.), Knowledge and Reality, 215-241.

part, but only part, of what many philosophers aim to express by phrases like 'moral properties are not mind-dependent' or 'moral facts are not theory-dependent'.[2]

Some naturalists already accept the conclusion that I want to defend here, but many continue to resist it. For reasons that will become clear below, those who resist have typically done so by arguing for one of the following claims:

(C1) Regardless of whether they are reducible to non-moral properties, objective moral properties play an indispensable role in the best causal explanations of at least some natural phenomena (e.g., moral beliefs and judgments, or morally significant behavior).

(C2) Moral properties are reducible to non-moral properties which, in turn, play an indispensable role in the best causal explanations of various natural phenomena.

Part of what I aim to show is that, contrary to widespread opinion, neither of these claims offers any promising line of resistance against the conclusion I'll be defending.

My argument will come in two parts. The first part aims to show that any plausible and naturalistically acceptable argument in favor of belief in objective moral properties will appeal in part to simplicity considerations (broadly construed)—and this regardless of whether moral properties are reducible to non-moral properties. By 'simplicity considerations (broadly construed)' I mean just those considerations that reflect our preference, *ceteris paribus*, for theories that are elegant, ontologically economical, mathematically simple, and consistent with our considered judgments, theoretical commitments, and other entrenched background presuppositions.[3] (Such considerations are often referred to as 'pragmatic' considerations; but I avoid that label because I do not want to presuppose that they are merely pragmatic and thus not indicative of truth.) Henceforth, I will speak of an appeal to such considerations just as an "appeal to simplicity".

The second part argues for the conclusion that appeals to simplicity justify belief in moral properties only if either those properties are not objective or something like theism is true. Thus, if my argument is sound, naturalists can reasonably accept moral realism only if they are prepared to accept something like theism. But, as will become clear, naturalists can reasonably accept theism or something like it only if belief in some such doctrine is justified by the methods of science. For present purposes, I'll assume (what I think virtually every naturalist will grant) that belief in theism and relevantly similar doctrines is not justified by the methods of science. Thus, I will conclude that naturalists cannot reasonably accept

moral realism. Before presenting the details of the argument, however, I'll first say a few words about the nature of naturalism.

1. NATURALISM

As I understand it, naturalism is not a view, or a philosophical thesis, but a research program. A research program is a set of methodological dispositions—dispositions to trust particular cognitive faculties as sources of evidence and to treat particular kinds of experiences and arguments as evidence. Naturalism, so I say, is a research program that treats the methods of science, and those methods alone, as basic sources of evidence (where a putative source of evidence is treated as basic just in case it is trusted in the absence of evidence in favor of its reliability).

In characterizing naturalism this way, I put myself at odds with many philosophers—naturalists and non-naturalists alike. But the philosophers with whom I am at odds are not at all unified in their views about what naturalism is. Some say that naturalism is primarily a metaphysical view (for example, the view that the universe is a closed causal system).[4] Others say that it is primarily an epistemological view (for example, the view that scientific inquiry is the only avenue to knowledge).[5] Still others say that it is primarily a view about philosophical methodology (for example, the view that philosophers ought to abandon traditional problems about skepticism and ontology and pursue their various projects in a way continuous with the methods of science).[6] Most naturalists would affirm Wilfrid Sellars's slogan that "science is the measure of all things: of what is that it is and of what is not that it is not" (Sellars 1963: 173); and many, no doubt, would say that this slogan captures the heart and soul of naturalism. But apart from that, there is little agreement about what, precisely, naturalism amounts to.

It is tempting, in light of the proliferation of different and conflicting formulations of naturalism, to say that naturalism comes in different varieties, each expressible by a different philosophical thesis. Those who give in to this temptation typically list three varieties—metaphysical, epistemological, and methodological—though once in a while other varieties are identified.[7] Different philosophers are then labeled not simply as naturalists but as metaphysical, epistemological, or methodological naturalists, depending on which of the relevant theses they seem to endorse.

But this is not the only way of accounting for the diversity of formulations of naturalism. Another possibility is that there is indeed only one version of naturalism, but many mischaracterizations of it. Given the current state of the literature, to embrace this possibility is to say that many naturalists have mischaracterized their own naturalism. Saying this might

seem uncharitable. It might also seem implausible. Nevertheless, I think that there are very good reasons for doing so.

Despite all the disagreement about how to formulate naturalism, almost every naturalist agrees that naturalism somehow involves deep respect for the methods of science above all other forms of inquiry. To the extent that one fails to manifest a disposition to follow science wherever it leads, one fails to count as a naturalist. But if we take this idea seriously, then we are led fairly directly to the conclusion that naturalism couldn't be a substantive philosophical thesis. For naturalists will agree that any substantive thesis that we might plausibly identify with naturalism is itself at the mercy of science. That is, any such thesis must be justified by the methods of science, if at all; and any such thesis can, at least in principle, be overthrown by scientific investigation. But no one seems to think that naturalism itself would be refuted if science were to produce evidence against some favored thesis of (e.g.) metaphysics, epistemology, methodology, or semantics. Again, the heart of naturalism is to follow science wherever it leads; but, clearly enough, one cannot be a naturalist and be disposed to follow science wherever it leads if naturalism itself is inextricably tied to some thesis that science might overthrow. To suppose that naturalism involves dogmatic adherence to a substantive philosophical thesis is, therefore, either to suppose that naturalists one and all have fallen into a rather elementary and uninteresting sort of incoherence or to suppose that, appearances to the contrary, naturalists are not really unified by a disposition to follow science wherever it leads. But neither of these alternatives seems plausible. Thus, in my view, it is much better (and, ultimately, more charitable) to say that naturalism is not a thesis, but something else.

I suppose there are many other things naturalism could be: an attitude, a value, a preference, etc. However, in light of what the most prominent 20th century naturalists have said about it, my own view is that naturalism is best characterized as a research program. Taking it this way fits very nicely with the characterizations (slogans aside) offered by its most prominent spokesmen in the 20th Century—John Dewey and W. V. Quine. Moreover, it faithfully captures what is common to virtually all of those who call themselves naturalists without falling prey to the problem (briefly described above) that besets any attempt to express naturalism as a thesis. As I see it, then, what unifies naturalists is not adherence to a philosophical position, but rather a disposition to conduct inquiry in a certain way—a way dominated by the methods of science.[8]

What are the methods of science? Notoriously, it is hard to say exactly what they are. But we can say very roughly that the methods of science are, at present anyway, those methods (including canons of good argument, criteria for theory choice, and the like) that are regularly employed and

respected in contemporary university science departments (e.g., departments of biology, chemistry, geology, physics, etc.). Reliance on memory and testimony is surely included among these methods, as are judgments about apparent mathematical, logical, and conceptual truths. Ruled out, on the other hand, are evidential appeals to ungrounded hunches, rational intuitions (conscious episodes in which a proposition seems to be necessarily true), putative divine revelations or religious experiences, manifestly unreliable sources of testimony, and the like.[9] Again, this characterization is rough; but it will do well enough for present purposes.

2. SCIENCE AND MORALITY

In light of the characterization of naturalism just given, it should be clear that a naturalistically respectable argument for any conclusion will be one that appeals only to premises that can be known by way of the methods of science. In this section, I will argue that any naturalistically respectable argument for belief in objective moral properties will have to appeal to simplicity.

I'll take as my point of departure Gilbert Harman's well-known argument for the general conclusion that moral realism is untenable. In short, Harman rejects moral realism on the grounds that objective moral facts have no role to play in our best causal explanations of natural phenomena. In response to his argument, those interested in defending *both* naturalism and moral realism have typically defended either C1 or C2:

(C1) Regardless of whether they are reducible to non-moral properties, objective moral properties play an indispensable role in the best causal explanations of at least some natural phenomena (e.g., moral beliefs and judgments, or morally significant behavior).

(C2) Moral properties are reducible to non-moral properties which, in turn, play an indispensable role in the best causal explanations of various natural phenomena.

Indeed, as I'll make clear below, there seems to be no naturalistically respectable way of resisting Harman's argument apart from defending C1 or C2. But I'll also argue that, if this is right, then even if C1 or C2 can be successfully defended, any naturalistic argument for belief in objective moral properties will have to make some appeal to simplicity.

2.1 Harman's Argument

In the opening chapters of *The Nature of Morality*, Gilbert Harman argues that ethics is problematic because it appears that "there can be no explanatory chain between moral principles and particular observings in the way that there can be such a chain between scientific principles and particular observings" (1977: 9). The "particular observings" for which moral facts are candidate explanations are just moral observations. For example, Harman points out that if we see some young hoodlums pour gasoline on a cat and ignite it, we do not need to *conclude* that their behavior is wrong; we can *see* that it is an instance of *wrong behavior* just as clearly as we can see that it is an instance of *cat-burning behavior* (1977: 4). But, he argues, moral facts have no role to play in explaining this sort of observation. More exactly: they have no role to play in the *best causal* explanation of this sort of observation (or of anything else). As he makes clear elsewhere (Harman 1985: 33-4; Harman 1986: 61-4) the point isn't that moral facts are never invoked in an explanatory way. The point, rather, is that the fact that the behavior of the cat-burning hoodlums is wrong, the fact that the hoodlums are depraved, and other such moral facts seem not to figure in any causal explanations of anything. Thus, Harman can grant that it makes perfect sense to say (e.g.,) that we are repulsed by the behavior of the cat-burning hoodlums because the behavior is wrong, and that the children are behaving in that way because they are depraved. The point is just that the wrongness of their behavior does not cause our observation that the behavior is wrong, nor does it cause our repulsion at that behavior; and the depravity of their character does not cause the hoodlums to burn the cat.[10] All of these events have perfectly natural, non-moral causes, and it is those causes, rather than any alleged moral facts, that seem to figure in the best (causal) explanations of their effects. But if that's right, Harman thinks, then, absent a reduction of moral facts to non-moral ones, we have no scientific reason—and hence, on his view, no reason at all—to believe in moral facts.

Officially, Harman's argument thus far is directed against belief in irreducible moral *facts*. He also expresses reservations about the possibility of offering a plausible and sufficiently detailed naturalistic reduction of moral facts. But it is clear from his presentation that his objections against belief in irreducible moral facts, as well has his reservations about the possibility of reducing moral facts to non-moral facts, apply equally to belief in objective moral properties and to the possibility of reducing such properties to non-moral ones. Thus, I will henceforth talk about Harman's argument and responses to it as if what is at issue is belief in objective moral properties rather than belief in moral facts. Also, unless otherwise indicated,

I will use the term 'moral properties' unqualified as shorthand for the term 'objective moral properties'.

I have already mentioned two ways of replying to Harman's argument: defend C1 or defend C2. Before discussing those replies, however, I want first to identify and set aside two other ways of replying. As I understand it, Harman's argument rests on four premises:

(P1) Irreducible moral properties have no (indispensable) role to play in our best causal explanations of any natural phenomena.
(P2) Moral properties are not reducible to non-moral properties that play indispensable roles in our best causal explanations of natural phenomena.
(P3) Scientific justification proceeds by way of inference to the best (causal) explanation.
(P4) If there is no scientific justification for believing in *x*s, then there is no justification at all for believing in *x*s.

C1 and C2 are attacks on P1 and P2 respectively. But one might also resist the argument by attacking either P3 or P4. Attacking P4 is unacceptable from a naturalistic point of view, however—not because science *couldn't* provide reason for rejecting P4, but because (as far as we know) science *hasn't* offered reason to reject P4.[11] Thus, short of defending C1 or C2, the only other avenue of reply is to attack P3.

For the most part, P3 has gone unquestioned in the literature;[12] and the importance of C1 in the moral realism debate is powerful testimony to the fact that naturalists, in general, have been prepared to accept it. P3 is not beyond question, of course. But I doubt that there are alternatives that stand a better chance of being compatible with both naturalism and realism about the entities posited in scientific theories. Bas van Fraassen (1989), for example, is a well-known critic of inference to the best explanation, but his own conception of scientific justification is explicitly anti-realist. Likewise, Richard Boyd (1988) urges the conclusion (superficially contrary to what philosophers like Harman seem to think) that the *method of reflective equilibrium* is the method of science. But Boyd does not deny that the method of reflective equilibrium as he understands it is equivalent to what Harman would call "the inference to the best explanation", and a look at Harman's explicit account of inference to the best explanation bears out the equivalence (cf. Harman 1965). Moreover, Boyd (1982, 1988) concedes what will be crucial for my point later on—namely, that by employing the method of reflective equilibrium as a method of theory choice, we inevitably choose theories in part on the basis of simplicity considerations.[13] This fact is all that I aim to establish by assuming with Harman that scientific justification proceeds by way of inference to the best

explanation.[14] But it is, I think, a fact that will be as easily established under any other plausible assumption (like Boyd's) about the process of scientific justification that purports to be compatible with scientific realism.

Granting P3 and P4, the only way to resist Harman's argument is to endorse either C1 or C2. As it happens, I think that both C1 and C2 are false; but in the remainder of this section my main concern will simply be to show that, even if one or the other is true, naturalistic arguments in support of moral realism must ultimately rest on an appeal to simplicity.

2.2 Inference to the Best Explanation

According to C1, objective moral properties play an indispensable role in the best causal explanations of at least some natural phenomena, and this regardless of whether they are reducible to non-moral properties. Plausible examples in support of C1 are hard to find; but three seem especially worthy of attention. First, one might note that we often regard the moral judgment of others as being more or less *reliable* than our own. But, one might think, one's moral judgment can be reliable only if the presence or absence of moral properties at least partly causally explains one's moral beliefs (Sturgeon 1986: 71-2; Adams 1999: 67-8). Second, one might note that we're inclined to believe that, say, moral depravity *leads* people to do terrible things, or that moral decency *keeps* people from doing such things. But this too might seem to make sense only if moral properties enter into causal explanations (Sturgeon 1986: 74-5). Third, one might think that "certain regularities—for instance, honesty's engendering trust or justice's commanding allegiance, or kindness's encouraging friendship—are real regularities that are unidentifiable and inexplicable except by appeal to moral properties" (Sayre-McCord 1988b: 276). Here too, then, we might seem to have a case of moral properties playing a role in causal explanations.

My own view is that naturalists should not put much stock in examples like these. Kurt Gödel's mathematical sensibilities were more reliable than my own; and both Fermat's Last Theorem and Goldbach's Conjecture have kept many a mathematician up late at night. Does it follow from any of this that mathematical propositions or properties enter into causal explanations? Ironically enough, Harman would probably concede that mathematical propositions and properties do enter into causal explanations, but that is only because they play an indispensable role in the sorts of causal explanations that constitute our best physical theories. It is emphatically not because they must be invoked as causes either of mathematical beliefs or of insomnia. As Harman points out, however, moral propositions do not enter into physical theory, or any other scientific theory, in the way that mathematical

propositions do. And I think that there is no more reason to think that they must be (or even can be) invoked as causes of moral beliefs or morally significant behavior than there is to think that mathematical properties or propositions can be invoked as causes of mathematical beliefs or of insomnia. Of course, if Sayre-McCord is right in thinking that there are at least some regularities in the world that are "unidentifiable and inexplicable" apart from an appeal to moral properties, then there *is* reason to think that moral properties enter into our best causal explanations of natural phenomena. But I see no naturalistically acceptable reason for thinking that Sayre-McCord's claim is true. Consider his first example: honesty's engendering trust. The clear, empirically detectable regularity here is a connection between a certain kind of truth-telling disposition and various other dispositions to believe and act on the things that are said by people with the first disposition. But why think that this regularity can't be identified or explained apart from an appeal to moral properties? Similar remarks apply to the other examples on Sayre-McCord's list.

I needn't press this point, however. For, as I will now argue, there's good reason to think that, regardless of whether C1 is true, any scientific justification we might have for belief in objective moral properties will depend on an appeal to simplicity. As Section 3 will make clear, this is all that is required to show that naturalists cannot accommodate belief in objective moral properties.

Suppose, as we have been, that scientific justification proceeds by way of inference to the best explanation. There are, very roughly speaking, two ways in which we can be justified by an inference to the best explanation in believing that properties of a certain kind are exemplified. The properties in question might be among the explainers, explicitly posited as salient causes of particular empirical phenomena. Or their existence might be implied by background presuppositions which are part of the theory because of their simplifying role (i.e., their presence in the theory helps to make it more elegant, more ontologically economical, less mathematically complicated, or more consistent with our considered judgments, theoretical commitments, or other entrenched presuppositions). I do not mean to suggest that there is any sharp distinction to be drawn between explanatory posits and background assumptions. But there is at least an intuitive, rough-and-ready distinction here that is worth attending to. So, for example, if belief in the fundamental, causally efficacious properties of protons is justified by an inference to the best explanation, it is so because those properties are posited by our best explanations of various empirical phenomena as causes of those phenomena. On the other hand, if belief in the kind-property *being a proton* is justified by an inference to the best explanation, it probably is so not because that property too is posited as a cause of various empirical phenomena, but rather because our theories are simplified by framing them in terms of an ontology

that includes protons rather than, say, in terms of an ontology that includes only mere bundles of the more fundamental properties, or aggregates of instantaneous proton-stages, or something else empirically but not metaphysically equivalent. I say this because, plausibly, there is nothing that would be causally explained by the property *being a proton* that isn't already causally explained by the more fundamental, intrinsic, non-sortal properties of protons. Likewise, I think, with properties like *being a material object*, *being an enduring particular*, and *being an intrinsic modal property*. Such properties are either causally inert or causally redundant. Thus, whatever scientific justification we have for believing in them would seem to come from the *simplifying* role they play in our theories, since whatever causally explanatory roles they might be thought to play are either spurious or else already being played by other, more fundamental properties.

Now, it is hard to take seriously the idea that moral properties are explanatory posits. That is, it is hard to take seriously the thought that our main reason for believing in moral properties is that our best scientific theories posit them as the salient explanatory causes of particular empirical phenomena. As we have seen, some do claim that moral properties are causally efficacious and that they play a role in our best explanations of natural phenomena. But no naturalist seems seriously to think that the explanations in question invoke moral properties to explain phenomena that are otherwise causally *unexplained.* To whatever extent moral properties are causally efficacious at all, from a naturalistic point of view they are either reducible to non-moral properties or else irreducible but causally redundant. In either case, all of the relevant explanatory work is already done by non-moral properties. Thus, there is no need to posit distinctively moral properties for explanatory purposes. So if belief in moral properties is justified by an inference to the best explanation, this must be because our theories are somehow simplified by framing them in terms of an ontology that includes moral properties rather in terms of one that doesn't.

Further evidence for this comes from the fact that none of the major defenders of the explanatory value of moral properties attempts to defend the claim that moral properties are explanatory posits. Nicholas Sturgeon (1985, 1986), for example, makes it his strategy to *assume* that there are moral properties and then to show that, on that assumption, such properties have a role to play in our explanations of various phenomena. Thus, rather than attempt to show that moral properties must be posited to explain various phenomena, he only aims to show that explanatory roles can be found for moral properties if we take for granted (presumably for other reasons) that there are such properties. Boyd (1988), Jackson (1998), Jackson & Pettit (1995), Railton (1986), and Sayre-McCord (1988b) among others all take similar strategies. And this is precisely the strategy we should expect to find

naturalistic proponents of C1 taking if, as I have argued, whatever scientific justification we have for belief in moral properties comes from the simplifying role of such belief.

There is another reason for thinking that if belief in moral properties is justified by an inference to the best explanation then it is justified in part on pragmatic grounds. It is widely believed that, in science, what counts as the best explanation of some phenomenon is determined in large part by what I have called simplicity considerations, broadly construed.[15] I will not attempt to defend this view here; but if it is true, then it follows directly that, if belief in moral properties is justified by an inference to the best explanation, it's justification depends ultimately upon an appeal to simplicity.

2.3 The Irrelevance of Reducibility

According to C2, moral properties are reducible to non-moral properties that figure in our best causal explanations of natural phenomena. I take it that, in the context of the moral realism debate, the project of reducing moral properties to non-moral properties is just the rather broad project of trying to show how moral properties might be identical with or in some sense composed of properties that are quantified over in paradigmatically scientific theories. (Thus, there is no reason to suppose that a reduction would have to provide "bridge principles" explicitly identifying specific properties mentioned in existing moral theories with specific properties mentioned in existing physical, chemical, or biological theories.) In the remainder of this section, I will argue that even if objective moral properties are reducible to non-moral properties, naturalists still must appeal to simplicity in order to justify belief in such properties. If I am right, then establishing the reducibility of moral properties to non-moral properties is of no use to a naturalist hoping to resist the overall conclusion of this paper.

The basic problem is just this: Demonstrating reducibility is not the same as demonstrating the truth of a particular reduction. Plausibly, one can demonstrate reducibility simply by showing that *if* we take moral realism for granted, and if we take for granted various assumptions about what non-moral properties are objectively good or bad, or about what non-moral states of affairs are objectively rational to promote or to avoid, then moral properties will be identical with or composed of the members of a certain class of non-moral properties. Demonstrating the truth of a particular reduction, however, requires one to demonstrate, in addition, the truth of moral realism and the correctness of one's various assumptions about what non-moral properties are objectively good or bad and about what non-moral states of affairs are objectively rational to promote or to avoid. Thus, even if we are presented with a perfectly compelling argument for the conclusion

that objective moral properties are reducible to non-moral properties, we are still left with the question of why we should believe that there are any objective moral properties. And here we are returned to the pair of options sketched in section 2.2: Assuming we are naturalists, we either posit moral properties as non-redundant causal explainers of natural phenomena (an option hardly worth taking seriously) or we presuppose their existence as a way of simplifying our theorizing.

To illustrate this problem, let me briefly sketch one well-known attempt to reduce moral properties to non-moral properties. In "Moral Realism" (Railton 1986), Peter Railton argues that facts about moral rightness are reducible to facts about what about what an impartial hypothetical observer would approve of under conditions of ideal information. These counter-factual facts, in turn, are supposed to be reducible to purely descriptive facts about the nature of the society in question, it's particular circumstances, and so on. As a first step into the task, Railton begins by showing how the *non-moral good* of an individual agent can be reduced to facts about what a cognitively idealized version of the agent would desire for his or her unidealized self. Crucial to his account is the idea of an agent's *objectified subjective interest*. Railton introduces that idea as follows:

> Give to an actual individual *A* unqualified cognitive and imaginative powers, and full factual and nomological information about his physical and psychological constitution, capacities, circumstances, history, and so on. *A* will have become *A*+, who has complete and vivid knowledge of himself and his environment, and whose instrumental rationality is in no way defective. We now ask *A*+ to tell us not what *he* currently wants, but what he would want his non-idealized self *A* to want—or, more generally, to seek—were he to find himself in the actual condition and circumstances of *A* (Railton 1986: 173-4).

What *A*+ would want *A* to want in *A*'s actual condition and circumstances is what is in *A*'s objectified subjective interest. By way of example, Railton invites us to consider a man who is dehydrated in the desert and finds himself desiring a glass of milk. In fact, a glass of water would be much better for the man from the point of view of improving his health; and, intuitively, it seems that a glass of water is what is objectively in his best interests (assuming, anyway, that he wants to survive and be healthy). Railton's account accommodates this intuition. On the assumption that the man desires to survive and be healthy, it turns out that drinking water is in the man's objectified subjective interest, since that is clearly what a cognitively idealized version of the man would desire his non-idealized self to desire in the man's actual condition and circumstances of dehydration. What is in a person's *objective interest* to do is just what he has an

objectified subjective interest in doing; and the *non-moral good* for a person is to do what it is in his objective interest to do. Moreover, the fact that it is in a person *A*'s objective interest to do something is supposed to supervene on "those facts about *A* and his circumstances that *A*+ would combine with his general knowledge in arriving at his views about what he would want to want were he to step into *A*'s shoes" (174-5). Thus, Railton's view rightly yields the judgment that it is objectively non-morally good for the dehydrated man to drink water even though he actually wants to drink milk; and, plausibly, these facts about the man's non-moral good supervene on purely descriptive, non-normative facts.[16]

From here, the account of moral rightness unfolds roughly as follows. Moral rightness is understood as rationality from a social point of view; rationality is understood as the pursuit of what it is in one's objective interests to do; and so *social* rationality is understood as pursuit of whatever is in the objective interests of society. Furthermore, the objective interests of society are characterized in a way analogous to the characterization of the objective interests of an individual: again, roughly, those interests are whatever would be approved of by an impartial observer under conditions of ideal information. Of course, one's own objective interests might not coincide with society's; but, Railton says, facts about social rationality can still ground ought claims that apply to individuals because the social point of view "includes but is not exhausted by" the individual's (1986: 201). Moreover, these ought claims will satisfy the two conditions I identified as necessary for objectivity since they are, in the relevant sense, theory-independent.

We may note in passing that, even if Railton's account thus far is true, it is not at all clear that it implies that moral facts are genuinely reducible to non-moral facts.[17] The reason is that it is not clear what non-moral facts are supposed to determine the desire structure of the hypothetical observer; hence, it is not clear what facts determine the relevant hypothetical reactions of approval and disapproval. In the case of an individual agent, Railton invites us to suppose that the desire structure of the agent's idealized self depends importantly upon the agent's *actual* desire structure. And we can see how the dependence would go: take that initial desire structure, and then suppose that it remains generally intact in the agent's idealized self except for whatever modifications would be induced by improving the agent's cognitive abilities and information base in the ways suggested. One might reasonably doubt that there are any facts about what modifications *would* be induced in an agent's desire structure by making the requisite cognitive improvements.[18] But even if there are such facts, the point is that in the case of social rationality, a story analogous to this one about how the hypothetical observer's desire structure is to be determined seems impossible to tell.

We might suppose that the hypothetical observer's desire structure would depend *in some way* upon the actual goals and desires of individual agents; but it is not at all clear how the dependence would go.

Let us leave this worry aside, however, and let us simply concede that Railton's account has shown us how moral facts might be reducible to non-moral facts. Still, Railton's account crucially depends on the assumption that one in some sense ought to act in accord with social rationality and that one ought to do what it is in one's objective interest (as defined by Railton) to do. Granted, we can see why, given a certain set of interests and desires, it would be *attractive* or *efficient* or *useful* to act in these ways, and that various tangible benefits would be produced by so acting. But Railton's reduction of non-moral goodness and moral rightness does not justify the claim that one objectively ought to pursue one's non-moral good and that one objectively ought to do what is morally right. As Railton himself points out, his defense of moral realism *presupposes* a particular understanding of morality and of rationality; and what he has shown is that *if* morality and rationality are to be understood in that way, then objective moral properties are reducible to non-moral properties. But what he has not shown (and has not purported to show) is that the methods of science do, or even could, reveal that morality and rationality are to be understood in the way that he understands them. In other words, Railton has shown, at best, that if there are objective moral properties, and if his assumptions about what non-moral states of affairs are objectively rational to pursue are correct, then objective moral facts are reducible to the sorts of facts he has described. He has not shown that his reduction is *true*.

One might think that we could go some distance toward showing that a particular reduction is true if we could show that the reduction in question has correctly identified non-moral properties (or clusters of properties) that are tracked by our actual use of the terms 'morally good' and 'morally right'.[19] But even if we could show this, we would still not have enough to show how belief in objective moral properties is justified. Consider the following two premises:

(1) If there are objective moral properties, and if theory T of the nature of morality, rationality, and related notions is correct, then moral properties are identical with or composed of natural properties N_1 – N_n.

(2) Our uses of words that allegedly refer to moral properties reliably track N_1 – N_n.

Perhaps some interesting conclusions follow from these premises. But clearly the conclusion that there are objective moral properties does not

follow from the premises.[20] Thus, even if C2 is true, and even if it can be shown that a particular reduction has correctly identified natural properties tracked by our moral terms, there is still work for a naturalist to do in showing how belief in objective moral properties could be justified by the methods of science. And, for precisely the reasons laid out in section 2.2, it seems that the only plausible stories to be told here are ones according to which belief in moral properties depends for its justification on considerations of theoretical simplicity.

3. PRAGMATIC ARGUMENTS

In Section 2, I argued that any naturalistically respectable argument for belief in objective moral properties will have to appeal to simplicity. In this section, I'll argue that appeals to simplicity justify belief in moral properties only if moral properties are not objective or something like theism is true.

Some philosophers make a distinction between pragmatic and epistemic justification. The distinction between the two parallels the distinction between pragmatic and epistemic rationality—i.e., the distinction between what is rational to do given the goal of furthering one's overall best interests and what is rational to believe in light of one's evidence given the goal of believing in accord with the truth. It is epistemic justification that we're interested in here. And the initially pressing question is whether an argument that invokes considerations of simplicity as reasons for belief can provide epistemic justification for its conclusion.

For reasons I won't get into here, I'm inclined to think that one is automatically epistemically justified in believing things that are sanctioned by sources of evidence that one treats as basic.[21] Insofar as naturalists treat the methods of science as basic sources of evidence, and insofar as simplicity considerations are (apparently, anyway) routinely invoked as reasons for belief in the natural sciences, I am prepared to assume for the sake of argument that naturalists are epistemically justified in believing propositions that are supported by appeals to simplicity (especially those that figure in inferences to the best explanation or the method of reflective equilibrium). If this assumption is false, then my ultimate conclusion follows directly: naturalists are not epistemically justified in believing propositions supported (only) by arguments that appeal to simplicity; from a naturalistic point of view, belief in objective moral properties is sanctioned (if at all) only by arguments that appeal to simplicity; therefore, naturalists cannot reasonably accept commonsense moral realism.[22] Thus, the initially pressing question—whether one can be epistemically justified in believing something partly on the basis of an appeal to simplicity—is resolved by stipulation.

But once the stipulation is granted, we are committed to thinking that there is some connection between simplicity and truth. The reason is that arguments appealing to simplicity can yield epistemic justification only if believing propositions on the basis of such arguments is a reliable way of believing in accord with the truth.[23] Let us suppose, then, that simplicity is somehow a reliable indicator of truth. The pressing question now is: What would be the best explanation for this fact?

One interesting suggestion that I'll set aside is that our preference for simplicity is just a disguised preference for truth. According to Richard Boyd (1980, 1985), for example, what often get described as considerations of simplicity are really nothing more than manifestations of a preference for theories that are relatively "simple" modifications of existing, evidentially supported theories. Thus, given that our existing theories are at least approximately true, the preference for simplicity turns out, on this view, to be little more than a preference for (approximate) truth.

There is a lot that is worth exploring in this view, but for now I'll simply observe that adopting it leaves the naturalist no better off with respect to belief in objective moral properties than I have so far taken her to be. Suppose we grant that "existing moral theory" (whatever exactly that would be) is approximately true. The fact is, this might be so whether or not there are objective moral properties, and whether or not existing moral theory quantifies over objective moral properties. Now, if Boyd's understanding of simplicity is correct, then one who believes in objective moral properties on the basis of such considerations believes in them either because so doing represents a simple modification of an existing theory, or because their existence is already implied by an existing theory. In light of the arguments of Section 2, it is hard to see what reason a naturalist could ever have for modifying an existing theory so that it quantifies over objective moral properties. An appeal to simplicity is ruled out because, on Boyd's view, that's not a reason for modifying a theory; it's a reason for preferring one modification rather than another. But the point of Section 2 was to show that, from a naturalistic point of view, there aren't any (evidential) considerations apart from simplicity that would lead one to posit objective moral properties. Thus, if Boyd's understanding of simplicity is right, then if existing moral theory quantifies over objective moral properties, it does so for no reason at all, or it does so simply because existing moral theory has always quantified over such properties. Thus, if his view is right, it looks as if a naturalist's belief in objective moral properties is either ungrounded or grounded simply in the fact that such belief is and always has been prescribed by existing moral theories. But even if we grant that believing something simply because you (or others) always have believed it is a reliable way of reaching the truth, nothing in Boyd's view explains *why* this should be a reliable way of reaching the truth.

It's easy to see how a preference for existing theories can reliably lead us to approximate truth in the selection of new theories, given that our existing theories are already approximately true. But it remains a mystery how believing a *specific proposition*—such as the proposition that there are objective moral properties—simply because you and others have always believed it should be a reliable way of reaching the truth. And I take it that any resolution of this mystery will roughly parallel answers to the more general question at issue here—namely, the question of what would explain the fact that simplicity considerations *as I understand them* are generally truth-indicative.

So what would explain the fact that simplicity is truth-indicative? One possibility is that someone or something in the universe is somehow benevolently guaranteeing that it will be. This, clearly enough, is in the neighborhood of theism. Another possibility is that a pragmatic theory of truth is correct: truth is, roughly, acceptability or assertibility under ideal conditions, where "ideal conditions" are spelled out partly in terms of simplicity considerations. A third possibility, *constructivism*, is that we *make it the case* that our theories are true by conceptualizing the world in whatever way we do.[24] Thus, so long as we conceptualize the world in a way that is empirically adequate (as our scientific theories aim to do) there is no real question whether the ontological commitments we thereby incur will be true.[25] On this view, simplicity isn't really an *indicator* of truth (truth is guaranteed by empirical adequacy); rather, it is just a constraint that happens to govern our theorizing.

It is hard to imagine (plausible) explanations other than these for why simplicity would be a reliable indicator of truth.[26] Of course, one can't infer much from a mere failure of imagination. But if, upon reflection, we simply can't see why theoretical virtues that we take to be truth-indicative should be truth-indicative, it is hard to see how we can be justified in continuing to treat them as truth-indicative. Thus, assuming it is non-negotiable for naturalists to continue treating simplicity as a reliable indicator of truth, and assuming that they (like me) have no other plausible story to tell about why it ought to be a reliable indicator of truth, it seems that the only reasonable option is to embrace one of the above three alternatives. As a theist, I am sympathetic to the first. Moreover, the second (as I shall argue) implies something very much like theism. Thus, on the assumption that the methods of science do not by themselves justify belief in God, or even belief in something very much like God, naturalists are committed to the third alternative. In what follows, I'll first explain why accepting constructivism commits one to the conclusion that moral properties are not objective. I'll then go on to argue that pragmatic theories of truth imply something very much like theism.

To see why constructivism requires us to give up the objectivity of moral properties, we must first get a clearer grasp on what the position amounts to. At first blush, it might seem to be incoherent. It is, after all, rather hard to see how we could accomplish the creative feats that constructivism seems to require. How could we make it the case that there are stars, or planets, or human organisms simply by theorizing about the world in a way that quantifies over stars, planets, and human organisms? More pressingly, how could we—by using our minds—make it the case that there are minds? These are serious questions; but I think that constructivists can provide answers, and a brief look at those answers will help to clarify the position as I understand it.

The second question can be treated quickly. As I see it, constructivists must simply deny that we make it the case that there are minds; thus, they must deny that minds are part of the material world that is constructed by our theories.[27] If this is right, then constructivists are committed to substance dualism. This is surely an interesting (and probably generally unwelcome) consequence; but it is not a refutation, and embracing it enables the constructivist to avoid the charge of incoherence.

The first question is more complicated. In response to it, I think that constructivists should articulate what many take to be a Kantian view of the world. Roughly, that view is as follows. None of the properties that appear to be sortal properties of non-abstract, non-mental objects are intrinsic to anything. Properties like *being an electron*, *being a horse*, *being a star*, *being a human organism*, and so on are all extrinsic. Notoriously, it is hard to say exactly how such properties could be extrinsic. The most intelligible versions of constructivism typically make it clear that the reason they are extrinsic is that whether they are exemplified depends importantly upon relations obtaining between our minds and the mind-independent world (i.e., whatever thing or things of a wholly unidentifiable sort exist independently of our minds).[28] Moreover, they make it clear that those relations involve, at least in part, our conceiving of the world in the ways that we do. But beyond this, it is hard to say exactly what the relevant relations consist in.

Be that as it may, some analogies may help to clarify the position a bit further. Consider some other properties that are often, even if not universally, regarded as "being in the eye of the beholder": properties like *being a work of art*, or *being a thing of great beauty*. The constructivist might say that, just as the matter in a region of spacetime counts as a work of art or a thing of great beauty only if we (or the members of some relevant group) think of it as a work of art or a thing of great beauty, so too whether the matter in a region of spacetime counts as a star, or a planet, or a human organism, depends upon our thinking of it as a star, or planet, or human organism. Likewise, she might say, just as there would be no art, or nothing

beautiful, if we regarded nothing as art or as beautiful, so too there would be no stars if we regarded nothing as a star. There would, of course, still be the stuff that causes our star-like sensations. That stuff is part of the mind-independent world.[29] But apart from our belief-forming activities, that stuff would not constitute a star.

Even with these analogies on hand, we are still a far cry from having answered all of the questions one might have about the intelligibility of constructivism. But we at least have enough of a picture to see clearly why *moral* constructivism is incompatible with commonsense moral realism. Quite simply, constructivism implies that goodness, like beauty or art, is in the eye of the beholder. Admittedly, matters will probably be a bit more complicated than this. Constructivism is, for example, compatible with the view that what's good is what the members of some salient majority take to be good, or what our most pragmatically virtuous theories identify as good. But regardless of the details, any constructivist theory will, by its very nature, make facts about goodness dependent upon our beliefs about goodness. Thus, a constructivist account of goodness will not be an account according to which goodness is a theory-independent property; hence, it will not be an account according to which goodness is an objective property; hence, it will be incompatible with commonsense moral realism.

All that remains, then, is to deliver on my claim that pragmatic theories of truth imply something like theism. I have defended this conclusion at length elsewhere (Rea 2000, Rea 2002), so for present purposes I will only provide a brief sketch.

My argument draws its inspiration from Alvin Plantinga's 1982 Presidential Address to the American Philosophical Association. In that address, Plantinga argues that a thesis about truth which he attributes to Hilary Putnam implies that, necessarily, there exists an ideally rational community. The Putnamian thesis about truth is as follows:

(HP) Necessarily: *p* is true ≡ if there were an Ideally Rational Scientific Community (IRS) that had all of the relevant evidence, it would accept *p*.

In short, Plantinga points out that, by substitution, we can easily obtain HP1:

(HP1) Necessarily: it is true that there is an IRS ≡ if there were an IRS that had all of the relevant evidence, it would accept that there is an IRS.

But, of course, it is eminently plausible that an IRS possessed of "all the relevant evidence" would accept the conclusion that there is an IRS. Thus, HP1 implies the "dismal conclusion" that, necessarily, there exists an IRS.

HP is what we might call an *epistemic truth equivalence* (or "ETE" for short). An ETE is any claim that asserts that there is a necessary equivalence between what is true and what would be believed by a rational agent or community of agents under certain specified conditions. More exactly, an ETE is any thesis that conforms to the following schema:

(E) Necessarily: *p* is true ≡ if there were a rational community that satisfied condition C with respect to *p*, then there would be a rational community that both satisfies condition C with respect to *p* and accepts *p*.

By 'rational community', I just mean 'a being or group of beings capable of thought and reasoning'. 'Condition C' refers to what we might call "the acceptance condition". It is a schematic term that takes as substitution instances descriptions of the conditions that must be satisfied by a rational community in order for its acceptance of *p* to be necessary and sufficient for the truth of *p*. The "with respect to *p*" qualifier is added to take account of the fact that what counts as satisfying the acceptance condition might vary from proposition to proposition. Such would be the case if, for example, the acceptance condition is satisfied only if the community in question possesses all *and only* the evidence relevant to *p*.

The first premise in my argument for the conclusion that pragmatic theories of truth imply something like theism is that pragmatic theories of truth entail epistemic truth equivalences. Below are some representative examples of claims that might be taken to express pragmatic theories of truth:

> "True ideas are those that we can validate, corroborate, and verify" (James 1907: 142).

> "The opinion which is fated to be ultimately agreed to by all who investigate, is what we mean by the truth…" (Peirce 1960: 407).

> "[T]ruth is an *idealization* of rational acceptability. We speak as if there were such things as epistemically ideal conditions, and we call a statement 'true' if it would be justified under such conditions" (Putnam 1981: 55).

> Truth is superassertibility, or "assertibility which would be durable under any possible improvement to one's state of information" (Wright 1992: 75).

Pretty obviously, each of these claims taken as a theory of truth is equivalent to a thesis that satisfies schema (E). Granted, one might argue (quite

convincingly in some cases) that these authors did not *really* mean to be giving a *theory* about what truth is. But each of these views is such that *if* it were a theory of truth, it would clearly be a pragmatic theory *and* it would clearly imply an ETE. Moreover, I see no way in which a theory of truth could plausibly count as pragmatic without implying an ETE; for what makes a theory of truth distinctively pragmatic is just its having as a consequence the claim that truth is importantly tied to what is useful (in some sense) for humans to believe.

Of course, there are theses that imply that truth is importantly tied to what is useful for humans to believe but that do not imply ETE's. For example:

(W) Were P to be appraised under (constructively specified) sufficiently good epistemic conditions, P would be true if and only if P would be believed (Wright 2000: 350).

As Crispin Wright points out, conditionals like W serve to constrain the notion of truth in the ways that pragmatists typically want; and, importantly, they do not suffer from many of the problems that plague ordinary ETE's. But, though W-style conditionals surely say something interesting and important about truth, they are not *theories* of truth. A genuine theory of truth will offer or imply, at the very least, a necessary equivalence of the form 'Necessarily, *p* is true ≡ _____'. And, again, it is hard to see how any such equivalence could constitute a *pragmatic* theory of truth without being or entailing an ETE.

The second premise in the argument is that every ETE implies something like theism. To establish this conclusion, I need two assumptions. The first is that it is possible that there are no contingent beings. The second is as follows:

(SC) For any true ETE: Let C be its acceptance condition and let α be the following proposition:

(α) There exists a rational community S such that, for every proposition *p*, S satisfies C for either *p* or the denial of *p*.

Then: Necessarily, if there is a rational community that satisfies C with respect to α, then α is true.

The first assumption isn't wholly uncontroversial; but I assume it will be granted by most naturalists. After all, naturalism typically (though not necessarily) goes hand-in-hand with atheism, and atheists are typically prepared to admit that there might have been nothing at all. Regarding SC,

the idea is roughly just that only a being ideally situated with respect to every proposition would be in an *ideal* position to evaluate a proposition like α. A less-than-ideally situated being (e.g., a being very much like one of us) might have less-than-ideal evidence for either α or its denial. But having ideal evidence in favor of α would guarantee it's truth (since ideal evidence must be infallible), and having ideal evidence against α seems to be impossible (since, plausibly, only a being ideally situated with respect to every proposition could infallibly rule out the truth of something like α).

Given these two assumptions, the second premise can be defended as follows. Let EC below be any *true* ETE (if such there be); let C be EC's acceptance condition; let α, β, and γ be propositions as follows:

(α) There exists a rational community S such that, for every proposition p, S satisfies C with respect to either *p* or its denial.
(β) There exists a rational community that satisfies C with respect to α.
(γ) There exists a rational community that both satisfies C with respect to α and accepts α.

We then have:

(EC) Necessarily: *p* is true ≡ if there were a rational community that satisfied condition C with respect to *p*, then there would be a rational community that both satisfies condition C with respect to *p* and accepts *p*. (Premise)

(6.1) Necessarily: α is true ≡ if β were the case then γ would be the case. (From EC, by substitution)
(6.2) $\beta \Rightarrow \alpha$ (From SC)
(6.3) Necessarily: α. (From 6.1, 6.2)

6.1 and 6.2 together entail 6.3 on the assumption that the correct modal system is S4 or stronger and that the correct semantics for counterfactuals guarantees that (i) a counterfactual conditional implies its corresponding material conditional, and (ii) a strict conditional implies its corresponding counterfactual conditional.[30] But from 6.3, it is a short step to the conclusion that, necessarily, there exists an omniscient community. 6.3 implies that it is necessarily true that there exists a rational community S such that, for every proposition *p*, S satisfies C with respect to either *p* or its denial. But this in conjunction with EC implies that it is necessarily true that, for every *true* proposition *p**, there is a rational community that both satisfies C with respect to *p** and accepts *p**. Hence, it follows that, necessarily, there is a rational community that accepts a proposition that tells the whole truth about

whatever world is actual.[31] Thus, necessarily, there exists an omniscient community. Moreover, if one is willing to grant that the correct modal system is S5, then 6.3 implies that there exists a necessarily existing rational community. Again, 6.3 implies that it is necessarily true that there exists a rational community; but, on the assumption that it is possible that there be no contingently existing beings, it follows that there is a possible world *w* that contains a rational community but no contingently existing rational beings. Thus, *w* must contain a necessarily existing rational community. But this implies that there in fact exists a necessarily existing rational community.[32]

If this argument is sound, then pragmatic theories of truth entail that (a) necessarily there exists an omniscient community, and (b) there exists a necessarily existing rational community. This isn't quite theism, but it is close. Theists, of course, will not be bothered by this conclusion, for their view already entails it and (typically) is motivated by considerations independent of a commitment to an epistemic account of truth. Naturalists, on the other hand, ought simply to reject epistemic accounts of truth; hence, they ought also to reject pragmatic theories of truth. However, as we have already seen, naturalists who reject a pragmatic theory of truth must either embrace theism or give up belief in objective moral properties. Assuming, as I have been, that belief in God is not justified by the methods of science, the first alternative is unavailable (short of giving up naturalism). Thus, we reach the main conclusion of this paper: naturalists must give up moral realism.

ENDNOTES

* I am grateful to Michael Bergmann, Jeff Brower, David Haslett, Trenton Merricks, Christian Miller, Mark Murphy, and Alvin Plantinga for generous and very helpful comments on earlier versions of this paper.

1 I take it that, even if the present understanding of moral realism is controversial, it is not idiosyncratic. (Cf. Boyd 1988, Brink 1989, Railton 1986, and Smith 1994.)

2 Thus, conditions (i) and (ii) provide necessary, but not sufficient, conditions for objectivity. To get in the neighborhood of a sufficient condition, we would have to add that the exemplification of a moral property *p* is in some relevant sense independent of actual human desires and attitudes.

3 For further, more detailed discussion of simplicity as I am understanding it here, see Koons 2000, Swinburne 2001: Ch. 4, and Weinberg 1994: Ch. 6.

4 See, e.g., Armstrong 2002: 35 and Danto 1967: 448.

5 See, e.g., Quine 1995: 257 and Devitt 1998: 45.

6 See, e.g., Leiter 1998: 81.

7 Indeed, giving into this temptation is now the standard way of characterizing naturalism. See, e.g., Schmitt 1995, Hampton 1998: 19-21, Katz 1998: xii.

[8] For a fuller defense of these claims, a fuller explanation of the term 'research program', a more thorough argument for the conclusion that naturalism is not a thesis, and for references to Dewey, Quine, and other prominent naturalists, see Part 1 of Rea 2002.

[9] But here we must add a caveat. Though it is surely right to say that rational intuition isn't *generally* treated as a source of evidence in science, there *might* be a case to be made for the conclusion that it is treated as a source of evidence in the ill-defined domain of mathematical, logical and conceptual truths. (See Rea 2002: 67, 199-210 for further discussion.) But even if this is right, it does not affect the present discussion; for moral truths clearly aren't mathematical or logical truths, and the phenomenon of widespread intractable disagreement is just one among several convincing pieces of evidence that they aren't sufficiently similar to paradigm cases of conceptual truths (e.g., 'All bachelors are male') to be treated as such.

[10] Note that '*x* because *y*' isn't equivalent to, and does not entail '*x* is a cause of *y*'. We might say 'The vase broke because it was fragile'; but in saying this we don't commit ourselves to the claim that the fragility of the vase *caused* its breaking.

[11] Scientific reasons for rejecting P4 would just be scientific reasons for believing that there are non-scientific sources of evidence (e.g., clairvoyance, rational intuition, etc.).

[12] For the most part, but not entirely. See, e.g., Sayre-McCord 1988b.

[13] As I'll note in Section 3, Boyd understands the notion of simplicity in a way different from the way I am understanding it here. But I'll also argue that understanding simplicity in his way won't help a naturalist to avoid the conclusion that I am defending.

[14] I also assume that, so far as *scientific* justification is concerned, there is no distinction to be drawn between inference to the best explanation and inference to the best *causal* explanation. Hence, I'll drop the qualifier here and, for the most part, in what follows.

[15] Cf. Koons 2000, Lipton 1991, and Swinburne 2001: Ch. 4. Koons 2000 argues that, because simplicity considerations play such an important role in scientific justification, naturalists cannot accommodate *scientific* realism. This is a conclusion that I am inclined to agree with, but adding to Koons's defense is not my purpose here.

[16] Railton's account of an agent's non-moral good is similar to the account of normative reasons offered in Smith 1994. Smith, however, does not take himself to be offering a fully reductive analysis of normative reasons. As he himself points out, normative concepts are employed in spelling out what it means for S to have a normative reason to φ (162).

[17] For the record, I do not believe that Railton's account thus far is true. The most compelling problem is that his account is unable to accommodate the fact that it might be in a person's objective interest to desire something but not to have it. Suppose it is a fact about Kevin that if he were to desire to go to medical school, he would embark upon a course of action that would very probably not result in his actually going to medical school but would result in his achieving something else that is very satisfying for himself (perhaps a career as a science teacher or some such thing). Suppose furthermore that if he were actually to go to medical school, he would be absolutely miserable. We may assume that Kevin himself does not know these facts, but that Kevin+ would know them. What then would Kevin+ desire to desire were he in the actual condition and circumstance of Kevin? Pretty obviously, he would desire to desire to go to medical school. But according to Railton's account, it does not follow from this that *desiring* to go to medical school is in Kevin's objective interest, which is true; rather, it follows that *going* to medical school is in Kevin's objective interest, which is false.

[18] As Mark Murphy (1999: 261-265) argues, there is also reason to doubt (a) whether such modifications would all count as improvements in the agent's desire structure, and (b) whether there's any good reason to think that the hypothetical second-order desires of an agent's cognitively idealized self are any more authoritative with respect to the agent's well-being than the agent's actual second-order desires.

[19] Boyd (1988) presses this point in his own attempt to show that moral properties are reducible to non-moral properties.

[20] A somewhat related point is made by Robert Adams (1999: 77-8).

[21] I defend this claim in Chapter 1 of Rea 2002.

[22] I assume that one can reasonably accept only what one is epistemically justified in believing. But this is just a terminological point—a point about how I am here proposing to use the word 'reasonably'.

[23] Or so I assume. But I acknowledge that the assumption is controversial.

[24] Here I am not using the term 'constructivism' in the way that Rawls (1980) does. Rather, the view I have in mind is primarily a view about ontology, and it often goes by labels like *conventionalism*, *(global) anti-realism*, *Kantian idealism*, and so on (though somewhat different views go by those labels too).

[25] It is, perhaps, tempting to conflate the third possibility with the second. But we can avoid the temptation if we attend to the fact that constructivism, insofar as it is coherent, is compatible with deflationism about truth—a rejection of more substantive theories of truth in favor of the view that Tarksi's T-schema says all there is to say about truth. For more on constructivism, see Chapter 1 of Rea 2002. For detailed arguments for the conclusion that constructivism does not imply a pragmatic theory of truth, see Alston 1996: Ch. 6.

[26] Koons (2000) discusses a suggestion by David Papineau and Ruth Millikan to the effect that perhaps evolutionary processes have "taught" us that there is a correlation between (e.g.,) simplicity and truth. Weinberg (1994) makes a similar suggestion. But, as Koons points out, accidental correlation isn't sufficient for reliable indication. The laws *might* have been complex; indeed, for all we presently know, the actual laws might (unexpectedly) *in fact be* complex. After all, we don't yet have the much sought after "final theory". So even if the Papineau-Millikan-Weinberg suggestion is true, it remains hard to see what would give us grounds for thinking that virtues like simplicity are reliable indicators of truth.

[27] I defend this conclusion in detail in Chapter 7 of Rea 2002.

[28] The thing or things belonging to the world as it is in itself must be of an *unidentifiable* sort because the constructivist's thesis is that all of the sortal properties we are familiar with are extrinsic; but if the thing(s) belonging to the world in itself is (are) to be truly mind-independent, it (they) must have its (their) sortal properties intrinsically.

[29] Note that 'stuff' is not being treated here as a sortal term. There is, in other words, no object kind (or even a particular stuff-kind) that is referred to by the word 'stuff'. (If there are stuff-kinds, then *stuff* is just whatever it is that stuff-sortal terms sort.)

[30] For proof, see Rea 2000: 296 or Rea 2002: 152.

[31] Or, if there is no such proposition, then at least this much follows: necessarily, for any true proposition that *approximates* telling the whole truth about the world, there is a rational community that accepts it.

[32] Here are the steps: Let W be a world with no contingently existing rational beings and let E_1 - E_n be the members of the rational community that exists in W. We then have:

(1) $\Diamond\Box P \Rightarrow \Box P$

(2) $\Diamond\Box$ (E_1 – E_n exist)

(3) Therefore: $\Box$ (E_1 - E_n exist).

REFERENCES

Adams, Robert. 1999. *Finite and Infinite Goods*. New York: Oxford University Press.

Alston, William. 1996. *A Realist Conception of Truth*. Ithaca, NY: Cornell University Press.

Armstrong, David. 2002. Naturalism, materialism, and first philosophy. In *Contemporary Materialism*, eds. Paul Moser and J. D. Trout, 35-46. New York: Routledge.

Boyd, Richard. 1988. How to be a moral realist. In *Essays on Moral Realism*, ed. Geoffrey Sayre-McCord, 181-228. Ithaca, NY: Cornell University Press.

Boyd, Richard. 1985. Observation explanatory power, and simplicity. In *Observation, Experiment, and Hypothesis in Modern Physical Science*, eds. P. Achinstein and O. Hannaway, 47-94. Cambridge, MA: MIT Press.

Boyd, Richard. 1981. Scientific realism and naturalistic epistemology. In *PSA 1980*, vol. 2, eds. Peter Asquith and Ronald Giere, 613-662. East Lansing, MI: Philosophy of Science Association.

Brink, David. 1989. *Moral Realism and the Foundations of Ethics*. Cambridge: Cambridge University Press.

Copp, David and David Zimmerman, eds. 1985. *Morality, Reason, and Truth*. Totowa, NJ: Rowman & Allanheld.

Craig, William Lane and J. P. Moreland, eds. 2000. *Naturalism: A Critical Approach*. London: Routledge.

Danto, Arthur. 1967. Naturalism. In *The Encyclopedia of Philosophy*, vol. v, ed. Paul Edwards, 448-450. New York: MacMillan and Free Press.

Devitt, Michael. 1998. Naturalism and the a priori. *Philosophical Studies* 92: 45-65.

Hampton, Jean. 1998. *The Authority of Reason*. Cambridge: Cambridge University Press.

Harman, Gilbert. 1977. *The Nature of Morality*. New York: Oxford University Press.

Harman, Gilbert. 1985. Is there a single true morality? In *Morality, Reason, and Truth*, eds. David Copp and David Zimmerman, 27-48. Totowa, NJ: Rowman & Allanheld.

Harman, Gilbert. 1986. Moral explanations of natural facts—can moral claims be tested against moral reality? *Southern Journal of Philosophy* 24 (supp.): 57-68.

Jackson, Frank. 1988. *From Metaphysics to Ethics*. Oxford: Oxford University Press.

Jackson, Frank and Philip Pettit. 1995. Moral functionalism and moral motivation. *Philosophical Quarterly* 45: 20-40.

James, William. 1907. *Pragmatism, A New Name for Some Old Ways of Thinking*. New York: Longmans, Green, and Co.

Katz, Jerrold. 1998. *Realistic Rationalism*. New York: MIT Press.

Koons, Robert. 2000. The incompatibility of naturalism and scientific realism. In *Naturalism: A Critical Approach*, eds. William Lane Craig and J. P. Moreland, 49-63. London: Routledge.

Leiter, Brian. 1998. Naturalism and naturalized jurisprudence. In *Analyzing Law: New Essays in Legal Theory*, ed. Brian Bix, 79-104. Oxford: Clarendon Press.

Lipton, Peter. 1991. *Inference to the Best Explanation*. London: Routledge.

Murphy, Mark. 1999. The simple desire-fulfillment theory. *Noûs* 33: 247-272.

Peirce, Charles Sanders. 1960. How to make our ideas clear. In *Collected Papers of Charles Sanders Peirce*, Vol 5., eds. Charles Hartshorne and Paul Weiss. Cambridge, MA: Harvard University Press.

Plantinga, Alvin. 1982. How to be an anti-realist. *Proceedings and Addresses of the American Philosophical Association*, September 1982.

Putnam, Hilary. 1989. *Reason, Truth, and History*. Cambridge: Cambridge University Press.

Quine, W. V. 1995. Naturalism; or, living within one's means. *Dialectica* 49: 251-262.

Railton, Peter. 1986. Moral realism. *Philosophical Review* 95: 163-207.
Rawls, John. 1980. Kantian constructivism in moral theory. *Journal of Philosophy* 77: 515-572.
Rea, Michael. 1998. Naturalism and material objects. In *Naturalism: A Critical Approach*, eds. William Lane Craig and J. P. Moreland, 110-132. London: Routledge.
Rea, Michael. 2000. Theism and epistemic truth equivalences. *Noûs* 34: 291-301.
Rea, Michael. 2002. *World Without Design: The Ontological Consequences of Naturalism*. Oxford: Clarendon Press.
Sayre-McCord, Geoffrey, ed. 1988a. *Essays on Moral Realism*. Ithaca, NY: Cornell University Press.
Sayre-McCord, Geoffrey. 1988b. Moral theory and explanatory impotence. In *Essays on Moral Realism*, ed. Geoffrey Sayre-McCord, 256-81. Ithaca, NY: Cornell University Press.
Schmitt, Frederick. 1995. Naturalism. In *Companion to Metaphysics*, eds. Jaegwon Kim and Ernest Sosa, 343-345. Oxford: Basil Blackwell.
Sellars, Wilfrid. 1963. Empiricism and the philosophy of mind. In *Science, Perception*, *and Reality*, 127-196. London: Routledge & Kegan Paul.
Smith, Michael. 1994. *The Moral Problem*. Oxford: Basil Blackwell.
Sturgeon, Nicholas. 1985. Moral explanations. In *Morality, Reason, and Truth*, eds. David Copp and David Zimmerman, 49-78. Totowa, NJ: Rowman & Allanheld.
Sturgeon, Nicholas. 1986. Harman on moral explanations of natural facts. *Southern Journal of Philosophy* 24 (supp.): 69-78.
Swinburne, Richard. 2001. *Epistemic Justification*. Oxford: Oxford University Press.
Van Fraassen, Bas. 1989. *Laws and Symmetry*. Oxford: Clarendon Press.
Weinberg, Steven. 1992. *Dreams of a Final Theory*. New York: VintageBooks.
Wright, Crispin. 1992. *Truth and Objectivity*. Cambridge, MA: Harvard University Press.

Chapter 11

A PROBLEM WITH BAYESIAN CONDITIONALIZATION

Richard Otte
University of California, Santa Cruz

1. INTRODUCTION

Bayesianism is often divided into a static and dynamic part. The static part is a theory of rational belief at an instant, and the dynamic part is a theory of rational belief change. In this paper I will begin by looking closely at the rule of conditionalization, which is central to the dynamic side of Bayesianism. I will argue that the dynamic side of Bayesianism needs to make distinctions central to traditional non-Bayesian epistemology; Bayesians may consider this unfortunate, since Bayesianism is supposed to be a way to avoid many of the problems that arise in classical epistemology. I will then give a counter-example that shows we can be rational even if we do not conditionalize in situations in which conditionalization is applicable. This counterexample also raises problems for the view that conditionalization is a way to manage our beliefs in pursuit of some ideal. Contrary to Bayesian doctrine, I will argue that we generally cannot use conditionalization to manage our beliefs. The reason for this is that conditionalization is an externalist requirement that aims at the goal of being in a good epistemic situation, but rationality contains only internal requirements; thus conditionalization is not required in order to be rational. I will then look at whether we can develop an internalist version of

T. M. Crisp, M. Davidson and D. Vander Laan (eds.), Knowledge and Reality, 243-255.

conditionalization which does not have these problems. It turns out that although an internalist requirement of conditionalization can solve some problems, it faces other serious problems.

2. THE REQUIREMENT OF BAYESIAN CONDITIONALIZATION

If as a result of experience we come to believe E to degree 1, conditionalization tells us how to modify the rest of our beliefs to take account of learning E. The requirement of conditionalization is very simple: our new probability function, P_n, is obtained from our old probability function, P_o, by conditionalizing on our new evidence. If E represents our new evidence, we can write the requirement of conditionalization as: $P_n(*) = P_o(*/E)$. This requirement of conditionalization only applies when E is everything we have learned from experience and when our new probability of E is 1.

Although Bayesians often claim that all rational belief change is by conditionalization, upon closer inspection we find that things are not as clear as first appeared. Contrary to what is usually said, Bayesianism must admit at least two different ways in which belief may rationally be changed. This distinction is seldom stated; perhaps this is due to relying too much on generalizations such as 'all belief change is via conditionalization.' First, one may change belief in direct response to experience. Although this is rarely discussed in connection to Bayesianism, we must allow for the rationality of changing belief as a direct result of having some experience; otherwise there would be nothing to apply the rule of conditionalization to. Some beliefs may be completely based on experience, and other beliefs may be partly based on experience and partly based on other beliefs. What is important is that beliefs can be rationally changed as a direct result of experience, and not as a result of other changes in belief. There are no Bayesian requirements or rules that govern belief change directly based on experience; everything is permitted.

The second way one may rationally change belief is in response to some other change in belief. Bayesianism provides rules, either traditional conditionalization or Jeffrey's generalization of conditionalization, that govern this type of belief change. These rules presuppose a distinction between beliefs that are changed as a result of experience and beliefs that are changed only in response to some other belief change, and these rules only govern belief change that is not directly affected by experience. It may be that there is rational change of belief not based on either experience or other

belief change; memory, and a priori beliefs may be examples of these. If so, Bayesianism will also have to let them be the basis of rational belief change.

Since conditionalization only applies to beliefs that are changed in response to other beliefs being changed by experience, in order to state when conditionalization is applicable, we need to first distinguish beliefs that are changed in direct response to experience from beliefs that are changed in response to other beliefs. We can then give the requirement of conditionalization:

ROC: If E contains all and only those beliefs that have been directly affected by experience, and if $P_n(E) = 1$, then $P_n(*) = P_o(*/E)$.

Notice that conditionalization will not be applicable to situations in which some beliefs are at least partly based on experience and are not raised to degree 1. Richard Jeffrey developed a version of conditionalization that did not require what was learned on the basis of experience to be believed to degree 1. Jeffrey noted that learning experiences often directly result in partial beliefs. Jeffrey conditionalization is a generalization of ROC that tells us how to modify the rest of our beliefs in response to forming these partial beliefs directly on experience:

ROJC: If partition $E = \{E_1, E_2, ...\}$ contains all of the beliefs that are changed in direct response to experience, then $P_n(A) = \sum_i P_o(A/E)P_n(E_i)$.

As with traditional conditionalization, Jeffrey's probability kinematics does not apply to beliefs that are directly based on experience; no restrictions are placed on belief change based on experience. According to ROC and ROJC, our beliefs are required to be in accord with conditionalization only in very specific circumstances.

3. A COUNTEREXAMPLE TO CONDITIONALIZATION

Conditionalization is normally presented as a requirement or necessary condition of rationality. A change of belief in a situation in which conditionalization is applicable cannot be rational if it violates conditionalization. In the following I will focus my discussion on ROC, but it should be clear that the problems I raise also apply to ROJC.

A problem arises because there are situations in which conditionalization is applicable according to ROC, but in which it is rational to not conditionalize. Consider examples with the following structure. Suppose an agent has a certain experience, and as a result comes to believe E to degree 1. The agent's old conditional probability of some proposition A given E is some number r. In this situation ROC is applicable, and would require that the agent's new degree of belief in A be r. But suppose the agent also believes (incorrectly) that belief A is directly affected by the experience. This false belief that A is based on experience may be a rational belief; it may be directly based on experience (and thus in E), or even in accord with conditionalization on previous belief changes. As a result, the agent feels free to rationally believe A to some degree other than r. In this situation the agent believes that the requirement of conditionalization does not place any restrictions on her belief in A. The agent is rational in these situations, even though she does not conditionalize and has violated ROC. She is not aware of having violated any rule of rationality and may be completely consistent; she has managed her beliefs well. Given other beliefs she may have, it may even be irrational for her to conditionalize. Examples of this sort are counterexamples to the requirement of Bayesian conditionalization, because the agents are rational in spite of not conditionalizing in a situation in which the rule of conditionalization is applicable.

Two examples with this structure involve Al the Bayesian rock climber. Suppose Al, a recent convert to Bayesianism, is rock climbing the Royal Arches route in Yosemite. Early in the climb Al looks at a piton in a crack and is thinking about whether it is trustworthy (would it hold him if he fell?). Al's subjective probability that the piton is trustworthy given that it is rusted is .7. Al looks at the piton, and as a result of experience believes to degree 1 that that it is rusted. Based on this belief change, Al modifies the rest of his beliefs and comes to believe to degree 1 that the piton is trustworthy; thus Al has violated ROC. However, based on experience Al mistakenly believes that his belief that the piton is trustworthy is directly based on experience. Wanting to be a good Bayesian, Al reflects on how to apply ROC to this situation. Since he mistakenly believes that his belief that the piton is trustworthy is directly based on experience, he mistakenly believes that his new beliefs are consistent with ROC. His mistaken belief about the source of his beliefs is rational, and thus it is rational for him to be certain the piton is trustworthy. But this rational change of belief violates ROC, and is thus considered irrational by traditional Bayesianism.

The above example does not depend on Al believing to degree 1 that the piton is trustworthy. He could have believed it to any lesser degree, as the following example illustrates. Later in the climb Al comes to a place where

he can not be certain where the route goes: he could climb a steep face with tiny holds to the left or a steep crack to the right. Al's subjective probability that the route goes left up the face given that he comes to such a fork on this route is .2 and he believes to .8 that the climb goes right up the crack given this fork. Since as a direct result of experience Al believes that he is at such a fork, in order to be consistent with ROC, Al would have to believe to .8 that the climb goes right and up the crack. However, after looking at the two options, Al finds himself believing to .9 that he should go left up the steep face instead of right up the horrendous looking crack. He is now convinced that the route cannot go up such a crack; it looks much too difficult for this route. As a direct result of experience, Al mistakenly believes that his belief to .9 that the climb goes left up the face is directly based on experience; in reality the belief is based only on his other beliefs and he is in a situation in which ROC is applicable. Since he is trying to manage his beliefs like a good Bayesian, he pauses to think about whether his belief that he should go left up the face is in accord with ROC. He immediately realizes that since that belief is directly based on experience, ROC does not place any restrictions upon it. Since Bayesianism permits any belief that is directly based on experience, he believes that his belief to .9 that the route goes left up the face is rational according to Bayesianism. Of course, since he believes this belief is directly based on experience, Al believes he will have to use ROJC to modify the rest of his beliefs to account for this, which he quickly does. Furthermore, Al believes that his belief to .9 that the climb goes left up the face is what any properly functioning mind would believe in these circumstances; the belief has warrant. As a result, it would be irrational for him to conditionalize and believe only to .2 that the climb goes left up the face.

These counterexamples are different from standard examples which show that Bayesian requirements are not sufficient for rationality. These examples show that the Bayesian requirement of conditionalization is not even necessary for rationality.[1] Furthermore, the last example shows that there are circumstances in which it may be irrational to form beliefs in accord with ROC.

4. MANAGING BELIEFS

The above counterexamples were based on the fact that we can be rational and yet mistaken about the basis of our beliefs. This possibility raises serious problems for the common view that we should manage our beliefs with Bayesian conditionalization. Bayesians often claim that conditionalization is

a requirement of ideal rationality; having our judgments in accord with ROC is an ideal that we should aim towards as rational beings.[2] Accordingly, we are supposed to use conditionalization to manage our beliefs in pursuit of this ideal.

However, there is a problem with this proposal. Even if conditionalization were a requirement of ideal rationality, we are generally unable to pursue it as a goal. In order to strive towards having beliefs in accord with ROC, we must be able to know in what situations we should apply it. We cannot use conditionalization to manage our beliefs and pursue the goal of ideal rationality without knowing which beliefs are at least partially directly based on experience (and are thus exempt from conditionalization) and which beliefs are based only on other beliefs (and are thus required to be in accord with conditionalization). But in general we are unable to reliably determine which beliefs are formed in direct response to experience and which ones are formed only on the basis of other beliefs. Through introspection it is very difficult to find beliefs that are based only on other beliefs. I have often attempted to determine which of my beliefs are directly based only on other beliefs and not on any experience, but I am unable to do so. For example, once a friend and I were lost on a dirt road in Joshua Tree while discussing Bayesianism. As we looked at maps and tried to figure out where we were, I attempted to determine if my beliefs satisfied ROC. To do so I had to determine which of the beliefs I was forming were based only on other beliefs and which beliefs were directly based on experience. Even though I was making a serious effort to determine what my beliefs were based on, I could not find a single belief that I was confident was directly based only on other beliefs and was not partly based on experience. Perhaps all of my beliefs in that situation were based on experience, but if so, it would be difficult to find many situations in which I have beliefs based only on other beliefs. As a result of my inability to determine which beliefs were based on other beliefs, I was unable to use conditionalization in that situation to manage my beliefs, even though I was consciously thinking about conditionalization. This inability seems to be the norm rather than the exception. For all we know, any of our beliefs could be directly influenced by experience and not based only on other beliefs. As a result, we cannot know if ROC applies to any of our beliefs, and thus we are unable to use ROC to manage our beliefs. If ROC is part of ideal rationality, then we have no idea how to begin pursuing this ideal.

This same problem arises if we attempt to manage our beliefs with Jeffrey's probability kinematics. As in the case with ROC, we generally are unable to know if the situation we are in is one in which ROJC applies. Like ROC, ROJC presupposes a distinction between beliefs that are directly based

on experience and beliefs that are based on other beliefs. For example, if a partial belief is directly based on experience, this degree of belief is rational and in accord with ROJC no matter what it is, even if it is different from what would be required if it were not directly based on experience. But if the belief is not based on experience, then it must be in accord with Jeffrey conditionalization. Since we cannot determine which of our partial beliefs are based on experience, we cannot use ROJC to manage our beliefs; for all we know almost all of our partial beliefs could be directly based on experience, and thus there would be no Bayesian requirements of belief change applicable to them.

These problems arise because ROC and ROJC are external requirements of rationality. In order to use ROC or ROJC to manage our beliefs we need access to the basis of our beliefs, which is not an internal matter. In general, we do not have access to whether our beliefs are based solely on other beliefs, or whether they are also partly based on experience.

Rationality is usually thought to involve only internal features of our cognitive life. We are not irrational for failing to account for things that we are unaware of. Even brains in vats and those whose brains are manipulated in various ways can be rational, if they properly deal with all that is accessible to them. For this reason some epistemologists classify Bayesianism as a version of coherentism, which is an internalist theory.[3] Although they may be correct about the static part of Bayesianism, this is not a correct description of Bayesianism that includes ROC or ROJC. Since ROC makes reference to the basis of our beliefs, ROC is an external requirement, and any theory which requires agents to use conditionalization to manage their beliefs is an externalist theory. Because of this ROC is very different from other requirements of Bayesianism, such as coherence or even reflection. The static part of Bayesianism is an internalist theory; the agent is not required to have access to any non-internal factors in order to satisfy coherence.[4] Similarly, an agent can satisfy reflection while having access only to internal factors. But when Bayesianism is expanded to a dynamic theory of belief change that includes ROC, then it is no longer an internalist theory. Since rationality requires only internal requirements, and since ROC is an external requirement, rationality does not require ROC. Our counterexample showed that one can satisfy all requirements of rationality and not satisfy ROC. Satisfying ROC may be necessary to be in a good epistemic situation, which externalists often cite as our epistemic goal, but it is not necessary for rationality.[5] Since Bayesianism is usually construed as being a theory of rationality, this is in conflict with ROC. Of course, if ROC is required in order to be in a good epistemic situation, and if an agent can see how to fulfill ROC, then the agent should do what is required to get in that good epistemic situation. Being in a

good epistemic situation (which may include satisfying ROC) may be part of an ideal we should pursue if we are able. My claim is that since we generally are unable to know how to even begin working towards this good epistemic situation, we can be rational without pursuing it. Rationality may require us to aim at that ideal when we have some idea of how to get there, but rationality does not require us to aim at something that we have no idea how to pursue.

At this point one might object and argue that various Dutch book arguments demonstrate that rationality requires us to satisfy ROC and ROJC.[6] But although these arguments may show that there is a problem in forming beliefs in accordance with some rule other than conditionalization, they do not reveal a problem for one whose violations of ROC do not follow any rule.[7] In order to create a Dutch book for someone who violates ROC, one needs to know if her new belief will be greater or less than what conditionalization would require. But a Dutch book cannot be made against someone who mistakenly believes that a belief is based on experience and as a result violates ROC. The agent is not following any rule that will tell her when she will mistakenly believe some belief is based on experience and what degree of belief she will then form. As a result, we don't know if the new belief will be greater or less than what ROC requires. Thus the Dutch book arguments for conditionalization do not help with our problem.

The difficulty in distinguishing between beliefs based directly on other beliefs and beliefs based on experience may be partly responsible for the appeal of accounts that postulate an observation language that adequately describes our experience. On these proposals, the agent is supposed to have a belief in an observation language that captures the content of his or her experience.[8] Other beliefs in a natural language are then arrived at by conditionalizing on beliefs in the observation language. This provides a natural way to classify the origin of a belief: beliefs in the observation language are directly based on experience, and beliefs not in the observation language are directly based on other beliefs (that are ultimately based on beliefs in the observation language). However, there is no reason to think humans must have such an observation language in order to be rational. And even if we did have this observation language, we would still need to require that only beliefs in the observation language be directly based on experience and not on other belief change, and that only beliefs in the natural language be directly based on other beliefs and not on experience; this is by no means implied by the existence of an observation language. Thus the same problem arises, since there is no convincing argument that experience only directly affects beliefs in the observation language.

5. INTERNAL RULES OF CONDITIONALIZATION

Although ROC is an external requirement, it may be possible to develop an internalist version of ROC that agents could use to manage their beliefs. ROC was an external theory because we are generally unaware of the basis of our beliefs. But we do have access to our beliefs about the basis of our beliefs, and thus beliefs about the basis of our beliefs are internal factors. One possible internalist requirement of conditionalization is the following:

IROC: If we are now certain that E contains all and only those beliefs that have been directly affected by experience, and if $P_n(E) = 1$, then $P_n(*) = P_o(*/E)$.

We can formulate an internalist version of ROJC as follows:

IROJC: If we are now certain that partition $E = \{E_1, E_2, ...\}$ contains all and only those beliefs that have been changed in direct response to experience, then $P_n(*) = \sum_i P_o(*/E)P_n(E_i)$.

Principle IROC refers to the belief that E contains all and only those beliefs that have been directly affected by experience. Unlike ROC, Principle IROC requires we conditionalize only when we believe that E contains all and only those beliefs that have been directly affected by experience. As a result, requirement IROC does not face the above problem that ROC faced. It can be used by agents to manage beliefs, and agents are able to strive towards it as an ideal. A rational person could satisfy IROC without satisfying ROC. An example is our original counterexample of Al the Bayesian rock climber in which a person falsely believes that a belief is directly based on experience and the new degree of belief is not in accord with conditionalization. Although this belief change violates ROC, it is consistent with IROC and is rational. Furthermore, a person can satisfy ROC without satisfying IROC, although these are cases in which the person is irrational. If a person mistakenly believes ROC is applicable and fails to conditionalize, then that person is guilty of not managing his or her beliefs properly, even though ROC may be satisfied. In these situations ROC makes no requirements on rational belief, but IROC does make requirements. The agent is irrational in these situations, even though ROC is satisfied. These differences between ROC and IROC arise because ROC is an external requirement, and since rationality only has internal requirements, ROC is not a necessary condition of rationality. But since IROC is an internal requirement, it is not ruled out as a necessary condition of rationality.

A serious problem with IROC is that it is a very weak requirement. IROC places no requirements on an agent who is not certain of the basis of his or

her beliefs. Since it is common to not be certain of the basis of our beliefs, IROC is a rather empty requirement of rationality. But Bayesians certainly don't think of conditionalization as an empty vacuous requirement.

One might respond by claiming that I have misconstrued the requirement of conditionalization and what it means to learn something as a result of experience. Suppose we temporarily ignore whether a belief is directly based on experience, and simply look at situations in which an agent changes certain degrees of belief to degree 1. One way of looking at this proposal would be to view it as assuming that all of the beliefs that I learn as a result of experience change to degree 1. Conditionalization would then require that the agent modify her other beliefs to take account of changing these beliefs to degree 1. Consider the following:

IROC*: If E contains all of the beliefs that change to degree 1 ($P_n(E) = 1$), then $P_n(*) = P_o(*/E)$.

This proposal has the significant advantage that the agent does not have to know which of her beliefs are directly based on experience, or even have any beliefs about the basis of her beliefs; any belief that is changed to degree 1 is conditionalized on. Upon this interpretation of conditionalization we can manage our beliefs with conditionalization, because we can know which of our beliefs change to degree 1.

A problem with this proposal arises when we consider situations like the following. Suppose E is all of my beliefs that change to degree 1, and that $P_o(A/E) = .8$. But instead of believing A to .8 as required by IROC*, I believe A to .4. Am I irrational? If I think that my new belief in A is directly based on experience, then I have no reason to apply any rule of conditionalization to it. This is a counterexample to IROC*, because I am rational, but don't conditionalize according to IROC*. The problem with IROC* is that it excludes the possibility of beliefs based on experience that are not changed to degree 1, but rationality does not require we believe that all beliefs based on experience are raised to 1. As a result, IROC* is not successful.

Let us also consider a version of this proposal that deals with partial belief:

IROJC*: If partition $E = \{E_1, E_2, ...\}$ contains all the beliefs that have been changed, then $P_n(*) = \sum_i P_o(*/E)P_n(E_i)$.

Principle IROJC* does not require that all learned beliefs or beliefs based on experience be raised to degree 1. However, this principle has a serious problem; it is trivially true. One always satisfies this rule, no matter what one's beliefs are. If I am not required to distinguish between what I learn by

experience and my other beliefs, in effect all of my new beliefs are based on experience. Thus IROJC* places no restrictions on rational belief change.

One might propose to avoid these problems by requiring that the agent have a rational belief about the source of her beliefs. We could thus modify IROC as follows:

IROC**: If E contains those and only those beliefs that S rationally believes are based directly on experience, and if $P_n(E) = 1$, then $P_n(*) = P_o(*/E)$.

Principle IROC** will not be vacuously true, because we may have rational beliefs about the basis of our beliefs.

However, a very serious problem now arises. Even though IROC** is an internal rule, an agent will not be able to use it to manage her beliefs. Suppose that for any belief, a belief about whether that belief is based only on other beliefs or is based on experience is a next higher order belief. If we let our ordinary beliefs be of level 1, then our beliefs about which of those beliefs are based on experience will be beliefs of level 2. Although this could be stated precisely, for our purposes we only need this rough intuitive notion. Of course, there will also be level 3 beliefs about which of our level 2 beliefs are directly based on experience, etc. Now suppose our agent decides to use IROC** to manage her beliefs. She must first form a rational second order belief about which of her first order beliefs are based on other beliefs and which are based on experience. But in order to rationally form that second order belief, she will have to form a rational third order belief about whether that second order belief is based on experience. One can easily see the regress our agent will fall into if she attempts to use IROC** to manage her beliefs. In order to use IROC** to rationally manage a belief of level n, the agent must first form a rational belief of level n +1 about the source of her level n beliefs. Thus in order to manage her beliefs at level n she will need to first manage her beliefs at level n +1. This is an instance of a general problem in applying any epistemic norm. It appears that IROC** is no better off than ROC; neither can be used by an agent to manage beliefs.

6. CONCLUSION

We began by clarifying the requirement of conditionalization and saw there are two basic ways that beliefs may be rationally changed: in response to experience and in response to other belief change.[9] We presented a counterexample to the requirement of conditionalization and discussed the consequences of this counterexample for the view that conditionalization is a way to manage our beliefs in pursuit of some ideal. We then argued that since we are usually not able to determine which of our beliefs are directly

based on other beliefs, we are unable to apply rules of conditionalization to manage our beliefs. Bayesian requirements of belief change are external requirements and appeal to factors that are not generally accessible to epistemic agents. Conditionalization is based on an internally inaccessible distinction between beliefs that are directly based on experience and beliefs that are based on other beliefs, and Bayesians will have to adopt this traditional epistemic distinction if they wish to require a rule of conditionalization. This is unfortunate, because one of the appeals of Bayesianism is that it claims to not require most of the traditional epistemic distinctions.

Since rationality is an internal notion, it does not rely on external factors such as those required by conditionalization. Internalist versions of the requirement of conditionalization can be developed, and these avoid some of the problems that the externalist versions faced. But we saw that these rules are vacuous, have counterexamples, or face a regress problem. Thus neither external nor internal versions of conditionalization are helpful in managing beliefs.

ENDNOTES

[1] Once one sees the structure of these counterexamples one finds that it is very easy to construct other counterexamples. For example, suppose Susan's subjective probability that there is a rose in a garden given there is an assortment of flowers in the garden is .8. She sees the garden, and believes to degree 1 that there is an assortment of flowers in it. Based on this new belief, Susan modifies her other beliefs and contrary to ROC comes to believe to degree .9 that there is a rose in the garden. Susan mistakenly believes that her belief that there is a rose in the garden is directly based on experience, and that her new beliefs are consistent with ROC. Susan is rational in her mistaken belief, and thus it is rational for her to believe to .9 that there is a rose in the garden.

[2] It is not at all clear what ideal rationality is; for example, some might think one was ideally rational only if one believes all truths, or all logical truths, or….Furthermore, it is difficult to see how conditionalization can be a requirement of ideal rationality given that it is irrational to conditionalize in some situations (such as the previous counterexample). However, we will ignore these problems for now.

[3] See Alvin Plantinga 1993.

[4] Logical omniscience is required, but let us assume that this is not an external requirement.

[5] See William Alston 1985 and 1988.

[6] See van Fraassen 1989 and Teller 1973.

[7] See Bas van Fraassen 1989.

[8] See J.H. Sobel 1990.

[9] Once this is seen, one realizes that some claims often made about Bayesianism are false. For example, the frequent claim that according to Bayesianism all rational belief change is by conditionalization is seen to be false; one can also change beliefs in response to experience. Another example is the claim that a rational agent should not assign a probability of 1 or 0 to any contingent proposition. The reason given is that if a proposition has a probability of 0 or 1, then it is impossible to change it by conditionalization; no matter what is conditionalized

on, the result will remain the same. In order to preserve the ability to change our minds about a contingent proposition it is claimed that we must not assign the proposition to a probability of 0 or 1. But from our description of Bayesianism it is easy to see the flaw in this argument. It is correct that by conditionalization an agent can never change a degree of belief in a proposition of degree 0 and 1. But it is also possible that as a result of experience the agent could change his belief from 0 or 1 to some other degree of belief. Nothing in Bayesianism restricts the beliefs we can form in direct response to experience. Thus the primary argument given to not assign degrees 0 or 1 to contingent propositions fails.

REFERENCES

Alston, William. 1988. An internalist externalism. *Synthese* 74: 265-283.
Jeffrey, Richard. 1983. *The Logic of Decision*, 2nd ed. Chicago: University of Chicago Press.
Plantinga, Alvin. 1993. *Warrant: The Current Debate*. Oxford: Oxford University Press.
Sobel, J.H. 1990. Conditional probabilities, conditionalization, and dutch books. In *PSA* 1990 Volume I, eds. Arthur Fine, Micky Forbes, and Linda Wessels, 503-515. East Lansing, MI: Philosophy of Science Association.
Teller, Paul. 1973. Conditionalization and observation. *Synthese* 26: 218-58.
Van Fraassen, Bas. 1989. *Laws and Symmetry*. Oxford: Oxford University Press.

Chapter 12

MATERIALISM AND POST-MORTEM SURVIVAL

Keith E. Yandell
University of Wisconsin, Madison

> Whether we are to live in a future state, as it is the most important question which can possibly be asked, so it is the most intelligible one which can be expressed in language. Yet strange perplexities have been raised about the meaning of the identity or sameness of person, which is implied in the notion of our living now and hereafter, or any two successive moments. And the solutions of these difficulties have been stranger than the difficulties themselves. For personal identity has been explained so by some as to render the inquiry concerning a future life of no consequence at all to us the persons who are making it.
>
> —Bishop Joseph Butler

> When we speak of the Soul as *naturally* immortal, we mean it is by the *Divine Pleasure* created such a Substance, not having in itself any Composition, or any Principles of Corruption, will *naturally*, or *of itself* continue for ever; that is, will not by any natural decay, or by any Power of Nature, be dissolved or destroyed; But yet nevertheless depends continually upon God, who has power to destroy or annihilate it, if he should so think fit.[1]
>
> —Samuel Clarke, in *A Letter to Mr. Dodwell*

T. M. Crisp, M. Davidson and D. Vander Laan (eds.), Knowledge and Reality, 257-298.

A PERSONAL NOTE

Professor Alvin Plantinga is not retiring. Hence these papers cannot constitute a *Festschrift*. A *Festschrift* celebrates *a professor's career at his retirement*; celebrating *the fact that he is retiring* is different, compatible with celebrating the relevant career, and irrelevant here. We get to celebrate both Al's career, and Al, and the fact that he is not retiring, all at once. Nothing that I can write is anywhere near adequate for such high things. But I'm glad for the chance to add my small celebratory sounds to the chorus.

INTRODUCTION

Defenses of the view that persons survive the death of their bodies are not likely to be heard these days. An interview paper, say, read with the purpose of eliciting an appointment to a job teaching philosophy is not likely to take this form. Belief in post-mortem existence is not widely popular in philosophical circles. One reason is that it is often associated with the view that persons have immaterial souls and thus are not identical to their bodies, or indeed to any bodies at all. That view, in turn, is associated with mind-body idealism, for which there aren't any bodies for anything to be identical to, or mind-body dualism, for which there are bodies though it isn't even possible that a mind or person be identical to any of them. Idealism and mind-body dualism are views that current philosophical mores tell you that you should be *embarrassed* to be caught holding, rather as you should feel real chagrin at being caught stealing money from the departmental coffee fund. The idea is that the refutation of these views has either already been done (though there is no universal agreement about exactly how, when, where, or by whom) or isn't necessary. Materialism is the truth, and the only remaining question is which variety thereof deserves one's allegiance.[2]

In this intellectual environment it isn't surprising that, philosophers who think that persons do survive the death of their bodies, even if they think that the single consideration in favor of this thought is that their religious tradition says it happens, should want to show that embracing this view does not prevent one from being a materialist. I have no interest in impugning anybody's motives. I simply note that there have been a number of Christian philosophers—philosophers who actually accept as true the doctrinal claims of what C. S. Lewis called "mere Christianity," which includes the claim that persons do survive bodily death—who recently have embraced one or another variety of materialism and argued that their doing so is perfectly compatible with their belief that they will survive bodily death. The sort of

materialism thus embraced is one regarding human persons, not (yet, at least) about God.

On the face of things, this is not the most encouraging view one could think of. It lacks the high plausibility of the claim that, say, your average squirrel is smaller than a typical baby elephant, or that seventeen is not an even number. Worse, it seems to fly in the face of what actually happens to bodies. Bodies die. They are then buried in (sometimes very expensive) boxes, cast into the sea, cremated, mummified, eaten by animals, or the like. On a materialist view, a person is identical to her body (or to some material thing that is a proper part thereof). So, on a materialist view, persons are buried in boxes, cast into the sea, cremated, mummified, eaten by animals, or the like. None of these things looks much like survival, let alone like going to heaven. Nonetheless, various contemporary philosophers have embraced one or another variety of materialism and maintained that persons survive the death of their bodies and live again in an after-life.

Things are complicated by there being at least two general sorts of materialism that are relevant here. One is that a person, as stated, is numerically identical to a body or a proper part of a body. On this view—on *Materialist Substantivalism*—one sticks to her guns about a person being identical to something physical and gives an account of the physical something in question that is intended to allow for survival. The other is that a person is not numerically identical to a body or a proper part of a body, but rather is identical to something that is closely related to a body but is perhaps somehow detachable from it, for example a series of psychological or conscious states; the person is a kind of sequence. This view—we will call it *Sequentialism*—can occur in a substantivalist context (a sequence requires some close association with a material body) or a non-substantivalist context (the conscious states that comprise a person are all identical to physical states, and so-called material things are really merely collections of physical states). If one holds that there are no material *substances* or physical *things*, one can still be a materialist in virtue of holding that only material *states* exist. (Some self-identified materialists also think that there are abstract objects, and are materialists about everything else.)

My take on these matters is that these Christian materialist maneuvers are unnecessary and that the efforts to develop a defensible view of this sort fail. Thus my purpose here is to suggest some of why the maneuvers are unnecessary and to present a fair rendering of some of the resulting views plus a successful critique thereof.[3,4]

1. IMMORTALITY AND RESURRECTION

The doctrine of the immortality of the soul is often contrasted to the doctrine of the resurrection of the body, with the suggestion that the two doctrines are logically incompatible, or at least that anyone who embraces the one will find it impossible to relate it to the other in a perspective that is coherent overall.[5] This seems to me false, and my purpose in this first Section is to argue that it is.[6]

1.1 A Minimal Condition of Being Immortal

One might, not implausibly, think that *being immortal* involves as least this much:

(I) X is immortal entails There is some time T at which X exists, and for any time later than T, X exists then too.[7]

By no means all those who believe that persons are immortal would accept even this seemingly minimal characterization. Among philosophers who think that persons are immortal, some believe that personal identity is possible over time gaps, so that a person can exist at time T1, then at T2 take a vacation from existence for however long a period, and then come back into existence or exist again at (say) T10,000.[8] On this sort of view, one could be immortal but not satisfy the conditions that (I) imposes. One would just need to satisfy the weaker:

(I*) X is immortal entails X exists at some time and there is no last time at which X exists.

On this account, one would be immortal even if she, say, popped into existence for a second every trillion years and then ceased to exist again until another trillion years went by. One could not exist at all most of the time and yet be immortal. One might, then, quite reasonably think that satisfying (I*) is not enough for us to say that a being that did so was immortal in any serious sense of the term. Even those who think that identity over time gaps is possible for persons are likely to require that immortal persons exist a decent amount of the time, though how long that might be is unclear. It is perfectly open to them to hold that there must sooner or later be a time after which one never again blips out of existence.[9,10] I confess to thinking that, regarding persons—as opposed to, say, wars and sporting events—existence again once one had altogether ceased to exist is impossible, and that the closest thing to

it would be a copy of the original person which would, as is the case for all copies, not be the original.

1.2 Natural Immortality: Two Views

The view that we are naturally immortal is often thought of as incompatible with a Christian view of immortality. Taking immortality to be partly characterized by (I) rather than (I*), the Indian religion and philosophy Jainism thinks of the immortality of persons as intrinsic, inherent, natural, and independent. Jainism is a philosophy or religion that believes in personal immortality but not in God. It holds that persons are in principle indestructible. No person, and nothing else, can put a person out of existence. In contemporary terms, Jainism in effect construes persons as enjoying logically necessary existence. In this regard, it is in agreement to what seems to be Plato's view—a soul is immaterial and cannot begin or cease to exist. For neither view is there any power in the world that could make a soul go out of existence. So understood, the claim that the soul is immortal is incompatible with the Jewish and Christian doctrine of creation.

There seems to be little reason to think that we are in principle indestructible, necessarily existent, safe no matter what, or existent in all possible worlds. It seems true that the world did without us for a very considerable period, may do so again, and might never have been graced by our presence.[11]

Natural immortality need not mean immortality that holds no matter what God does. Christianity has typically believed in personal immortality and in God. It has typically and traditionally held that persons are souls that are so created by God as not to be open to annihilation by anything but God. They are created by God to exist forever; this is part of their nature. But what God can create, God can annihilate. Continued existence is grace as much as beginning to exist is grace. This, I take it, is also a sort of natural immortality—immortality by nature, which involves immunity from destruction by fellow creatures, graciously given and sustained by God.

In contrast to Jainism, Christianity characteristically embraces a robust doctrine of creation which holds something along the lines of:

(C1) If X could have failed to exist, then X depends for its existence on God.[12]

This claim arguably is to be taken in conjunction with:

(C2) For X to depend on God for existence is for God to sustain X in existence for every moment at which it does exist, and for this divine sustaining to be necessary for X's existence.

A Jain understanding of natural immortality is incompatible with what the Christian tradition takes to be true of persons. "Natural immortality" means different things in a Jain versus a Christian context.

Further, Christianity holds to a doctrine of the resurrection of the body. This entails, or at least is best understood as, involving the view that souls of the sort we are flourish best and fully only as embodied. There is nothing contradictory in the idea of something being immaterial and yet its being the case that it has important capacities most fully developed and most satisfactorily exercised when it is embodied and operating in a physical environment. There is also an idealist version of this idea, one that holds that human persons are souls that best flourish when they have conscious experiences that possess sensory content.

It is true that, so far as I am aware, this notion of a soul that flourishes best and most fully only as embodied has not received the articulate and detailed formulation one might reasonably expect would have been provided, given that the traditionally widely accepted notion of human persons as embodied souls so strongly suggests it. Whatever the cause for this, it is not the internal inconsistency of the notion or its inherent lack of coherence with Christian theology or defensible metaphysics.[13]

What belief in personal immortality amounts to depends on what persons are, and what immortality is. *Being immortal* is not an inherently good thing. One could be immortal and have one's immortality consist in one's endlessly seeming to experience being boiled in oil as one's eyes seem to be pierced with hot needles and one's ears seem to have hot wax poured into them. These things might seem that way because they were that way. One who turned down the offer of such an immortality would not necessarily be irrational in so doing. The notion of an immortality worth the having is obviously more complex than the simple notion of *being immortal*. It is *personal* immortality that Jainism and Christianity value, an immortality that includes the retention of personal identity over time and after the death of one's body, and under conditions that a rational person could welcome.[14] Christian immortality involves more, of course, than the retention of existence by a person. The person will be conscious and self-conscious, in a community with God and other persons, free from pain and sorrow, and so on. But the notion of retention of personal identity will be plenty to occupy us here.

That a soul is naturally immortal at least entails that there is nothing that can cause it to cease to exist except God, and that not even God can cause it

to cease to exist by causing something else to cease to exist or by bringing something else into existence or by bringing it to pass that some event occurs in some other created thing. God's annihilating a soul's accompanying body, or bringing it to pass that everything in a soul's environment—trees, houses, other souls, and the like—were annihilated would not suffice to annihilate it, nor would creating nuclear fields or atomic explosions, if it has natural immortality.

That a soul is naturally immortal also entails that it lacks two sorts of internal defects. Perhaps God could create something in such a manner that it would, after some time had passed, simply altogether cease to exist—something that had built-in strong obsolescence. Presumably God could create something in such a manner that, over time, it progressively came apart, becoming scattered and no longer existing though the stuff it was made of hung around—something that had built-in milder obsolescence. Possessing natural immortality would preclude a thing's being strongly obsolescent or its being mildly obsolescent.

1.3 Simplicity

Clarke emphasizes simplicity. Part of Clarke's claim is that the soul is simple, partless, "not having in itself any composition." We can distinguish this claim from the claim that it is naturally immortal. Consider two criteria for distinctness:

D1. X and Y are distinct beings if and only if it is logically possible that one exist and the other not.

D2. X and Y are distinct beings if and only if it is logically possible that one have a property that the other lacks.

These criteria are not extensionally equivalent; you can get distinctness given a proper application of the one and lack of distinctness given a proper application of the other. For example, God the Father and God the Son are, on any proper rendition of the doctrine of the Trinity, such that neither depends for existence on any being that is not a member of the Trinity and such that both are members of the Trinity essentially—neither can exist without the entire Trinity existing and the entire Trinity cannot exist without either of them. Neither, then, can exist without the other; by D1, they are not distinct. But the God the Father is not the member of the Trinity who became incarnate in Jesus Christ and God the Son is the member of the Trinity who became incarnate in Jesus Christ, so by D2 they are distinct. The Father and Son are one God and two Persons.

Suppose item X has two members by criterion D2 but is one item by D1 since what are distinct members by D2 are one item by D1. Roughly, X has two members, distinct on a property criterion, that cannot exist, either without the other. Then, for present purposes, we shall take X to be simple, to be partless, to lack composition. To fail to be simple, on this account, is to have at least two parts such that it is logically possible that one part exists in the absence of the other.

It is possible that something that is, in this sense, simple also exists at some time T such that never after T does it cease to exist. A simple thing necessarily lacks mild obsolescence and can lack strong obsolescence. Thus it can be naturally immortal. Clarke's view is that souls satisfy this description. The notion is, so far as I can see, both logically consistent and consistent with a robust doctrine of divine creation. Further, if persons are souls or minds, then presumably they are simple; they have no proper parts that are persons, and they have no proper parts at all. This is especially clear if one thinks of parts as being able to exist under circumstances in which they are not parts—if one holds that *If X is a part of whole Y, then X can exist even if Y does not*. Bodies without minds or souls are corpses; they are former human bodies. So bodies are not, on this construal, parts of human persons.

The claim that persons are immortal, then, is not incompatible with the Christian doctrine of the resurrection of the body. Persons or souls being naturally immortal and simple is compatible a robust reading of the Christian doctrine of creation. These notions are traditionally associated with mind-body dualism (or idealism) and are relevant to the recent and rather extraordinary claim that the view that a person is concrete but immaterial is inconsistent with Christian theology.

2. FISSION AND REFUTATION

Suppose that on a view of what a person is, the following is possible: that a person X exists at time T, and at time T1 there exist two persons, Y and Z. Each of Y and Z have equal metaphysical claim to be X. But Y is distinct from Z, and so X cannot be identical with both of them. Further, were Y to exist and Z not exist then X would be identical to Y, and were Z to exist and not be accompanied by Y then X would be identical to Z. It is the presence of both Y and Z that produces the existential loss of X. Under these circumstances, to use the usual parlance, X has undergone fission. It isn't unclear what has happened to the person X, provided what it is to be a person is what the view in questions says—the view on which X's fissioning can occur. X has ceased to exist, for had X continued to exist, given some

view on which fission is possible, then X would have to be identical to two non-identical things, and that is logically impossible. But, on such a view, it is also true that X is identical to both Y and Z, because, on the view on which fission is possible, X would meet the conditions for being numerically identical to Y and would also meet the conditions for being numerically identical to Z.

The idea can be put in these terms:

> *Definition of the Possibility of Fission*: It is possible, according to a view V of what a person is, that fission occur to a person X if and only if, [if according to V, X is a substance, then it is possible that X exist at time T, and that there exist at time T1 two distinct substances Y and Z such that Y and Z are persons according to V, that Y meets the conditions that V holds to be sufficient for Y being numerically the same person as X, and that Z meets the conditions that V holds to be sufficient for Z being numerically the same person as X] or [if according to V, X is a sequence, then it is possible that there exist at time T1 two sequences Y and Z such that Y and Z are persons according to V, Y meets the conditions that V holds to be sufficient for Y continuing the sequence that is X, and Z meets the conditions that V holds to be sufficient for continuing the sequence that is X].

On Sequentialist terms, if (i) a person-constituting sequence X exists up through time T, (ii) a person-constituting sequence Y begins to exist at time T1, and (iii) Y continues X, then the person that X (partly) constitutes and the person that Y (partly) constitutes are the same person.

Consider the following propositions:

a. Person X exists at time T

b. Person Y exists at time T1

c. Person Z exists at time T1

d. Person Y is numerically the same as, or continues the sequence that is, person X

e. Person Z is numerically the same as, or continues the sequence that is, person Z

f. Person Y is not numerically identical to, or does not (partly) constitute the sequence that is, person Z.

The set a-f is logically inconsistent. Hence it is not possibly true. Any view that entails that it is possibly true entails a contradiction. Any view that entails a contradiction is false, and necessarily so. The argument:

1. View V entails Fission is possible.

2. Fission is not possible.

Hence:

3. View V is false.

is obviously valid. Premise 2 is true. So if one shows that, regarding some view, it entails that Fission is possible, that view has been refuted.

The tactic of saying "So long as there never is a case in which two persons meet the conditions specified for Y and Z in relation to person X, the fact that a-f cannot be true together does not matter" fails utterly. One may as well say that it does not matter if a theory of squirrels—an account of what it is to be a squirrel—entails that a thing can be a squirrel even though it has an outside but no inside, so long as no such squirrel turns up. Whatever entails that what is impossible is possible is itself impossible—necessarily false.

It is sufficient, then, to refute a view of what a person is that one show that it entails that Fission is possible. Our concern here is to do this regarding Materialist Substantivalism and Sequentialism. The critique of these positions is best developed in a context that is clear regarding identity, particularly regarding its absoluteness and necessity.

3. IDENTITY AS ABSOLUTE AND NECESSARY

3.1 Concerning the Absoluteness of Identity

Professor David Wiggins rejects the relativity of identity. The view that identity is relative takes sentences of the form "*X = Y*" to be strictly ill-formed. Where "T" is some sortal term, a sentence of the form "X = Y" is at best shorthand for sentences of the form *X is the same T as Y*. For it, identity is sortal-relative, so a sentence of the form *X is the same T1 as Y* can be true and a sentence of the form *X is the same T2 as Y* can be false, where "X" is replaced by a term that refers to the same item in both of its occurrences, "Y" is replaced by a term refers to the same thing in both of its occurrences,

and "T1" and "T2" are distinct sortal terms. There is, on this account, no such thing as plain old identity; there is only identity as being the same so-and-so which is compatible with difference regarding being a such-and-such. The same lump of clay can be a statue in the morning, a heap in the afternoon, and a plate in the evening. The lump itself can be the same lump before some coloring is added and after, but not the same chemical mixture later as earlier. According to an identity relativist, even an incomposite item can be the same T1 and not the same T2 as another incomposite item.

In *Sameness and Substance* (1980: 19ff.), Professor Wiggins offers an argument against the relativity of identity. Let T be some sortal term—some term T such that if T is true of X, then X belongs to some sort or is some sort of thing—a goat, a tree, or a board, for example. Wiggins' view is that if some X is the same goat, tree, or board, as Y then it follows that, for any property Q, X has Q if and only if Y has Q. If X and Y are the same item of one sort, then there is no sort of item that the one is and the other is not. It is impossible that:

a. X is a T

b. Y is a T

c. X is the same T as Y

d. X is a T*

e. Y is not the same T* that X is.

That propositions expressed by sentences that result from replacing the variables in propositional functions a-e with appropriate constants can form a consistent set is precisely what the relative identity theorist claims is possible, and often true.

Professor Wiggins argues for the falsehood of the relativity of identity view as follows. Suppose that:

1. X is a T.

2. Y is a T.

3. X and Y are the same T.

Then:

4. If X and Y are the same T, then for any property Q, X has Q if and only if Y has Q.

So:

5. For any property Q, X has Q if and only if Y has Q. (1-4)

Now consider the necessary truth that:

6. If X is a T, then X is the same T as X.

We can then infer:

7. X is the same T as X. (1,6)

Suppose further that:

8. X is a T*.

Consider the necessary truth that:

9. If X is a T*, then X is the same T* as X.

We can then infer:

10. X is the same T* as X. (8,9)

The relative identity theorist claims that all of this is compatible with:

8a. Y is not a T*, even though X is, and even though Y is the same T as X.

Consider this value for Q: *being the same T as X*. Given 1-4, X has this property if and only if Y has it. But the same holds for this value of Q: *being the same T* as X*; X has this property if and only if Y has it. By 10, X has it. So from 5 and 10, it follows that:

11. Y is the same T* as X.

Of course if Y is the same T* as X, then Y is a T*, and 8a is false.

Or, if you prefer, consider this value for Q: *being a T**. Given 8, X has this property. Given 5, X has this property if and only if Y has it. So, given 5 and 8, it follows that Y has it. It follows, that is, that:

11*. Y is a T*.

But then it follows that 8a is false.

This is, I think, a fair presentation of Wiggins' argument. Further, I take that argument to have two pleasing features: it is valid, and its premises are true. Hence its conclusion is true. Its conclusion is that 8a cannot be true if [1 through 10] are true. The relative identity theory entails that 8a is compatible with the truth of [1 through 10]. So the relative identity theory entails a falsehood. Hence the relative identity theory is false. Further, of course, having this entailment is no accidental feature of the relative identity theory such that one could repair the defect by removing the elements in relative identity theory that entail the falsehood. It is propositions essential to, and constitutive of, relative identity theory that entail the falsehood. So the relativity of identity thesis is false and irreparably so. Hence identity is absolute.

We sometimes speak of qualitative identity, which is perfect qualitative similarity. If items X and Y are strictly qualitatively identical, then X has a quality Q if and only if Y has its twin. Nonetheless, if X and Y are distinct, then X and Y are not numerically identical. Whether perfect qualitative identity is consistent with numerical distinctness is controversial. If it is, perfect similarity is one thing and identity another. Whether it is possible or not, strictly speaking "qualitative identity" is a misnomer. As Professor David Lewis used to say, all identity is numerical identity.

3.2 Concerning the Necessity of Identity

An important thesis in metaphysics is:

> *The Principle of the Necessity of Metaphysical Identity* (PNMI)*:* for any items X and Y, if X is identical to Y then necessarily, X is identical to Y.

This claim does not entail that if some phrase designates X, then it necessarily designates X, or that if two phrases A and B both designate X then necessarily both designate X or both designate the same thing, or that if X has a property then it necessarily has that property. It is not a thesis about how language relates to the world. It does entail that if Karen has a dog named Max, Max could not have been some other dog than the one he is. It entails that if the person Karen existed a year ago, and exists now, there isn't

any other person that exists now that Karen could have been numerically identical to other than the one that she is identical to.

More abstractly, (PNMI) entails:

> *The Principle of the Necessity of Synchronic Identity* (PNSI): for all X and T, if x exists at T then X is necessarily self-identical at T, and if X is identical to Y at T, X is necessarily identical to Y at T, and necessarily there is nothing Z that is distinct from Y and exists at T such that X could have been identical to Z at T.

Further, (PNMI) entails:

> *The Principle of the Necessity of Diachronic Identity* (PNDI): for all X, Y, Z, T1, and T2, if X at T1 is identical to Y at T2, then necessarily X at T1 is identical to Y at T2, and if X at T1 is identical to Y at T2, then necessarily there is nothing Z that exists at T2 such that Z is distinct from Y and it is possible that X at T1 be identical to Z at T2.

3.3 Metaphysical Distinctness Conditions and Identity

A bit of reflection on distinctness can be enlightening concerning numerical identity.

> D. If X and Y are possibly distinct, then X and Y are (actually) distinct.

This entails its contrapositive:

> CD. If it is not the case that X and Y are (actually) distinct, then it is not the case that X and Y are possibly distinct.

If it is not the case that X and Y are distinct, then X and Y are numerically identical. If it is not the case that X and Y are possibly distinct, then X and Y are necessarily numerically identical. So CD is identical to:

> PNMI. If X and Y are numerically identical, then necessarily X and Y are numerically identical.

Metaphysical distinctness conditions concern that in virtue of which some item X is distinct from some item Y, at some time T and at (say) sequential times T1 and T2. The distinctness conditions of most importance are those true of incomposite items. Metaphysical identity conditions of an

item X concern that in virtue of which X is what it is and not another thing at some time T, and that in virtue of which X at T1 is the same item at (say) sequential times T1 and T2. The identity conditions of most importance are those true of incomposite items. The adequacy of the metaphysical distinctness and identity conditions we are concerned with here does not depend on how they relate to, or how well they might serve as clues to, epistemological distinctness and identity conditions—conditions by reference to which *we can tell* what is and is not distinct from what else.

There are various criteria for lack of identity or distinctness. Let "quality" include spatial and temporal properties, *X conditions Y* be true if X's failing to exist is sufficient for Y's failing to exist, and *X conditions Y with respect to Q* be true if either X's failing to exist is sufficient for Y's failing to have Q or if X's failure to have some quality Q* is sufficient for Y's failing to have Q. Then some criteria for distinctness follow.

1. X is distinct from Y if it is logically possible that X exist and Y not exist or that Y exist and X not exist.

2. X is distinct from Y if there is some property Q such that X has Q and Y lacks Q or that X lacks Q and Y has Q.

3. X is distinct from Y if there is some property Q such that it is logically possible that X has Q and Y lacks Q or that X lacks Q and Y has Q.

4. X is distinct from Y if there is a Z such that Z existentially conditions Y but not X or Z existentially conditions X but not Y.

5. X is distinct from Y if it is logically possible that there is a Z such that Z existentially conditions Y but not X or Z existentially conditions X but not Y.

6. X is distinct from Y if there is a Z such that, for some property Q, Z conditions X with respect to Q but does not condition Y with respect to Q or Z conditions Y with respect to Q but does not condition X with respect to Q.

7. X is distinct from Y if it is logically possible that there is a Z such that, for some property Q, Z conditions X with respect to Q but does not condition Y with respect to Q or Z conditions Y with respect to Q but does not condition X with respect to Q.

If we are in doubt whether X and Y are the same (is the white rat who runs the complex maze so well the same one who won't eat cheese?) and we learn that the talented maze-runner could be eaten by a cat while the cheese-refuser was still refusing cheese, or the maze-runner could turn left and the cheese-refuser turn right—the one could have a property the other lacked, or a future the other did not, and so on—then we will have learned that they are distinct. If these things are possibilities, the rats are distinct, whether we ever learn that or not. Thus 3, 5, and 7 are properly included in our list.

Among the distinctness conditions concerning X and Y, then, is this: that it is *possible* that X and Y are distinct. It is true that:

8. If it is logically possible that X is distinct from Y, then X is distinct from Y.

Now make one unproblematic simple substitution in 8: for "X is distinct from Y" substitute "X is not numerically identical to y". The result is:

8*. If it is logically possible that X is not numerically identical to Y, then X is not numerically identical to Y.

Line 8, of course, is identical to its contrapositive:

C8. If it is not the case that X is not numerically identical to Y, then it is not the case that it is logically possible that X is not numerically identical to Y.

For clarity, make one unproblematic substitution in C6: for "it is not the case that X is not numerically identical to Y" substitute "X is numerically identical to Y" so that we get:

C8*. If X is numerically identical to Y then it not the case that it is logically possible that X is not numerically identical to Y.

What C8* amounts to is simply the Principle of the Necessity of Metaphysical Identity.

4. A WORRY ABOUT INDIVIDUATING IMMATERIAL PERSONS

One significant worry that philosophers have had about immaterial minds is that there are no possible metaphysical identity conditions for them.

Insofar as the worry has been about epistemological identity conditions, or has confused epistemological identity conditions with metaphysical identity conditions, the dualist or idealist need not be much concerned. The basic issue here concerns metaphysics. The issue is simply whether one can state what it is that might make it the case that there were as many as two minds. The demand is that one say what a mind is, and say what would have to hold in order for there to be more than one of them. The proposition *There cannot be more than one mind* is plainly false, but what—for an idealist or a dualist—would it be like for there to be two of them? Insofar as philosophers have thought there to be no good answer to this question, they have also thought that dualism and idealism didn't so much as have an ontology of minds on offer.[15] If this is so, then however promising or unpromising materialism may be as a basis for thinking of persons as surviving the death of their bodies, it will be understandable if believers in personal survival do their philosophical exploring there. In particular, the worry rests on the feeling that if X and Y are distinct individuals, they must be in different places; so only persons that are bodies or are embodied in bodies (or the like) can be distinct. But is it right to follow this feeling?

Suppose we are told that in a locked box there either there are two items or else there is one. If we learn which, we get a million dollars, but we must know and not guess. The names "Kim" and "Mik" both refer to the one item in the box or to the two items in the box, as the case may be. Now consider two continuations of the story.

> *Continuation One*: We are told that it is possible that Kim come to have a property that Mik lacks; it is possible that they not be qualitatively identical.
>
> *Continuation Two*: We are told that it is not possible that Kim come to have a property that Mik lacks; it is not possible that they not be qualitatively identical.

Given Continuation One, we may conclude that there are two items in the box; Kim and Mik are distinct. The reason behind this is simple: possible difference in properties entails distinctness in entities. Formally:

> For all X and Y, if it is possible concerning some property Q that, of X and Y, one has Q and one lacks Q, then X is not numerically identical to Y.

What about Continuation Two? Here, I think, we must know more. Consider the notion of *twinettes*. Items X and Y are twinettes if and only if

(a) X and Y each depends on the existence of the other, (b) neither can exist unless for any property Q, X has it if and only if Y has it. If this is true, then perhaps what Continuation Two tells us is compatible with X and Y being distinct. Co-dependent twins each of whose existence depends on its being just like the other seem still two in number. Or perhaps one should conclude that the "twins" really are a single; "Kim" and "Mik" name different aspects of a single thing. They aren't parts of that thing if a part of a whole must be able to exist independent of being a part. One question all this raises is: just how are we to understand "having a property"?

The *Identity of Indiscernibles*, regarding any X and Y there are, says this:

> [For no property Q is it the case that, of X and Y, one has it and the other lacks it] entails [X = Y].

Or (to put it affirmatively):

> [For every property Q, it is the case that, of X and Y, one has it if and only if the other has it] entails [X = Y].

In contrast, *Leibniz's Law*, regarding any X and Y there are, says this:

> [X = Y] entails [For no property Q is it the case that, of X and Y, one has it and the other lacks it].

Or (to put it affirmatively):

> [X = Y] entails [For every property Q, it is the case that, of X and Y, one has it if and only if the other has it].

Leibniz's Law seems perfectly secure; the Identity of Indiscernibles is controversial. It is asked: why can't there be, say, two identical disks—two disks that are made of the same sort of plastic, always equal in weight and dimensions, of the same color throughout, each two feet apart, of exactly the same duration or both eternal, alone in the universe? Or why can't God decide to create two persons—two self-conscious beings capable of thought and whatever else being a person involves—who will have twin intellectual biographies? At the moment of their creation, Bob1 and Bob2 reflect that 7 and 2 are nine and hope for more interesting thoughts. Their thoughts do become more complex, and for every time T at which Bob1 and Bob2 exist, the description that is true of the thought-life of Bob1 at T is also true of the thought-life of Bob2 at T. One can think of this situation as one in which every mental state S1 of Bob1 corresponds to a twin mental state S2 of

Bob2, or in which every mental property M1 of Bob1 corresponds to a twin mental property M2 of Bob2. If the two disks or the two minds scenario is possible—if it is logically possible that they obtain—then this raises questions concerning the Identity of Indiscernibles. So far as I can see, the questions properly concern what the Identity of Indiscernible means—how it is to be understood. The disks and minds examples help clarify this matter rather than providing a refutation of the Identity of Indiscernibles.

Consider the two minds case. Here is one way of seeing it. Bob1 and Bob2 are distinct because, at any time T you like, it is the case that Bob1 has some mental state or property and Bob2 has some mental state or property. One of these states or properties is (say) Bob1's thought that it is logically impossible that a being be omnipotent but not omniscient and another is Bob2's (simultaneous) thought that it is logically impossible that a being be omnipotent but not omniscient. Bob1's-thought-that-P is distinct from Bob2's thought-that-P. Either can exist in the absence of the other, and if either can exist in the absence of the other, they are distinct. In seeing it in this manner, one can have in mind any of at least three views of properties. "Property," as it appears in the statement of the Identity of Indiscernibles, can be understood in various ways. For example:

> *Exemplification*: a property is an abstract object, and what it is for something to have a property is for it to exemplify an abstract object; two items have the same property in virtue of each exemplifying one abstract object, so that there are two exemplifications.

I do not pretend to know what "exemplification" means here, but others do claim to. If there is some possible relation between an abstract object (roundness, say) and a circular spatial item such that the item's bearing that relation to the abstract object explains its being round, then the Identity of Indiscernibles says this:

> (II:E) For any X and Y, if X's property exemplifications are identical to Y's property exemplifications, then X is numerically identical to Y.

This is different from saying that if the abstracta that X exemplifies are identical to the abstracta that Y exemplifies, then X is numerically identical to Y. The disks and the minds exemplify the same abstracta, but they differ in that the exemplifications that belong to or obtain in (or whatever) one disk or mind differ from those that belong to or obtain in (or whatever) the other.

> *Features*: a property is properly described, not (say) as *pain*, but as *Tom's being in pain*, not as *the thought that utilitarianism is indefensible* but as *Donagan's thought that utilitarianism is indefensible*; the entities in question are not properties *simpliciter*, but properties-of-something; properties are neither exemplifications of abstracta nor could they exist on their own nor are they parts of anything; things are not collections of properties, but bearers of properties.

It would be mistaken to say that the distinctness of Bob1 and Bob2 is a matter simply of the distinctness of two mental substances or the distinctness of two mental states or properties. Their distinctness is a matter of the distinctness of a substance-with-its-properties from another substance-with-its-properties. This sort of distinctness is primitive; there is no further *in virtue of which* in virtue of which Bob1 and Bob2 differ beyond or beneath *Bob1's-having-property-Q* and *Bob2's-having-property-Q* (even where these are twin properties). What the Identity of Indiscernibles says, on a Features reading, is this:

> (II:F) For any X and Y, if X's features are identical to Y's features, then X is numerically identical to Y.

> *Tropes*: a property is not an abstract object, an exemplification of an abstract object, or a feature; it is an entity that exists on its own, can with other properties make up a thing, and has second-order properties which are also tropes; the relation (say) between a red ball and its redness is that redness is a member of a collection of tropes and "ball" refers to that collection, not to something that has the property of being red or that exemplifies the abstract object redness.

A substance is typically understood along the lines of being something that is not itself a property or a collection of properties, has properties, has an essence, and can endure over time. Tropes are not substances. What the Identity of Indiscernibles says, on a Trope reading, presumably is something like this:

> (II:T) For any X and Y, if X is composed of a one-or-more-membered bundle of tropes and Y is composed of a one-or-more-membered bundle of tropes, and for any trope Z, Z is contained in the X-bundle if and only if Z is contained in the Y-bundle, then X is numerically identical to Y.

It seems clear that if properties are exemplifications, then (II:E) is true; if properties are features, then (II:F) is true; if properties are tropes, then (II:T) is true.

I shall assume here that one or the other of these views of properties is correct, or that at least if there is some other defensible view of the nature of properties, the Identity of Indiscernibles is true if that view of properties is correct. On that assumption, the Identity of Indiscernibles, for present purposes, can be understood as follows:

> [For every property Q, it is the case that, of X and Y, one has it if and only if the other has it] entails [X = Y] = [(II:E) or (II:F) or (II:T)]

and, I take it, is true.

If both Leibniz's Law and the Identity of Indiscernibles are true, so is their conjunct, which we can call *The Law of Metaphysical Indiscernibility* (it has been the having of properties, not the accessibility of properties had, that has been relevant to the discussion):

> (LMI) [X is numerically identical to Y] if and only if [For any property Q, X has Q if and only if Y have Q].

Leibniz's Law, the Identity of Indiscernibles, and hence the Law of Metaphysical Indiscernibility, are either true in all possible worlds, or under all possible conditions, or in all possible circumstances, or they are false. Thus, if they are true, they are necessarily true. So if LMI is true there is a logically necessary connection between lack of property difference between X and Y and presence of numerical identity between X and Y: necessarily, X = Y if and only if for all Q, X has Q if and only if Y has Q.

Distinctness between two items X and Y, then, will for an Exemplarist be a matter of X and Y having different exemplifications, whether of the same or of different abstract properties. For a Featurist, it will a matter of X and Y having different features—i.e. of X being one substance-with-property and Y being another substance-with-property. A Tropist will ground distinctness in bundles of tropes in the bundles having different members; difference among the tropes themselves presumably will be basic. There won't be something further, beyond different in Exemplifications, Features, or Tropes, in virtue of which two distinct items are distinct. Identity conditions are absence-of-distinctness conditions. There is no reason why minds X and Y cannot differ in their exemplifications of conscious properties, or the conscious features that they have, or the mental tropes of which they are made. I don't believe that minds can be made up of tropes, or that properties

are tropes. But if I am wrong about that, then presumably a Tropist can distinguish between minds in the manner indicated.

We started this section with one or two items in a box. But nothing in the development of our discussion of the proposed doctrine of metaphysical identity conditions requires that distinct things need be in boxes or capable of being in boxes. Nothing in it requires that distinct items be physical.

The net result of our discussion in this section can be put as follows. We define the term "characteristic" as follows: Q is a characteristic if and only if Q is an exemplification, a feature, or a trope. Then person, soul, or mind X is metaphysically distinct from person, soul, or mind Y if and only if, for some characteristic Q, it logically possible that, of X and Y, only one has Q. Depending on what the characteristics are in a particular case, there may be entailments concerning presence or absence of difference in spatial location. The conclusion to be drawn is that there are various ways in which to construct a doctrine of the metaphysical identity of minds on which multiple minds is a legitimate option.

If one should reject even this doctrine of metaphysical individuation, there remains this doctrine: person, mind, or soul X is metaphysically distinct from person, mind, or soul Y if and only if there is a time T such that, of X and Y, one began to exist at T and one did not begin to exist at T. This doctrine entails that it is impossible that two minds come to exist at the same time, and it seems to me quite possible that two minds do come to exist at the same time. My point is simply that what we might call the doctrine of the *Uniqueness of First Time Slots for Minds* is an epistemically possible alternative account of metaphysical individuation among minds.[16]

My answer, then, is that it is not wise to follow the feeling that only physical or physically-based items can be distinct—not even if for some reason one thought one could limit this doctrine to things other than God. The metaphysical doctrine(s) of individuation on which the feeling rests, or which articulate that feeling, are false in virtue to failing to provide all of the sufficient conditions for items X and Y being distinct. But there is an interesting epistemological doctrine that one might think of even after rejecting the metaphysical view that only occupying different points in a spatial or spatio-temporal matrix can be the basic ground of distinctness.

Consider this claims concerning epistemological identity conditions, where "X" and "Y" range over what theists call created persons.[17]

> EIY. For any persons X and Y, if X's cognitive capacities relevant to person individuation [discerning that there is a person distinct from oneself] are finite, then X can identify Y as a person only if Y is embodied.

> EIX. For any persons X and Y, if X's cognitive capacities relevant to person recognition [discerning what person another person is] are finite, then X can recognize Y as the person Y is only if X is embodied.
>
> EIXY. For any persons X and Y, if X's cognitive capacities relevant to person individuation and recognition are finite, then X can identify Y as a person, or as the person Y is, only if X and Y are embodied.

The varieties of EI are relevant to communication between persons. If some variety of EI doctrine is true, that has epistemological relevance. Only insofar as (created) persons are embodied can they become aware of their kith and kin. (Perhaps some a priori argument could persuade them that other persons exist, but this would not be enough to bring about societies.) This would be relevant to theodicy; a reason would thereby be given for God creating *embodied* persons. Alternatively, there are weaker claims that are in some ways similar to the varieties of EI doctrines. Perhaps only certain kinds of communication, certain ways of relating, are possible given embodiment; that much could be true without any of the EI doctrines being exactly right. If there was significant value in persons communicating and relating in those ways—in particular if certain kinds of immaterial souls flourished most fully if they engaged in them—then there would be theological relevance and significance for theodicy in embodiment. These things, of course, have been suggested though I'm not aware of their having been rigorously explored. But we have moved too far from the main line of argument.

5. MATERIAL SUBSTANTIVALISM

Materialism regarding persons is the ontology of choice these days in philosophy. Some philosophers who have embraced it also hold that persons are immortal.

One version of this perspective holds that a human person simply is her body. Of course it isn't clear that she is identical to all of her body. Haircuts and manicures don't do much to threaten personal identity. But at least there is some (proper) part of her body—perhaps her brain in a vat, or her central nervous system, or whatever—such that its existence is necessary and sufficient for, and identical to, her existence.

Another articulation of this perspective affirms that a person is a material life. Being a material life is being a (typically complex) physical substance that is alive in the right way. A person is a material thing that is alive in the right way for it to be a person. A person at time T1 is numerically identical to person S* at time T* if and only if the living material substance that S is

at T is identical to the living material substance that S* is at T*. This requires that S exist continuously from T through T*. (Such phrases as "Person S at time T" are *not* to be read as "The stage or slice of S that exists at T" or as "S-at-T"—here or below in this section.)

As we all know, what appears to happen to human bodies is that they die—they cease to support the sort of life that is the life of a person. This makes the prospects for immortality seem bleak for a material life theory of persons. Bodies might never have died, but the fact is they do die. So things might have been such that persons both were material beings that were alive in the right way and lived forever. But things aren't that way.

It is *logically possible* that things seem as they do but bodies alive in the sort of way that makes them persons do not die. It is logically possible that God whisks away our bodies while we are still alive and replaces them with nonliving bodies, so that in fact we don't ever die. It is logically possible that God does this with such eminent skill that we never discover this curious fact about our world. This possibility is intended to show that if the material life theory of persons is true, and one believes that persons are immortal, one's position is not logically inconsistent with how things appear to go with regard to human bodies. There is maneuvering room for the no-identity-over-time-gap material lifer even given the fact that most bodies appear to suffer cremation or decay, a few are mummified, and at least one has the status of Jeremy Bentham's body. Not a lot of room, but a little. But this is a "just so" story, showing only sheer logical consistency with things appearing as they do and material lives being immortal.[18] The evidence is that things are as they appear, and if they are as they appear, the prospects of immortal material life look magnificently bleak.

Alternatively, perhaps every human life is preserved dormant in a tiny material speck—a seed or pellet—in which that life resides dormant until resurrection day. A material substance exists continuously that once participated in an active life of the sort that one must have in order to be a person. After death the speck carries that life dormantly within itself, and finally that life is awakened again. Here there is no time gap—neither the life nor the material substance carrying it ever ceases to exist, though the former is dormant and the latter is minuscule. The material life is suspended, but not disrupted.[19]

Something at least along these lines is suggested in this passage from *Faith and Philosophy*:

> Suppose that a thousand years from now it is Time and God brings the present order of things to an end and inaugurates the new age. But how shall omnipotence bring *me* back ...? This question does not confront the dualist who will say that there is no need to bring me back because I have

> never left. But what shall the materialist say? ... what can even omnipotence do but *reassemble*? What else is there *to* do? And reassembly is not enough, for I have been composed of different atoms at different times. If someone says, "If, in a thousand years, God reassembles the atoms that are going to compose you at the moment of your death, those reassembled atoms will compose you," there is an obvious objection to his thesis. If God can, a thousand years from now, reassemble the atoms that are going to compose me at the moment of my death—and no doubt He can—He can also reassemble the atoms that compose me right now. In fact, if there is no overlap between the two sets of atoms, He could do both, and set the two resulting persons side by side. And which would be I? Neither or both, it would seem, and, since not both, neither. "God wouldn't do that." I daresay he wouldn't. But if he were to reassemble either set of atoms, the resulting man would be who he was, and it is absurd, it is utterly incoherent, to suppose that his identity could depend on what might happen to some atoms other than the atoms that compose him. In the end, there would seem to be no way around this requirement: if I am a material thing, then, if a man who lives in the future is to be *I*, there will have to be some material and causal continuity between this matter that composes me now and the matter that will then compose that man (van Inwagen 1995: 486).

The argument here seems to go like this. Suppose there is a man Irving whose body at time T1 is composed of the IT1 particles and whose body is composed at time T2 of the IT2 particles and that, by the time T2 rolls around, the IT1 particles are scattered around the world. Irving dies at T2 and after while judgment day comes. It is logically possible that God on that day brings together all the IT1 particles in such a manner that they compose one living body and also brings together all the IT2 particles so that they compose another living body. According the Materialist Substantivalism, Irving is identical to the body composed of the IT1 particles but also identical to the body composed of the IT2 particles, but the body composed of the IT1 particles is distinct from the body composed of the IT2 particles. Hence Materialist Substantivalism entails that a proposition of the form [(X= Y) and (X = Z) and it is false that (Y = Z)] is true. No such proposition can be true, so Materialist Substantivalism is false (indeed, it is a necessarily false doctrine).

I happily assent to the critique offered in this passage against the view that a person is identical to all or some privileged part of her body. This view entails that Fission in possible, and Fission isn't possible. So the view entails that something is possible which isn't, and so the view itself isn't possibly true. So be it.

So we are offered the somewhat different view that a person is a material life—a living material substance whose life is of the right sort for her being a person. The material substance can be (and is) composed of different atoms at different times, but the life is always supported by[20] (more properly, constituted by the activities of) a continuously existing and typically complex material substance.

Consider, then, the person Sue. Suppose Sue is born in 1928 and dies in 2000. Shortly after her death, Sue is cremated, and the resulting dust is sifted through a fine sieve. That, in turn, is run though a grinding machine that reduces everything put through it to a very fine powder. Still, on a material life view, if Sue is immortal, there must be at least one material seed or pellet in which Sue's life is dormant. This is at least less plainly against the appearances than that her body does not die or decay. We don't see it happen, but we don't see it *not* happen either, as we do see it not happen that the ice cube stays frozen when dropped into boiling water.

This view is problematic in a familiar manner. If one such seed or pellet is possible—a being that survives Sue's sifting—so are two, or some larger number. On a *keep it as simple as possible* rule, suppose there are only two. Then Sue's life is dormant in two seeds or pellets. If the life in one of these can be awakened, so can the life in the other. If the life of one of these being awakened results in Sue living again, then the being awakened of both results in something one might be tempted to call two living Sues existing—Sue* and Sue**.

The same life that is dormant in Sue* is dormant in Sue**. The seed that is Sue* and the seed that is Sue** have equal metaphysical claim to continue the existence of the material life that was Sue. Each was part of her body; each was produced by the same cremation, sifting, and grinding as produced the other working on the same body; in each Sue's life lies dormant. Call this *Sue followed by Sue* and Sue** scenario* the *2S-scenario*.

This scenario is possible if the similar single seed scenario is possible—if the sort of process described can occur (the death, the cremation, the grinding and sifting) in such a manner as to result in just one seed preserving Sue's life dormant. That scenario is possible. So we have this argument.

1. On a material life view of persons, the 2S-scenario is possible.

2. On the 2S-scenario, given the material life view of persons, Sue is identical to both S* and S**.

3. It is not possible that Sue is identical to both S* and S**.

4. It is not possible that both the 2S-scenario is possible and the material life view is true.

Hence:

5. The material life view is false.

An easy answer is that God, wanting to preserve Sue, does not allow there to be two such seeds. The 1-5 argument does not deny this—it simply requires that the 2S-Scenario is *possible*, not that God would (or would not) allow it to obtain.

One can put the point in a paraphrase of a part of the passage most recently quoted. Concerning the surviving of two seeds, in both of whom Sue's life is dormant, we can say:

> "God wouldn't do that." I daresay He wouldn't. But if He were to preserve either seed, the resulting woman would be who she was, and it is absurd, it is utterly incoherent, to suppose that her identity could depend on what might happen to some other seed other than the seed that composes her.

There is a further argument in the neighborhood.

1. On the material life view, it is possible that Sue's life continues in (the life dormant in) seed Sue*.

2. On the material life view, it is possible that Sue's life continues in (the life dormant in) seed Sue**.

3. If Sue's life is continued in (the life dormant in) Sue*, but could have been continued in (the life dormant in) some other seed Sue**, then the relation of identity between Sue and Sue* is contingent.

4. If Sue's life is continued in (the life dormant in) Sue**, but could have been continued in (the life dormant in) some other seed S*, then the relation of identity between Sue and Sue** is contingent.

5. On the material life view, it is possible that the relation of identity between Sue and her life-continuing seed be contingent.

6. On the material life view, the relation of identity between Sue and her life-continuing speck is that of personal identity.

7. Metaphysical identity is necessary, not contingent, and personal identity is necessary, not contingent.

Hence

8. The material life view is false.

The 1-8 argument is perfectly compatible with God never allowing Sue to give rise to two dormant-life-supporting seeds. Professor Van Inwagen is fully aware of this objection. He gives us what may be, in good spirits, be called the Petrine Assurance that Fission of human persons, if they are as he tells us they are, is impossible. Here (but not in the similar case noted in his own critique of another version in the passage quoted) our modal intuitions fail us. I understand the assurance. I just don't believe it.

The sum of this section, then, is that material life Materialist Substantivalism entails that Fission is possible, and it still isn't, and any view that entails that a proposition that cannot be true can be true cannot itself be true.

It is perhaps worth putting this last point as follows. The argument form *P entails Q, and Q is false; so P is false* deservedly is widely recognized as a paradigm of validity. It doesn't matter whether or not one of the propositions that replace "P" or "Q" is the argument form is a modal proposition—whether or not one is of the form *Possibly, Q*. Put in other terms, *P entails Possibly Q*, and *Possibly, Q is false; so P is false* is as splendidly valid a form of reasoning as is *P entails Q, and Q is false; so P is false*. One may as legitimately offer it one's full assent.

6. SEQUENTIALISM

Sequentialism is the view that a person at a time is composed of a collection or bundle of one or more states, and over time is composed of a sequence of such bundles. The basic idea here is problematic. A person Kim, who spends the morning baking bread, can either go to the movies or organize her library this afternoon; either way she will be the same person. But a series of states composed of bread-baking-movie-going (BBMG, for short) is not the name sequence as a series of states composed of bread-baking-library-organizing (BBLO, for short). Sequentialism says that Kim is identical to a sequence of states. She can numerically identical either to a BBMG sequence (S1) or to a BBLO sequence (S2), but not both. But by PNMI, if a sequence S1 is identical to a sequence S2, then necessarily S1 is identical to S2. If Kim can be identical to S1 and Kim can be identical to S2,

then S1 can be identical to S2. But S1 and S2 have different members, and so they cannot be identical. A sequence just is a series of members; add or subtract members and you have another sequence. The reason Kim as a person can either go to the movies or organize her library is that she is a substance that can retain substantival identity throughout either activity; it is the fact that her identity resides elsewhere than in numerical identity with a sequence that gives rise to the fact that her future can go either way. Sequentialism is not entitled to borrow possibilities that arise only on non-sequentialist ontologies.

On a Sequentialist type view, the states that compose a sequence may be said to be mental, physical, or some of each. The states may be viewed as states of things, or things may be viewed as themselves composed of states. On the Sequentialist view, each bundle or collection is a part or stage or momentary constituent of a person. A person is the whole series of bundles or collections, and at no one time is the whole person present. In contrast, on a view which takes a person to be a substance, all of a person exists at each time at which the person exists.[21] The states in a bundle at a time must satisfy some relation R in order for them to compose a bundle,, and the bundles over time must satisfy some relation R* in order for them to comprise a sequence of bundles. Sheer simultaneity, for example, is not sufficient as a value for "R" nor is temporal succession sufficient as a value for "R*". Varieties of Sequentialism differ as they offer differing accounts of what should give concrete sense to "R" and "R*". An alleged charm of Sequentialist views is that if some version of Sequentialism is true, a person can continue to exist after the demise of her body. The person is alleged to be identical to a sequence of states, and thus the person continues to exist so long as the sequence does. If the sequence must be associated with a substance, then when a body with which it is associated dies, the person continues if the sequence is somehow switched to another substance. If substances or things are just sequences of states, then perhaps the series that constitutes the person can continue even when it no longer contains states that compose a body. A Sequentialist will give some account or other of how a series of states must be constituted in order for it to be a person-constituting sequence. The following remarks are intended to give some notion of how such accounts can go.

6.1 Sequentialism and Structure

Let us say that if an item X is more determinative than is any other entity concerning the existence of an item Y then it is Y's *main existential determiner (MED)*; if X is more determinative of Y's sortal (kind-determining) properties than is any other entity, it is Y's *main sortal*

determiner (MSD); if X is more determinative of Y's individuating properties (properties that make Y an individual member of its kind) than is any other entity then it is Y's *main individual determiner (MID)*. This suggests a variety of doctrines regarding what makes a series of states a person-constituting sequence.

A temporal series is an MED series if its successive members (two or more) are such that each of its members after the first (if it has a first) is the MED of its immediately succeeding member. A temporal series is an MSD series if its successive members (two or more) are such that each of its members after the first (if it has a first) is the MSD of its immediately succeeding member. A temporal series is an MID series if its successive members (two or more) are such that each of its members after the first (if it has a first) is the MID of its immediately succeeding member. Then we get these suggestions as to what makes a temporal series a sequence.

iia) a series is a sequence if it is an MED series.

iib) a series is a sequence if it is an MSD series.

iiic) a series is a sequence if it is an MID series.

iv) a series is a sequence if it is an MED and MSD series.

v) a series is a sequence if it is an MED and MID series.

vi) a series is a sequence if it is an MSD and MID series.

vii) a series is a sequence if it is an MED and MSD series.

viii) a series is a sequence if it is an MED, MSD, and MID series.

Perhaps the strongest candidates for being a sequence will be those that are MED, MSD and MSI.

6.2 Content

These structural considerations presumably are central to what makes a series be a sequence. Another consideration is also relevant, namely content—what *sorts* of states are included in a person-constituting sequence? must the members of a person-constituting sequence relate together to comprise a coherent narrative content? assuming that one can talk about a

sequence possessing dispositions, what sort of dispositions (if any) must be included, and with what sort of stability?; and so on. One might hold that:

(S1) S is a person-constituting sequence only if it is an MED, MSD, and MID series that contains conscious states.

We can fine-tune things further. An intention to order tea will require, on Sequentialism, that there be a series of successive and properly related conscious states. Each state must have the right features for it to be intention-constituting, and for the states to constitute a tea-ordering intention rather than some other; call these *constituting features*. The same sort of thing will be true for hopes, fears, desires, pleasures, pains, and so on, as well as for thoughts and choices. Further, intentions and other mental states play causal roles; they have *causal features*.

The relevant idea, then, is something like this. Let KM be a list of all the sorts of mental states there are. Let KM1 be one such sort. That a series of states constitute a mental state of a KM1 sort will require that each have the right sort of constituting features and causal features. The same, of course, holds for every other member of KM. A person-constituting sequence will have to contain sub-sequences whose members satisfy the condition of having the right sorts of constituting features and the right sorts of causal features to make up conscious states of various kinds. A person-constituting sequence must contain various sub-sequences that satisfy the conditions for being various mental states, its composing states and sub-sequences having appropriate constituting and causal features; it must be *mentally composed*. We can thus fine-tune things a bit by replacing (S1) by:

(S2) S is a person-constituting sequence if it is an MED, MSD, and MID series that is mentally composed.

By now, I hope, the rough idea behind Sequentialism is clear—something like as clear anyway as it can get without some actual detailed version being presented. Sequentialists hold that there is *some account or other* of relations that can be used to replace "R" and "R*" so that the result is a defensible, even correct, account of what a person is.

Suppose, then, some Sequentialist person X exists at time T. At T1, persons Y and Z exist such that Y is one bundle and Z is another bundle, both of the sequentialist person-at-a-moment-constituting sort—both Y and Z can be members of a mentally composed sequence. Suppose, further, that both Y and Z are related to the bundle that constitutes X at T in such a fashion that, on Sequentialist grounds, both are continuations of the person that X was at T. Further still, with the addition of Y and Z, the sequence that

is identical to X branches into two distinct sequences, one new sequence beginning with Y and one new sequence beginning with Z. Thus both the Y-sequence and the Z-sequence have, on Sequentialist terms, equal metaphysical claims to be continuations of X (i.e., of the sequence with which X, according to Sequentialism, is numerically identical).

The situation now is this. Sequence Y is a person, and meets the conditions for being a continuation of person X. Sequence Z is a person and meets the conditions for being a continuation of person X. X is a single person, and Y is not the same person as Z. Thus it cannot be the case both that Y is a continuation of X and Z is a continuation of X. Were Y to exist and Z not exist, Y would be the continuation of X as a person. Were Z to exist and Y not exist, Z would be the continuation of X as a person. As things stand, with both Y and Z in existence, X ceases to exist at time T.

The result is that Sequentialism entails that Fission is possible, and hence Sequentialism is false. If Sequentialism is false, then Materialist Sequentialism, Idealist Sequentialism, and Dualist Sequentialism are false—the argument offered here is anti-Sequentialist, and is not based on some feature peculiar to a materialist, idealist, or dualist version. The point here, however, is that Sequentialism in its Materialist mode provides no refuge for the materialist.

The obvious suggestion is that R* be so described that does not allow for this. But R* is supposed to be some combination of causal and psychological relations such that, by satisfying them, two successive bundles related by R* become part of the same sequence. Such causal and psychological relations do not include *not bearing R* to more than one bundle*. "R*" is supposed to be comprised by relations that link bundles in a person-constituting-over-time way. Nothing in a value for "R*" so comprised will include *not bearing R* to more than one bundle*. We spent some time in characterizing the sorts of considerations relevant to constructing a version of Sequentialism that appealed to sequentialist-internal constraints—that developed what Sequentialists think of as the sequentialist insight. Call the property *not bearing R* to more than one bundle* the *uniqueness property*. The uniqueness property is not included in those properties. It is a negative property, a matter of a bundle not bearing a certain relation to two other bundles.

If there are no negative properties then reference to such supposed properties can play no part in an ontology of persons. Negative properties seem especially unwelcome in a robustly materialist world, but let that pass. The point, then, is that in adding reference to a negative property to the Sequentialist account of the metaphysical identity conditions of persons, one adds a new sort of element to the account. It does not arise intrinsically from the other elements in the account. (This is so, whether one likes my

description of the addition in terms of the positing of negative facts or not.) It is pasted on in order to fend off an objection.

Is the addition legitimate? To tell, begin by asking whether, given the original Sequentialist account of the metaphysical identity conditions of a person, there can be Fission of a person. The answer is that plainly Fission is possible, given the original Sequentialist account. The addition does not make Fission of a person impossible, as if one could save a view from refutation by adding to the view the phrase "this view can't be refuted." Adding the suggested qualification is tantamount to granting that the Fission objection is devastating to the original view. It does nothing, of course, to show that Fission cannot occur. What it admits is that if Fission does occur, you don't have a continuation of the original person, even though, on Sequentialist terms, you should. Further, the addition adds no further positive element to the view, and in that sense it doesn't revise the view in the light of the criticism. One is not thereby presented with a new, more illuminating account. It simply admits that the account offered gives us conditions for numerical identity, or the continuation of, persons that are not sufficient—because they can all be satisfied and there not be any person who is numerically identical to, or the continuation of, an earlier person. It says that person X at T will be numerically identical to, or will have her life as a person continued by, someone that meets certain conditions, unless two persons meet those conditions, in which case neither later person is numerically identical to, or a continuation of, the earlier person. Usually, overdetermination makes the result that is overdetermined more sure to occur. Here, it guarantees that the result does not occur.

There are two further objections to Sequentialism to be considered, and both apply even if one does not press the objection noted in this section. One objection is presented in the next section. The other is addressed in the latter part of our final section.

7. ANOTHER OBJECTION

Suppose person X exists at time T. At time T1, God creates a person Y, following this recipe: let Y at T1 be a copy of what X will be at T1. God can make Y in accord with this recipe, even if God has no interest in destroying X.

Consider again, however, a curious consequence of both Material Substantivalism and Sequentialism. On both views, it is the case that a person X can exist at time T, two distinct persons Y and Z can exist at time T1, it can be the case that were Y and not Z to exist at T1 the person Y would be numerically identical to (or would be a continuation of) the person

X, and were Z and not Y to exist the person Z would be numerically identical to (or would be a continuation of) the person X. Nonetheless, the existence of both Y and Z prevent anything that exists at T1 being numerically identical to (or a continuation of) X. Y's existence prevents Z form being numerically identical to (or a continuation of) X. Z's existence prevents Y from being numerically identical to (or a continuation of) X. Yet Y does nothing whatever to Z or to X—not even think an unkind thought about either of them. Further, Z does nothing whatever to Y or X—not even wish either of them the slightest misfortune. While it would be too much to claim that Y or Z, singly or together, murdered X or otherwise harmed him, their sheer mutual existence annihilates X. Their co-existence removes X from existence. The objection is: that cannot happen. It can't be the case that the existence of two persons—who do not in any fashion, directly or indirectly, causally impact or act upon another person—is such as to bring that other's person's existence to a dead end.

Consider, then, this claim:

> *The Causal Inefficacy Principle*: if X exists at T, whether X also exists at T1 can be determined negatively by the sheer existence of two other persons Y and Z at T1, even though it is false that either Y or Z, singly or together, exercise the slightest causal impact or act in any way whatever regarding X, X's existence, or X's well-being.

The objection we are concerned with here is this.

1. The Causal Inefficacy Principle is false.

2. Both Material Substantivalism and Sequentialism entail that the Causal Inefficacy Principle is true.

Hence:

3. Both Material Substantivalism and Sequentialism are false.

8. THE FUTILITY OF CORCORAN'S AXIOM

In Corcoran 2001, Professor Corcoran raises an objection to the sort of argument we have been developing here. A bit of background will be helpful. A proposition P is *narrowly logically necessary* if and only if the denial of P is, or is reducible to, a proposition of the form [Q and not-Q]; such propositions are syntactically or structurally necessary, or formally necessary. A proposition P is *broadly logically necessary* if and only if,

while P is not formally logically necessary, nonetheless it is false in every possible world. Such propositions are necessarily true in virtue of their meaning; they are semantically or informally logically necessary. *If Sarah smiles, then her smile has shape* is informally necessary. Our objection to Material Substantivalism, for example, has been that it entails, regarding a narrowly logically necessary truth, that it is false. It entails that *It is impossible that X be identical to both Y and Z, where Y is distinct from Z* is false, whereas that proposition is a necessary truth.

Corcoran, in effect, appeals to a third brand of logical necessity which we can call *theistic necessity*. Suppose that it is a necessary truth that God exists. Suppose there is also a set of propositions about God that are necessary truths. Finally, suppose that some proposition P is such that it is true only if it is false that God exists or if some member of the set of necessary truths about God is false. Then P is *theistically necessarily false* and not-P is *theistically necessarily true*.

Corcoran's own view is expressed as follows:

> ... here is a way resurrection could go. It seems possible that the causal paths traced by the simples composing my body just before death can be made by God to fission such that the simples composing my body are then causally related to two different, spatially segregated sets of simples. One of the two sets of simples would immediately cease to constitute a life and come instead to compose a corpse, while the other would either continue to constitute a body in heaven or continue to constitute a body in some intermediate state. In other words, the set of simples along one of the branching paths at the instant after fission fails to perpetuate a life while the other set of simples along the other branch does continue to perpetuate a life. If this is at least possible, as it seems to be, then we have a view of survival compatible with the joint theses that human persons are essentially physical objects and that such objects cannot enjoy gappy existence [i.e., existence over time gaps] (2001: 210).

Here is another variety of Material Substantivalism, though Corcoran says that if this view won't work, he can switch to a closest continuer view—that the existence of a person X at T is continued by whatever person at T1 most resembles X at T (or what X at T would become at T1 if something numerically identical to X were to have survived), provided there is only one closest resembler and provided some person closely enough resembles X. Closest continuer theories are varieties of Sequentialism.

Materialist Substantivalists are generous with their stories. This time we are told that perhaps God brings it about that the body one is identical with just before death at time T causes two bodies to exist at T1, one of which is a corpse and the other of which is numerically identical to one's body at T.

Here, the body is not whisked away and replaced by a corpse and one does not have to bring in seeds or pellets. But of course a familiar objection looms. If God can cause two bodies as suggested by Professor Corcoran, God can cause three bodies, one corpse, one body Y that goes to heaven, and one body Z that goes to an intermediate state, or perhaps Y and Z both go to heaven, or Y and Z both go to an intermediate state. But then both Y and Z will be the person X who existed at T, but they are not identical, and so X cannot be identical to both of them.

Professor Corcoran wants to somehow eliminate this objection. To turn to the specifics of how he endeavors to do this, suppose it is necessarily true that *(K) God exists, God is essentially omnipotent and omniscient, God is essentially morally perfect, If a person exists then God created her, and It is necessarily wrong to create a person and then allow that person to cease to exist; So God would never create a person and allow that person to exist*. This complex proposition combines the claim that God exists with the relevant supplemental set of propositions needed for Corcoran's Axiom. If possible world talk is permissible here, the idea is that in no possible world can it be true both that God exists and that a created person ceases to exist. The conditional *If God creates a person then that person never goes out of existence* follows from (K), and hence is necessarily true if (K) is necessarily true.

We can now state Corcoran's Axiom:

(CA) (K) is true, and it is logically impossible that (K) is true and Fission is possible.

That Fission occur is theistically impossible—inconsistent with (K)'s truth, where (K) is proposed as a necessary truth that involves certain claims concerning God, God's nature, and the moral propriety of letting a person cease to exist. The claim is that if God has logically necessary existence, if God is essentially omnipotent and omniscient, if God is essentially morally perfect, and if there are no possible conditions under which it would be permissible for God to allow a person to cease to exist, then it is not possible that Fission occur.

These are controversial claims. The view that God has logically necessary existence is currently popular among philosophical theists, but it is also generally agreed that the ontological argument is unsuccessful. The controversial basis of the popularity of the view seems to reside in the idea that theism is best articulated along the lines of Anselmian or perfect being theology, a belief neither silly nor self-evident. A robust libertarian will deny that any being can be essentially morally good, though an omnipotent, omniscient being can decide to never choose or act wrongly and know the

she will never go back on her choice. It isn't clear that there are no possible conditions under which God could not rightly cause a person's existence to cease—suppose Irving will be in such agony that he cannot do anything other than exist with a pain-filled consciousness forever if he continues to exist at all. Presumably the reply will be this: either there are no possible conditions such that, if a person is in them, it would be proper for God to cause or allow him to cease to exist, or else there are such possible conditions and it would never be proper of God to allow any of those conditions to obtain, so (necessarily) God never allows them to obtain. (This entails that those who accept an annihilationist view of hell embrace a view that is theistically necessarily false.)

Two points are obvious. One is that (K) contains a set of controversial claims. I'm dubious about all of them, save the thesis that God is essentially omnipotent and omniscient. But this isn't the place to discuss them. Second, Materialist Substantivalism and Sequentialist accounts of persons are such that, were they true, it would be possible that an X be identical to (or continued by) both Y and Z, though Y and Z are distinct. That is not possible—it is formally or narrowly logically impossible. If we then add (K), this does not change what Materialist Substantivalism, for example, entails. So Materialist Substantivalism (MS) entails that Fission is not formally or narrowly logically impossible (i.e., entails that it is formally or narrowly logically possible) and (K) entails that it is theistically impossible. The conjunct [(K) and (MS)] entails all of what each of its members entails, and so entails both that:

(KMS1) Fission is not formally or narrowly logically impossible.

(KMS2) Fission is theistically impossible.

Since (KMS1) is still false, and (MS) entails (and thus [(K) and (MS)] entail) (KMS1), [(K) and (MS)] is false.

There is another fatal flaw. Fission is not possible if CA is true. But something else is. Consider again the scenario in which we have X at T, and two possibilities, namely that Y at T1 is numerically identical to, or continues the personal life of, X at T, and that Z at T1 is numerically identical to, or continues the personal life of, X at T. If CA is true, it is impossible that both Y and Z exist. If CA is false, Y and Z can both exist, and were Materialist Substantivalism or Sequentialism true, both Y and Z would be numerically identical to, or continue the personal life of, X. Since that is impossible, Y being distinct from Z, Materialist Substantivalism and Sequentialism entails that something that cannot be true is true. The idea is

that CA comes to the rescue by making it impossible that Y and Z both inhabit the same possible world.

Note, however, what even the truth of CA does not prevent. Even if CA is true, (i) it is possible that Y at T1 be numerically identical to (or continue the existence of) X at T and (ii) it is possible that Z at T1 be numerically identical to (or continue the existence of) X at T. CA rules out [(i) and (ii)] but it does not rule out [either (i) or (ii)]. God can decide which way things go—whether (i) or (ii) obtains—and God can pick either (i) or (ii). So either (i) or (ii) can obtain. If (i) obtains, then X is numerically identical to (or X's personal life is continued by) Y. But it is contingent that X is identical to (or X's personal life is continued by) Y rather than Z. If (ii) obtains, then X is numerically identical to (or X's personal life is continued by) Z. But it is contingent that X is identical to (or X's personal life is continued by) Z rather than Y. So the identity of X with Y, or of X with Z, whichever happens to obtain, is contingent on a Materialist Substantivalist or Sequentialist view. This is inconsistent with PNMI.

Of course there are more objections and replies, but perhaps this is enough for the present. To provide a challenge to the view that a turn to materialism is necessary or appropriate on the part of philosophers who embrace the doctrine of the resurrection of the body.[22]

ENDNOTES

[1] Clarke 1978: 722. Clarke is opposing a view on which God takes a naturally mortal soul and in saving it grants to it a superadded immortality not originally present.

[2] This isn't my idea, or Professor Plantinga's, but this essay isn't a defense of mind-body dualism or idealism.

[3] In what follows, references to "logical necessity" and "necessity", unless otherwise specified, will be to what many call "broad logical necessity" (which I will typically construe for convenience to include narrow logical necessity). I take these to be, as Aristotle remarked regarding the Principle of Non-contradiction, laws of both thought and things. Those skeptical in this regard may consult Plantinga 1974.

[4] My concern here will be only with Materialist Substantivalism and with Sequentialism. This is quite enough for one paper. I will not discuss the view that a person can exist at time T, cease to exist at time T1, and then exist again at some time later than T1. I take it that persons cannot take vacations from existence and then return. I will not discuss Constitutionalism, which holds that when one buys a statue of David that is made of marble, one buys two things—the statue "David" and the composing material "Marble." I deny that one gets a bargain, two things for the price of one, in such circumstances.

[5] Stendahl 1965, which contains Oscar Cullman, *Immortality of the Soul or Resurrection of the Body*, which was published in 1956. "Or" is not here used in what logicians call the inclusive sense.

[6] Of course there are also exegetical concerns. Regarding these, see Cooper 2000.

[7] Those who believe that God is both immortal and eternal can add "or X is eternal" at the end of (I) and the later (I*). Those who hold that God is everlasting will view divine immortality as a matter of God's existing without beginning or end, and add that to (I), which is fine with me. There are Hindu monotheists who hold that God created all non-divine persons, that they wholly depend for their existence on God's sustaining them, and that God has always been sustaining them in existence—that non-divine persons exist dependently and beginninglessly as well as endlessly, all by divine courtesy. They too would count as immortals.

[8] Two examples: Reichenbach 1978, and Merricks 2001.

[9] If satisfying (I*) is not enough for the satisfier to be immortal, or even if it is barely enough, it is more than is required by talk of having immortality in the memories of others, or in the traditions of a country or some smaller community, or in the effects of one's actions, or in the endurance of the elementary particles of which one's body was made. Anyone having only such "immortality" has no immortality at all—nothing that does not satisfy something at least as strong as (I*) has immortality.

[10] As we have noted, discussions of life after death often suggest that there is an incompatibility between resurrection and immortality. Here is another consideration in favor of the conclusion that there is none. Consider two notions of immortal life.

L1. If X has immortal life then X lives forever and there is a time T such that, at T and after, X is not embodied.

L2. If X has immortal life then X lives forever and there is a time T such that, at T and after, X is embodied.

If we specify that "X" ranges over only items that have been embodied and whose body has died, then L1 does not require a resurrection and L2 does require a resurrection. One could not have immortal life in virtue of satisfying both of the conditions—there would have to be a time T at and after which one was, and was not, embodied. But it is perfectly consistent to offer this characterization:

L3. X has immortal life if either X lives forever and there is a time T such that, at T and after, X is embodied, or there is a time T such that, at T and after, X is not embodied.

Being immortal is a matter (roughly) of living forever, whether embodied or not. The Christian doctrine of the resurrection is a doctrine to the effect that we live forever, and that there is a time T such that, at T and after, we are embodied. This is a version of the view that we are immortal, not a denial of our immortality.

[11] Of course a defender of the opposite view could appeal here to an alleged weakness in our powers to see what is really possible, suggest that while the earth may have done without us for millions of years, the universe is far larger than the earth, and otherwise make room for the idea that each of us enjoys logically necessary existence after all. Or one could hold that we are extremely fortunate in that, while our non-existence is logically possible, nothing in the actual world could stamp us out. Neither Jain nor Platonic texts contain, however, anything like a satisfactory argument for this position, and so far as we can see, our non-existence is logically possible. Contingently natural immortality is not something whose presence or absence, I take it, can be discovered by reflection alone.

[12] Philosophically, the backing for (C1) seems to be something like these claims:

(C1a) Necessarily, if X might not have existed, then X exists dependently.

or, more exactly:

(C1a') Necessarily, if X might not have existed, and it is possible that X be dependent for existence on something else, then X is dependent for existence on something else.

combined with:

(C1b) The only being on which *there being things that might not have existed and can have a cause* could depend is a being that either cannot not exist or at least cannot depend for existence on another, and has the power to make it the case that *there being things that might not have existed and can have a cause*.

and:

(C1c) A being that either cannot not exist, or at least cannot depend for existence on another, and has the power to make it the case that *there being things that might not have existed and can have a cause* has properties sufficient to make it God.

Whatever is the case with any necessarily existing abstract objects there may be, concrete items that do exist might not have existed. By (C1), they exist dependently. One could hold that at least some of them exist dependently only contingently—they happen to exist dependently, but they might not have existed dependently. Thus runs counter to the claim that:

(C1d) If X exists dependently, then necessarily, X exists dependently.

or:

(C1d') If x exists dependently, then it is an essential property of X that it exists dependently if at all.

which connects with (C1a-C1c).

[13] But see Taliaferro 1994.

[14] A person cannot be immortal without retaining personal identity over time. Any doctrine of persons that is compatible with persons being immortal must be compatible with persons retaining their identity as such over time. Of course retaining identity as *some person or other* over time would not be enough. One would have to be the same person each time as the person was previously in order for one to enjoy personal immortality. There is no such thing as *retaining identity as some person or other* save by retaining identity as the same person.

[15] A recent distinguished possessor of something in the neighborhood of such worries is Professor Jaegwon Kim. See his contribution to Corcoran 2001. His worries concern the possibility of causal interaction between non-spatially located items, and he thinks that this isn't possible. Individuation worries are distinct, but can be extrapolated from causal worries.

[16] One could try out doctrine of spatial location if one could make sense of spatially located immaterial objects. Or one could develop an account in terms of location of action (one's location of action being the set of spatial locations at which one can bring things about or exercise causal powers), and see if this worked. I see no need for such efforts.

[17] I hope that I may be forgiven here for not pursuing how this might affect the doctrine of angels.

[18] The phrase "just so story" is used on page 51 of van Inwagen 1998, where it is also said that the body-snatching account expresses a very abstract truth about the resurrection—at least, a "some important but very abstract feature of the real thing." I'd be more inclined to say that, at best, it perhaps points to some proposition that a materialist who accepts what orthodox Christian belief concerning the resurrection entails is committed to, insofar as one can be committed to a proposition the identity of which is quite obscure, as a way of putting together in a logically consistent package the apparent facts about dead bodies, materialism, and what the doctrine referred to entails. Perhaps the story about the seed is intended to identify the very abstract feature in question, or to take a further step in that direction.

[19] Here is a passage from van Inwagen 1990 that contains the suspended/disrupted distinction: "If we use the word 'alive' in such a way that a frozen but undamaged (and revivable) organism is not alive, then I will distinguish two sorts of ways in which a life may cease: a life may be *disrupted* and a life may be *suspended.* ... a life has been suspended if it has ceased and the simples that were caught up at the moment it ceased retain—owing to the mere absence of disruptive forces—their individual properties and their relations to one another.... A life that has ceased but not suspended has been disrupted. We may be confident that the life of an organism which has been blown to bits by a bomb or which has died naturally and has been subject to the normal "room temperature" processes of biological decay for, say, fifteen minutes has been disrupted.... If a life has been suspended, it can begin again ...."

[20] "Supported by" is no more than a placeholder here; consider the following quotation: "If an organism exists at a certain moment, then it exists whenever and wherever—and only when and where—the event that is its life at that moment is occurring; more exactly, if the activity of the *x*s at *t1* constitutes a life, and activity of the *y*s at *t2* constitutes a life, then the organism that the *x*s compose at *t1* is the organism that the *y*s compose at *t2* if and only if the life constituted by the activity of the *x*s at *t1* is the life constituted by the activity of the *y*s at *t2*. This is from van Inwagen 1990: 145.

[21] The argument that God must be eternal because otherwise God never exists at once seems simply inapplicable to the view that God is an everlasting substance.

[22] Perhaps I should note that I have not assumed that we are immortal or will be resurrected. What I have assumed is something like this: for any person S and time T, if T is the time of S's death, it is logically possible that S survive her (bodily) death, or for any person S and time T before S's death, it is logically possible that S exist at T+1. A view that denies this is, for that reason, false.

REFERENCES

Clarke, Samuel. 1978. *The Works, in Four Volumes, Volume Three*. New York: Garland Publishing, Inc.

Cooper, John. 2000. *Body, Soul, and Life Everlasting*. Grand Rapids, MI: Eerdmans.

Corcoran, Kevin, ed. 2001. *Body, Soul, and Survival*. Ithaca, NY: Cornell University Press.

Merricks, Trenton. 2001. *Objects and Persons.* Oxford: Oxford University Press.

Plantinga, Alvin. 1974. *The Nature of Necessity*. Oxford: Clarendon Press.

Reichenbach, Bruce. 1978. *Is Man The Phoenix*. Grand Rapids, MI: Eerdmans.

Stendahl, K., ed. 1965. *Immortality and Resurrection*. New York: Macmillan.

Taliaferro, Charles. 1994. *Consciousness and the Mind of God.* Cambridge: Cambridge University Press.

Van Inwagen, Peter. 1995. Dualism and materialism: Athens and Jerusalem? *Faith and Philosophy* 12: 475-488.
Van Inwagen, Peter. 1998. *The Possibility of Resurrection*. Boulder, CO: Westview Press.
Van Inwagen, Peter. 2000. *Material Beings*. Ithaca, NY: Cornell University Press.
Wiggens, David. 1980. *Sameness and Substance*. Cambridge, MA: Harvard University Press.

Chapter 13

SPLIT BRAINS AND THE GODHEAD*

Trenton Merricks
University of Virginia

> Suppose one were to argue as follows: 'This thing is the Father; this thing is the Son; therefore, the Son is the Father.' It seems as good an inference as possible. And no other precisely analogous counterexample can be found in all the world.
>
> —Robert Holcot (d. 1349), *Quodlibet* 1, q. 2

1.

I believe in the Holy Trinity. So I believe that there are three divine persons—Father, Son, and Holy Spirit—and one God. Now the mere claim that there are three of one thing and one of another is logically unproblematic. After all, there is no problem with the claim that, for example, there are three musketeers and one Eiffel Tower. But the Doctrine of the Trinity says more than just that there are three divine persons and one God. It seems to say that each of these three persons *is* this one God. And so it seems to imply that each person is the same God—the one and only God—as each of the others. Thus the Athanasian Creed:

> ...there is one Person of the Father, another of the Son, and another of the Holy Spirit...the Father is God; the Son is God; the Holy Spirit is God. And yet there are not three Gods, but one God.

T. M. Crisp, M. Davidson and D. Vander Laan (eds.), Knowledge and Reality, 299-326.

So the Doctrine of the Trinity involves something like the following:

(1) The Father is a person, the Son is a person, and the Spirit is a person.

(2) The Father is not the same person as the Son.

(3) The Son is not the same person as the Spirit.

(4) The Spirit is not the same person as the Father.

(5) The Father is the same God as the Son.

(6) The Son is the same God as the Spirit.

(7) The Spirit is the same God as the Father.

There is more to the Doctrine of the Trinity than (1) through (7). (For example, (1) through (7) are silent on Who proceeds from Whom.) Nevertheless, I shall use 'the Doctrine of the Trinity'—or just 'the Doctrine'—to refer to the conjunction of (1) through (7). For my only aim is to defend the Doctrine from the charge that it entails a contradiction. And that charge is inspired by (1) through (7).

The charge is easy to motivate. Taken most straightforwardly and naturally (and given (1)), (2) implies that the Father is a person and the Son is a person and the Father is not identical with the Son. From this we get:

(8) It is false that the Father is identical with the Son.

The most straightforward and natural reading of (5) entails that the Father is God and the Son is God and the Father is identical with the Son. This implies:

(9) The Father is identical with the Son.

Obviously, (8) and (9) are contradictory. And similar reasoning easily generates contradictory statements about the identity of the Son with the Spirit and of the Spirit with the Father.

I shall defend the Doctrine of the Trinity from the charge that it is contradictory. But before presenting my own arguments, I shall examine two other ways one might try to defend the Doctrine, one involving "relative identity" and the other "social trinitarianism." My own defense does not require that these familiar defenses fail. But—I shall argue—they do fail. And, in the course of arguing for this, it will become clearer what a successful defense must do.

2.

The claim that the Father is the same God as the Son seems to entail that the Father is identical with the Son. This entailment seems to hold because, in general, A's being the same F as B seems to entail that A is identical with B.

More carefully, this entailment seems to hold for relations like *being the same dog as* or *being the same tree as* or *being the same human as* or *being the same God as*. But it does not seem to hold for relations like *being the same shape as* or *being the same size as* or *being the same height as*. We are happy to accept, for example, that A is the same height as B while denying that A is identical with B.

Indeed, A's being the same height as B not only fails to imply that A is identical with B; it also fails to imply that A is a height and that B is a height. A's being the same dog as B, however, seems to imply not only that A is identical with B but also that A is a dog and that B is a dog. From now on, when I make a claim about A's being the same F as B, I shall have in mind only those cases where this entails that A is an F and B is an F. And in those cases, it seems obvious that A's being the same F as B entails that A is identical with B.[1]

At least, it seems obvious to me. Defenders of relative identity, however, deny just this entailment. They typically insist that, for example, A's being the same tree as B does not imply that A is identical with B. (Paradigmatic relative identity theorists insist on this because, they say, there is no such thing as absolute identity to be entailed; more on this below.) Relative identity is most closely associated with Peter Geach (1972: 238-249 and 1973). But it may not have originated with him. Geach himself claims to find it in Aquinas (Anscombe and Geach 1961: 118). Moreover, Richard Cartwright reports finding relative identity endorsed by both Anselm and the Eleventh Council of Toledo (Cartwright 1987: 193).

Whatever its provenance, relative identity promises to free the Doctrine from contradiction. For relative identity tells us that the Father's being the same God as the Son does not entail that the Father is identical with the Son. If this is right, then obviously the Doctrine does not imply the contradiction noted in the previous section. And of course relative identity offers a way out of similar contradictions regarding the Son's identity with the Spirit and the Spirit's with the Father.

This theological benefit notwithstanding, I think we should reject relative identity. To begin to see why, note that John Perry (1970: 185) compares the view that identity is relative to the thesis that *being a left-handed brother of* does not entail *being a brother of*. That thesis seems flatly false. But pretend for a moment that it is true. Then, I think, we would have to admit that we

have no idea what the relation of *being a left-handed brother of* is supposed to be. Similarly, pretend for a moment that the thesis of relative identity is true. So let's pretend, for example, that *being the same dog as* does not entail *being the same as* (i.e., *being identical with*). But then we must admit that we have no idea what the relation of *being the same dog as* is supposed to be. And it seems that all alleged "relative identity relations" are likewise unintelligible. That is the first objection to relative identity.

Believers in relative identity do not typically think that something special about, say, trees precludes an analysis of *being the same tree as* in terms of being a tree and *being the same as*. Rather, they think that all identities—*being the same tree as*, *being the same dog as*, *being the same electron as*, and so on—are relative and so fail to entail *being the same as*. For they typically deny that there is any such thing as *being the same as* to be entailed. In other words, and as noted above, they typically deny there is any such thing as absolute (i.e., classical, non-relative, plain old) identity.

Insofar as relative identity implies that there is no absolute identity, then it is false. For surely there is absolute identity. Surely there is something that is identical with itself. This is my second objection to relative identity. Of course, this is no objection to relative identity on its own terms. Geach would not take the rejection of absolute identity to be a *reductio* of his view; rather, he takes it to be his central insight. Nevertheless, I think this second objection is decisive. And, at any rate, it is the principal reason I (and I think many others) reject the view that all identity is relative. So I conclude that no defense of the Doctrine of the Trinity is successful if it requires denying that there is something that is identical with itself.

But suppose someone claimed only that identity was *sometimes* relative. So suppose he conceded that there is such a thing as absolute identity and there is something that is identical with itself. But suppose he went on to insist that not every identity implies absolute identity; some identities are relative. Suppose he said, for example, that while *being the same tree as* entails *being the same as*, *being the same God as* does not.

This attenuated version of relative identity is immune to my second objection. And this attenuated version may seem more attractive than full-throttle relative identity, especially if it postulates relative identity only in very unusual cases, cases where absolute identity might seem more trouble than it is worth. For example, one might claim that the logic of absolute identity—which is good enough for everyday purposes—"breaks down at the quantum level" or "breaks down when it comes to the very nature of God."

Peter van Inwagen (1995) presents something like an attenuated version of relative identity in defending the Doctrine of the Trinity. He takes a relative identity reading of the relevant trinitarian claims. But he is careful to

add: "...I shall assume neither that classical identity exists nor that it does not exist" (1995: 241). So van Inwagen's solution, which invokes relative identity, is intended to be consistent with (but not entail) the existence of absolute identity. And so it is meant to be consistent with the claim that, for example, *being the same tree as* is analyzed as being a tree and being the same as (i.e., being identical with).

When first motivating the charge of contradiction, I said that, read most naturally and straightforwardly, claim (5)—the Father is the same God as the Son—entails that the Father is identical with the Son. Now those who (like Geach) insist that *all* identity is relative will disagree. They will object that the most natural and straightforward reading of (5) does not entail the Father's identity with the Son. For they would say that the relative identity reading of (5) is the most natural and straightforward. After all, they will insist, in every paradigm case of "identity," we have only one or another kind of relative identity, never absolute identity. And so Geach can, by his own lights, plausibly maintain that his reading of the Doctrine is the default one.

Van Inwagen endorses a relative identity reading of (5). Yet he cannot agree with Geach that that reading of (5) is the most natural and straightforward. For if—like van Inwagen—we do not deny that there is such a thing as absolute identity, we should say that the following is a perfectly intelligible reading of (5): The Father is God and the Son is God and the Father is identical with the Son. And we should surely say that that reading—again, assuming we do not reject absolute identity out of hand—is the most natural and straightforward.

The defender of attenuated relative identity cannot plausibly maintain that her reading is the default one. Rather, she recommends that we take a less-than-most-natural reading. But once we open the door to less-than-most-natural glosses of (1) through (7), there is—absent further argument—no reason to accept the relative identity gloss as opposed to some other.

Now perhaps the defender of attenuated relative identity will reply that no other gloss is as compelling as hers. Fair enough. But in order to make that point, she will have to do more than present her reading of the Doctrine; she'll have to say something about how it is better than its competitors. And this shows that van Inwagen's approach faces a hurdle that Geach's does not. For, as we have seen, Geach can claim that his reading of the Doctrine is the default reading; nothing similar can plausibly be claimed of any of the "glosses," including the gloss suggested by attenuated relative identity.

As noted above, some object that alleged kind-relative identity relations are unintelligible. But at least Geach can reply that, definitions of those relations aside, we are acquainted with kind-relative identity all the time.

With respect to the relativity of identity, Geach would say, *being the same God as* is just like *being the same tree as*.

The attenuated relative identity theorist says that identity is relative only with respect to the Trinity—or only in cases far removed from common experience. So she cannot say that *being the same God as* is anything like *being the same tree as*. And so *being the same God as*, besides being undefined, turns out to be unlike paradigm cases of *being the same F as,* all of which involve absolute identity. In light of this, the objection that relative identity relations are unintelligible is even more compelling when made against attenuated relative identity than when made against Geach's view.

Geach would say that, because there is no such thing as *being the same as*, *being the same God as* does not entail it. This would render the relation of *being the same God as* mysterious enough. But I think the mystery is increased if there is indeed the relation of *being the same as*, but *being the same God as* is allegedly too weak to entail it. After all, given the existence of *being the same as*, surely there is *some* relation that entails it and *being God*. If *being the same God as* is not *that* relation, then which relation is it? (And what are we supposed to call the relation that entails absolute sameness and otherwise looks for all the world like it is *being the same God as*?) Again, the charge that relative identity relations are unintelligible gets a leg up if relative identity is attenuated.

Attenuating relative identity exacerbates worries about the intelligibility of the relative identity relations, which worries were serious enough to begin with. This in turn makes it harder for attenuated relative identity to answer adequately the first question asked about it. That question was why—if we are to depart from the most natural and straightforward reading of the Doctrine—should we depart in the relative identity way. For this particular departure, of course, can be no more attractive than it is intelligible.

I say that the attenuated relative identity theorist cannot overcome these challenges. She cannot make the relevant relative identity relations intelligible and so she cannot persuade us that the right reading of the Doctrine invokes them. So I conclude that we should reject her defense of the Doctrine. [2]

My conclusion is based, in part, on the idea that if attenuated relative identity relations are unintelligible, a defense of the Doctrine in terms of such relations is unacceptable. The final move open to the advocate of this defense is to question that idea. So I close my discussion of attenuated relative identity by considering the following speech:

> There is such a thing as absolute identity. So, to avoid contradiction, we must depart from the most natural and straightforward reading of some part of the Doctrine. Let's depart from the most natural reading of claims

> invoking "being the same God as." I depart by saying that such claims assert a relation—call it 'relation X'—between the divine persons that does not entail absolute identity. I add that, whatever X is, it doesn't result in a heretical reading of the Doctrine. But that is all I add. Note, specifically, that I don't purport to make X "intelligible." Now for some nomenclature: I call X a 'relative identity relation' and my view 'attenuated relative identity'.

The view expressed in this speech is immune to my objections above. But, its "nomenclature" notwithstanding, this speech does not contain an attenuated relative identity defense of the Doctrine. Indeed, it contains no defense of any sort. Instead, it merely expresses confidence that there is some (non-heretical) defense or other. I think this confidence is praiseworthy. Nevertheless, to express such confidence is not the same thing as defending the Doctrine. (That is why someone can, without contradicting himself, say he has no defense of the Doctrine but is confident that some defense or other is out there.) And it is a defense we are after in this paper.

3.

Social trinitarianism emphasizes the interpersonal (or social) relationships among the divine persons. Social trinitarianism has many contemporary advocates. Moreover, its advocates credit it with a venerable history, finding its roots in the Cappadocian Fathers, including Gregory of Nyssa and Gregory Nazianzus (Morris 1986: 212; Plantinga 1989: 32; Brown 1989: 55).

Its recent popularity and rich history notwithstanding, social trinitarianism is sometimes accused of falling into tritheism, one of the two principal heresies regarding the Trinity. Tritheism, obviously enough, says that there are three Gods. Tritheism does not do justice to claims (5), (6), and (7) of the Doctrine, claims like the Father *is the same God as* the Son. (The other principal heresy here is modalism, which denies that there really are three distinct divine persons. Modalism does not do justice to claims (2), (3), and (4) of the Doctrine, claims like the Father *is not the same person as* the Son.) As noted above, I want to defend the Doctrine from the charge that it is contradictory; let me now add that I won't count as successful any heretical defense.

It is hard to know how to evaluate the charge that social trinitarianism is tritheistic. This is primarily because social trinitarianism itself is hard to define. Sometimes its defenders seem to equate it with the utterly

unobjectionable claim that there really are three persons in the Trinity. Thus Cornelius Plantinga, Jr., says:

> So the first defense of social trinitarianism against the charge of tritheism is this: to say that Father, Son, and Spirit are the names of distinct persons in a full sense of *person* scarcely makes one a tritheist (1989: 34).

Or consider this from David Brown:

> The most common objection raised against defenders of the social model for the Trinity like myself is that it must inevitably lead to tritheism, given its understanding of Father, Son, and Holy Spirit as three distinct persons (1989: 48).

These comments (and others like them; see also Morris 1986: 212-213) make social trinitarianism sound equivalent to the thesis that the Doctrine of the Trinity is true but modalism is false. There is nothing in this thesis to suggest tritheism, unless one has already cast one's lot with the heretics by claiming that modalism and tritheism are the only options.

Sometimes, however, social trinitarians do seem tritheistic, their protests to the contrary notwithstanding. For example, C. Stephen Layman (1988) argues that because each divine person is itself a distinct substance, there are three divine substances. And Thomas Morris at least leans in this direction when he glosses claims about three persons as claims about three "divine beings" (1986: 217-218). Moreover, Richard Swinburne claims that "there are three and only three Gods," saying that this way of putting things avoids the "traditional terminology" (1988: 234).[3]

Given the purposes of this paper, it does not matter what exactly social trinitarianism amounts to or whether that theory—once clearly defined—is tritheistic. Instead, I want only to explore whether something like social trinitarianism offers an orthodox defense of the Doctrine against the charge of contradiction. Thus social trinitarianism is of interest to us only if it (or something like it) can block the charge that (1) through (7) lead to contradiction.

Let's start by considering a version of social trinitarianism no one explicitly endorses. (Discussing this version will set up the main point I want to make about social trinitarianism as it is actually defended.) So consider absolutely pure social trinitarianism. It is "pure" because it claims that the unity among the divine persons is purely social. It claims that harmonious social relationships exhaust that unity. Let's assume that perfect love encompasses every such relationship. Thus pure social trinitarianism asserts that:

(5) The Father is the same God as the Son

means only that the Father and the Son love each other perfectly.

Pure social trinitarianism's reading of (5) is surely consistent with (2), the claim that the Father and the Son are not the same person. And since the pure social trinitarian will read (6) and (7) along the same lines as (5), on her reading the Doctrine of the Trinity is definitely not contradictory.

Pure social trinitarianism renders the Doctrine non-contradictory. Nevertheless, we should reject it. For it is tritheistic. To begin to see why I say this, note that the pure theory implies that A's *being the same God as* B is analyzed as A's being divine, B's being divine, and A and B's loving each other perfectly.[4] This understanding of *being the same God as* implies that two or three or ten humans, when able to love each other perfectly, will be one in the same way that the Father and the Son are one. The relata will of course differ: in the one case it will be humans, in the other divine persons. But the relation—the oneness, the unity—will be the same. Surely, something has gone wrong. Surely, the sense in which the divine persons are one is stronger than the sense in which, once freed from sin and its effects, you and I shall be one.[5]

I believe that each divine person—a divine relatum—loves each of us perfectly. (So the pure theory implies that each of us is thereby "halfway" to being the same God as the Father, the same God as the Son, and the same God as the Spirit!) Given the pure theory, the only thing keeping us from being one with the Father in just the sense that the Son is one with Him is a failing of love on our part, a failing due to sin. But, again, something has gone wrong. For surely it is false that each of the redeemed in Heaven will enjoy exactly the same unity with the Father as that enjoyed by the Son.

And imagine Apollo, Zeus, and Ares resolving their differences, making amends, mending fences and so finally loving each other perfectly, loving each other just as the Father loves the Son. What you are imagining, I insist, is a species of tritheism. Yet we have the relationship of perfectly loving holding among divine relata. Given the pure theory, each of these divine beings would "be the same God as" each of the others. That is, Apollo would be the same God as Zeus (and so on) *in exactly the sense* in which the pure theory says that the Father is the same God as the Son (and so on). And I think this shows that the pure theory is tritheistic (cf. Leftow 1999: 232).

As noted above, no actual social trinitarian is pure. For example, Cornelius Plantinga, Jr. thinks that not only love unifies the divine persons, but also (among other things) the impossibility of each person's existing without the other two (1989: 37). Presumably, there is no limit to the "unifying factors" the social trinitarian can add, just so long as she remains true to her central claims: modalism is false and the divine persons love each

other perfectly. Indeed, she can even add that the divine persons are unified by *being the same God as*. Looked at in this light, it's hard to see how any orthodox believer could fail to be a sullied social trinitarian.

Social trinitarians need not—must not—defend the pure theory. So they must allow that more than love unites the divine persons. But social trinitarianism *as such* does not say what this more is. As a result, social trinitarianism is not really a theory about what (1) through (7) mean.[6] As I noted above, it is hard to say exactly what social trinitarianism is. But I think that the following is in the ballpark and at least makes impure social trinitarianism more than the claim that modalism is false and the divine persons love each other. Social trinitarianism is the view that it is important, for theological and pastoral purposes, to articulate and emphasize the love and other social relationships among the persons of the Trinity. Thus understood, I hope it is clear that whatever its insights, social trinitarianism is not the place to look for a way to block the charge of contradiction.

4.

At one time, a treatment for serious attacks of epilepsy was brain bisection or commisuratomy. (This may seem like a change of subject, but its relevance to the Doctrine will become clear below.) Brain bisection is the severing of the patient's corpus callosum, a band of nerve fibers through which the brain hemispheres communicate directly with each other. Cutting the corpus callosum limits the spread of a seizure to one half of the brain. Yet it has a side effect, well known and beloved among philosophers of personal identity. After brain bisection, a distinct "sphere of consciousness" seems to be correlated with each brain hemisphere.

Moreover, the evidence indicates that each sphere has its own ways of getting information. Here are some examples. The left half of the patient's visual field is accessible only to the sphere of consciousness associated with the right hemisphere (and vice versa). Similarly, the right-hemisphere-sphere gets tactile input from—and for the most part has motor control over—the left hand (and conversely). And each sphere of consciousness enjoys its very own nostril.

The evidence for these and similar claims is the strange behavior that can be elicited, in experimental situations, from patients who have had the surgery. Here is a representative account:

> What is flashed to the right half of the visual field, or felt unseen by the right hand, can be reported verbally. [Typically, the left hemisphere of the brain controls speech.] What is flashed to the left half field or felt by

> the left hand cannot be [verbally] reported, though if the word "hat" is flashed on the left, the left hand will retrieve a hat from a group of concealed objects if the person is told to pick out what he has seen. At the same time he will insist verbally that he saw nothing. Or, if two different words are flashed to the two half fields (e.g., "pencil" and "toothbrush") and the individual is told to retrieve the corresponding object from beneath a screen, with both hands, then the hands will search the collection of objects independently, the right hand picking up the pencil and discarding it while the left hand searches for it, and the left hand similarly rejecting the toothbrush which the right hand lights upon with satisfaction (Nagel 1975: 232).

Moreover, "if a split-brain monkey gets hold of a peanut with both hands, the result is sometimes a tug of war" (Nagel 1975: 231). And a physician told me that, when he extended his hand to a split-brain patient, the patient responded by reaching out to shake with his right hand—and also with his left! (For detailed experimental data, see Gazzaniga 1970.)

I shall go along with the general consensus that brain bisection results in "two spheres of consciousness." But I add that brain bisection does not result in two persons. When one person lies down on the table for the surgery, that same person (and she alone) gets back up. I think this is the right thing to say, in part, because of reflecting on the possibility of a temporarily disabled corpus callosum.

Derek Parfit asks us to suppose that he has:

> …been equipped with some device that can block communication between my hemispheres. Since this device is connected to my eyebrows, it is under my control. By raising an eyebrow I can divide my mind. In each half of my divided mind I can then, by lowering an eyebrow, reunite my mind (1984: 246).

Parfit then imagines availing himself of this device while taking a physics exam. He imagines "dividing his mind" so that he can—in one sphere of consciousness and with one hand—work out one way of solving a problem and—in the other sphere and with the other hand—work out another. He "reunites his mind" ten minutes later.

Suppose this were to happen. Then it seems that Parfit—one person—would acquire a novel psychological ability. But it does not seem that he would acquire a novel way of reproducing; nor does it seem that Parfit would, ten minutes after thus reproducing—and with impunity—annihilate one of his recent offspring (or himself). So I conclude that only one person is involved in the case Parfit imagines. And I think that if there is one person in a case of a temporary division of consciousness, then there is one when the division is

more lasting. After all, the number of persons involved should be fixed once the division occurs; whether there is (say) one person *right now* should not be a matter of what will happen in the future. Thus whether a division will be temporary, or instead be long lasting, is irrelevant to the number of persons resulting when it occurs.

And consider this:

> …if the patient is permitted to touch things with both hands and smell them with both nostrils, he arrives at a unified idea of what is going on around him and what he is doing, without revealing any left-right inconsistencies in his behavior or attitudes. It seems strange to suggest that we are not in a position to ascribe all those experiences to the same person, just because of some peculiarities about how the integration is achieved. The people who *know* these patients find it natural to relate to them as single individuals (Nagel 1975: 238).

(According to Nagel, the "most notable deviation in ordinary behavior was a patient whose left hand appeared to be somewhat hostile to the patient's wife" (1975: 233).) Brain bisection is not a philosopher's fantasy. It really occurs. And we really treat those with split brains as a single person. And I think this is the right thing to do.

Anyone who denies that each split-brain patient is a single person must say that those closest to such patients are deeply confused. And I suppose that if there are two persons associated with each split brain, their friends and family (and the law?) should treat them as two persons, radically altering current practice. But I don't think anyone would seriously recommend changing our practice in this way. I think all will agree that it is false that we should treat each actual split-brain patient as two persons. Yet the claim that two persons result from bisection implies this falsehood. Thus we have a second reason to reject that claim.

Neither of the considerations just noted, however, is the main reason that I say that brain bisection does not multiply persons. The main reason is that—so I say—each of us is a human organism. And I deny that brain bisection results in two human organisms where once there was one. So, I conclude, brain bisection does not result in two of us where once there was one. Instead, brain bisection divides the consciousness of a single human organism; that is, it divides the consciousness of a single person.

The substance dualist could defend a similar argument. Suppose each of us is a simple, immaterial soul. Suppose further, as substance dualists typically do, that each of us (in this life at least) is associated with a particular body and brain. And suppose that bisecting a brain does not produce a new soul. Then brain bisection would not make a new one of us; it

would not produce a new person. Instead, bisecting a brain would split the consciousness of its associated soul, the one soul that was there all along.[7]

I conclude that brain bisection does not give us two persons where before there was one. Instead, one person remains, but with two spheres of consciousness. This conclusion—and my argument for it—is controversial. For I admit that part of my argument relies on one or another disputed assumption about the nature of human persons. And I concede that I have gone along, uncritically, with the received view that brain bisection results in two spheres of consciousness. Thus there are a number of objections one might raise to my claims above. But I don't need to respond to these objections. For, as we shall see below, all that matters for my purposes is that, in making my controversial claims, I have not contradicted myself.[8]

Perhaps some will charge that what I say above is in some sense contradictory. Perhaps they will say that I have contradicted some necessary or conceptual truth. Thus one might claim that it is a matter of necessity—or even analyticity—that one person cannot have two spheres of consciousness. I reply that this claim itself is controversial. And so its denial cannot be contradictory in the most straightforward way. It is at least a live philosophical option that a single human person can have two spheres of consciousness. After all, it is at least a live philosophical option that a human person is either an organism or a substantial soul; and, as noted above, if we are organisms or souls, bisecting a brain would produce a single person with two spheres of consciousness rather than two persons with one sphere each.

Henceforth, when I consider whether something is contradictory, the issue is not whether it merely contradicts a substantive metaphysical thesis. Rather, it is whether it is contradictory in the sense in which the Doctrine of the Trinity is charged with being contradictory. The idea behind that charge, as we saw at the start of this paper, is not merely that the Doctrine contradicts a contentious metaphysical thesis. It is instead (and very roughly) that any clear thinking person can see that the Doctrine leads to a formal contradiction. My reading of split brain cases is not thus contradictory. My reading of these cases is philosophically defensible. My reading is a live philosophical option. And, as we shall see, this is all that my defense of the Doctrine requires of my reading.

5.

S's corpus callosum is severed. S is one person. But she now has two spheres of consciousness, named (for obvious reasons) 'Lefty' and 'Righty'. S decides that she might as well have a little fun with her condition. And so

she engages in written correspondence in such a way that if Lefty is involved, Righty is not; and vice versa.

For example, she makes sure that letters sent to her are presented to only one side of her visual field, accessible to only one sphere of consciousness. And she replies to letters read by Lefty with a hand under only Lefty's control, likewise for letters read by Righty. Moreover, she signs her letters not with her own name, but either as 'Lefty' or 'Righty', depending of course on the responsible sphere of consciousness. Lefty and Righty take turns on correspondence duty, alternating daily.

You enjoy a lengthy correspondence with S. Because your letter might be received on a day when (for example) Lefty is in charge of correspondence, you can't assume that anything read by Righty will be available to S as she reads your letter. After all, if S reads the letter "as Lefty," she won't read it in the full knowledge of what she previously read "as Righty." So your letters are to a large extent redundant. Just in case.

In what follows, it is important that, in the story just told, you correspond with S. This is unproblematic. After all, S is a person and persons correspond. Yet it is also important (in what follows) that you correspond—in some sense—with Lefty and with Righty. It is important that, in corresponding with S, you thereby correspond with Lefty or with Righty. And, so I say, this is what happens. After all, it seems clear from the story I have told that Lefty and Righty take turns corresponding.

But someone might object as follows:

> In your story Lefty and Righty do *not* take turns corresponding. Only *persons* correspond and Lefty and Righty, according to your story, are not persons. And the way S signs her letters—'Lefty', 'Righty', 'S', 'Willard van Orman Quine'—is irrelevant to who actually wrote them. Let me drive this point home with a story of my own: S* has two eyes, Lefty* and Righty*. Sometimes she tapes Righty* shut, reading and responding to letters making use of only Lefty*; she then signs her letters 'Lefty*'. I hope you agree that this does not imply that sometimes one corresponds with S's left eye.

My reply begins by returning to some of the evidence for the claim that brain bisection results in two spheres of consciousness. The patient is told to search for whatever is flashed on the screen. On one side the word 'pencil' is flashed, on the other 'toothbrush'. One of the patient's hands searches for the pencil but not the toothbrush, the other for the toothbrush but not the pencil.

The natural assumption here—the assumption embodied in the claim that brain bisection results in two spheres of consciousness—is that one sphere of consciousness knows that the word 'pencil' was flashed on the screen but not that the word 'toothbrush' was; the other sphere knows that the word

'toothbrush' was flashed, but not 'pencil'. The natural interpretation of the experimental data is that each sphere knows something the other doesn't.

Moreover, the spheres could communicate with each other. If the sphere controlling speech shouted out "I saw the word 'toothbrush' on the screen," then both spheres would know that 'toothbrush' was flashed on the screen. Indeed, just as Lefty could communicate with Righty by shouting, so Lefty could communicate with Righty by writing. Lefty could use the right hand to write a letter, a letter which is then projected on the side of the visual field accessible to Righty. And if Lefty can write letters to Righty, she can write them to you.

Nothing remotely like all of this is true of Lefty* and Righty*. It is false that Lefty*, a mere left eye, knows something that Righty* does not. Nor could Lefty* correspond with Righty*. The story of Lefty* and Righty* was intended to support the objection that, in corresponding with S, you do not in any way correspond with Lefty or with Righty. Instead, that story illustrates the failure of that objection. Its failure is illustrated by the fact that the case of Lefty* and Righty* is simply not analogous to the case of Lefty and Righty.

We can reinforce the relevance of the disanalogy by noting that no one defends the claim that an eye is a person. So no one would say that Lefty* and Righty* are persons. And anyone who did would be beyond the philosophical pale. But some insist that spheres of consciousness are persons. (We shall see later that some social trinitarians seem to say this.) So some will insist that Lefty and Righty are persons. For reasons noted in Section 4, I disagree with them. But their view is not crazy. They are not beyond the pale. Unlike eyes, spheres of consciousness are at least somewhat *person-like*.

They are "person-like." For, as noted above, there seems to be some sense in which spheres of consciousness, again unlike eyes, know things and can correspond. Indeed, let's add that S's corresponding is somehow *analyzed in terms of* Lefty's corresponding or Righty's corresponding. (Or perhaps vice versa.) This analysis is most plausible—and most clearly non-circular—if the sense in which S corresponds is a different sense from the sense in which Lefty or Righty corresponds. So let's add that. Let's add that although there is a sense in which S corresponds and a sense in which Lefty corresponds, there is no univocal sense of 'correspond' in which both S and Lefty correspond.[9] We can also add that Lefty's and Righty's failing to correspond—or to know or to think or to love or to hope or to believe—in the sense in which S does partly explains why Lefty and Righty fail to be persons.

Recall your lengthy correspondence. The letters from S pile up on your desk. A colleague leafs through them and asks: "Who has been writing to

you?" You say: "S; remember her?" He says: "Sure. But"—glancing at the signatures on the letters—"who is Lefty? Who is Righty?" You think to yourself that letters from Lefty just are letters from S, likewise letters from Righty. You think that to write to S is nothing other than to write to Lefty or to write to Righty. You think that to hear from S just would be to hear from Lefty or to hear from Righty. And so you say: "Lefty is S. Righty is S. They are both our friend S." To clarify that there is just one person (i.e., S) authoring the letters, you add the following:

(A) Lefty is the same person as Righty.

Read in the most natural and straightforward way, (A) is false. For thus read, (A) implies that Lefty is a person and Righty is a person and Lefty is identical with Righty. Yet neither Lefty nor Righty is a person; each is, instead, a sphere of consciousness; nor is Lefty identical with Righty. Nevertheless, (A) seems like an appropriate thing to say. You may not have told your colleague the whole story; but you haven't been obscurantist, either. (A) is a pretty good first stab at the situation.

You aren't sure exactly how to go beyond the first stab. For you aren't quite sure what more to say about Lefty and Righty. It might help if they were physical objects, like brain hemispheres. But you know that can't be right. For (we now add to our story) substance dualism is true. So rather than physical objects like brains or brain hemispheres, it is immaterial objects—souls—that have mental properties. You know all this. So you know that each "sphere of consciousness" is not "had" by a brain hemisphere, but by S's soul (i.e., by S herself). So you can safely rule out the possibility that Lefty and Righty are themselves brain hemispheres. You can likewise rule out the possibility that Lefty and Righty are proper parts of S. For S has no proper parts at all.

S is a soul. So she has physical properties only (at best) in an extended and relational sense. For example, the claim that she is over five feet tall is (at best) a shorthand way of saying that she is appropriately related to a body that is over five feet tall. But not all of S's properties are thus extrinsic; not all of her properties are relations to her body. Her mental properties are intrinsic.

Because S's mental properties are intrinsic, it is possible for S to have her mental properties without standing in relations to contingent things outside of herself, to things like her body. This is not to deny that S's intrinsic mental properties are typically related in important ways to her body. For example, stimulate her body in the right way, and S's soul will feel pain. Sever the corpus callosum in her body's brain, and S's consciousness will be divided. But any relation here is presumably causal and so contingent. So if

S is a soul, it is possible for her to feel pain even if disembodied. And if S is a soul, her consciousness could be divided even if no split brain belongs to her.

6.

Let's alter the story a bit. S has not undergone brain bisection. Indeed, S could not undergo brain bisection. No brain belongs to her. For, we are now imagining, S is a disembodied soul who has never had a body or a brain. Nevertheless, S has a divided consciousness of the sort typically induced by brain bisection. And S is somehow able to communicate with the embodied. She can somehow control a pen so as to write letters. And she can somehow read letters written to her.

You didn't know anything about S until you saw her ad in the personals. You like what the ad says (no picture, though), so you begin to correspond. Or at least you try to. You are frustrated to find that your letters are sometimes answered by Lefty and sometimes by Righty but—so it seems to you—never by S. (Lefty and Righty presume to speak for S—indeed, they write as if they were S—so you assume that they are secretaries, acting under S's direction.)

You are annoyed by the—so it seems to you—impersonal nature of this arrangement. And, to make matters worse, apparently Lefty and Righty don't communicate with each other or with S very well. For example, you tell S in one letter that you like long walks on the beach and fruity rum drinks. But, in a letter S writes (by way of Lefty) several weeks later, she asks if you are a teetotaler. You begin to wonder whether your letters are reaching S at all.

You write several times demanding to correspond with S *directly*, not *via* Lefty or Righty. But S (in a letter from Lefty and also in a letter from Righty) replies that you demand the incoherent. Your demand, she says, presupposes that Lefty and Righty are intermediaries between you and her, intermediaries that can somehow be circumvented. But that presupposition, she continues, is all wrong. Rather, to write to Lefty or to write to Righty *just is* to write to her. To correspond with her is nothing other than to correspond with Lefty or with Righty. S tells you that your asking to correspond with her but with neither Lefty nor Righty is like her asking to correspond with you but not with your mind.

S realizes that these claims will seem odd to you. So she tries to cast light on them by explaining her somewhat peculiar nature. She says things like: "I am one immaterial person but two spheres of consciousness." She is careful to insist that she is not two immaterial persons. And she emphasizes that Lefty and Righty are not merely roles she occupies.

Now reconsider the following:

(A) Lefty is the same person as Righty.

Taken most naturally and straightforwardly, (A) is false; for thus taken it entails that Lefty is a person, Righty is a person, and Lefty is identical with Righty. But, in light of the story I have just told, I think there is a fairly natural reading of (A) that comes out true.

Above I tried to motivate the way in which (A) seems true. But let me say more. Note that if you want a relationship with S, you have a relationship with Lefty or Righty. To interact with Lefty or Righty is to interact with S. And for S to love you just is for Lefty to love you or for Righty to love you. Likewise for S's hating you or talking to you or issuing a command to you or…for the purposes of friendship and interaction—indeed, for all practical purposes—Lefty is the same person (that is, S) as Righty.

Moreover, when (for example) S issues a command and Righty issues that command, the command is not issued twice over. For S's commanding does not duplicate Righty's commanding. Rather, S's acting in any way at all is somehow analyzed as either Righty's acting or Lefty's acting. (Or perhaps it goes the other way, and Righty's and Lefty's acting are analyzed in terms of S's acting.) Thus Lefty's acting is the same person's acting—S's acting—as is Righty's acting. Lefty is the same actor—that is, S—as Righty. (That is, Lefty is the same actor as Righty in the sense of 'actor' in which we have one actor: S. In another sense of 'actor', we have two: Lefty and Righty. But in no univocal sense of 'actor' do we have three.)

Similarly, for S to believe a proposition (for example, the proposition that the word 'toothbrush' is flashed on the screen) just is for Righty to believe that proposition or for Lefty to believe it. Righty's believing something is the same person's believing it—S's believing it—as is Lefty's believing something. Righty is the same believer—when by 'believer' we mean person who believes—as is Lefty.

S's believing something just is Righty's believing it or Lefty's believing it. But there is no contradiction if Righty believes that the word 'toothbrush' is flashed on the screen but Lefty does not. For this does not imply both that S believes this and also that it is false that S believes this. Rather, it implies that S believes in one way that the word 'toothbrush' is on the screen and fails to believe this in some other way.

Not only is it possible for S to believe one thing by way of Righty but not by way of Lefty, sometimes this is how it should be. If 'toothbrush' is only in the left half of her visual field, S should believe by way of Righty that 'toothbrush' is on the screen; but S would be unjustified in believing this by way of Lefty's believing it. More interestingly, there may be some things that S would be unjustified in believing by way of Lefty's believing it no

matter what. These might include claims like "this sphere of consciousness controls my left hand" or "this sphere of consciousness is Righty."

Indeed, we can even suppose some ambiguity in the word 'I' so that on one reading it refers to the relevant person, on another to the relevant sphere of consciousness. Then S might truly and justifiedly believe, by way of Righty, "I am Righty"; but it would always be amiss for S to believe this by way of Lefty's believing it. Similar comments apply, of course, to the ways S can justifiedly believe "I am Lefty" and "this sphere controls my right hand." And given this ambiguity of 'I', we can see why S would find herself saying things like "I am exactly one person but I am a first sphere of consciousness and I am also a second sphere of consciousness."

Above I noted that there is a sense in which Lefty is the same believer as Righty; and I also noted that there is a sense in which Lefty is the same actor as Righty. For reasons like these, I conclude that (A)—Lefty is the same person as Righty—is an appropriate and fairly direct way to express a truth. Indeed, I don't know of any better way to express that truth. We can get at it by multiplying examples like those above. Yet all those examples seem to support or indicate or gesture at a peculiar relationship between Lefty and Righty, a relationship that I can't better express than by saying that they are, in some very important sense, the same person.

Given the story I have told, the following is true taken straightforwardly and naturally. Lefty and Righty are both spheres of consciousness and:

(B) Lefty is not the same sphere of consciousness as Righty.

In light of the above, I say that our story about disembodied S shows that (A) and (B) are non-contradictory when appropriately understood.[10] Moreover, I say, (A) and (B) are not obscurantist or misleading. They do as good a job as any pair of claims could at getting at what is going on in the story. And if all that is right, then I suggest we should—because of the obvious analogies—say something similar about the following two claims. I say we should conclude that the following need be neither contradictory nor obscurantist and misleading:

(5) The Father is the same God as the Son.

(2) The Father is not the same person as the Son.

Indeed, I think the analogy between the story of S and the Doctrine of the Trinity is even stronger than (A) and (B) and (5) and (2) suggest. To begin to see why I say this, note that orthodoxy requires us to say that the three divine persons are not three substances. What, then, are they? In presenting his social theory of the Trinity, Plantinga says:

> [A social theory of the Trinity] must have Father, Son, and Spirit as distinct centers of knowledge, will, love, and action. Since each of these capacities requires consciousness, it follows that, on this sort of theory, Father, Son, and Spirit would be viewed as distinct centers of consciousness or, in short, as *persons* in some full sense of that term (1989: 22; emphasis in original).

Similarly, Morris, another social trinitarian, often uses 'centers of consciousness' as a synonym for 'persons' (1986: 210-218).

The social trinitarian cannot accuse us of modalism if we defend the claim that there are three divine persons in what she takes to be the relevant sense of 'person'. And so even modalism's most emphatic opponents should have no objection to our glossing (2) as:

(2*) The Father is not the same center of consciousness as the Son.

Of course, something similar can be said about Lefty and Righty. So let's say it:

(B*) Lefty is not the same center of consciousness as Righty.

One might worry that the persons of the Trinity cannot be separate centers of consciousness in the way that Lefty and Righty are. For, so the worry goes, Lefty's knowing something unknown to Righty—such as that the word 'toothbrush' is flashed on the screen—is crucial to their being two centers rather than one. And even if Lefty and Righty just happen to know all the same things, there *could* be a difference in knowledge between them. But not so for the persons of the Trinity, who are all omniscient. Furthermore, even if Lefty and Righty willed the same things, there *could* be a difference in volition between them. But, again, not so for the divine persons, who essentially will the same things. Thus, one might charge, the divine persons cannot be different centers of consciousness in the way in which Lefty and Righty are.

This objection presupposes that what makes Lefty a different center of consciousness from Righty is that they possibly differ with respect to some mental attribute such as knowledge or will. But I think this gets it backwards. I say that mental differences between Lefty and Righty are possible only if it is already the case that Lefty and Righty are two centers. (If they were one and the same center, such differences would be impossible.) So I say that a difference (or possible difference) in mental attributes is not what makes centers of consciousness distinct. Thus a lack of mental difference among the members of the Trinity would not preclude their being distinct centers of consciousness.

Besides, there almost surely are mental differences among the divine persons. Presumably, one of them has the belief "I am the Father" and the others do not; one of them believes "I proceed from the Father through the Son" but not the others; and so on. Moreover—and more speculatively—perhaps they differ in how things seem to them (that is, in terms of their subjective phenomenological experience) or in what they consciously entertain. And so even if a difference (or possible difference) in mental attributes were required for centers of consciousness to be distinct, this would be no threat to the analogy suggested by (2*) and (B*).

The analogy doesn't end with (2*) and (B*). As already noted, the three divine persons are not three divine substances. There is only one such substance, God. And the claim that each divine person is God is standardly taken to be equivalent to the claim that each is this one divine substance. With this in mind, we can endorse:

(5*) The Father is the same substance as the Son.

The divine persons are not substances. But S is. Being a soul, she is an immaterial substance. Thus—in the same sense in which we endorsed (A) above—we can add:

(A*) Lefty is the same substance as Righty.

It should be obvious that there is a striking analogy between (A*) and (B*) and (5*) and (2*). And note that the story of S and Lefty and Righty contains no analogue of modalism or tritheism; that is, that story does justice to the diversity of Lefty and Righty and to the unity of S. Moreover, (A*) and (B*), as I have explained them, are not contradictory. So I conclude that—even if we insist on orthodoxy—we are not compelled to think that (5*) and (2*) are contradictory; likewise for the Doctrine as a whole.

7.

Someone might object that I don't *really* endorse (A*), since I reject its most natural and straightforward reading, since I deny the claim that Lefty is a substance and Righty is a substance and Lefty is identical with Righty. And so, someone might suspect, since I take (5*) to be analogous to (A*), I don't *really* endorse (5*) either.

In reply, return to the point that opened this paper. If each of the Doctrine's claims is read in the most straightforward and natural way possible, the Doctrine is contradictory. So, assuming that the Doctrine is non-contradictory, at least some of its claims should not be read in the most straightforward and natural way possible. Nevertheless, to borrow language

used earlier, a defense of the Doctrine must "do justice" to all its claims, even the ones that are not read in the most straightforward way possible.

Like the pure social trinitarian, I reject the most natural and straightforward reading of the claim that the Father is the same substance as the Son. For that reading entails that the Father is a substance, the Son is a substance, and the Father is identical with the Son. But I think that, unlike the pure social trinitarian, I "do justice" to that claim. After all, I have argued, my story "does justice" to the claim that Lefty is the same substance as Righty (although of course I reject that claim's most natural and straightforward reading). And I say that the Father's being the same substance as the Son is analogous to Lefty's being the same substance as Righty.

To get a better sense of the analogy here, consider that corresponding with Lefty is corresponding with the same substance as is corresponding with Righty. Similarly, praying to the Son is praying to the same substance as is praying to the Father. Also, Lefty's commanding is the same substance's commanding as is Righty's commanding. Similarly, the Son's commanding is the same substance's commanding as is the Father's commanding. And so on.

It should be clear that my treatment of (5) and (5*) is not that of the pure social trinitarian. But some might worry that it falls into relative identity. After all, so this worry goes, I read 'is the same God as' and 'is the same divine substance as' in such a way that each fails to imply absolute identity. And merely adding—as I most certainly do add—that there is such a thing as absolute identity is not enough to fend off this worry. For this addition leaves open the possibility that my view is a species of attenuated relative identity.

My first reply is that endorsing "the Father is the same God as the Son" while denying that the Father is identical with the Son does not automatically make one a relative identity theorist. If it did, pure social trinitarianism—and probably any way of rendering the doctrine non-contradictory—would count as a version of relative identity.

My second reply is to highlight one of the more important ways my view differs from every version of relative identity. The relative identity theorist says that the Father's "being the same substance as" the Son does not entail their absolute identity but does entail that each is a divine substance. I agree that the Father's "being the same substance as" the Son does not entail their identity. But, unlike the relative identity theorist, I say this fails to entail that each is—in the most straightforward sense—a divine substance.

To clarify this feature of my view, return to Lefty and Righty. While I affirm a properly interpreted (A*)—Lefty is the same substance as Righty—I deny that Lefty is a substance in the most straightforward sense. Note,

however, that in just the same sense that Lefty "is the same substance as" Righty, so Lefty "is the substance S" and thus "is a substance." Understood in a certain way, then, Lefty is a substance. Nevertheless, Lefty is not a substance in the straightforward way that S is. Rather, Lefty is a substance only in virtue of her peculiar relationship to S.

Lefty is a substance in a less-than-straightforward sense. Likewise, I'd say that each divine person is a substance in a less-than-most-straightforward sense. The Father is a substance; the Son is a substance; the Spirit is a substance; and yet—speaking most straightforwardly—there are not three divine substances but one. For each is a substance in virtue of being the same substance as each of the others. Similarly, the Father is God; the Son is God; the Spirit is God; and yet there are not three Gods but one God.

Recall again the story of your correspondence with Lefty and Righty and S. Obviously, it would be a mistake to say that you correspond with three persons. It would be a mistake even to say that you correspond with three things. For, to return to a theme developed earlier, S and the two spheres of consciousness do not correspond in the same sense. S's corresponding in one sense is somehow analyzed in terms of—or is itself part of the analysis of—Righty's corresponding or Lefty's corresponding in another sense. So in one sense of 'correspond', you correspond with one thing: S. In another sense of 'correspond', you correspond with two things: Lefty and Righty. But in no single sense of 'correspond' do you correspond with Lefty, Righty, and S.

With all this in mind, consider the objection that the believer in the Trinity worships *four* things, believes that there are *four* things that are omniscient, omnipotent, and perfectly good: Father, Son, Holy Spirit, and God. I suppose the pure social trinitarian can block this objection right away, insisting that there are exactly three such things, end of story. And the relative identity theorist might insist that this objection presupposes something unacceptable about *not being the same thing as*, a relation suspiciously like *not being absolutely identical with*. But I think the analogy between (A*) and (B*) and (5*) and (2*) suggests a reply that is better than the two just mentioned, a reply that in turn casts light on the depth of the analogy itself.

The analogy suggests that in one sense of, for example, 'is perfectly good', there are three such things (the divine persons), and in another sense there is one (God). But—just as there is no single sense of 'correspond' in which Lefty, Righty, and S all correspond—there is no single sense of 'is perfectly good' in which the Father, the Son, the Spirit, and God are all perfectly good. (Presumably, the persons' being perfectly good in one sense is somehow analyzed in terms of God's being perfectly good in another sense, or vice versa.) More generally, we could say that, taken one way, there are three things worthy of worship, taken another just one; but in no

univocal sense of 'worthy of worship' are there four such things. Thus the analogy developed in this paper not only defends the Doctrine from the charge of contradiction, but also suggests a reply to the objection that trinitarians worship four things.

8.

The story of disembodied S includes something controversial about the nature of human beings. Perhaps the claim that human persons are immaterial souls is false. (I think it is.) And perhaps if it is false, it is necessarily false. And so perhaps the story I told is impossible. And so—one might object—nothing I have said suggests that the Doctrine of the Trinity is possibly true.

If a triune God exists, then this is (presumably) a matter of necessity. And so to show that (1) through (7) are possibly true would be tantamount to showing that they are in fact true. Yet surely we do not need to show that the Doctrine is true to defend it from the charge that it is contradictory. And so surely we do not need to show that the Doctrine is possibly true to defend it from that charge. So although I have not shown that the Doctrine is possibly true, I do think the analogy defended above shows that we are not forced to conclude that the Doctrine is contradictory. As I put a similar point earlier in the paper, the Doctrine of the Trinity is at least a live philosophical option.

I claim only that (A) and (B) and (A*) and (B*) are appropriately analogous to the Doctrine of the Trinity. I do not claim to have presented a theory of the Trinity. I have not defended a gloss or account or analysis of 'being the same God as' as it is used in formulating the Doctrine. I do not claim that each divine person is a center of consciousness exactly like Lefty and Righty and that God is an immaterial substance akin to S's soul.

One reason that I don't claim this is that, even if there were human souls, God and souls would not be kindmates. Of course, God and souls would be alike in being immaterial. But this alone does not make them members of the same kind, lest that kind include, for example, abstract objects as well. And the other obvious way in which God is like a human soul—having mental properties—does not suggest a theory of the nature of God, lest it suggest that all beings with mentality (humans, God, angels, demons, dolphins, dogs) have the same nature.

Similarly, I do not think that reflections on Lefty and Righty yield an analysis of what it is to be a divine person. One reason is that, beyond the sorts of things already said in this paper, I don't know what a center of consciousness is. And I certainly do not purport to have an informative

account of *being a center of consciousness* that applies univocally to Lefty and to the Father.

Nor do I claim to have an account of *being a person* that applies univocally to S and to the Father. Indeed, I think it is an open question whether there is any such account. And because it is a difficult question with a long history (e.g., Boethius explicitly addresses it in *Contra Eutychon*, III), I don't want anything I say here to turn on how that question is answered.

And nothing I say does turn on how it is answered. It should be clear, for example, that nothing I say here requires that S and the Father are persons in exactly the same sense of 'person'. Moreover, nothing I say requires that S and the Father are not persons in just the same sense of 'person'. My argument does require that I can tell a non-contradictory story according to which S is a person and also a substantial human soul. But that does not so much as suggest that 'person' *means* substantial human soul. (If it meant that, the Father would not be a person.) Nor does it suggest any other reason to deny that 'person' applies to the Father in just the same sense that it applies to S.

My argument requires that Lefty and Righty are not persons. So I must insist that 'person' does not mean a center of consciousness in just the way Lefty is a center of consciousness. (If it meant that, Lefty would be a person.) But this does not entail that the Father is not a person, since the Father is not like Lefty in every way. Again, as far as this paper goes, it is an open question whether 'person' is predicated univocally of S (and of you and of me) and of the Father.

My defense of the Doctrine leaves us wondering what it is to be a person. And we are left wondering what it is to be a center of consciousness. And we are left wondering whether we are persons in exactly the sense that the Father is a person. Moreover, we are left wondering how deep the analogy between Lefty—who is "person-like"—and the Father—who is a person—goes.

There is no clear answer to any of these questions. At least, my defense contains no clear answers. Moreover, there are no answers from the logic of identity or the law of non-contradiction or merely knowing how to count. And so believers in the Trinity can happily admit not knowing how to answer these questions. Happier still, we can do so without feigning ignorance about *one*, *three*, or *identity*.

I claim that I have defended the Doctrine from the charge of contradiction. But I also deny having a theory of the nature of the Trinity. Someone might object that I can't make my defense without such a theory. For, so this objection goes, unless we know exactly how to interpret (1) through (7), we have no right to say that (1) through (7) are non-contradictory. And unless we

can rightfully say that, we have no defense against the charge that the Doctrine is contradictory.

I concede that I have not proven that the Doctrine, rightly interpreted, is not contradictory. But such a proof is not the only way to defend the Doctrine from the charge of contradiction. One could, instead, argue that there is no compelling reason to believe the Doctrine is contradictory. It is this sort of defense I have presented.

Let's return to my defense one last time. So suppose the story I told about disembodied S is true. Even in such a (comparatively) mundane case, mysteries persist. We don't know the nature of S's soul, other than its being non-physical and mental. (So we do not know what makes S a different kind of thing from angels and demons and God.) Nor do we know what exactly spheres of consciousness are.

Suppose, again, that the story about disembodied S is true. But suppose further that we (like the Church Fathers) are unfamiliar with brain bisection and its odd effects. Suppose moreover that the ideas of a soul (much less a disembodied one) and of a sphere of consciousness have never occurred to us. And finally suppose that S decides to reveal her nature to us so as to help us interact with her rather than to teach us metaphysics. Then (A) and (B) and (A*) and (B*) would get at something non-contradictory in about as clear and direct a way as we could hope for. Then (A) and (B) and (A*) and (B*) would be in equal measure appropriate and puzzling and—when rightly interpreted—non-contradictory.

And so, I say, it goes for claims (1) through (7). A full-blown theory of the metaphysics of the Trinity would tell us what the divine substance is and what the divine persons—Father, Son, and Holy Spirit—are. I don't have that. But I do think we have seen enough to conclude that (1) through (7) could be wholly appropriate and puzzling and—when rightly interpreted—non-contradictory. And so we are not forced to conclude that they are in fact contradictory. The charge of contradiction fails to stick.

ENDNOTES

*Thanks to Bob Adams, Mike Bergmann, Jeff Brower, Jim Cargile, Tom Crisp, Chuck Mathewes, Dan Moseley, Mark Murphy, Mike Murray, Al Plantinga, Phil Quinn, Mike Rea, Ted Sider, Donald Smith, Eleonore Stump, Richard Swinburne, Peter van Inwagen, Thomas Williams, and Dean Zimmerman. This paper was part of the 2002 Workshop on the Metaphysics of the Human Person, sponsored by the Pew Charitable Trusts. I also presented this paper at Purdue University and at the University of Virginia.

[1] I suppose that we could use (for example) 'being the same dog as' to express a relation that holds between A and B if and only if A is a dog and B is a dog and A is heavier than B. If we adopted this usage, then of course 'A is the same dog as B' would express a proposition entailing that A and B are dogs but not entailing that A is identical with B. (Indeed, since

nothing is heavier than itself, it would entail that A is not identical with B.) But, obviously, this has nothing to do with the question of whether A's being the same dog as B entails that A is identical with B.

[2] Van Inwagen (1995) meticulously presents the formal properties of some relative identity relations; but I believe this falls short of telling us what those relations are.

Someone might claim that the explicitly stated formal properties are all there is to the relevant relation. That is, someone might claim, for example, that (5) says only that the Father is related to the Son by a relation with the relevant explicitly stated formal properties. But, obviously, she must defend this highly technical gloss on (5), this departure from the most natural and straightforward reading of (5). That is, she must give a reason to believe this is the *right* way to understand (5).

[3] Of course, Swinburne puts his view this way tongue in cheek, not meaning to be explicitly heretical; a more recent discussion of the Trinity can be found in Swinburne (1994: 170-191).

[4] The only other way to take the pure theory is as identifying the relation of *being the same God as* with the relation of *loving each other perfectly*. Thus taken, the pure theory implies that, once conformed to the image of Christ, you will be the same God as I. And I shall be the same God as you. And each of us will be the same God as the Father. This is a *reductio*.

[5] Perhaps some social trinitarians would object. Morris offers the following in partial support of the social theory of the Trinity:

> When Jesus...is represented in the gospel of John (17:21) as praying to the Father concerning his disciples and other followers "that they may all be one; even as thou, Father, art in me, and I in thee," he was surely not asking that there be only a single, solitary Christian. He was asking for unity among numerically distinct individuals, not for numerical identity here, and thus he was implying that he perceived the oneness between himself and the Father not to be that of numerical identity, as that between, say, Cicero and Tully, but rather to be that of some sort of harmonious unity between ontologically distinct individuals (Morris 1986: 209-210).

[6] Of course, some particular social trinitarian may have a theory about what (1) through (7) mean. That's a different point.

[7] The (misguided) claim that a soul "explains the unity of consciousness" is inconsistent with a soul's having two spheres of consciousnesses. But that claim is not an essential part of substance dualism. Substance dualism itself is consistent with one soul's having more than one sphere of consciousness.

[8] For the record, I am inclined to deny that, as far as human persons are concerned, there are any objects or events or whatever that are "spheres of consciousness." So I would be inclined to object to my description of split brain cases above, if that description needed to be true in every detail. But—as is further discussed in this section and in §8—all that matters is that that description is not contradictory. And it is not contradictory.

[9] Compare: The baseball causes the window to shatter and the baseball's striking the window causes the window to shatter. Both the baseball and the striking really do cause this, but in different senses of 'cause'. And perhaps the baseball's "object-causing" an effect is analyzed as the ball's being the constituent object of an event that "event-causes" that effect.

[10] Some might object that there is a contradiction in my story's claim that disembodied S corresponds with embodied humans. In reply, if it is contradictory for a non-physical thing to interact in these ways with the physical world, then theism itself—with its creator God—is itself contradictory. But theism is not contradictory. And if it were, there would be no point in defending the trinitarian species of theism.

REFERENCES

Anscombe, G.E.M. and P.T. Geach. 1961. *Three Philosophers*. Ithaca, NY: Cornell University Press.

Brown, David. 1989. Trinitarian personhood and individuality. In *Trinity, Incarnation, and Atonement: Philosophical & Theological Essays*, eds. Ronald J. Feenstra and Cornelius Plantinga Jr. Notre Dame, IN: University of Notre Dame Press.

Cartwright, Richard. 1987. On the logical problem of the trinity. In *Philosophical Essays*. Cambridge, MA: MIT Press.

Davis, Stephen T., Daniel Kendall S.J., and Gerald O'Collins S.J., eds. 1999. *The Trinity*. New York: Oxford University Press.

Feenstra, Ronald J. and Cornelius Plantinga, Jr., eds. 1989. *Trinity, Incarnation, and Atonement: Philosophical & Theological Essays*. Notre Dame, IN: University of Notre Dame Press.

Gazzaniga, Michael. 1970. *The Bisected Brain*. New York: Appleton-Century-Crofts.

Geach, P.T. 1972. *Logic Matters*. Berkeley, CA: University of California Press.

Geach, P.T. 1973. Ontological relativity and relative identity. In *Logic and Ontology*, ed. Milton K. Munitz. New York: NYU Press.

Layman, C. Stephen. 1988. Tritheism and the trinity. *Faith and Philosophy* 5: 291-298.

Leftow, Brian. 1999. Anti Social Trinitarianism. In *The Trinity*, eds. Stephen T. Davis, Daniel Kendall S.J., and Gerald O'Collins S.J. New York: Oxford University Press.

Morris, Thomas V. 1986. *The Logic of God Incarnate*. Ithaca, NY: Cornell University Press.

Nagel, Thomas. 1975. Brain bisection and the unity of consciousness. In *Personal Identity*, ed. John Perry. Berkeley, CA: University of California Press. (Originally published in *Synthese* 22 (1971): 396-413.)

Parfit, Derek. 1984. *Reasons and Persons*. Oxford: Clarendon Press.

Perry, John. 1970. The same *F*. *Philosophical Review* 79: 181-200.

Plantinga, Jr., Cornelius. 1989. Social trinity and tritheism. In *Trinity, Incarnation, and Atonement: Philosophical & Theological Essays*, eds. Ronald J. Feenstra and Cornelius Plantinga Jr. Notre Dame, IN: University of Notre Dame Press.

Swinburne, Richard. 1988. Could there be more than one God? *Faith and Philosophy* 4: 225-241.

Swinburne, Richard. 1994. *The Christian God*. Oxford: Clarendon Press.

Van Inwagen, Peter. 1995. And yet they are not three Gods but one. In *God, Knowledge, and Mystery*. Ithaca, NY: Cornell University Press.

Index

PHILOSOPHICAL STUDIES SERIES

1. Jay F. Rosenberg: *Linguistic Representation.* 1974 ISBN 90-277-0533-X
2. Wilfrid Sellars: *Essays in Philosophy and Its History.* 1974 ISBN 90-277-0526-7
3. Dickinson S. Miller: *Philosophical Analysis and Human Welfare.* Selected Essays and Chapters from Six Decades. Edited with an Introduction by Lloyd D. Easton. 1975 ISBN 90-277-0566-6
4. Keith Lehrer (ed.): *Analysis and Metaphysics.* Essays in Honor of R. M Chisholm. 1975 ISBN 90-277-0571-2
5. Carl Ginet: *Knowledge, Perception, and Memory.* 1975 ISBN 90-277-0574-7
6. Peter H. Hare and Edward H. Madden: *Causing, Perceiving and Believing.* An Examination of the Philosophy of C. J. Ducasse. 1975 ISBN 90-277-0563-1
7. Hector-Neri Castañeda: *Thinking and Doing.* The Philosophical Foundations of Institutions. 1975 ISBN 90-277-0610-7
8. John L. Pollock: *Subjunctive Reasoning.* 1976 ISBN 90-277-0701-4
9. Bruce Aune: *Reason and Action.* 1977 ISBN 90-277-0805-3
10. George Schlesinger: *Religion and Scientific Method.* 1977 ISBN 90-277-0815-0
11. Yirmiahu Yovel (ed.): *Philosophy of History and Action.* Papers presented at the First Jerusalem Philosophical Encounter (December 1974). 1978 ISBN 90-277-0890-8
12. Joseph C. Pitt (ed.): *The Philosophy of Wilfrid Sellars: Queries and Extensions.* 1978 ISBN 90-277-0903-3
13. Alvin I. Goldman and Jaegwon Kim (eds.): *Values and Morals*. Essays in Honor of William Frankena, Charles Stevenson, and Richard Brandt. 1978 ISBN 90-277-0914-9
14. Michael J. Loux: *Substance and Attribute.* A Study in Ontology. 1978 ISBN 90-277-0926-2
15. Ernest Sosa (ed.): *The Philosophy of Nicholas Rescher*. Discussion and Replies. 1979 ISBN 90-277-0962-9
16. Jeffrie G. Murphy: *Retribution, Justice, and Therapy.* Essays in the Philosophy of Law. 1979 ISBN 90-277-0998-X
17. George S. Pappas (ed.): *Justification and Knowledge.* New Studies in Epistemology. 1979 ISBN 90-277-1023-6
18. James W. Cornman: *Skepticism, Justification, and Explanation.* With a Bibliographic Essay by Walter N. Gregory. 1980 ISBN 90-277-1041-4
19. Peter van Inwagen (ed.): *Time and Cause.* Essays presented to Richard Taylor. 1980 ISBN 90-277-1048-1
20. Donald Nute: *Topics in Conditional Logic*. 1980 ISBN 90-277-1049-X
21. Risto Hilpinen (ed.): *Rationality in Science.* Studies in the Foundations of Science and Ethics. 1980 ISBN 90-277-1112-7
22. Georges Dicker: *Perceptual Knowledge.* An Analytical and Historical Study. 1980 ISBN 90-277-1130-5
23. Jay F. Rosenberg: *One World and Our Knowledge of It.* The Problematic of Realism in Post-Kantian Perspective. 1980 ISBN 90-277-1136-4
24. Keith Lehrer and Carl Wagner: *Rational Consensus in Science and Society.* A Philosophical and Mathematical Study. 1981 ISBN 90-277-1306-5
25. David O'Connor: *The Metaphysics of G. E. Moore.* 1982 ISBN 90-277-1352-9

PHILOSOPHICAL STUDIES SERIES

26. John D. Hodson: *The Ethics of Legal Coercion.* 1983 ISBN 90-277-1494-0
27. Robert J. Richman: *God, Free Will, and Morality.* Prolegomena to a Theory of Practical Reasoning. 1983 ISBN 90-277-1548-3
28. Terence Penelhum: *God and Skepticism.* A Study in Skepticism and Fideism. 1983 ISBN 90-277-1550-5
29. James Bogen and James E. McGuire (eds.): *How Things Are.* Studies in Predication and the History of Philosophy of Science. 1985 ISBN 90-277-1583-1
30. Clement Dore: *Theism.* 1984 ISBN 90-277-1683-8
31. Thomas L. Carson: *The Status of Morality.* 1984 ISBN 90-277-1619-9
32. Michael J. White: *Agency and Integrality.* Philosophical Themes in the Ancient Discussions of Determinism and Responsibility. 1985 ISBN 90-277-1968-3
33. Donald F. Gustafson: *Intention and Agency.* 1986 ISBN 90-277-2009-6
34. Paul K. Moser: *Empirical Justification.* 1985 ISBN 90-277-2041-X
35. Fred Feldman: *Doing the Best We Can*. An Essay in Informal Deontic Logic. 1986 ISBN 90-277-2164-5
36. G. W. Fitch: *Naming and Believing.* 1987 ISBN 90-277-2349-4
37. Terry Penner: *The Ascent from Nominalism.* Some Existence Arguments in Plato's Middle Dialogues. 1987 ISBN 90-277-2427-X
38. Robert G. Meyers: *The Likelihood of Knowledge.* 1988 ISBN 90-277-2671-X
39. David F. Austin (ed.): *Philosophical Analysis.* A Defense by Example. 1988 ISBN 90-277-2674-4
40. Stuart Silvers (ed.): *Rerepresentation.* Essays in the Philosophy of Mental Representation. 1988 ISBN 0-7923-0045-9
41. Michael P. Levine: *Hume and the Problem of Miracles.* A Solution. 1989 ISBN 0-7923-0043-2
42. Melvin Dalgarno and Eric Matthews (eds.): *The Philosophy of Thomas Reid*. 1989 ISBN 0-7923-0190-0
43. Kenneth R. Westphal: *Hegel's Epistemological Realism.* A Study of the Aim and Method of Hegel's *Phenomenology of Spirit*. 1989 ISBN 0-7923-0193-5
44. John W. Bender (ed.): *The Current State of the Coherence Theory.* Critical Essays on the Epistemic Theories of Keith Lehrer and Laurence BonJour, with Replies. 1989 ISBN 0-7923-0220-6
45. Roger D. Gallie: *Thomas Reid and 'The Way of Ideas'*. 1989 ISBN 0-7923-0390-3
46. J-C. Smith (ed.): *Historical Foundations of Cognitive Science*. 1990 ISBN 0-7923-0451-9
47. John Heil (ed.): *Cause, Mind, and Reality.* Essays Honoring C. B. Martin. 1989 ISBN 0-7923-0462-4
48. Michael D. Roth and Glenn Ross (eds.): *Doubting*. Contemporary Perspectives on Skepticism. 1990 ISBN 0-7923-0576-0
49. Rod Bertolet: *What is Said*. A Theory of Indirect Speech Reports. 1990 ISBN 0-7923-0792-5
50. Bruce Russell (ed.): *Freedom, Rights and Pornography.* A Collection of Papers by Fred R. Berger. 1991 ISBN 0-7923-1034-9
51. Kevin Mulligan (ed.): *Language, Truth and Ontology.* 1992 ISBN 0-7923-1509-X

PHILOSOPHICAL STUDIES SERIES

52. Jesús Ezquerro and Jesús M. Larrazabal (eds.): *Cognition, Semantics and Philosophy*. Proceedings of the First International Colloquium on Cognitive Science. 1992 ISBN 0-7923-1538-3
53. O.H. Green: *The Emotions*. A Philosophical Theory. 1992 ISBN 0-7923-1549-9
54. Jeffrie G. Murphy: *Retribution Reconsidered.* More Essays in the Philosophy of Law. 1992 ISBN 0-7923-1815-3
55. Phillip Montague: *In the Interests of Others.* An Essay in Moral Philosophy. 1992 ISBN 0-7923-1856-0
56. Jacques-Paul Dubucs (ed.): *Philosophy of Probability.* 1993 ISBN 0-7923-2385-8
57. Gary S. Rosenkrantz: *Haecceity.* An Ontological Essay. 1993 ISBN 0-7923-2438-2
58. Charles Landesman: *The Eye and the Mind.* Reflections on Perception and the Problem of Knowledge. 1994 ISBN 0-7923-2586-9
59. Paul Weingartner (ed.): *Scientific and Religious Belief.* 1994 ISBN 0-7923-2595-8
60. Michaelis Michael and John O'Leary-Hawthorne (eds.): *Philosophy in Mind.* The Place of Philosophy in the Study of Mind. 1994 ISBN 0-7923-3143-5
61. William H. Shaw: *Moore on Right and Wrong*. The Normative Ethics of G.E. Moore. 1995 ISBN 0-7923-3223-7
62. T.A. Blackson: *Inquiry, Forms, and Substances*. A Study in Plato's Metaphysics and Epistemology. 1995 ISBN 0-7923-3275-X
63. Debra Nails: *Agora, Academy, and the Conduct of Philosophy.* 1995 ISBN 0-7923-3543-0
64. Warren Shibles: *Emotion in Aesthetics*. 1995 ISBN 0-7923-3618-6
65. John Biro and Petr Kotatko (eds.): *Frege: Sense and Reference One Hundred Years Later.* 1995 ISBN 0-7923-3795-6
66. Mary Gore Forrester: *Persons, Animals, and Fetuses.* An Essay in Practical Ethics. 1996 ISBN 0-7923-3918-5
67. K. Lehrer, B.J. Lum, B.A. Slichta and N.D. Smith (eds.): *Knowledge, Teaching and Wisdom*. 1996 ISBN 0-7923-3980-0
68. Herbert Granger: *Aristotle's Idea of the Soul.* 1996 ISBN 0-7923-4033-7
69. Andy Clark, Jesús Ezquerro and Jesús M. Larrazabal (eds.): *Philosophy and Cognitive Science: Categories, Consciousness, and Reasoning*. Proceedings of the Second International Colloquium on Cogitive Science. 1996 ISBN 0-7923-4068-X
70. J. Mendola: *Human Thought.* 1997 ISBN 0-7923-4401-4
71. J. Wright: *Realism and Explanatory Priority*. 1997 ISBN 0-7923-4484-7
72. X. Arrazola, K. Korta and F.J. Pelletier (eds.): *Discourse, Interaction and Communication*. Proceedings of the Fourth International Colloquium on Cognitive Science. 1998 ISBN 0-7923-4952-0
73. E. Morscher, O. Neumaier and P. Simons (eds.): *Applied Ethics in a Troubled World*. 1998 ISBN 0-7923-4965-2
74. R.O. Savage: *Real Alternatives, Leibniz's Metaphysics of Choice*. 1998 ISBN 0-7923-5057-X
75. Q. Gibson: *The Existence Principle.* 1998 ISBN 0-7923-5188-6
76. F. Orilia and W.J. Rapaport (eds.): *Thought, Language, and Ontology.* 1998 ISBN 0-7923-5197-5

PHILOSOPHICAL STUDIES SERIES

77. J. Bransen and S.E. Cuypers (eds.): *Human Action, Deliberation and Causation*. 1998 ISBN 0-7923-5204-1
78. R.D. Gallie: *Thomas Reid: Ethics, Aesthetics and the Anatomy of the Self*. 1998 ISBN 0-7923-5241-6
79. K. Korta, E. Sosa and X. Arrazola (eds.): *Cognition, Agency and Rationality*. Proceedings of the Fifth International Colloquium on Cognitive Science. 1999 ISBN 0-7923-5973-9
80. M. Paul: *Success in Referential Communication*. 1999 ISBN 0-7923-5974-7
81. E. Fischer: *Linguistic Creativity*. Exercises in 'Philosophical Therapy'. 2000 ISBN 0-7923-6124-5
82. R. Tuomela: *Cooperation*. A Philosophical Study. 2000 ISBN 0-7923-6201-2
83. P. Engel (ed.): *Believing and Accepting*. 2000 ISBN 0-7923-6238-1
84. W.L. Craig: *Time and the Metaphysics of Relativity*. 2000 ISBN 0-7923-6668-9
85. D.A. Habibi: *John Stuart Mill and the Ethic of Human Growth*. 2001 ISBN 0-7923-6854-1
86. M. Slors: *The Diachronic Mind*. An Essay on Personal Identity, Psychological Continuity and the Mind-Body Problem. 2001 ISBN 0-7923-6978-5
87. L.N. Oaklander (ed.): *The Importance of Time*. Proceedings of the Philosophy of Time Society, 1995–2000. 2001 ISBN 1-4020-0062-6
88. M. Watkins: *Rediscovering Colors.* A Study in Pollyanna Realism. 2002 ISBN 1-4020-0737-X
89. W.F. Vallicella: *A Paradigm Theory of Existence.* Onto–Theology Vindicated. 2002 ISBN 1-4020-0887-2
90. M. Hulswit: *From Cause to Causation*. A Peircean Perspective. 2002 ISBN 1-4020-0976-3; Pb 1-4020-0977-1
91. D. Jacquette (ed.): *Philosophy, Psychology, and Psychologism*. Critical and Historical Readings on the Psychological Turn in Philosophy. 2003 ISBN 1-4020-1337-X
92. G. Preyer, G. Peter and M. Ulkan (eds.): *Concepts of Meaning*. Framing an Integrated Theory of Linguistic Behavior. 2003 ISBN 1-4020-1329-9
93. W. de Muijnck: *Dependencies, Connections, and Other Relations*. A Theory of Mental Causation. 2003 ISBN 1-4020-1391-4
94. N. Milkov: *A Hundred Years of English Philosophy*. 2003 ISBN 1-4020-1432-5
95. E.J. Olsson (ed.): *The Epistomology of Keith Lehrer*. 2003 ISBN 1-4020-1605-0
96. D.S. Clarke: *Sign Levels*. Language and Its Evolutionary Antecedents. 2003 ISBN 1-4020-1650-6
97. A. Meirav: *Wholes, Sums and Unities*. 2003 ISBN 1-4020-1660-3
98. C.H. Conn: *Locke on Esence and Identity*. 2003 ISBN 1-4020-1670-0
99. J.M. Larrazabal and L.A. Pérez Miranda (eds.): *Language, Knowledge, and Representation*. Proceedings of the Sixth International Colloquium on Cognitive Science (ICCS-99). 2004 ISBN 1-4020-2057-0
100. P. Ziff: *Moralities*. A Diachronic Evolution Approach. 2004 ISBN 1-4020-1891-6
101. J.A. Corlett: Terrorism: *A Philosophical Analysis*. 2003 ISBN 1-4020-1694-8; Pb 1-4020-1695-6

PHILOSOPHICAL STUDIES SERIES

102. K. Korta and J.M. Larrazabal (eds.): *Truth, Rationality, Cognition, and Music*. Proceedings of the Seventh International Colloquium on Cognitive Science. 2004 ISBN 1-4020-1912-2

103. T.M. Crisp, M. Davidson and D. Vander Laan (eds.): Knowledge and Reality. 2006 ISBN 1-4020-4732-0