Christian Schmidt
Pierre Livet

Comprendre nos interactions sociales

Une perspective neuroéconomique

Odile Jacob
économie

Christian Schmidt
Pierre Livet

COMPRENDRE NOS INTERACTIONS SOCIALES

Une perspective neuroéconomique

Odile Jacob

15, RUE SOUFFLOT, 75005 PARIS

www.odilejacob.fr

ISBN : 978-2-7381-6860-3

Ce document numérique a été réalisé par Nord Compo.

Introduction

Pourquoi un économiste théoricien des jeux et un philosophe épistémologue ont-ils formé le projet de rédiger à deux voix un livre sur les interactions sociales, alors qu'aucun des deux n'est sociologue ou psychologue social ? Et pourquoi ont-ils adopté une perspective de neuroéconomie, puisqu'ils ne sont ni l'un ni l'autre experts en neurosciences ?

Il est clair, à première vue, que les phénomènes étudiés par l'économiste sont toujours les résultats d'interactions sociales, soit directes (échanges marchands, compétitions, répartitions, négociations), soit indirectes (redistributions, organisations). L'objet de la théorie des jeux qui représente l'une des approches privilégiées pour étudier ces phénomènes consiste à formaliser les interactions entre des agents supposés rationnels. Plus récemment, une économie expérimentale s'est développée, qui étudie au moyen d'expériences les décisions effectivement prises par les individus, en réaction aux actions des autres. Mais les comportements ainsi observés restaient inexpliqués tant que l'on ne disposait pas d'informations plus précises sur le fonctionnement du cerveau dans ces situations. Tel n'est plus le cas aujourd'hui où beaucoup des résultats obtenus au cours de ces protocoles expérimentaux peuvent être maintenant complétés par des informations sur le fonctionnement cérébral, grâce, en particulier, aux différentes techniques de l'imagerie cérébrale. C'est ainsi que s'est développée, depuis quelque temps, une branche sociale des neurosciences dans laquelle la neuroéconomie se trouve aux avant-postes.

Le philosophe est aussi dans son rôle en défendant la position philosophique suivante : les sujets humains se constituent dans leurs interactions avec leurs semblables, qui orientent aussi leurs perspectives sur leur environnement. Les manières dont les capacités physiques de notre corps et de son système nerveux facilitent ou limitent ces interactions ne sont pas sans intérêt pour le philosophe, ne serait-ce que dans la mesure où mieux les connaître peut l'amener à réviser ses « intuitions » sur les processus cognitifs et affectifs humains, voire à modifier les catégories qui lui permettent de penser ces processus dans leurs interactions, ou encore à ne pas proposer d'idéal qui ne tienne pas compte des limitations humaines.

Voilà pour nos justifications. Mais qu'en est-il des interactions *sociales* ? L'économiste ne va-t-

il pas être tenté de les réduire à des échanges entre les agents et aux décisions qui les précédent en se concentrant exclusivement sur l'étude de leur ancrage « rationnel » ? Le philosophe, de son côté, pourrait soit, en soutenant l'individualisme, être tenté de réduire les interactions sociales aux capacités interactives des individus, soit au contraire, en soutenant le « holisme », la prééminence du tout social, vouloir montrer qu'on doit toujours présupposer un social déjà constitué, pour comprendre comment les décisions des individus manifestent une sensibilité à des normes collectives et ne se réduisent pas à suivre leur intérêt personnel.

Dans ce livre, le théoricien des jeux justifiera l'emploi du qualificatif « social », en se montrant attentif aux résultats obtenus lors des nombreuses expériences qui montrent que les individus réels diffèrent de l'individu *self-interested* qu'est supposé être l'agent économique. Il s'efforce de comprendre les racines de ces différences, à la lumière des premiers résultats mis en évidence par ces neurosciences sociales. Cette investigation lui enseigne que nous sommes plus largement dépendants de la coordination de nos actions avec autrui et plus sensibles aux opportunités de coopération, et aux normes sociales qui les accompagnent, que ne le suppose le schéma « autiste » auquel se réfère l'économie classique. Les interactions sociales, entendues en ce sens, ouvrent ainsi le champ à une analyse renouvelée de nombreux phénomènes économiques, concernant notamment les décisions et les anticipations qui les précèdent, ainsi que les actions et les négociations qui souvent les suivent, sans oublier l'organisation qui les accompagne.

Le philosophe, en cette affaire, ne se veut ni individualiste ni holiste. Les processus d'interaction sont pour lui les constituants ontologiques fondamentaux et de l'individu et des collectifs sociaux. Ces interactions ne s'établissent pas seulement à des niveaux séparés – processus infra-individuels, individuels, interindividuels, collectifs – mais aussi entre des niveaux différents. Ainsi, les limitations de certains processus infra-individuels peuvent trouver des compensations dans des interactions interindividuelles et des organisations collectives – songez à la communication des savoirs. Inversement, les collectifs ont des limitations propres – ainsi l'anonymat relatif de leurs membres ne permet pas à un individu, ni même à quelques-uns, de contrôler directement le degré d'engagement de tous dans des coopérations. Plutôt que d'opposer l'individu et le tout social, il est préférable d'étudier précisément comment interagissent ces différents types de processus, comment ils constituent et maintiennent des structures qui sont d'échelles différentes mais qui peuvent aussi entrecroiser différents niveaux – un organisme, des relations parentales, des collectifs, des institutions – et comment les différentes structures d'interaction qui se mettent en place se combinent, se soutiennent mutuellement, ou, au contraire, s'imposent les unes aux autres des transformations.

Nous suivrons à la trace ces processus d'interaction en examinant successivement les niveaux différents où ils se manifestent.

Dans une première partie, intitulée « Interactions et intersubjectivité », nous analyserons les interactions entre des processus infra-individuels qui sont nécessaires à la constitution du sujet. Mais ces processus infra-individuels ne se bornent pas à construire un sujet isolé, ils l'engagent déjà dans des interactions, si bien que ce sujet est nécessairement intersubjectif, comme nous le montrerons au

chapitre 1. Pour autant, cette dynamique d'intersubjectivité ne réduit pas la spécificité de chaque sujet puisque elle contribue, au contraire, à sa construction. Nous étudierons au chapitre 2 comment la figure d'autrui se constitue. Nous avons posé cette intersubjectivité comme fondatrice, et nous analyserons, sur cette base, la portée de la référence à « un autre moi-même » et discuterons ses limites. Cette recherche suggère, qu'au rôle de l'autre qui nous fait face, il nous faut ajouter celui du (des) tiers. Une fois cette structure interactive de base constituée, nous pouvons revenir sur les processus de décision, et montrer au chapitre 3 comment ils sont le fruit de l'interaction entre différentes perspectives, par exemple, entre des perspectives temporelles à court et à long terme, entre des interactions à courte et à longue portée. Ce chapitre nous a conduits à repenser, dans cette optique, comment les sujets conçoivent et traitent les relations d'intertemporalité, avec toutes leurs conséquences, parfois inattendues et tout au moins différentes de celles le plus souvent enseignées, sur les anticipations des agents. Nous aurons ainsi, au cours des trois chapitres qui forment cette première partie, établi que les thématiques qu'on associe généralement à une focalisation sur les seuls individus pouvaient avantageusement être repensées, dans une perspective interactionniste de part en part.

Nous aborderons ensuite, dans une deuxième partie, intitulée « Coordination et coopération » ce que l'on considère communément comme des interactions sociales, au sens de rapports interindividuels. Sur ce terrain, les distinctions introduites par la théorie des jeux, entre les jeux de pure coordination (reposant sur l'identification de points focaux), les jeux de coordination à équilibres multiples (lorsqu'une multiplicité de possibilités aboutit à des résultats différents) et les jeux de confiance (lorsque la coopération s'avère risquée pour chaque joueur considéré individuellement), nous ont servi de premiers repères. Les problèmes de coordination, lorsque certaines solutions satisfont l'intérêt de chacun des joueurs, suggèrent l'émergence d'un point de vue du groupe, dont l'adoption par les individus conduit au mode de coordination le plus efficace. On mettra en évidence, au chapitre 4, les différentes difficultés auxquelles se heurte cette coordination, en signalant, au passage, les problèmes posés par la notion de confiance, face au risque inhérent à certaines formes de coordination. Cette analyse permettra notamment de dégager la notion clé d'« interintentionnalité ». Le chapitre 5 expose les voies de passage d'une simple coordination à une véritable coopération. Il discute les différents modes de coopération et montre que le succès de leur fonctionnement reste dépendant du nombre des candidats à cette coopération et de leurs modes d'organisation en réseaux. Une analyse plus poussée des modes d'accès à la coopération révèle ainsi l'importance des formes de coopérations conditionnelles, souvent négligées par les approches traditionnelles.

Notre but n'est pas ici d'exposer des résultats déjà bien connus en théorie des jeux, mais plutôt de les mettre en perspective, et de proposer de nouvelles interprétations des distinctions que suggère leur confrontation aux données mises en évidence par la psychologie expérimentale et la neuroéconomie. En amont de la coordination il faut prendre en compte cette « interintentionnalité » qui se manifeste entre les sujets. Ce travail aura été facilité par l'analyse préalable, développée dans la première partie, concernant la structure du sujet, qui intègre déjà autrui, les tiers, sans oublier l'émergence d'un point de vue du groupe. Cela nous a permis, à propos de la notion de confiance, qui occupe une position centrale dans ce dispositif, d'introduire une distinction éclairante entre deux de

ses modalités différentes, une « confiance-cadre » et une « confiance-pari ».

Nous traiterons enfin dans une troisième partie, intitulée « Règles et normes », le niveau des interactions que l'on considère d'ordinaire comme paradigmatique du social : celui qui concerne les conventions et les règles et renvoie à des normes. On a d'abord saisi dans les « conventions » (au sens de Lewis : des conduites que chacun a intérêt à tenir s'il sait que les autres vont faire de même, et si cela est de savoir commun) une manière de réduire le social à la capacité des individus à se représenter les états mentaux des autres. Mais la connaissance commune qui serait exigée, en bonne logique, pour réussir cette opération (savoir que tu sais que je sais, que je sais que tu sais, etc., jusqu'à l'infini) dépasse, dans la réalité, nos capacités mentales. Un problème qui a été clairement identifié par les théoriciens des jeux, mais qu'ils n'ont pas réussi, jusqu'à présent, à résoudre de manière satisfaisante. Le chapitre 6 dégage des conditions plus réalistes de développement de pratiques qui installent de manière, au moins implicite, des règles de conduite communes débouchant sur des « standards de comportements » répondant aux intuitions des fondateurs de la théorie des jeux. Comment rendre compte de l'émergence et du maintien de ces règles ? Pour y parvenir, il est nécessaire d'adopter une perspective dynamique, d'où le recours aux théories des jeux évolutionnaires. Pour expliquer, en outre, le maintien et la reproduction de comportements empathiques, si ce n'est altruistes, au cours de l'évolution, nous nous devons, de fait, de raisonner à une échelle plus globale que celle d'un individu – à l'échelle de sa parenté, mais aussi de ceux qui ressemblent à ses parents, etc.

Une majorité des travaux expérimentaux dérivés de la théorie des jeux portent sur différentes modalités de partage d'un bien entre les joueurs (jeu du dictateur, jeu de l'ultimatum, jeu de la confiance…). On a cherché, de cette manière, à dégager les conditions qui président concrètement à l'émergence de normes d'équité. Des travaux combinant l'expérimentation et l'imagerie cérébrale ont, dans cette perspective, mis en évidence des comportements, à première vue surprenants, qualifiés de « punition altruiste », dont l'interprétation a suscité, et suscite encore, d'âpres discussions. Plusieurs sujets observés dans cette expérience préfèrent, en effet, punir ceux dont le comportement n'a pas respecté les règles tirées de ces normes d'équité, même si cette punition s'accompagne, pour eux, d'un coût financier. Certaines activations neuronales observées à cette occasion portent à penser que ces sujets éprouvent un plaisir singulier à punir. Mais on peut aussi, soutenir, au vu d'autres zones cérébrales qui sont également activées, qu'ils tiennent compte de l'opinion des tiers. On peut même élargir cette perspective et imaginer que ces individus veulent acquérir une réputation dans le groupe, et sont ainsi, peut-être, à la recherche d'une reconnaissance sociale. De telles interprétations n'étant pas nécessairement contradictoires.

Le chapitre 7 s'attache aux normes proprement dites, celles par lesquelles le social se manifeste formellement, à travers des règles explicites. Les théories qui assimilent les normes à des signaux indicateurs ne sont pas suffisantes ici, pas plus que celles qui les font émerger, plus ou moins spontanément, dans des jeux évolutionnaires, ou celles qui les réduisent à des connaissances partagées sensibles aux différents contextes. Elles peuvent, certes, rendre compte de l'émergence et du maintien

de règles implicites, mais elles ne parviennent pas à dégager les effets spécifiques propres à l'explicitation des normes.

Celle-ci fait, en effet, intervenir des activités de métacognition qui nous informent du mode sous lequel nous nous rapportons à une représentation, et nous permettent ainsi de distinguer des niveaux d'explicitation. Les normes explicitées dans cette perspective ouvrent deux possibilités d'enquête. La première porte sur la question de savoir si les autres se rapportent bien à une règle donnée de la même manière que nous – pour savoir, par exemple, non seulement s'ils coopèrent, mais s'ils coopèrent dans le sens où nous pensons coopérer. La seconde concerne les conditions qui peuvent nous inciter à réviser ces règles. Elle conduit au problème posé par le contrôle des règles elles-mêmes, et touche ainsi au fondement du social et du politique. On songe évidemment d'abord au droit, mais on peut également considérer, dans cette perspective, d'autres types de normes, comme les normes religieuses et les normes scientifiques, voire des normes esthétiques. Nous verrons que l'analyse de ces extensions fournit des aperçus stimulants. Cet accent placé sur le contrôle nous rappelle, enfin, les implications directes de ces règles explicites sur le contrôle exercé par les sujets sur eux-mêmes.

Voici les principaux points sur lesquels nous pensons avoir apporté des éclairages susceptibles d'étendre et d'enrichir la compréhension de nos interactions :

- La spécification de la variété des capacités qui doivent se mettre en place et se combiner de manière dynamique et complexe pour assurer le fonctionnement de nos interactions.
- La nécessité de penser l'intentionnalité subjective comme une interintentionnalité.
- L'exigence de prendre en compte dans les interactions intersubjectives les références aux perspectives de tierces personnes, qu'elles soient présentes ou absentes, voire anonymes.
- L'identification de la différence entre interactions à courte et à longue portée, déjà mise en jeu pour déterminer ce qui est pour nous perceptivement saillant, et ses diverses implications sur les choix intertemporels.
- L'importance de la notion de coopération conditionnelle, et la nécessité de différencier les différents types de coordination et les différents modes de coopération, en replaçant les modèles de théorie des jeux dans les contextes d'interactions sociales qui les ont inspirés.
- La distinction entre, d'une part, le fonctionnement implicite de règles dans des pratiques et des usages, et d'autre part, le rôle des normes explicitement formulées, en la reliant à différents niveaux de connaissance et à différents niveaux de contrôle des conduites, qui se superposent les uns aux autres.

L'ouvrage est le fruit d'un long et stimulant dialogue entre l'économiste et le philosophe. Chaque chapitre est rédigé par l'un des auteurs et suivi d'un commentaire, parfois tout aussi développé, écrit par l'autre. Les chapitres 1, 3 et 7 ont été écrits par Pierre Livet, les chapitres 2, 4, 5, et 6, par Christian Schmidt.

P

Pour mener à bien leur investigation sur l'intelligence de nos interactions, l'économiste et le philosophe ont utilisé des travaux réalisés par une nouvelle discipline, la neuroéconomie. Ils pensent tous les deux que ces recherches permettent de faire avancer nos approches sur les conduites humaines et qu'elles sont porteuses d'informations prometteuses, et parfois même inédites. Certes, il s'agit d'une discipline dont tous les résultats ne sont pas encore établis, et où l'enthousiasme des auteurs de ces travaux les a parfois conduits à en surévaluer l'importance. Il nous a donc fallu réévaluer les résultats des différents travaux de neuroéconomie qui ont servi de base à nos réflexions à leur juste mesure, en discutant leur portée et en revenant parfois sur les interprétations qui ont été données de leurs résultats. Nous aborderons de manière plus spécifique ces questions de méthode lorsque nous nous référerons à telle ou telle expérience, mais nous pouvons d'ores et déjà indiquer au lecteur ce que nous entendons ici par « juste mesure ».

Les économistes qui s'investissent aujourd'hui dans les programmes de recherche de neuroéconomie sont souvent tentés de développer, en leur faveur, l'argument schématique suivant : grâce aux différentes données recueillies sur le fonctionnement du cerveau au cours des expériences auxquelles sont soumis les sujets, nous allons enfin savoir comment raisonnent et décident réellement les êtres humains. Nous pourrons alors ajuster à ces comportements effectifs des modèles de décision et d'interactions qui renouvelleront les modèles formels construits par la théorie économique et par la théorie des jeux pour en rendre compte. Cet argument, dans sa simplicité, peut susciter des illusions. L'utilisation de l'imagerie cérébrale fournit des résultats passionnants, mais l'interprétation de ses résultats se révèle souvent problématique. Ce ne sont pas, du reste, les seules sources où puisent aujourd'hui les neurosciences. La biochimie, avec la sécrétion des neurotransmetteurs qui activent les différentes régions cérébrales, complète les informations fournies par l'imagerie. La génétique, avec l'étude de certaines spécificités cellulaires, se trouve également parfois sollicitée.

En nous centrant ici sur le dispositif le plus souvent utilisé par les chercheurs en neuroéconomie, qui consiste à appareiller un protocole expérimental avec un procédé technique d'imagerie cérébrale, nous avons identifié quatre types principaux de problèmes méthodologiques.

1) Le premier type de problème est d'ordre technique, et c'est aux chercheurs en neurosciences de le résoudre. Il tient à ce que les « images » ainsi fournies sont très dépendantes des procédés utilisés pour traiter les données. Rappelons, tout d'abord, que ce que transmet l'imagerie cérébrale correspond le plus souvent aux variations de la pression sanguine dans les régions activées. L'intelligence de ces données exige alors une interprétation qui est fournie par des modèles statistiques. Cette interprétation statistique peut susciter des discussions à ce niveau. En outre, les différentes techniques d'imagerie utilisées n'ont pas le même degré de finesse dans leurs définitions spatiales et temporelles. Il faut en tenir compte au moment d'interpréter leurs données respectives. De manière plus générale, le problème se pose de définir le seuil à partir duquel on décide que telle activation dans telle zone cérébrale est considérée comme significative (ou plus précisément, de la différence entre les activations de deux tâches, mais aussi entre les activations des tâches et l'activation du cerveau « au repos »).

Deux observations complémentaires doivent être faites à ce sujet. En premier lieu, il faut distinguer, aussi précisément que possible, l'activation de certaines régions cérébrales de la désactivation qui correspond à l'inhibition d'autres régions. Or cette inhibition est plus délicate à observer, dans la mesure où, même au repos, le cerveau entretient toujours une activité. Cela conduit à une seconde observation. L'état de repos appelé par commodité « mode par défaut » présente ses particularités, qui correspondent sans doute également à des tâches propres. Peu de travaux lui ont encore été consacrés, mais nous verrons, dans la suite des développements de cet ouvrage, que de premiers résultats sont porteurs d'informations intéressantes.

Il faut, par conséquent, croiser les techniques et comparer les procédés pour utiliser au mieux les ressources mises à notre disposition par l'imagerie cérébrale.

2) Le deuxième type de problème est relatif à l'identification des fonctions des zones activées (et désactivées). Les neurosciences progressent, en détectant, d'une part, des coactivations de plusieurs zones diversement localisées, que l'on peut penser fonctionner en réseau – on vient de citer le réseau du « mode par défaut » – et en identifiant, d'autre part, avec plus de précisions des différences locales d'activation, selon les variations des tâches et des conditions. Ainsi, certains neuroscientifiques pensent, par exemple, avoir montré que troubler le fonctionnement de la partie droite du cortex préfrontal dorsolatéral réduit la tendance à rejeter des offres inéquitables, mais que ce n'est pas le cas pour la partie gauche (Knoch *et al.*, 2006). Chacune de ces deux approches soulève de nouvelles questions.

La première approche confronte les neurosciences à une question familière aux économistes. Comment faut-il modéliser les coactivations observées dans certaines aires cérébrales, de manière à les relier à certaines fonctions ? Les chercheurs en neurosciences sont ainsi passés de l'étape initiale, qui consistait simplement à identifier les régions cérébrales activées lors de la réalisation de certaines tâches incorporées dans un protocole expérimental (choisir telle option plutôt que telle autre, accepter ou refuser un partage ou une offre, etc.), à une étape plus avancée, qui vise à repérer un ou plusieurs systèmes d'activations. Des débats sont alors apparus entre ceux qui pensent que les mécanismes de la décision mettent en jeux plusieurs systèmes cérébraux différents (un système des émotions, un système de la cognition,… ou même plusieurs systèmes distincts pour chacune de ces fonctions), qui pourraient être concurrents, et ceux qui soutiennent, au contraire, que ces différents systèmes font partie d'un seul système d'ensemble (Rustichini, 2008). Derrière cette question se profile une interrogation plus générale qui rapproche curieusement les neurosciences de la science économique : le cerveau fonctionne-t-il comme un système décentralisé, dans lequel ses différentes composantes sont, en quelque sorte en concurrence, ou est-il, au contraire, régi de manière centralisé ? Par-delà cette analogie, la compréhension de la régulation des activités cérébrales constitue un véritable défi pour les neurosciences.

La deuxième approche se concentre sur les activations locales et s'efforce de dégager l'organisation hiérarchique des réseaux identifiés, en la décomposant en modules et en sous-modules de réseaux. On pourrait parler à son sujet de microneuroscience. Son développement dépend

étroitement de la puissance technique des appareils d'imagerie, mais il est également tributaire d'un bon usage de modèles de systèmes hiérarchiques à niveaux multiples. Là encore plusieurs formules peuvent être avancées.

Il reste, alors, à trouver les voies permettant d'articuler entre eux les résultats mis en évidence par chacune des deux approches et l'on retrouve, à ce niveau, la quête d'une modélisation appropriée.

3) Le troisième type de problème concerne les relations entre, d'une part, les hypothèses neuronales formulées par les chercheurs en neurosciences, sur la base de connaissances qui restent assez dépendantes des techniques de l'imagerie, et, d'autre part, le recours à des protocoles expérimentaux, plus ou moins directement dérivés de la théorie des jeux. C'est en empruntant, et parfois, en simplifiant, des protocoles de jeux conçus par l'économie expérimentale qu'une véritable neuroéconomie des jeux s'est développée. Ainsi, par exemple, le jeu du dictateur, le jeu de l'ultimatum, le jeu de la confiance dans leurs différentes variantes servent aujourd'hui de soubassements expérimentaux à la majorité des recherches neuroéconomiques portant sur les interactions. À l'origine, ces protocoles expérimentaux avaient été imaginés par certains théoriciens des jeux, en vue de tester la valeur empirique d'hypothèses et de résultats énoncés par la théorie des jeux, comme la rationalité individuelle des joueurs, ou l'équilibre de Nash vers lequel tendraient, ou tout au moins devraient tendre, leurs transactions. Forts de résultats, le plus souvent négatifs, obtenus au cours de ces expériences, les chercheurs en neurosciences s'en sont servis, avec ou sans la complicité d'économistes, pour valider, avec l'aide de l'imagerie, de tout autres hypothèses, relatives, par exemple, à la prépondérance des émotions dans les prises de décision, ou à la domination de l'empathie, dans les relations à autrui. Un tel exercice s'expose à la critique. La théorie des jeux est d'abord une construction logique. L'interprétation des résultats obtenus en exposant certains de ces segments à des tests empiriques soulève déjà des difficultés, qui ont été discutées par les théoriciens des jeux eux-mêmes (Aumann, 1999). Réutiliser ces résultats pour les réinterpréter dans un cadre conceptuel tout à fait différent, comme celui développé par les neurosciences, risque de se révéler hasardeux, ou, à tout le moins, arbitraire, lorsque l'on ne l'accompagne pas de précautions méthodologiques. Nous verrons, par exemple, que le sujet qui fait confiance à l'autre joueur plutôt que de maximiser individuellement ses gains immédiats n'est pas nécessairement un altruiste empathique étranger au calcul qui suit les intérêts du sujet.

En prolongeant cette perspective on se heurte à une difficulté supplémentaire. Chacun des protocoles correspondant à ces jeux expérimentaux renvoie à une situation spécifique. La tentation est grande, toutefois, de regrouper plusieurs d'entre eux, sous une même rubrique, du fait que l'on s'attend à des activations neuronales semblables, ou tout au moins voisines. Il en va ainsi, par exemple, des jeux associés aux notions de coordination et de coopération. Ils constituent, chaque fois, comme on le montrera, un ensemble assez disparate de situations qu'il faut ensuite déshomogénéiser pour analyser la signification que chacune apporte. Car le risque serait alors d'associer trop rapidement à l'ensemble d'une catégorie de situations (coordination, coopération) les caractéristiques neurophysiologiques et neurobiologiques mises en évidence dans une seule, ou seulement quelques-

unes de ces situations. La comparaison des résultats obtenus dans deux ou plusieurs jeux différents, qui mettent chacun en évidence certaines manifestations distinctes d'une même notion requiert, pour cette raison, un soin particulier.

4) Le quatrième type de problème est plus général, bien qu'il s'applique ici au cas particulier des neurosciences. Les catégories d'arrière-plan qu'on ne peut se passer de présupposer pour imaginer des expériences et pour les interpréter, et les catégories utilisées pour mettre en place ces expériences et en tirer des différences significatives ne sont ni de même structure ni de même grain ; d'où des difficultés pour passer des unes aux autres. L'expert en neurosciences, pour déterminer quelles sont les fonctions liées à l'activation d'une zone, va présupposer que la tâche qu'il propose au sujet dans son expérience induit un certain état mental, qu'il va, par exemple, nommer « plaisir », ou « peine », ou « perplexité » (dans le cas où il pèse le pour et le contre). Ce sont là des catégories de la psychologie ordinaire, qui n'ont pas le même degré de finesse que les distinctions qui permettent au neurologue de repérer des différences d'activation et de différencier des zones (en l'occurrence, respectivement, le striatum, l'insula antérieure, et l'ACC, cortex cingulaire antérieur) En se référant à la chimie et au fonctionnement de neurotransmetteurs, comme la dopamine et la sérotonine, le neuroscientifique peut également utiliser le terme de « circuit de la récompense ». Il renvoie alors à des catégories de la psychologie expérimentale – où l'on a souvent amené des rats à répéter une action qui leur procurait en récompense de la nourriture. Mais pour un humain, la catégorie de la récompense renvoie à des scénarios qui mettent en jeu des interactions et plusieurs personnes, tandis que le plaisir peut aussi être solitaire et passif. Il devient dès lors délicat d'utiliser un terme pour l'autre.

Au fur et à mesure, l'accumulation des expériences diverses qui rendent active telle ou telle zone cérébrale amène le neuropsychologue à superposer différentes fonctions dans une même zone, avec la tentation de distinguer des sous-zones pour mieux s'y repérer. Il devient donc plus sensible à la complexité du fonctionnement cérébral. Mais il pourrait aussi avoir tendance à ajouter une nouvelle fonction chaque fois qu'une tâche différente est proposée, ou inversement à ranger sans trop d'examen plusieurs tâches dans la même classe, alors qu'on pourrait, tout aussi bien, insister sur les capacités différentes qu'elles mettent en jeu. Cela rajoute des suppléments de complexité à l'analyse du système déjà compliqué qui caractérise le fonctionnement cérébral. Pour ne pas les multiplier, il faut éviter, d'un côté, de projeter sur des localisations cérébrales des fonctions définies trop grossièrement, et, de l'autre, d'invoquer pour toute différence observée une différence de scénario cognitif. Dans ces deux cas l'interprétation des résultats serait suspecte de circularité, puisque sous une catégorie trop vague, on pourra toujours subsumer des résultats variés, et qu'il sera toujours possible de partir de différences dans les résultats pour trouver des différences dans le contexte de l'expérience.

Le problème posé par la superposition de différentes grilles catégorielles qui n'ont, ni les mêmes origines, ni les mêmes inspirations et ne sont donc pas assurées de situer leurs différents résultats dans un espace commun est encore plus délicat. Comme souvent en sciences expérimentales, il est nécessaire d'avancer en parallèle sur deux voies. D'une part, il faut construire des protocoles d'expériences dont les interprétations sont les moins équivoques possibles et qui se différencient donc

entre elles, et rechercher alors des résultats d'expériences conformes aux hypothèses et qui confirment ainsi ces différenciations. D'autre part, en sens inverse, on risque d'obtenir beaucoup d'informations alors qu'on s'attendait pour deux tâches jugées semblables à observer des zones et types d'activations similaires, l'expérience peut montrer que ces activations sont différentes. Cela permet ainsi de remettre en cause les catégories d'arrière-plan que l'on avait présupposées.

Il y a donc deux manières de briser la circularité qui posait initialement problème : rendre les conditions d'interprétation de l'expérience plus restrictives que les présuppositions initiales, ou obtenir des résultats contre-intuitifs. Ces deux voies ne sont, du reste, pas mutuellement exclusives.

Les difficultés méthodologiques, qui ont été répertoriées ici pour les besoins de l'approche que nous avons privilégiée dans notre enquête sur les interactions, doivent être situées dans le cadre plus large des problèmes d'ordre épistémologique posés par la neuroéconomie[1]. À l'origine, la neuroéconomie se présente comme une manière d'hybridation scientifique entre, d'un côté, plusieurs disciplines qui participent aux neurosciences (neurophysiologie, neurochimie, neurobiologie et neuro-imagerie) et, d'un autre côté, différentes composantes de l'analyse économique (théorie de la décision, théorie de l'équilibre, théorie des jeux). Le terme même de neuroéconomie a été, du reste, popularisé par l'ouvrage d'un spécialiste des neurosciences consacré à l'analyse neuronale des mécanismes décisionnels (Glimcher, 2003). L'intuition qui guidait Glimcher était de retrouver dans le fonctionnement du cerveau décideur certaines propriétés qui avaient été formalisées par les modèles de la théorie économique de la décision. Ainsi a-t-il traqué la signature cérébrale de plusieurs concepts familiers aux économistes, comme la maximisation, l'optimalité et l'équilibre, notamment dans l'acception de Nash en théorie des jeux. Selon cette perspective, il reviendrait aux neurosciences de fournir le substrat matériel de l'objet étudié, par l'intermédiaire de la physiologie et de la biochimie du cerveau, tandis que la modélisation de l'activité cérébrale serait apportée par la théorie économique.

En dépit de son apparente clarté, ce partage simplificateur des tâches entre neurosciences et économie se révèle inexact et même trompeur. Il souffre d'une absence. Entre la physiologie et la chimie du cerveau, d'une part, et les modèles mathématiques de la décision, d'autre part, il faut nécessairement prendre en compte les comportements des agents décideurs, qui font d'abord l'objet d'observations, avant de pouvoir être expliqués. Or, indépendamment de toute référence au cerveau des neurosciences, le développement de la psychologie expérimentale a révélé aux économistes que, dans bien des cas, leurs modèles théoriques de prise de décision se trouvaient invalidés dans un grand nombre de leurs expériences. Ce constat a, du reste, provoqué dans les rangs des économistes, un mouvement en sens inverse. Il a poussé certains d'entre eux à rechercher dans les informations fournies par les sciences du cerveau, des matériaux leur permettant de construire des modèles alternatifs plus réalistes, au sens de leur congruence avec les résultats observés expérimentalement. Il y aurait ainsi deux voies différentes et à certains égards opposées, qui ont conduit des chercheurs des deux disciplines à se rapprocher, en vue d'élaborer un programme de recherche commun ; d'où les malentendus qui entourent, parfois même encore, ce programme neuroéconomique.

Une illustration en est fournie par les ambiguïtés suscitées par les utilisations qui ont été proposées, de part et d'autre, des modèles bayésiens du traitement de l'incertitude. La théorie économique eut, tout d'abord, recours à une modélisation d'inspiration bayésienne pour formaliser l'inférence statistique qui préside (ou devrait présider) à la logique des choix dans le cadre d'un calcul de probabilités dites « subjectives ». Ainsi, dans le modèle classique de l'utilité espérée, les probabilités associées aux anticipations des agents sont conditionnelles aux conséquences attendues des choix qui seront arrêtés. Si ces conséquences sont identiques, quelle que soit l'option choisie, on en déduit, conformément à la règle bayésienne, que la probabilité qui est associée à son occurrence est indépendante. Comme elle n'est, en l'occurrence, porteuse d'aucune information nouvelle pour le décideur, elle est alors considérée sans pertinence pour son choix (Savage, 1954). Depuis la célèbre expérience d'Allais (1953), de très nombreuses expériences ont cependant montré que les choix raisonnés des agents s'écartaient souvent sur ce point de ce modèle. De fait, cette première version du modèle de l'utilité espérée ne cherche pas à rendre compte des probabilités subjectives des agents obtenues par induction de valeurs observées, mais propose seulement un mode opératoire pour un choix « rationnel » dans un univers probabilisé. Il s'agit par conséquent d'une version statique du bayésianisme. Les véritables difficultés apparaissent avec l'introduction de la dynamique, lorsque la perception du temps par les agents transforme et complique, comme on le verra, la logique de leurs choix. L'hypothèse simplificatrice tirée de Bayes, selon laquelle la distribution de probabilités *a posteriori* (*Posterior distribution*) serait la même que la distribution initiale (*Prior distribution*), au nouveau facteur près, semble alors souvent remise en cause expérimentalement dès que l'on fait intervenir la temporalité dans ce schéma. Il se confirme aujourd'hui qu'entre le temps *t* de la *Prior distribution* et le temps *t'* de la *Posterior distribution*, d'autres éléments que la simple information fournie par cette nouvelle donnée interviennent dans la décision réfléchie des agents.

De leur côté, plusieurs chercheurs en neurosciences, estimant plus récemment que la méthodologie dominante dans leur discipline restait principalement descriptive, ont utilisé une approche bayésienne, pour formaliser dans des modèles calculables (*computational*) les processus de transmission de l'information au sein des multiples niveaux des réseaux neuronaux. Cette approche consiste à traiter le système nerveux comme un mécanisme qui réagit aux informations fournies par son environnement, selon un schéma bayésien d'inférence statistique qui traduit mathématiquement une actualisation permanente des informations mémorisées à la lumière des résultats obtenus. La dynamique de cet « apprentissage renforcé » fournirait alors une base théorique pour comprendre le phénomène central du codage neuronal (Dayan et Abbot, 2001 ; Doha, Ishi, Pouget, Rao, 2007). Si cette formulation bayésienne d'une neuroscience théorique n'est pas partagée par tous les chercheurs, elle constitue néanmoins, aujourd'hui, ce que l'on pourrait appeler le noyau dur de cette nouvelle discipline.

On serait tenté de voir, dans ces deux démarches opposées, une manière de contradiction. D'un côté, les expériences de neuroéconomie portant sur des choix en avenir incertain, en combinant des protocoles expérimentaux à des explorations d'imagerie, semblent infirmer l'idée selon laquelle le

travail cérébral, suivrait, en ces circonstances, une simple logique d'inférence bayésienne. En sens inverse, la référence à un mécanisme stimulus – réponse qui guiderait les tâtonnements de nos croyances, selon une règle bayésienne, semble aujourd'hui en phase avec le fonctionnement cérébral observé à l'occasion des différentes tâches motrices élémentaire. Plusieurs travaux expérimentaux réalisés sur la perception et, en particulier sur la perception visuelle, paraissent valider ce schéma d'un fonctionnement cérébral qui suivrait cette dynamique d'inférence bayésienne (Knill et Pouget, 2004 ; Simonelli, 2009). Comment, dès lors, concilier les données contradictoires transmises par ces deux séries d'informations ?

Une première réponse s'appuie sur le fait que les expériences évoquées dans les deux cas ne se rapportent pas exactement aux mêmes fonctions cérébrales. Il s'agit, dans le premier cas, de choix conscients de la part de sujets placés dans différentes situations sur lesquelles ils disposent d'informations objectives calibrées selon un répertoire précodé (conséquences, probabilités…). Le second cas concerne le comportement de sujets confrontés à différents stimulants visuels qu'il convient d'abord d'encoder pour leur permettre de donner sens à leur perception. On serait tenté d'avancer ici une manière de paradoxe. Le mécanisme largement inconscient de la perception visuelle en avenir incertain suivrait ainsi une loi proche de la simple logique statistique, que l'on ne vérifie plus lorsqu'il s'agit d'un choix raisonné. Ce paradoxe cependant n'est qu'apparent. Si, en effet, les choix raisonnés en incertitude qui ont été observés s'écartent de cette formulation élémentaire de la logique bayésienne, il serait inexact d'en tirer une preuve de leur irrationalité. Il semblerait plutôt qu'ils s'inscrivent dans une logique plus flexible et sans doute plus subtile, qui a été formalisée mathématiquement par des fonctions cumulatives de rang, ou, plus récemment, par une généralisation de fonctions logarithmiques bien connues (Takahashi, 2009, 2013). Cet écart par rapport au modèle simple de la logique bayésienne élémentaire pourrait dès lors s'expliquer par le fait que les choix risqués sont des opérations beaucoup plus complexes que ne le pressentaient initialement les économistes, parce qu'elles mobilisent, à la fois, plusieurs fonctions cérébrales différentes. Or ce sont précisément les travaux de chercheurs en neurosciences qui ont mis en évidence que ces choix réfléchis font nécessairement intervenir, outre le raisonnement, une relation émotionnelle à la dimension temporelle, qui caractérise précisément les anticipations. En sens inverse, il n'est pas surprenant que les systèmes neuronaux qui régulent l'adaptation non consciente de nos perceptions à un environnement inconnu fonctionnent selon les règles statistiques plus rudimentaires de l'apprentissage bayésien. Il serait pour autant téméraire d'induire de ce constat, comme sont tentés de le faire aujourd'hui certains spécialistes des neurosciences, que ces règles bayésiennes peuvent servir de paradigme pour modéliser le fonctionnement de l'ensemble des activités cérébrales. Le vrai problème qui nous est posé est donc plutôt celui de comprendre comment les mécanismes premiers d'apprentissage de l'incertitude se trouvent transformés dans les projections intertemporelles conscientes qui guident nos choix.

Une seconde réponse fait intervenir une autre distinction, de caractère épistémologique cette fois. Les résultats mis en évidence dans le cas de la perception ne prouvent pas que le fonctionnement du cerveau soit bayésien, mais seulement que ces résultats correspondent à ceux que fournirait un

mécanisme bayésien. En d'autres termes, le bayésianisme invoqué à cette occasion par les chercheurs en neurosciences est de caractère *instrumentaliste* et ne concerne pas le *réalisme* d'une telle hypothèse, qui impliquerait, cette fois, que le mécanisme cérébral qui donne naissance à ces résultats soit lui-même bayésien (Colombo et Series, 2012). Des modalités du fonctionnement cérébral qui ne correspondent pas à un mécanisme bayésien peuvent, en effet, fort bien aboutir à des résultats interprétés, ou seulement même interprétables, en termes bayésiens. On l'observe notamment dans certains jeux répétés, lorsque, par exemple, un mécanisme mental d'aversion au regret qui détermine les choix des joueurs peut être traduit dans les termes d'un algorithme d'apprentissage d'inspiration bayésienne (Foster et Young, 2003, 2006). Il n'en résulte pas, pour autant, que le mode de fonctionnement du cerveau de ces joueurs soit, lui-même, bayésien.

La posture épistémologique adoptée aujourd'hui par les tenants d'une théorie bayésienne des neurosciences évoque curieusement un *instrumentalisme* assumé il y a plus d'un demi-siècle par l'économiste Milton Friedman, lorsqu'il défendait ce qu'il appelait une théorie économique *positive*, contre une théorie *réaliste* (Friedman, 1953). Les choses ont changé, et c'est en quête d'une théorie réaliste que les économistes engagés dans la neuroéconomie s'interrogent aujourd'hui sur le fonctionnement du cerveau. Il serait pour le moins troublant que les chercheurs en neurosciences se prévalent maintenant de ce type de positivisme que les économistes intéressés aux processus réels qui guident les décisions des agents et leurs interactions ont, justement, abandonné. Nous prendrons, pour notre part ici, le pari que cette quête de réalisme se trouve également partagée par une majorité des scientifiques des neurosciences qui participent à ce programme de recherche commun.

Cette évocation des ambiguïtés qui accompagnent le recours à des modèles bayésiens dans les neurosciences de la décision a été choisie ici, parce qu'elle nous a paru emblématique des obstacles épistémologiques rencontrés par l'approche interdisciplinaire qui a été choisie dans cet ouvrage. On les rencontre dans d'autres questions également traitées dans cet ouvrage (l'appréhension de l'autre, la référence à des normes…).

Notre ambition n'est pas ici de résoudre toutes ces questions, mais plutôt de les éclairer, en nous interrogeant sur la pertinence des articulations entre neurosciences et théorie des jeux, et d'en dégager les interprétations qui nous paraissent les plus instructives pour mieux comprendre les interactions sociales.

---

1. Sur cette question C. Schmidt, *Neuroéconomie*, Paris, Odile Jacob, 2010.

PARTIE 1

INTERACTIONS ET INTERSUBJECTIVITÉ

CHAPITRE 1

L’intersubjectivité

Qui dit intersubjectivité dit subjectivité. Cependant il serait difficile de rendre compte des interactions sociales en partant d’une conception solipsiste du sujet. Les relations et les processus d’interaction apparaîtraient alors extérieurs aux sujets. Le simple amorçage des relations entre des sujets enfermés chacun dans leur bulle demanderait des constructions compliquées. Chaque sujet devrait en lui-même trouver tous les ingrédients de l’interaction et en construire sa version, puis il devrait aussi trouver en lui-même les moyens de raccorder cette version à celle d’autrui, un autrui qu’il aurait dû aussi construire par ses propres moyens.

On peut s’épargner une partie de ces efforts en admettant que toutes les capacités du sujet ne sont pas pleinement disponibles en lui-même pris isolément et qu’elles ont besoin pour se déployer de s’appuyer sur des situations d’interaction déjà agencées par ses congénères. Pour naître à la subjectivité, nous avons besoin d’interactions qui nous transforment, par lesquelles nous transformons les autres, qui nous relient à des partenaires mais aussi qui nous plongent dans une collectivité.

Le sujet est donc interactif. Il est d’usage de distinguer trois dimensions de ces interactions, même si dans la pratique il est difficile de les séparer : les interactions avec l’environnement, celles avec des partenaires individuels, et celles avec des membres de notre groupe d’appartenance ou d’un autre collectif – les interactions proprement sociales. Mais ces interactions ont évidemment des incidences sur le sujet lui-même, et y déclenchent d’autres interactions, qu’on aura tendance à dire intrasubjectives par simplification, entre les processus sensorimoteurs, cognitifs, affectifs, et actionnels. Le sujet se constitue à la charnière entre ces interactions intrasubjectives et ces interactions environnementales et intersubjectives.

I

Cette perspective intersubjective traverse toutes les sciences sociales. Nous la rencontrons déjà chez l'un des fondateurs de l'économie politique, Adam Smith, qui analysait dans sa *Théorie des sentiments moraux* comment, en combinant des interactions entre partenaires et les perspectives d'observateurs, on pouvait passer des émotions aux évaluations morales. Plus proche de nous Peirce a développé une théorie de la signification qui exige, en plus d'un signe et d'un objet dont il dépend, un interprétant – en fait une chaîne de tels interprétants. Cela va donc au-delà d'une relation duale comme celle d'une causalité ou d'une action et réaction, et requiert une relation qui renvoie à une autre relation, comme « A donne C en présent à B ». Passer ainsi à trois termes peut impliquer des interactions qui renvoient à des tiers, ce qui permet à Peirce de relier la recherche du savoir à la succession de la dualité entre propositions et critiques dans une communauté toujours ouverte sur le futur, donc sur de tiers points de vue[1]. Dewey a proposé le concept de « transaction », qui fait de toute action une interaction avec les autres et avec notre environnement : le sujet humain n'est plus observateur extérieur d'interactions entre objets de la nature, mais impliqué en elle, et il ne réagit pas simplement à des stimuli, mais il les interprète selon ses attentes et tendances en cours.

On sait que le sociologue Simmel avait fait de l'interaction le pivot de ses travaux. Il montrait que toute interaction implique des effets en retour sur les partenaires, et qu'elle n'est socialement définie qu'une fois intégrés ces effets de rétroaction. Il insistait lui aussi sur le rôle, dans des interactions entre partenaires, du rapport au tiers – qui a un lien avec la position d'observateur chez Smith[2]. Citons encore Goffman, qui a développé une sociologie des interactions qui étudie les diverses modalités de mise en présence des acteurs sociaux. Il adopte une conception de l'interaction compatible avec celle de Simmel, mais analyse très précisément la manière dont les acteurs sociaux définissent le type de situation dont est censée relever leur interaction. Un type de situation indique des règles de déroulement de l'interaction et assigne des rôles à chacun. Goffman reconstitue les processus par lesquels ces sujets tentent de construire des situations dont ils puissent tirer parti pour poursuivre le jeu social sans « perdre la face », à condition de respecter ces « rites » du type d'interaction engagée. La psychologie sociale à la Moscovici a étudié la manière dont ces interactions diffèrent selon qu'elles ont lieu entre membres du même groupe ou entre membres de groupes différents. Aujourd'hui, une bonne partie des recherches en psychologie cognitive est consacrée à l'étude des modalités d'installation de la distinction entre soi et environnement et de la mise en place du rapport soi-autrui par contagion, imitation, simulation, ou encore constitution d'une théorie « naïve » de l'esprit des autres. Et les neurosciences, nous y reviendrons, sont mises à forte contribution dans ces analyses.

L'analyse de la dimension collective des interactions s'est partagée entre l'économie et la sociologie. La première l'aborde par la théorie de l'équilibre général d'un marché et surtout par celle du bien-être collectif et celle du choix social. Elle signale les difficultés que l'on a à passer de l'individuel au collectif : même en n'étant pas trop éloigné d'un état collectif d'équilibre dans un marché, on ne sait pas forcément comment changer les comportements individuels pour y parvenir. Il n'existe pas de fonction d'agrégation des choix individuels en un choix collectif qui puisse satisfaire

parfaitement toutes les exigences d'une rationalité sociale (nous reviendrons plus loin sur les perspectives ouvertes par la théorie des jeux). De son côté, la sociologie a souvent pris comme point de perspective le collectif, pour penser la pression sociale (Durkheim), les phénomènes de fusion et ceux de réciprocité (Tarde, Mauss), l'intériorisation des normes collectives (Parsons), et l'inscription des stratégies individuelles dans un cadre symbolique collectivement partagé (Bourdieu). En sens inverse, les tenants de l'individualisme méthodologique ont tenté de reconstruire les interactions sociales à partir de stratégies individuelles (Coleman, Boudon). Qu'on pense les interactions à partir des relations intersubjectives ou du collectif, elles restent au centre des études en sciences sociales.

D

Notre hypothèse est que le sujet se constitue dans des interactions de différents types, les suivantes s'appuyant sur les acquis des précédentes, selon plusieurs étapes. Nous pouvons recourir aux neurosciences pour mieux documenter les spécificités de ces étapes, en suivant le programme que nous nous sommes donné dans ce livre : proposer deux lectures de ces interactions, l'une par un philosophe, l'autre par un économiste théoricien des jeux, qui toutes deux s'appuient sur des résultats récents en psychologie et économie expérimentales et en neurosciences. Notre but, dans cette partie, est de rendre le lecteur sensible à la diversité des processus qu'il faut mettre en place pour assurer la constitution interactive de la subjectivité.

On pourrait penser que cette constitution exige seulement l'existence préalable de capacités de perception, cognition, motivation, motricité et communication propres au sujet, puis leur application aux autres humains, soit de manière individualisée, soit de manière collective. Or d'une part la constitution des capacités du sujet implique déjà celle de différences entre soi et non-soi, ainsi que de celles entre soi et autres soi ; d'autre part une application des facultés du sujet à autrui qui se ferait d'un bloc ne correspond pas à la réalité. Différentes modalités de rapport à autrui se mettent en place, chacune modifiant les capacités d'interaction du sujet avec autrui. Ces diverses capacités sont irréductibles les unes aux autres et elles mettent en jeu des versions différentes d'autrui, mais aussi des versions du sujet. Leur synergie est nécessaire pour notre intersubjectivité, elle-même nécessaire pour la constitution de notre subjectivité, contrairement au schéma d'une simple application de capacités de base à l'objet « autrui », que nous venons d'évoquer.

Nous allons seulement indiquer des modalités qui ont été bien distinguées par les travaux expérimentaux. L'ordre de cette exposition n'est pas forcément celui du développement chronologique. On est très tôt sensible aux expressions, par exemple, alors que nous n'en parlerons pas immédiatement. Certaines de ces capacités peuvent se développer en parallèle. Simplement celles de la fin de la liste présupposent celles de son début. Non seulement ces modalités sont différentes, mais la capacité de les faire fonctionner de manière distincte et articulée est aussi nécessaire.

A

Ces différentes modalités du sujet de l'intersubjectivité se constituent toujours à travers des interactions variées. Nous allons analyser chacune d'elles, pour montrer qu'aucune de ces modalités n'est une capacité sortie tout armée de notre cerveau avant des interactions. Les interactions sont nécessaires pour leur émergence. La preuve en est que les différents types d'interactions dans lesquelles nous sommes engagés font chacun apparaître un problème, et que chaque modalité est une réponse à un de ces problèmes.

1) Partons du système sensorimoteur[3]. Nos mouvements changent notre environnement – quand nous déplaçons des objets – et ils changent les dispositions de notre corps ainsi que notre position dans cet environnement. Nous rencontrons alors un problème : il se produit d'autres mouvements que les nôtres dans l'environnement. Nos mouvements et ceux d'autres mobiles produisent tous deux des mouvements apparents sur notre rétine. Nous disposons la plupart du temps d'un critère simple pour distinguer nos mouvements de ceux des autres mobiles : le mouvement de notre tête qui tourne, par exemple, crée une rotation apparente de tout notre environnement, alors que le mouvement d'un autre mobile se détache sur l'environnement, que celui-ci soit immobile ou qu'il soit animé d'un mouvement homogène. Mais si nous bougeons un objet, il se comporte comme un mobile qui se détache sur l'environnement. Il nous faut avoir enregistré que nous étions l'origine de ce mouvement pour pouvoir le distinguer des autres mobiles de l'environnement. On suppose alors que nous disposons d'une « copie d'efférence », donc d'une information sur la sortie (l'efférence) de la commande motrice que nous avons déclenchée, même si c'est là une dénomination ramassée pour des processus neuronaux qui semblent complexes. Notre interaction avec les objets que nous déplaçons se détache ainsi des interactions d'autres mobiles avec notre environnement. La comparaison de ces deux modes d'interaction est ainsi à l'origine d'une première distinction entre soi et non-soi. Nous retrouverons sans cesse cette caractéristique : nous sommes constitués non pas seulement dans des interactions, mais aussi dans la comparaison entre des interactions différentes, par une sorte d'interaction au carré.

2) Si nos actions atteignaient toujours leurs cibles, et nous permettaient toujours d'atteindre des états de satisfaction, il nous serait difficile de construire une subjectivité. Nous ne disposerions pas, en effet, de ce signal de distinction entre soi et non-soi qui tient à ce qu'entre nos actions et nos buts peuvent se dresser des obstacles. Inversement, si tout obstacle nous contraignait à changer de buts, nous ne pourrions pas maintenir face aux obstacles de l'environnement une constance dans les motivations du soi, et il perdrait de sa consistance. C'est donc en fait dans l'interaction entre une activité motivée, orientée vers un état de satisfaction qui passe par l'atteinte d'une cible dans l'environnement, et les obstacles qu'elle rencontre, que peuvent à la fois se maintenir les motivations du soi et s'ajuster nos activités aux contraintes de l'environnement, ce qui manifeste la stabilité et la résistance à la fois du soi et du non-soi[4].

3) Les neurones miroirs (découverts par Rizzolatti et son équipe) ont été observés d'abord chez des singes. Ils ont la propriété de s'activer aussi bien quand un singe déclenche un type de mouvement

– orienté vers un but – que lorsque l'expérimentateur (un animal humain) active un mouvement similaire vers un but similaire. Nous en possédons aussi. Le problème est qu'un tel dispositif pourrait nous amener à confondre le soi et le non-soi, si, quand nous voyions un tel mouvement chez autrui, nous étions poussés à déclencher de notre côté le même mouvement. Mais précisément, nous sommes capables, ainsi que les singes et d'autres animaux, d'activer nos neurones miroirs tout en inhibant le déclenchement du même mouvement (certains schizophrènes ont bien du mal à résister, et à ne pas déclencher les mouvements qu'ils voient accomplis par d'autres). La conjonction de ce couplage (activation des neurones miroirs) et de ce découplage (inhibition des neurones moteurs liés au même mouvement) nous fournit la structure paradigmatique d'une interaction qui permet de lier deux sujets tout en les distinguant. Nous allons retrouver bien des fois par la suite ce dispositif de couplage d'une activation et d'une inhibition, couplage qui permet d'entretenir une certaine activation (ici, celle du neurone qui interviendrait dans le mouvement s'il était déclenché) tout en inhibant certaines activations conséquentes (ici, le déclenchement du mouvement).

4) Même une fois autrui distingué du soi par l'inhibition du mouvement imitatif, nous pouvons encore rencontrer le problème de distinguer les mouvements de deux sujets qui visent une même cible selon des motivations qui ne visent pas seulement la cible, mais déclenchent une compétition entre eux, voire avec nous. Pour résoudre le problème de notre orientation dans cette structure triangulaire, le suivi du regard d'autrui, étudié par Baron-Cohen, nous est très utile. Il nous permet de découvrir l'objet de l'intérêt de chacun (par exemple le premier s'intéresse à un objet et le second à l'intérêt du premier pour cet objet). Nous pouvons alors choisir ou non de nous intéresser au même objet, d'entrer en compétition avec l'un pour posséder cet objet, d'aider l'autre, etc. Ce suivi des regards est d'ailleurs aussi nécessaire pour coordonner nos actions quand nous devons nous mettre à deux pour mouvoir un objet lourd.

5) Nous sommes très tôt sensibles aux expressions faciales d'autrui et aux affects qu'elles manifestent (les bébés tirent la langue quand on leur tire la langue, puis savent modifier leurs propres expressions en résonance avec celles de leur entourage). Ces expressions nous aident à résoudre un problème fondamental, celui d'identifier l'intentionnalité d'autrui, alors même que cette orientation mentale n'est pas directement observable. Ce que les philosophes appellent « intentionnalité », c'est la visée d'un référent, d'une cible, en tant que présentant un certain aspect. Si les mouvements visent la cible, l'expression peut indiquer l'aspect, par exemple qu'autrui veut se garder des risques que présente la cible ou bien bénéficier de ses attraits. Nous pourrons ainsi éviter un danger si nous détectons de la peur dans la posture d'autrui et sur son visage, ou un aliment toxique si nous y voyons du dégoût.

Nous ne serons cependant sensibles aux mouvements qui peuvent traduire des intentions que s'ils sont portés par des êtres que nous rangeons dans la catégorie des humains ou plus généralement des vivants animés d'intention. Nous les reconnaissons précisément au fait que chez eux se combinent mouvements orientés vers des cibles et expressions faciales qui révèlent à quel aspect de la cible il y a réaction. Nous n'attribuons pas d'intention nocive au robot qui a déclenché un mouvement qui se

trouve contrecarrer le nôtre (l'imagerie cérébrale le confirme : nos réactions neuronales envers les robots ne sont pas les mêmes qu'envers les humains), mais nous sommes tentés de nous mettre en colère quand un humain à l'expression hostile ou indifférente a déclenché un mouvement similaire (les robots « expressifs » n'auront donc pas que des avantages). Ainsi nous nous construisons comme sujets en apprenant à distinguer ceux qui sont des sujets de ceux qui ne le sont pas. Et nous les distinguons là encore en comparant deux modes d'interaction, celui des mouvements orientés et celui des expressions.

6) Examinons maintenant la différence entre imiter autrui et coordonner avec lui nos efforts pour une tâche commune, par exemple déplacer un meuble lourd. Certes il est toujours utile dans cette tâche de pouvoir identifier l'intention des mouvements d'autrui, par une sorte d'imitation interne, mais le problème est que cette tâche exige souvent que nous ayons des mouvements dissemblables et complémentaires – si vous avancez, et que j'aie pris le meuble face à vous, je dois reculer, donc aller à l'encontre de ma tendance à l'imitation. Dans cet exemple, l'objet commun et ses « réactions » à nos mouvements opposés sont là pour nous contraindre à la complémentarité. Un problème plus délicat se pose lorsque nous devons faire des mouvements préparatoires pour rendre possibles des mouvements futurs d'autrui. Pour le résoudre, il nous faut alors définir nos mouvements en fonction de la trajectoire dont nous *anticipons* qu'elle va être la sienne (par exemple amorcer une rotation du meuble transporté de l'horizontale à la verticale de manière à pouvoir plus loin passer par une porte étroite).

L'efficacité de ces coordinations est d'ailleurs surprenante. Dans certaines expériences, chacun des deux partenaires doit agir sur le levier à sa disposition pour influencer la position d'un pointeur sur un écran (Reed *et al.*, 2006). Or les deux leviers sont joints par une tige ; on pourrait donc penser que chacun entrave le mouvement de l'autre. C'est le cas au début, mais ensuite et très vite chacun spécialise son action, l'un contrôlant l'accélération du mouvement du pointeur et l'autre sa décélération (cela sans pouvoir nécessairement expliciter cette dualité de rôles). Cette coordination est même plus efficace que les mouvements d'un sujet qui doit accomplir à lui seul la tâche.

On peut se demander à partir de quel stade de ces développements on peut considérer comme directement pertinent le type d'analyse propre à la théorie des jeux[5]. Elle exige d'un joueur la capacité d'identifier les diverses intentions possibles d'autrui en réponse à ses propres intentions (une stratégie en théorie des jeux est une liste complète qui indique, pour chaque coup possible de l'autre joueur, le coup choisi en réponse par le joueur considéré). Les capacités données par les neurones miroirs ne suffisent pas, parce qu'elles ne permettent que des anticipations limitées – le mouvement n'est identifié qu'une fois déjà entamé. Le suivi des regards et des expressions peut nous permettre d'anticiper certains mouvements d'autrui, mais il ne nous indique pas de manière claire une multiplicité de différents mouvements possibles. Quand nous devenons capables d'imiter autrui pour l'imiter et donc d'activer des mouvements comme n'étant pas de notre chef – nous nous comporterions spontanément autrement – et que nous combinons cette capacité avec les précédentes, on peut penser que nous sommes capables d'anticiper différents comportements d'autrui. Il reste à savoir ajuster nos comportements sur ces différents comportements possibles, ce dont nous assure la capacité de nous coordonner avec autrui (cela n'implique pas que la théorie des jeux ne traite que de

jeux dits de coordination – les joueurs peuvent être en compétition – mais qu'elle implique une capacité de chaque joueur à évaluer chaque coup en corrélation avec tel coup d'un autre joueur). Il s'agit là de conditions minimales pour pouvoir tenir les sujets d'une interaction pour des joueurs de théorie des jeux. Nous verrons plus loin qu'on pourrait aussi exiger davantage.

7) Nous imitons aussi bien les mouvements d'autrui orientés vers une cible que ses expressions faciales, ses gestes expressifs ou ses postures, et ces partages peuvent être le fait de plusieurs partenaires, si bien qu'un « nous » commence à se constituer. Mais cela introduit une nouvelle complexité : dans un « nous », le rapport entre deux partenaires est observé par des tiers qui sont membres du même groupe. Or nos tendances à l'imitation peuvent nous amener à imiter des gens qui ne font pas partie de ce groupe. Une telle imitation doit-elle automatiquement impliquer, dans l'esprit de tiers qui sont membres du groupe, que nous partageons les visées et sentiments d'un extérieur au groupe ? En fait ce n'est pas forcément le cas, et même nous utilisons la réponse à ce problème comme un moyen de marquer les frontières du groupe. Notre imitation va s'adresser non pas à celui que nous imitons, mais aux tiers – ainsi reconnus comme référents du groupe. C'est ce qui se passe quand, au lieu que notre imitation soit un partage de l'activité de ce partenaire et de ses sentiments, nous imitons autrui pour l'imiter. C'est autrui qui devient la cible, et nous imitons autrui pour un public de tiers dont il est exclu, ce qui est une manière d'afficher notre appartenance au groupe en en excluant celui que nous imitons.

8) Constituer un groupe, c'est à la fois privilégier certaines interactions – entre les membres du groupe – et les distinguer d'autres interactions avec des personnes qui n'en font pas partie – là encore, on retrouve la comparaison entre deux modes d'interaction. Il faut pour cela résoudre un nouveau problème, celui de pouvoir distinguer trois sortes de tiers : ceux qui ne sont pas membres de mon interaction locale avec un autre, face à face, mais qui sont bien membres du groupe ; ceux qui sont bien en tiers par rapport au groupe, mais parce qu'ils sont membres de l'autre groupe ; enfin ceux qui ne sont membres ni de l'un ni de l'autre – les tiers « indépendants ». Certains travaux philosophiques sur les « nous », ou l'être ensemble, comme ceux de Searle ou de Margaret Gilbert, se concentrent sur la formation du groupe à partir d'individus qui sont ses membres, sans prêter attention au fait que se constituent en même temps ces différences entre tiers. Or on ne peut former un groupe à deux, sans le soutien de tiers qui font aussi partie du groupe, et sans se distinguer des tiers qui n'en font pas partie pour une des deux raisons que nous venons d'évoquer.

Nous avons vu qu'au sein d'un groupe, les expressions émotionnelles devaient pouvoir se propager. Or si nous sommes attentifs aux expressions de nos proches comme membres du groupe, nous sommes particulièrement attentifs à leurs expressions quand ils aperçoivent des visages inconnus de nous. Leur expression de méfiance suffit à nous faire rejeter le contact avec ces nouveaux venus, et cela plus encore si nous observons qu'il y a sur ce point contagion entre les expressions de nos proches. On a ici une structure au moins quadrangulaire : nous observons la convergence des expressions entre des proches qui observent un tiers inconnu. Leurs attitudes deviennent contagieuses, contagion d'autant plus influente qu'ils sont nombreux à converger ainsi.

On peut penser que quand un groupe a des réactions de rejet induites par des faciès supposés prototypiques, elles sont déclenchées par ces phénomènes de contagion expressive corrélés à quelques traits de faciès – ces traits ne sont pas nécessairement toujours les mêmes, ce peuvent être chaque fois quelques traits parmi ceux qui appartiennent à une gamme donnée. Ces réactions de contagion, comme celles que nous avons précédemment évoquées, sont au cœur des processus qui nous permettent de nous reconnaître « entre nous » comme faisant partie d'un groupe, par opposition à ceux que notre groupe n'admet pas en son sein. Elles ont pour fonction, plus peut-être que de constituer un antigroupe, de constituer le groupe et de renforcer ses liens. C'est parce qu'il ne nous est guère possible, dans nos réactions de défiance, de vérifier que nos convergences de réaction reposent bien sur des identifications d'un unique et même trait d'un individu, que nous en arrivons à définir caricaturalement les traits de membres des groupes adverses. La rapidité avec laquelle nous allons nous défier de ceux envers qui les membres de notre groupe expriment de la méfiance, et avec laquelle nous allons accorder notre confiance à ceux à qui ils l'accordent est sans doute encore une manière de démontrer notre attachement au groupe.

9) À ce stade, nous pouvons déjà esquisser une typologie des différents rôles assignés aux tiers. Nous avons vu que des objets pouvaient être en tiers entre deux personnes, l'une suivant le regard de l'autre – voire, selon René Girard, concluant au désir de l'autre pour cet objet. Mais un objet en tiers n'implique pas forcément de rivalité : un objet à manipuler en commun peut servir de médiateur de coordination. Les tiers, ce sont aussi et surtout les autres, ceux de notre groupe, ceux du groupe adverse, et les indépendants. Il faut y ajouter le rôle d'autorité de référence que peuvent jouer des tiers pour régler les disputes entre moi et un autre membre de ma parenté ou de mon groupe. Ce rôle est à distinguer de celui de simple observateur non participant (clé de la construction d'Adam Smith), puisque nos parents ou les membres référents de notre groupe peuvent participer à nos interactions. Mais même comme observateurs, voire *in abstentia*, les tiers peuvent jouer un rôle décisif, du moins si les partenaires des interactions se réfèrent à leur opinion supposée pour réorienter leurs actions. Selon Simmel, une interaction ne trouve véritablement son statut social que si d'une part la réinterprétation par chaque partenaire de l'action de l'autre a produit son effet rétroactif sur l'interprétation de l'interaction par l'autre partenaire, et que d'autre part ce statut peut être supposé admis par des tiers référents (il n'est pas nécessaire qu'ils montrent leur approbation, il suffit qu'ils ne manifestent pas de désapprobation).

10) Ajoutons à présent à notre édifice de constitution interactive des sujets un élément qui pourrait renforcer la pertinence d'une analyse en termes de théorie des jeux. Jusqu'ici, nous avons construit le socle de l'intersubjectivation sur des mouvements, des regards, des expressions, des réactions conjointes. Si intentions d'autrui il y avait, elles étaient toujours gagées sur les cibles visées, les mouvements pour les atteindre, ou les expressions observables. Tout cela pouvait jouer sans même que les sujets en aient une conscience suffisante pour qu'ils puissent en expliciter le contenu. Nous avons vu que les expressions nous donnaient une idée des intentions d'autrui, qui ne pouvaient être observées de manière directe. Mais un problème supplémentaire se pose : autrui pourrait avoir un état

d'information différent du nôtre, ce qui exige de nous, pour comprendre autrui, soit de ne pas tenir compte d'observables dont nous sommes seuls à disposer, soit de supposer chez autrui des informations issues d'observables auxquels nous n'avons pas accès.

La première situation a été stylisée dans la tâche dite de la « fausse croyance ». On fait voir à un enfant – qui a le rôle d'un observateur en tiers – un autre enfant qui voit un adulte mettre des bonbons dans une boîte. Puis cet enfant sort, et l'adulte déplace à la vue du premier les bonbons dans une deuxième boîte. On demande à notre observateur dans quelle boîte l'enfant qui va revenir cherchera les bonbons. Avant 4 ans, il répond qu'il s'agit de la deuxième boîte. Plus âgé, il reconstruit les représentations de l'enfant qui est sorti, et répond que c'est la première boîte.

L'interprétation usuelle de cette expérience est un peu similaire, bien qu'inverse, à celle de l'inhibition du mouvement pourtant préactivé dans l'aire de préparation motrice par les neurones miroirs : l'enfant observateur, pour donner la bonne réponse, doit inhiber sa propre représentation de la véritable place des bonbons, et reconstruire la représentation d'autrui, alors même qu'il la sait fausse (dans le cas des neurones miroirs, le sujet doit inhiber le mouvement induit dans sa zone de préparation motrice par le mouvement d'autrui, non parce qu'il est faux, mais parce que ce n'est pas le sien).

Cependant, on a observé que des bébés peuvent déjà être sensibles aux éléments de cette situation puisqu'ils sont surpris – ils sucent plus vigoureusement leur tétine – quand ils voient l'enfant revenir et chercher les bonbons dans la deuxième boîte (Frith C. et Frith U., 2007) ! Il semble donc que tous les enfants, même avant 4 ans, soient sensibles au fait que se pose pour l'enfant qui est sorti un problème de localisation des bonbons, mais qu'ils ne pensent pas, avant 4 ans, que la « bonne réponse » qu'on attend d'eux puisse être autre que la « vraie » localisation.

Le problème ne serait donc pas simplement de constituer une capacité à penser en changeant nos propres représentations et en les remplaçant par celles d'autrui – capacité que nous n'arrivons à posséder, même adulte, que de manière limitée. Il serait d'arriver à identifier quelle est la bonne incidence du point de vue d'autrui et de ses représentations sur notre réponse à l'expérimentateur – sur notre interaction avec un tiers référent –, dans le cas où les informations d'autrui et ses observables diffèrent des nôtres et des tiers référents, donc des propositions sur le monde que nous pouvons partager avec ces tiers référents.

Il faut ici distinguer la capacité d'imaginer chez autrui des intentions alors qu'aucune expression ou comportement ne les indique directement, et la simple sensibilité aux différences d'information sur le monde entre nous et autrui. Car cette sensibilité, là encore, semble déjà présente chez des bébés. Ils sont sensibles dès 12 mois à la différence entre des nouvelles fraîches et des nouvelles déjà connues par ceux qui les entourent. Ils marquent une préférence pour leur dire ce que ceux-ci ne savent pas (une tendance qui nous poursuit une fois adultes, et qui fait que bien des informations supposées confidentielles sont diffusées). On peut en inférer que lorsqu'ils communiquent, ils tiennent compte de ce que les autres ne savent pas. C'est d'ailleurs en étant sensibles à des signaux de l'intention des adultes de leur communiquer quelque chose (au départ à des signaux non sémantiques) qu'ils peuvent apprendre leur langue en repérant des mots nouveaux comme tels et en en mémorisant une dizaine par

jour.

Là encore, on retrouve ces processus d'inhibition qui permettent de mieux constituer les différences entre divers soi. Nous inhibons la communication des informations quand nous pensons que l'état de savoir des autres est le même que le nôtre. Là encore il s'agit d'une comparaison entre deux modes d'interaction, celui de cette inhibition d'une communication en cas de savoir commun, et celui de l'inhibition de cette inhibition. Cette dernière déclenche la communication, et permet par là même de rendre saillante ce qu'on appelle l'intention de communication. On ne produit donc pas de signal de communication si l'on pense l'information déjà partagée. Dès lors qu'un signal s'adresse à nous, nous présupposons qu'il nous apportera une information que nous n'avons pas[6].

On sait que cette tâche de la fausse croyance a pu servir d'argument aux partisans de la *theory theory* contre ceux de la « simulation ». Les premiers pensent qu'il est nécessaire, pour résoudre cette tâche, d'avoir constitué une théorie (naïve) de la manière dont fonctionne l'esprit de nos semblables, ce qui nous permet de faire des inférences sur les croyances d'autrui en fonction des informations dont nous savons qu'ils en disposent. Les seconds jugent qu'il suffit d'avoir nous-même des dispositions à tirer des inférences des observables, et de nous mettre avec nos propres dispositions dans la situation d'autrui – donc d'identifier les observables accessibles à autrui – pour la résoudre. La discussion menée par C. et U. Frith, leur rappel des observations sur les bébés et notre prise en compte du rapport au tiers (l'expérimentateur) nous permettent de prendre une position intermédiaire. Dans ses réponses à l'expérimentateur, l'enfant penserait devoir utiliser une théorie du monde qui soit partageable avec des tiers, et une telle théorie n'inclurait pas encore, avant 4 ans, le monde inobservable des représentations d'autrui. Mais, dans sa vision de la situation, l'enfant serait déjà capable, par « simulation », en se mettant « dans les chaussures de l'autre », de voir la situation à partir d'un point de vue qu'il n'occupe pas maintenant mais qu'il pourrait occuper sans difficulté.

Par parenthèse, la théorie de l'alter ego que l'on trouve chez Husserl est plus proche de la deuxième position que de la première. Pour lui, c'est de notre point de vue de sujet, d'ego, que nous pouvons penser la possibilité d'un point de vue qui est celui d'Alter. Il faut pour cela pouvoir assigner comme origine à la perspective d'Alter un *illic*, un là-bas tel qu'il puisse se transformer en *hic*, ici, et que la transformation réciproque soit aussi possible, tout en respectant la priorité de l'origine propre à ego, celle du *hic*. Or si ces transformations réciproques sont relativement aisées quand elles consistent simplement à interchanger les points origines des perspectives, elles sont plus difficiles quand on en vient au niveau du contenu des représentations, donc des « théories » au sens utilisé ici. Je ne suis pas capable de complètement activer les représentations d'un schizophrène ou celles d'un mathématicien de génie, pas plus que celles d'un chasseur Nambikwara émérite. Husserl ne recourt donc pas à une « théorie de l'esprit ». Il soutient que cette réciprocité des transformations, qui conserve la priorité de l'ego, est assurée de manière immédiate, par une *Einfühlung*, une sorte de syn-esthèse voire d'uni-esthésie, un concept qu'il a emprunté à Lipps : je me déplacerais dans la situation d'autrui avec mes propres capacités sensibles.

11) Jusqu'à cette dixième étape, nous avons bien vu se constituer un sujet et ses différentes

capacités, un sujet qui peut faire la différence entre lui et l'environnement, lui et les autres, lui et d'autres groupes, etc. ; mais cela n'implique pas nécessairement qu'il ait une conscience de soi qui soit autre que la conscience de ces différences.

Les neurosciences nous apprennent que lorsque nous laissons notre esprit errer, sans fixer notre attention sur un contenu, nous entrons dans un état que les chercheurs nomment le « mode par défaut », et ils suggèrent que ce pourrait être là une forme de conscience de soi. Mais c'est bien là un « soi » par défaut, donc obtenu comme le résidu qui demeure quand on a mis de côté les fixations sur les objets de l'environnement et sur autrui. C'est encore un soi par différence avec le non-soi, un soi produit par une comparaison entre deux modes d'interaction, celui entre soi et non-soi, et celui qui persiste quand on a inhibé ce premier type d'interaction (là encore, nous retrouvons le rôle du couple activation/inhibition).

Nous pouvons à présent en venir à un mode plus positif du soi, si nous apercevons que ce problème des inobservables auquel nous nous heurtons à propos d'autrui, nous le rencontrons aussi, sous une autre forme, à notre propre propos. Nous ne disposons pas de manière observable de tous les éléments qui nous concernent ; nous devons les compléter par des inobservables, en utilisant les capacités à les inférer qui ont été mises au point pour autrui. Le sujet qui s'applique à lui-même cette capacité va alors avoir une image de soi donc une conscience *de* soi au sens plein du terme, une conscience *du* soi qu'il est. Non seulement il donne un contenu à la différence entre non-soi et soi, non seulement il se saisit relationnellement en donnant un contenu au non-soi (objet ou autrui), ou encore il existe résiduellement (dans le mode par défaut), mais encore il se donne à lui-même un contenu. Ainsi, à la racine de la conscience de soi, on trouve un effet en retour sur le sujet de la différence soi/autrui.

Les traitements cognitifs qui concourent aux particularités de notre subjectivité fonctionnent d'abord dans l'ombre. Ils ne transmettent à notre conscience que ce qui fait saillance, cohérence et pertinence. Or cela seul peut rentrer dans le champ de ce qui est communicable. La conscience serait alors une sorte de champ intermédiaire entre le purement privé – qui travaille en nous et déclenche bon nombre de nos réactions, mais qui n'est pas partageable – et le vraiment public – ce qui est communiqué. Ce serait le champ de ce que nous pouvons inhiber au lieu de le communiquer, mais aussi de ce dont l'inhibition en question peut être levée (par une inhibition d'inhibition). Le sujet conscient se construirait donc dans les interactions intersubjectives et selon leurs dispositifs, celui de la comparaison différentielle entre deux modes d'interaction, et celui du couplage/découplage que forme le duo activation/inhibition.

12) Non seulement nous devons deviner ces inobservables que sont les intentions d'autrui, mais il nous serait utile d'en imaginer plusieurs versions, de manière à ne pas être pris au dépourvu dans nos réactions. Cependant, pour résoudre ce problème, nous n'avons pas besoin de créer un monde irréel à côté du monde réel. Il nous suffit d'inhiber notre prise directe sur le monde réel et, parallèlement, de libérer les tentatives de représentations qui ne pouvaient ou ne pourraient fonctionner dans le monde réel actuel, et cela en inhibant partiellement leurs inhibitions. Cette capacité, nous la trouvons déjà dans les jeux des enfants qui ne sont pas « pour de vrai », en anglais les

activités de « *pretense* ». Elles leur permettent d'évoquer entre eux des mondes certes incomplets mais dont le partage est plus attractif que celui du monde réel.

13) Il devient alors utile de pouvoir indiquer ces inobservables aux membres de notre groupe alors même que nous ne sommes pas en contact direct avec eux. La solution est d'utiliser des symboles, qui assurent des liaisons *in absentia*. Nous ne serions sans doute pas capables d'activités symboliques si nous n'étions pas capables de fictions, car les deux activités se renforcent l'une l'autre. Les jeux d'enfants combinent des comportements réels et des fictions, si bien que ces comportements réels sont censés avoir des résultats qui sont imaginaires, hors de proportion avec leurs effets réels. De même, la transmission de symboles ne vaut pas par ses effets directement observables, puisque nos communications par symboles sont surtout censées modifier les représentations d'autrui, qui ne sont pas directement observables. Comme les jeux, les activités de production et de manipulation de symboles permettent de construire des mondes – très incomplets, mais qui peuvent traiter des aspects principaux de nos existences sociales.

14) L'usage des symboles pose un problème supplémentaire : il exige une communication dédiée aux instructions concernant cet usage. Nous devons apprendre leurs codages comme tels et ce dispositif supplémentaire est le prix à payer pour un meilleur partage des connaissances. Qu'il y ait un niveau où les différences intersubjectives soient réduites autant qu'il se peut, non pas dans l'idéal mais dans la pratique effective des apprentissages, est paradoxalement la condition pour avoir un accès au summum de l'intersubjectivité, le monde culturel.

La capacité de passer de l'actuel au fictif et au monde symbolique est donc liée à la possibilité d'apprendre en suivant des instructions explicites. Un tel apprentissage n'est efficace que dans un cadre interactif, en pouvant bénéficier de commentaires sur nos essais qui nous expliquent en quoi soit le résultat, soit la méthode étaient déficients. Pour ce faire, il faut que nos partages portent non seulement sur la description de ce que nous avons effectivement accompli, mais aussi sur celle de ce que nous aurions *dû* accomplir. Si bien des effets de l'apprentissage fonctionnent selon des processus dont nous n'avons qu'une conscience implicite, ces derniers effets exigent une conscience *explicitement* partagée. Ce mode de conscience partageable semble bien requis pour que le sujet lui-même puisse parvenir à un mode de conscience réflexive explicite, et reprendre à sa manière, en les réélaborant, des représentations créées par ces communications ou instructions qui étaient de la part d'autrui délibérées.

Le sujet ne se construit donc que dans des interactions, et cherche comme guides de ses interactions avec son environnement celles qu'il peut avoir avec d'autres sujets. Il se construit par la combinaison de deux processus qui, chacun, mettent en jeu une forme de dualité : activer des imitations ou simulations d'autrui mais pouvoir les inhiber ; comparer des modes d'interaction différents et les faire interagir. Cette combinaison finit par lui permettre de vivre non seulement sur un mode actuel, mais sur un mode virtuel, parce qu'il met entre parenthèses certains effets sur le monde réel de ses actions, et qu'il laisse cours à d'autres possibilités qui ne pourraient pas s'incarner dans des réalités présentes. Il peut s'agir d'anticipations, de reconstructions partageables des

représentations d'autrui, de signalisations d'intentions de communication, ou encore de partages de fictions et de constructions symboliques.

Il semble que trois strates principales soient nécessaires dans cette édification de la subjectivité : la différenciation soi/non-soi ; les mises en résonance avec d'autres soi (imitation des mouvements, contagion des émotions, coordination d'actions communes) ; la différenciation entre des transformations qui portent directement notre environnement et d'autres qui transforment d'abord les relations intersubjectives – la communication, l'apprentissage, le partage de fictions et de symboles. On notera que si la théorie des jeux trouve déjà un terrain d'analyse pertinent dès la deuxième strate, ses concepts et ses modélisations offrent un cas paradigmatique de développement des potentialités offertes par la troisième.

Commentaire L’intersubjectivité des agents économiques

La science économique est souvent présentée comme l’héritière directe de l’économie politique, une approche des phénomènes sociaux apparue dans la seconde moitié du XVIII^e^ siècle, à la faveur des travaux de philosophes sensualistes appartenant à l’École écossaise, dont le plus célèbre est, aujourd’hui, Adam Smith[7]. Une telle présentation fait fi d’une rupture intervenue, près d’un siècle plus tard dans le développement de la discipline, au point qu’à certains égards, il n’est peut-être pas exagéré d’avancer l’idée d’un changement de paradigme et, à tout le moins, de perspective. L’agent économique de Smith est un être social qui, au cours d’activités d’échange, et de production, arbitre entre différentes passions, au sens où l’entend Hume. L’agent économique conceptualisé à la fin du XIX^e^ siècle est un sujet individuel dont toutes les activités économiques sont les résultats de décisions émanant d’un calcul raisonné. Il n’est guère surprenant, dans ces conditions, que l’intersubjectivité ne soit pas appréhendée de la même manière dans l’économie politique d’Adam Smith et dans les sciences économiques contemporaines.

Selon la première perspective, ce qui a été nommé « intersubjectivité » est, comme Pierre Livet l’a rappelé dans ce chapitre, indissociable du champ d’action des agents économiques, puisque le fonctionnement des interactions économiques dépend, en amont, d’un jeu mental de caractère social (l’autre, les autres, la position du tiers). Ce n’est pas, dès lors, l’intersubjectivité qui pose problème à Smith et à ses continuateurs directs, mais plutôt l’identification de la subjectivité. Elle se révèle à l’occasion d’une opération économique, comme l’échange. Smith rappelle, dans une formule devenue célèbre et souvent citée, que ce qui pousse le boucher et le boulanger à mettre leur production sur le marché n’est pas le désir de nourrir la population, mais, bien plutôt, leur intérêt personnel, entendons ici, la manifestation de leur subjectivité intéressée. Avec l’intérêt personnel on tient donc une propriété singulière du sujet. Cette mise en évidence de la subjectivité s’effectue, ici, dans le cadre de l’intersubjectivité, puisque nos commerçants savent parfaitement que, sans clients, ils ne pourraient pas satisfaire leurs intérêts subjectifs. Il reste toutefois à comprendre comment fonctionne cette intersubjectivité.

Dans la seconde perspective, qui domine aujourd'hui encore l'analyse économique, les actions des hommes qui déterminent les phénomènes économiques sont les choix réfléchis décidés, dans la majorité des cas, par des individus. Une référence essentielle pour effectuer ces choix est celle des préférences strictement subjectives dont sont dotés tous les individus, au moyen desquels ils peuvent hiérarchiser les conséquences possibles des actions qu'ils envisagent d'entreprendre. C'est ainsi que chaque individu est supposé construire un ordre (ou, tout au moins, un préordre) de préférence sur les conséquences anticipées de ses choix, auquel renvoie sa décision raisonnée. On part donc, cette fois, d'un sujet, caractérisé par ses préférences individuelles. La difficulté réside, ici, à l'opposé de la perspective précédente, dans la recherche des voies de passage entre cette subjectivité des agents économiques et l'intersubjectivité.

D

Plusieurs approches ont été suivies par la science économique contemporaine pour parvenir à cette fin. La première, et la plus classique, consiste à chercher à agréger les préférences individuelles, en vue d'identifier des préférences collectives. Mais si l'agrégation est déjà une opération statistique délicate, son application aux préférences des individus ne se réduit pas à une simple opération statistique. Elle nécessite d'introduire des procédures supplémentaires, dont les contraintes se révèlent souvent contradictoires, d'où les impasses logiques mises en évidence dans la théorie par une série de théorèmes d'impossibilité[8].

Le recours à l'agrégation des préférences ne se limite pas, en économie, aux domaines des choix collectifs et à l'évaluation du bien-être social. On l'utilise également en économie expérimentale. Ses chercheurs s'efforcent d'induire certains traits de comportement de portée générale, en partant d'observations sur des échantillons d'individus le plus souvent très réduits. Comme les résultats obtenus sont presque toujours affectés par une plus ou moins grande dispersion statistique dans les comportements individuels observés, ils évaluent les pourcentages des différents comportements observés. Cette pratique courante entraîne cependant fréquemment un biais d'interprétation lorsque l'on cherche, notamment, à comparer les choix respectivement effectués par des sujets confrontés à deux groupes d'alternatives distinctes, mais logiquement liées. C'est le cas notamment lorsque les options du premier groupe portent sur des gains et celles du second groupe sur des pertes, les uns et les autres étant seulement probables. On calcule alors, pour chaque alternative, la proportion de l'ensemble des sujets ayant choisi l'une ou l'autre des deux options, avant d'en déduire des pourcentages appliqués à l'ensemble des sujets interrogés. Ce résultat agrégé est moins significatif que ne le serait l'évaluation du nombre des sujets ayant choisi la même paire dans chacun des deux groupes. Le but de cette opération vise, le plus souvent, à dégager, à partir de ces données, ce que les économistes appellent « un agent représentatif » dont l'interprétation reste assez problématique. Dans certains cas, le comportement de cet agent représentatif a même pu se révéler différent de celui de

chacun des individus à partir desquels il a été construit. Ainsi, lorsque tous les agents optimisent leur utilité individuelle, comme le voudrait la théorie économique classique, l'agent représentatif fictif qui résulte de leur agrégation va, dans bien des cas, révéler les anomalies mises en évidence par Kahneman et Tversky, en raison même de la diversité de leurs préférences. Il s'agit, en définitive, d'une fiction statistique conçue sur la base de l'observation des comportements individuels dans des situations données impliquant des choix décisionnels. Il serait sans doute plus exact de considérer cet agent représentatif comme le reflet d'une manière de « trans-subjectivité », plutôt que comme la traduction statistique d'une intersubjectivité. De façon plus ambitieuse, un programme de recherche, développé dans le milieu des années 1980, visait à établir les fondements microéconomiques de la macroéconomie. Son échec, aujourd'hui reconnu, est sans doute en partie imputable à l'inadaptation des différentes approches et techniques d'agrégations à rendre compte des phénomènes d'intersubjectivité.

P

Une autre démarche suivie par les économistes et certains théoriciens de la décision consiste à enrichir le niveau des préférences individuelles. Les économistes ne se sont guère penchés sur les origines des préférences individuelles et sur les modalités selon lesquelles les agents économiques les construisent. Ils considèrent, en règle générale, les préférences des agents économiques comme des données premières, extérieures à leur champ d'étude, laissant à d'autres disciplines (psychologie, sociologie…) le soin de dégager leurs caractéristiques et leurs origines. Ils se contentent donc, le plus souvent, de leur associer différentes contraintes logiques (réflexivité, transitivité, antisymétrie…).

Plusieurs logiciens de la décision qui ont prolongé l'analyse économique des préférences se sont néanmoins efforcés d'enrichir son domaine, en introduisant des « métapréférences » (Davidson, 1980). Ces métapréférences renvoient, par exemple, à des prescriptions, d'ordre moral ou social. Un exemple plus trivial souvent cité est celui d'un alcoolique qui préfère boire (préférence de premier degré), mais préférerait ne pas préférer boire (métapréférence). Il lui faut, dès lors, au moment de prendre sa décision, arbitrer entre deux préférences contradictoires, ce qui complique l'analyse de son choix (Jeffrey, 1974, 1983). On peut expliquer, en ces termes, certains choix inattendus, observés au cours de fréquentes expériences, qui seront analysées plus en détail dans la suite de cet ouvrage. Le fait, par exemple, qu'un sujet, plutôt que de s'approprier la presque totalité d'une somme mise à sa disposition, choisit le plus souvent de la partager de manière moins inégale, pourrait s'expliquer parce que sa métapréférence sociale l'emporte, en l'occurrence, sur sa préférence égoïste. C'est ce qu'ont soutenu, comme nous le verrons, plusieurs économistes expérimentalistes. On ne peut cependant parler ici d'intersubjectivité que pour autant que ces métapréférences renvoient à des croyances partagées ou, tout au moins, à des croyances que le sujet croit qu'il partage avec les autres. Il faut alors s'interroger sur les origines de telles croyances. On abordera cette question dans la dernière

partie de cet ouvrage, à propos des règles et des normes.

Par ailleurs, des phénomènes aujourd'hui bien connus de renversement des préférences ont également été mis en évidence. Il s'agit le plus souvent de phénomènes provoqués par des effets de *framing* (cadrage) dans la présentation des données. Ainsi ont été observées des différences entre les préférences des consommateurs lorsqu'elles sont définies sur les biens, auxquels est associé un prix, ou sur les prix à partir desquels ils seraient prêts à acquérir ces biens (Lichtenstein et Slovic, 1971). De même, une information identique sur une situation de désastre conduisant à plusieurs décisions verra les préférences des individus changer, selon que cette information se trouve présentée en termes de personnes survivantes, ou en termes de personnes décédées (Tversky et Kahnemann, 1981). Par-delà l'effet de *framing*, ces exemples montrent que la définition des préférences subjectives des individus est sensible à leur interaction avec des informations concernant le monde extérieur. Certaines d'entre elles peuvent provenir d'autres sujets, ce qui suggère une certaine interdépendance des préférences, ouvrant ainsi la voie à l'hypothèse de leur intersubjectivité.

L

Il existe une troisième approche de l'intersubjectivité en économie, où l'intersubjectivité est, cette fois, consubstantielle de l'objet étudié. Son introduction a correspondu à l'émergence de la théorie des jeux, dont l'objet est précisément l'analyse des interactions interindividuelles. Les agents économiques y sont appréhendés comme les joueurs d'un jeu qui représente l'espace dans lequel se déroulent leurs interactions. On notera que le terme de jeu n'est pas ici une simple métaphore. Dans le prolongement de ce qui a été écrit dans ce chapitre, l'espace d'un jeu, au sens de la théorie, est composé d'éléments réels (les joueurs, les règles…), mais également d'éléments virtuels (ce que chaque joueur imagine sur les autres joueurs et ce qu'ils imaginent que ceux-ci imaginent sur eux-mêmes). Les actions des joueurs s'entendent, dès lors, comme des stratégies. Elles sont analysées en termes de réponses (stratégiques) du joueur aux actions des autres joueurs. Le joueur, qui est un sujet du jeu, se trouve donc directement confronté à un autre sujet. Cette confrontation implique, en particulier, la possibilité donnée à chaque joueur d'imaginer la stratégie que l'autre (les autres) décidera (décideront), de son (leur) côté, de mettre en œuvre, puisqu'au moment où il choisit sa stratégie il ne connaît pas celle que l'autre (les autres) a (ont), ou aura (auront), choisi. Un joueur peut, par ailleurs, vouloir, s'opposer (jeux non coopératifs), coopérer, voire s'allier avec un autre joueur pour former avec lui une coalition (jeux coopératifs). Ces différents éléments expliquent pourquoi la théorie des jeux offre, en son principe, un cadre formel privilégié pour saisir les structures logiques de l'intersubjectivité.

En reprenant les catégories élaborées par l'analyse économique, la théorie des jeux dispose d'un moyen simple de distinguer ce qui relève de la subjectivité des joueurs de ce qui appartient à leur intersubjectivité. Conformément à la convention économique précédemment rappelée, la subjectivité

des joueurs se traduit dans leurs préférences individuelles. Quant à leur intersubjectivité, elle se trouve associée à la propriété de rationalité qui, dans le contexte des jeux classiques, est supposée être une connaissance commune entre eux.

Cette manière de distinguer la subjectivité de l'intersubjectivité présente l'avantage d'une clarté logique, au moins en apparence. Elle soulève toutefois de sérieux problèmes. Les préférences d'un sujet révèlent sa subjectivité dans ses différences avec autrui. Rien n'empêche toutefois plusieurs sujets d'avoir des préférences identiques. En outre, comme on l'a indiqué, la formation des préférences d'un sujet n'est pas indépendante de sa rencontre avec les préférences d'autres sujets, Ainsi le phénomène bien connu du mimétisme a-t-il été souvent interprété comme une manifestation d'intersubjectivité. De plus, et par hypothèse, un sujet est supposé connaître ses propres préférences, et en être conscient, quel que soit, par ailleurs, le statut associé à la conscience de cette connaissance (fausse conscience, connaissance non consciente, voire inconsciente…), ce qui pose déjà problème. Il en va, évidemment autrement de sa connaissance des préférences des autres sujets. Si le jeu se déroule en plusieurs coups, ou s'il se répète, le joueur pourra induire certaines informations des mouvements stratégiques observés chez les autres joueurs aux phases antérieures. Mais dans les cas de jeux en un coup, qui servent souvent d'archétype à ces situations, une telle démarche n'est plus possible. On notera qu'à l'inverse, le théoricien des jeux, ou le modélisateur du jeu, est supposé connaître, lui, les préférences individuelles de tous les joueurs dont il étudie le jeu. La tentation existe alors d'assimiler le joueur, comme sujet du jeu, au théoricien (ou au modélisateur) qui en est l'observateur. C'est ce qu'ont fait, plus ou moins implicitement, les fondateurs de la théorie des jeux.

L'analyse de la multitude de ces différents points de vue sur l'autre développée dans ce premier chapitre permet de préciser les différents contours de cette assimilation abusive. Indépendamment du problème posé par l'information sur les préférences de l'autre, l'approche d'un joueur sur un autre joueur n'est jamais réductible à l'approche d'un tiers, en l'occurrence, ici, le théoricien, sur les joueurs qu'il observe. Ce constat reste valable lorsque le terme d'observation désigne, d'une manière plus générale, toute forme d'appréhension par un sujet d'autres sujets, quel que soit le processus cognitif retenu. Nous verrons dans le chapitre suivant, qu'en suivant une voie un peu différente de celle esquissée dans ce chapitre, les théoriciens des jeux ne sont plus aujourd'hui tout à fait victimes de cette méprise. Que peut, dans ces conditions, connaître un joueur des préférences d'un autre joueur, au moment où s'engage le jeu ? Une telle information reste indispensable pour permettre au sujet de choisir une stratégie par laquelle s'enclenche l'interaction entre les joueurs.

À l'autre extrémité du spectre balayé par la théorie des jeux est posée l'hypothèse que la rationalité est une propriété intersubjective, au sens où sa connaissance se trouverait partagée par tous les joueurs, ainsi, du reste, que par le théoricien. Cette hypothèse forte sur l'intersubjectivité permet de compenser, d'une certaine manière, l'opacité qui caractérise la connaissance des préférences des autres. Elle offre ainsi une possibilité à chaque joueur d'anticiper la stratégie qui sera choisie par l'autre joueur, en lui permettant de s'imaginer à sa place, à partir des informations objectives qui lui sont fournies par le jeu (règles, valeurs des paiements…). Le schéma retenu est ici le suivant : je sais que je suis un sujet rationnel, comme je sais que l'autre est également un sujet rationnel, je sais donc

qu'il fera le choix stratégique que je ferais moi-même si je me trouvais dans sa situation que je connais, grâce aux règles de ce jeu. Pas si sûr…

Un tel schéma comporte plusieurs failles. Il ne suffit pas de savoir l'autre rationnel pour en déduire son choix, qui dépend étroitement de l'objectif qu'il poursuit, dépendance qui est un exemple de ce que les philosophes désignent sous le terme d'intentionnalité. Un sujet, par exemple, peut évidemment chercher à maximiser son profit égoïste, mais il peut, tout aussi bien, chercher à satisfaire un penchant altruiste. Beaucoup d'expériences l'ont montré. Réduire la rationalité à une maximisation égoïste correspond, au mieux, à une extrême simplification, et relève, au pis, d'une confusion. Plus subtilement, la satisfaction d'un même objectif général, comme celui de son intérêt égoïste, peut emprunter des voies logiques différentes, qui aboutissent à des résultats, eux-mêmes, distincts. Les travaux économiques de logique de la décision ont mis en évidence la nécessité d'introduire des critères supplémentaires permettant de relier cet objectif à la rationalité. Ainsi, pour satisfaire son intérêt égoïste, un sujet peut fort bien chercher à maximiser son gain attendu, mais il peut également, tout aussi légitimement, chercher à minimiser le regret que pourrait entraîner un choix malencontreux. Plus simplement encore, certains sujets sont plus averses au risque de pertes, d'autres plus attirés par des gains même incertains : leurs choix rationnels seront nécessairement différents. Or maximiser ses espérances de gains et minimiser ses risques de pertes sont deux comportements tout aussi rationnels du point de vue de l'intérêt égoïste en cohérence logique avec des préférences définies subjectivement. On voit donc que les préférences individuelles des sujets s'infiltrent ici, subrepticement, au cœur de la définition de ces différents critères de rationalité. Enfin, et indépendamment de ces inflexions subjectives, ces différents choix, tous rationnels, font intervenir des procédures de raisonnement formellement distinctes. Un constat qui se trouve aujourd'hui corroboré par ce que l'on connaît du fonctionnement du cerveau. Ainsi a-t-on montré que la minimisation du regret, qui implique un raisonnement contrefactuel, entraîne une activation spécifique du cortex orbitofrontal, que l'on ne retrouve pas lorsque le sujet se contente de minimiser son risque de perte (Coricelli, 2007).

Supposer que la référence à la rationalité, en tant qu'elle serait commune aux sujets en interactions, permettrait, à elle seule, à un joueur de deviner le choix stratégique d'un autre joueur, en se mettant à sa place, apparaît finalement illusoire au vu de ces objections. Comme on le développera dans le chapitre suivant, la fonction véritable de cette hypothèse de rationalité pourrait être, en définitive, de fournir simplement aux théoriciens une manière d'heuristique pour les aider à comprendre, de leur point de vue, les comportements des joueurs en interactions.

Ce qui fait la fécondité des modèles de théorie des jeux pour appréhender l'intersubjectivité est donc ailleurs. Elle réside dans la stricte interdépendance des sujets qui caractérise toute situation de jeu. Dans une telle situation, le joueur (sujet) se définit, en action, par rapport à un autre sujet, conformément à l'analyse développée dans ce chapitre par Pierre Livet. Mais, il est, en même temps, lui-même, un autre sujet pour cet autre joueur. Il lui faut donc, pour penser stratégiquement cette situation, imaginer comment il est perçu comme un autre sujet par cet autre joueur. Le problème se

complique encore, du fait que la réciproque existe pour cet autre joueur. Il en résulte une imbrication provoquée par cette réciprocité des positions respectives des sujets, clairement mise en évidence dans les diverses situations étudiées par la théorie des jeux.

Cette épaisseur de l'intersubjectivité ainsi repérée et ses contours mieux cernés posent aux joueurs et à ses analystes de redoutables problèmes, au premier rang desquels celui de la coordination de leurs stratégies, qui sera analysé dans la seconde partie de cet ouvrage. Cette question trouve des illustrations très concrètes dans de nombreuses situations de la vie courante. L'exemple emblématique du jeu du rendez-vous, lorsqu'il existe plusieurs points de rencontre, a servi de support à Schelling pour en discuter les ressorts. Il sera analysé en détail au chapitre 4. Posés en termes plus abstraits, on retrouve ces problèmes de coordination au niveau macroéconomique, lorsqu'il existe plusieurs points d'équilibre sur un marché, une situation assez couramment observée sur les marchés financiers. Les économistes s'interrogent alors pour savoir sur lequel de ces équilibres, les joueurs, c'est-à-dire ici les opérateurs de marchés, vont faire converger la rencontre de leurs stratégies.

Fidèle à leur approche par l'abstraction, les théoriciens des jeux cherchent patiemment des solutions logiques aux mystères de la coordination intersubjective ainsi rigoureusement formulée. Dans le même temps, cependant, de nombreuses expériences révèlent qu'une telle coordination s'effectue souvent avec succès dans la réalité, indépendamment des procédures logiques imaginées par les théoriciens des jeux. Pour combler le fossé entre les impasses rencontrées dans la modélisation logique de cette coordination et les succès de coordination réussie empiriquement observés, il nous paraît nécessaire de faire un détour par une exploration du fonctionnement des cerveaux des joueurs. Le développement rapide des neurosciences permet et encourage aujourd'hui cette entreprise que l'on poursuivra dans les chapitres suivants.

---

1. Ce qui permet à Claudine Tiercelin de dire que chez Peirce, « le mental n'est plus la référence obligée à un esprit, une conscience ou une pensée : c'est la mise en œuvre de la tiercéité » (*La Pensée-signe*, Paris, Jacqueline Chambon, 1993, p. 197). Nous suivrons cette voie selon laquelle la notion d'interaction, qui peut impliquer trois termes (et deux relations, la seconde qualifiant la première, comme le présent qualifie le transfert d'un objet), est un concept qui vaut aussi bien pour des processus « mentaux » intra-individuels que pour des processus sociaux.

2. Nous reviendrons sur la construction que Smith donne de la sympathie et de l'observateur impartial dans le chapitre 2.

3. Un point de vue déjà bien étudié par Dorothée Legrand dans sa thèse sur les problèmes de la conscience de soi.

4. Ce lien mouvement-motivation peut être très basique : conserver la tête stabilisée pendant un mouvement permet d'orienter notre regard sur nos cibles ou trajets durant le mouvement (Berthoz, 1997).

5. On peut utiliser la théorie des jeux pour modéliser des interactions très variées, y compris non intentionnelles (théorie des jeux évolutionnaires), mais on procède alors par analogie entre des comportements sélectionnés par l'évolution et des stratégies raisonnées.

6. Exception faite des « amorces » de communication, comme « ça va ? », qui vérifient simplement la disponibilité d'autrui pour cette communication et donc n'apportent pratiquement aucune information non partagée.

7. Il faut souligner ici les influences de Hume, et plus encore de Hutcheson qui fut son maître à Glasgow, sur la formation de la pensée de Smith sur cette question.

8. K. Arrow a démontré que les hypothèses nécessaires pour transformer les choix individuels en choix collectifs, tout en respectant

les préférences individuelles des agents et l'indépendance de leurs choix, aboutissaient toujours à des contradictions (Arrow, 1951). Ce résultat puissant a été complété et étendu par différents travaux, au premier rang desquels les contributions de Sen (1970).

CHAPITRE 2

L’appréhension de l’autre

L’histoire des disciplines réserve des surprises. Tandis qu’Adam Smith apparaît rétrospectivement comme un précurseur des théories contemporaines de l’intersubjectivité, avec toute son épaisseur et sa complexité, la science économique s’est développée, au contraire, dès la seconde partie du XIX^e^ siècle, comme on l’a rappelé, en partant de l’analyse d’un sujet isolé, proche du solipsisme. Plusieurs explications peuvent être formulées pour en rendre compte. Il a souvent été avancé qu’une telle opposition se trouvait déjà dans l’œuvre de Smith[1]. Ce n’est pas, en effet, dans les *Recherches sur l’origine et les causes de la richesse des nations* que Smith met l’intersubjectivité au centre de son approche des phénomènes sociaux, mais dans la *Théorie des sentiments moraux*. Or les deux ouvrages n’ont pas le même objet. Dans la *Théorie des sentiments moraux*, Smith recherche les fondements moraux des systèmes sociaux, avec les *Recherches*, il propose une théorie économique de leur fonctionnement. Pour autant, Smith lui-même relie étroitement ces deux travaux, puisqu’il prend soin de noter, dans un avertissement à la dernière édition de la *Théorie des sentiments moraux* (1789) qu’il a intégré dans son *Essai sur la richesse des nations* les développements consacrés aux principes généraux des lois et du gouvernement, initialement annoncés dans la première édition de la *Théorie des sentiments moraux* (1759). Ces deux ouvrages relèvent donc, pour leur auteur, d’une même œuvre.

Une seconde explication met en avant les transformations, si ce n’est même une rupture, précédemment évoquée, intervenue dans l’appréhension de l’action et dans sa théorisation économique. Jusqu’à Smith et selon une tradition qui remonte aux philosophes écossais et plus particulièrement à Hume, l’analyse des actions des hommes prend sa source au niveau des émotions et, plus précisément des différentes passions qui les guident. Or, chez Hume, la définition des passions, telle qu’il en entreprend l’analyse, fait intervenir l’intersubjectivité au sens du chapitre précédent. À cette approche sensualiste de l’action se substitue, à partir de ce que l’on a coutume d’appeler la révolution marginaliste de la fin du XIX^e^ siècle, une perspective strictement rationaliste dominée par un cerveau calculateur. Ces deux traitements de l’action ne sont, du reste, pas

mutuellement exclusifs. Dans une tradition utilitariste héritière de Bentham, le but ultime de l'action est le plaisir, c'est-à-dire la recherche d'une émotion anticipée et éprouvée. Certains auteurs, comme Edgeworth, iront jusqu'à imaginer une *hédionométrie*. Cependant l'intersubjectivité n'a, ni la même importance, ni la même formulation dans chacune de ces deux approches de l'action. Ainsi l'estime de soi qui intervient dans l'intentionnalité d'une action appréhendée dans la perspective sensualiste est difficilement séparable de l'estime des autres, ce qui implique, dès l'origine, de poser une hypothèse sur la nature des relations entre soi et autrui. La perspective du calcul rationnel suppose seulement l'existence d'un moi décideur de l'action, à la fois indivisible, permanent et clairement différencié. Quant à autrui, il n'apparaît qu'au niveau des conséquences de l'action. Il est alors possible de distinguer, comme le fait, par exemple Edgeworth, le calcul égoïste (ou calcul économique), dont autrui se trouve exclu, du calcul altruiste (ou calcul utilitariste), où les autres sont intégrés comme argument dans le calcul[2]. C'est pourquoi, le problème de l'intersubjectivité coïncide avec l'identification de l'autre dans la première perspective, alors qu'il se confond avec celui de l'agrégation des autres dans la seconde.

Ce n'est pas ici l'objet d'expliquer pourquoi et comment la pensée économique a changé, sur cette question, de paradigme dominant. Pour être plus complet, il faudrait cependant mentionner l'existence de différents courants de pensée, qui se sont inscrits en faux par rapport à cette dichotomie. Leurs auteurs, qui refusent d'adopter l'une ou l'autre de ces deux perspectives, ont critiqué leur développement et ont recherché des approches alternatives de la réalité économique. On les regroupe un peu hâtivement, sous l'appellation ambiguë d'institutionnalisme. Le cas le plus intéressant pour notre propos est sans doute celui de Veblen, qui reste, aujourd'hui encore, difficilement classable dans les différentes écoles de pensée qui ont marqué l'histoire de la discipline. Élève du philosophe Pierce, sa pensée est imprégnée d'un mélange de pragmatisme et d'évolutionnisme darwinien. Il conçoit l'action des hommes comme un processus dynamique complexe, tout à la fois organique, parce que dépendant de nos instincts, et social, en raison du poids de nos habitudes de pensée et de nos expériences antérieures au cours de nos interactions avec nos divers environnements. Cela le conduit à déceler dans le processus économique un jeu au résultat incertain entre des instincts d'adhésion au groupe et des instincts contraires de prédation (Veblen, 1898). Mais le fait que la position soutenue par Veblen ait été, et reste encore, marginalisée au moins chez les économistes, peut précisément s'interpréter comme une manifestation de son caractère étranger à la pensée économique, confirmant ainsi le choix de notre présentation.

Une autre filiation, moins connue, relie certains membres de l'École autrichienne, et en particulier Hayek, à une approche phénoménologique de l'intersubjectivité inspirée d'Husserl, par l'intermédiaire du philosophe et sociologue Schütz. Schütz élabora sur cette base une sorte de dynamique naturelle de la « typification » qui, au travers des expériences tirées de confrontations quotidiennes aux autres, permet, dès l'enfance, à chacun d'induire une manière de schème social de référence, plus ou moins partagé par les autres (Schütz, 1935, 1953). On sait que Schütz entretint des contacts étroits avec les économistes autrichiens, et on retrouve une idée voisine de celle de Schütz

dans la série des essais consacrés par Hayek à cette question et publiés sous le titre *Individualism and Social Order* (1949)[3]. Mais, ici encore, il s'agit d'une approche singulière des relations sociales entre les sujets individuels, qui n'a guère été suivie par la majorité des économistes contemporains.

L'évolution de la discipline rapidement retracée éclaire sur la manière dont l'intersubjectivité s'est invitée, plus récemment, dans la problématique économique contemporaine, à la faveur d'un questionnement sur l'autre.

J

Tout commence réellement avec l'apparition de la théorie des jeux. Dans le paysage antérieur de la théorie classique, les rôles et les fonctions répondent à un découpage clair. Le décideur en est l'unique sujet avec ses goûts et ses préférences. Il est confronté à des états du monde qui lui sont extérieurs, mais sur lesquels il dispose d'informations. Point n'est besoin, dans ces conditions, de faire intervenir un autre sujet pour procéder à un choix rationnel. Il n'en va évidemment plus de même lorsque le sujet se trouve confronté stratégiquement à un, ou à plusieurs, autre sujet. Non seulement il lui faut connaître, ou tout au moins imaginer, ce que fera cet autre sujet, mais c'est à partir de cette connaissance qu'il va maintenant choisir son action, dans une perspective que l'on appelle stratégique, en tant qu'elle est une réponse au choix stratégique supposé de l'autre. Comme il en va de même pour cet « autre », en raison de la réciprocité, il en résulte une imbrication entre le sujet et l'autre sujet (les autres sujets), qui implique la prise en compte de l'intersubjectivité. Les situations de conflits militaires en fournissent sans doute l'illustration la plus saillante. Une négociation diplomatique met en œuvre ce même registre d'intersubjectivité.

Pour répondre aux nouvelles interrogations qu'elle suscite, les théoriciens des jeux ont d'abord pensé qu'il suffirait de compléter, ou de renforcer, des hypothèses internes ou externes, contenues de manière souvent implicite dans la théorie économique. La relecture des pères fondateurs est instructive sur les difficultés rencontrées dans ces différentes voies. von Neumann et Morgenstern développèrent ainsi une interprétation sociale des jeux à partir du concept de coalition (jeux coopératifs). Il reconnaît que pour y parvenir il est nécessaire de présupposer que les joueurs partagent des « standards de comportements acceptés » (*Theory of Games and Economic Behavior*, 1946). Mais en quoi consistent précisément de tels « standards de comportements » et, surtout, comment les joueurs les acquièrent-ils ? Sur ces points, pourtant fondamentaux pour leur nouvelle théorie, von Neumann et Morgenstern restent peu explicites. Ils se contentent de poser que leur investigation est extérieure à la théorie qu'ils construisent. Nous verrons, par la suite, comment cette notion de standard de comportement peut retrouver aujourd'hui une nouvelle place en théorie des jeux, entendue dans une perspective élargie, lorsque nous analyserons, au chapitre 6, les règles implicites d'un jeu.

Nash, de son côté, préféra privilégier une interprétation strictement individualiste des jeux, qui domine aujourd'hui encore l'analyse économique. À la recherche d'une définition purement formelle

de la solution d'un jeu, il esquissa cependant deux types différents d'interprétation dans sa thèse de doctorat (Nash, 1950). La première, et la plus connue, consiste à supposer que les joueurs formulent les uns sur les autres des hypothèses de choix rationnels, parce qu'ils savent que tous sont rationnels et qu'il n'y a dans le meilleur des cas qu'une seule manière de se comporter rationnellement. Nash précisait toutefois qu'une telle interprétation exigeait que tous les joueurs disposent également d'une connaissance parfaite de l'ensemble des structures du jeu (règles, environnement…). La mobilisation de toutes ces informations par les joueurs lui semblait suffisamment problématique pour qu'il ait également esquissé une seconde interprétation de son concept de solution. Selon cette autre interprétation, la convergence des stratégies requise par la solution d'équilibre pourrait résulter de l'accumulation des informations empiriques recueillies par les joueurs sur les autres joueurs, grâce à leurs observations au cours du déroulement du jeu. Pour qu'il en aille ainsi, selon un processus que Nash qualifie d'« actions de masse » (*Mass actions*) parce qu'il concerne des populations de joueurs, il faut maintenant que le jeu comprenne un nombre suffisamment élevé de séquences. Nash est cependant resté assez vague sur cette seconde interprétation du fonctionnement des interactions. Il concentra l'essentiel de son effort intellectuel à développer les conditions formelles de la première interprétation hyperrationaliste de sa théorie, en dépit de son irréalisme.

Par-delà leurs différences, ces différentes approches des jeux d'interactions partagent un point en commun. Que leurs auteurs abordent l'interaction par la coopération (alliances, coalitions), par la confrontation et la compétition, ou par la négociation (engagements non exécutoires), elles postulent toutes que les joueurs, en tant que sujets du jeu, détiennent une forme de connaissance qui leur est commune. Il est intéressant de noter à ce propos que certains modèles de négociation intègrent, avant le début du jeu lui-même, un échange informel entre les joueurs, baptisé *cheap talk*, qui leur permet précisément de révéler quelques éléments de cette base d'intelligibilité commune. Quant au contenu de cette connaissance prêtée aux joueurs, il varie d'une interprétation à l'autre. Les « standards de comportements » de von Neumann peuvent s'entendre comme des conventions sociales au sens de Lewis. La « rationalité forte », développée par Selten à partir de Nash, dans la perspective de ce qu'il a appelé des « équilibres purs » (Selten, 1975), ouvre la voie à une « rationalité connaissance commune » analysée par les logiques épistémiques. Quant à l'interprétation en termes d'actions de masse, développée par les théoriciens des jeux évolutionnistes, elle renvoie à des mécanismes d'apprentissage et de transmission, dont l'origine pourrait être d'ordre génétique. En clair, la relation à l'autre (aux autres), et donc l'intégration de l'autre (des autres), repose, en théorie des jeux, sur l'hypothèse qu'il existe, en amont du jeu lui-même, un support commun aux représentations des joueurs.

L

L'hypothèse d'un support commun des représentations réciproques des joueurs s'est d'abord

imposée à la théorie des jeux de manière implicite, comme on l'a montré. Ce n'est que plus récemment qu'elle a été explicitée.

Trois questions se posent désormais à son sujet :

- Sur quoi repose sa validité ?
- Si elle est confirmée, suffit-elle à garantir l'intelligibilité de l'interaction, telle qu'elle se manifeste au cours d'un jeu ?
- Comment identifier et caractériser l'altérité de l'autre (des autres), en partant de cette hypothèse ?

De fait, l'essentiel du travail des théoriciens des jeux a porté sur la résolution de la seconde question conjointement à son explicitation. Ils l'ont entrepris en privilégiant la rationalité comme support des représentations communes aux sujets, et en approfondissant sa dimension logique ; d'où leur rencontre avec les logiques épistémiques à partir du milieu des années 1990. Nous examinerons les possibilités, mais également les limites auxquelles se heurte aujourd'hui cette orientation, et suggérons d'autres directions pour appréhender ce support.

La rationalité, dans son acception cognitive la plus large, constitue une clé générale destinée à rendre intelligibles les phénomènes. Elle joue un rôle plus précis lorsqu'il s'agit de comprendre les phénomènes sociaux émanant d'actions intentionnelles individuelles, comme en économie. La rationalité supposée des acteurs y constitue une référence commune qui permet l'interprétation, par un observateur ou par un modélisateur, d'une action observée chez un autre sujet, ou même seulement imaginée effectuée par lui. Pour reprendre une formulation que nous avons utilisée dans un autre contexte, lorsque l'économiste énonce que « les agents sont supposés agir rationnellement », il sous-entend, en réalité, qu'« il est rationnel pour lui de construire un modèle où les agents agissent rationnellement » (Schmidt, 1996). Cela signifie que le premier mérite de cette hypothèse est de fournir une passerelle pour rendre intelligible une intersubjectivité fondatrice, puisque c'est elle qui permet à un agent d'analyser les actions d'un autre agent, qui lui est, en principe, extérieur. D'une certaine manière, par conséquent, c'est au moyen de cette rationalité que la théorie économique a cherché, d'une façon qui peut paraître à première vue paradoxale, à échapper au risque de solipsisme dénoncé au chapitre précédent.

Les choses se compliquent avec les jeux, puisque, ce support supposé commun ne relie plus seulement l'observateur et l'observé ; il doit également rendre compte des relations entre les joueurs, qui sont, eux-mêmes des observés du point de vue de l'observateur. On comprend mieux pourquoi la seule hypothèse de rationalité devient alors insuffisante pour saisir, en ces termes, les interactions entre deux joueurs. Si l'hypothèse que chaque joueur est supposé agir rationnellement peut garantir l'intelligibilité de leurs choix considérés séparément, elle ne suffit plus, en revanche, à rendre intelligible leur interaction. Celle-ci renvoie, en effet, à une intersubjectivité de second degré. De ce fait, c'est l'objet même de la théorie des jeux qui se révèle ainsi différent de celui de la théorie économique de la décision individuelle.

La théorie des jeux s'est d'abord contentée d'étendre aux interactions interindividuelles un modèle de type « problème-solution » conçu par la théorie économique pour traiter des choix

individuels. Chaque joueur se trouve confronté à un problème de choix optimal, la solution du jeu se déduit alors de la résolution de ce problème par chacun des joueurs. Cette approche simplificatrice réduit l'intersubjectivité propre aux relations interindividuelles à une simple juxtaposition des sujets. Les nombreuses difficultés qui sont apparues du fait de ce traitement, tant au niveau théorique (paradoxes, indécidabilité…) qu'au plan expérimental (invalidations répétées des résultats théoriques) ont conduit plusieurs théoriciens des jeux à dégager une nouvelle approche de l'intersubjectivité dans les situations d'interactions interindividuelles. Elle consiste à intégrer dans le processus décisionnel des sujets un niveau supplémentaire correspondant aux connaissances des joueurs sur les connaissances des autres joueurs. Comme ces connaissances sont incertaines, nous préférons parler ici de « croyances », dans l'acception logique où ce terme a été utilisé (Monderer, Samet, 1989). Plus précisément, considérons la proposition « le joueur 1 croit que le joueur 2 croit que… ». Cette proposition n'implique pas que le joueur 1 sache ce que le joueur 2 croit, mais seulement qu'il sait ce que veut dire « croire quelque chose » pour le joueur 2. C'est sur cette base épistémique qu'Aumann a, par exemple, construit un système logique cohérent de connaissances communes. Il en propose une interprétation intuitive, en recourant à l'image d'un code lexical, ou d'un dictionnaire (Aumann, 1999). Grâce à un tel dictionnaire les joueurs pourraient accéder à cette intersubjectivité, sans pour autant pénétrer dans la subjectivité de l'autre. Ce dictionnaire commun fournit ainsi un fondement à l'existence logique d'une intersubjectivité à l'œuvre dans toutes les situations d'interaction analysées par la théorie des jeux. On notera toutefois que cette approche syntaxique n'est pas tout à fait suffisante. Pour pouvoir se comprendre, les joueurs doivent également pouvoir disposer d'un moyen de traduire le contenu des catégories utilisées par l'autre (les autres) dans leurs propres catégories

Les théoriciens des jeux ont réinterprété dans ces termes l'hypothèse de rationalité, comme une croyance partagée par les joueurs relevant, de ce fait, de l'intersubjectivité ainsi définie. Rappelons seulement à ce propos que le concept de rationalité, tel que l'entend l'analyse économique et, plus largement, ses différentes théories de la décision, comporte deux composantes distinctes : une relation de cohérence strictement logique, et la référence à un (ou plusieurs) critère(s) que doit satisfaire le choix de la décision pour valider cette cohérence. Or, si on peut considérer la cohérence comme une connaissance commune aux hommes pourvus d'un cerveau raisonneur, il n'en va pas de même des critères retenus par les décideurs, qui impliquent une part d'arbitraire et donc de subjectivité. Dès qu'il existe plusieurs critères possibles, on peut choisir de manière cohérente selon des critères différents. Le choix du critère retenu porte nécessairement la marque d'une subjectivité propre au décideur. La référence à un critère fait certes partie du dictionnaire commun, mais elle n'est pas suffisante pour permettre au joueur d'en déduire le critère retenu par l'autre (les autres) joueur(s).

L'exemple classique du jeu de la chasse au cerf que nous analyserons plus en détail au chapitre 5, lorsque nous traiterons de la coopération, en fournit une illustration. Deux joueurs qui ne se connaissent pas sont disposés à chasser dans un terrain giboyeux. Chacun de leur côté a une très grande probabilité d'attraper un lièvre, mais ils peuvent, également, s'organiser ensemble pour capturer un cerf. De manière très schématique, on considère que chacun d'eux est assuré d'obtenir un

lièvre s'il choisit la première option, indépendamment de celle qui aura été choisie par l'autre. Tous les deux peuvent également gagner une proie plus avantageuse en choisissant la seconde, mais cette fois la réussite de l'opération dépend pour chacun du choix de l'autre. Ce dilemme met en lumière deux critères de rationalité différents qu'Harsanyi et Selten ont baptisés respectivement *Risk dominance*, pour justifier le choix rationnel de la première option, et *Payoffs dominance*, celui de la seconde option (Harsanyi et Selten, 1988). Le risque attaché à la seconde option, pourtant la plus avantageuse pour les deux, est ici que l'autre ne la choisisse pas. Dans cette hypothèse celui qui l'aurait retenue se trouverait sans rien. Dès lors, les deux joueurs, s'ils sont averses au risque, choisiront *rationnellement* la première option. On observera qu'en l'occurrence, le risque provient ici d'une défiance envers le comportement de l'autre joueur. Comme la prise d'un lièvre est la solution la plus intéressante pour les deux joueurs, certains sont tentés de soutenir qu'il suffit alors que la rationalité soit une connaissance commune entre eux pour que ce risque disparaisse, rendant ainsi la seconde option seule rationnelle pour chacun des deux. Cette interprétation est certes possible, mais son argumentation repose moins sur la rationalité elle-même, que sur le fait qu'elle est, dans cet exemple, le support d'une convention communément acceptée par les deux joueurs. N'importe quelle règle conventionnelle, dès lors qu'elle est une connaissance commune entre les joueurs, peut, en effet, leur permettre de se garantir contre ce type de risque. Le problème posé dans ces situations est alors de déterminer dans quelles conditions tel critère de rationalité est susceptible de se transformer en une connaissance commune[4].

Indépendamment de l'hypothèse de rationalité, cette approche des interactions stratégiques, analysée en termes de système de croyances des joueurs, se révèle de toute façon insuffisante pour comprendre comment fonctionne réellement l'intersubjectivité. Aumann et Brandenberger le reconnaissent eux-mêmes en ces termes :

« Comme nous l'avons indiqué précédemment, les systèmes de croyances sont des représentations commodes pour nous permettre, nous, les analystes, de discuter des choses dont nous voulons discuter : des actions, des paiements, des croyances, de la rationalité, de l'équilibre, etc. En ce qui concerne les joueurs eux-mêmes, il n'est pas clairement établi qu'ils aient besoin de recourir à cette structure de modèle. Mais si c'est le cas aussi, et s'ils veulent parler des mêmes choses que celles dont nous discutons, alors d'accord, cette représentation est aussi pertinente pour eux qu'elle l'est pour nous. Dans le même ordre d'idée, nous noterons que le système de croyance, lui-même, doit toujours être considéré comme une connaissance commune entre les joueurs » (Aumann et Brandenberger, 1995).

Cette formulation a le mérite de mettre en évidence une dimension essentielle de la difficulté à appréhender l'autre, qui tient à la pluralité des différents points de vue d'où l'autre (les autres) peut (peuvent) être considéré(s). Ce n'est pas, en effet, la même chose d'interagir avec un autre (point de vue des joueurs les uns sur les autres) et d'appréhender de l'extérieur le comportement des joueurs (point de vue des analystes sur les joueurs). De même un joueur n'appréhende pas de la même manière, son partenaire/adversaire de jeu, les joueurs d'un autre jeu qu'il observe, ou encore un sujet extérieur en dehors de toutes interactions. Ces différences se trouvent aujourd'hui corroborées au

niveau cérébral par les résultats de travaux récents de neurosciences, qui seront analysés et discutés dans les deux derniers chapitres de cet ouvrage, lorsque sera abordée la question de la tierce personne.

Transposer directement le schéma logique de partage des croyances, tel qu'il a été présenté, à l'analyse des mécanismes qui guide l'intersubjectivité effectivement à l'œuvre dans les interactions, se révèle insuffisant, en raison de deux difficultés principales.

Il est établi, en premier lieu, que la connaissance commune d'un système de croyances entendu au sens strict par les joueurs dépasse la compétence cérébrale de chacun. Cet argument, fréquemment avancé, n'est pas cependant dirimant. Sans pouvoir contrôler consciemment les itérations infinies qu'exige formellement une connaissance commune, les joueurs disposent néanmoins d'une idée intuitive de sa signification, comme lorsqu'il s'agit d'idéalités mathématiques (0, …, ∞). Une telle exigence n'est pas, en outre, toujours indispensable. Lorsqu'il s'agit d'une règle explicite de connaissance publique, par exemple un code. Puisque nul n'est censé ignorer la loi, chacun est alors censé connaître la règle, savoir que chacun la connaît et savoir que chacun sait qu'il sait que chacun la connaît.

Une seconde difficulté, plus sérieuse porte sur le contenu exact de ces croyances. Dans les exemples de la théorie des jeux, les systèmes de croyance des joueurs sont étudiés du point de vue de ceux qu'Aumann et Brandenberger appellent les analystes, c'est-à-dire les théoriciens des jeux. À ce titre, ils se trouvent appréhendés par ces derniers comme des états du monde des joueurs qui leur sont, par hypothèse, extérieurs. Certes, de tels états intègrent maintenant les actions des joueurs, leurs conséquences et les transformations du monde qu'elles entraînent pour eux. Mais leur point de vue sur ces états reste différent de l'appréhension que peuvent en avoir les joueurs eux-mêmes, car ils ne sont pas les acteurs du jeu. Pour les joueurs, ces états sont des états mentaux. Comprendre autrui en situation d'interaction, ce n'est pas seulement, en effet, connaître ce qu'il fait pour en déduire une croyance sur ce qu'il va faire, ni même prendre en compte dans son raisonnement un degré supplémentaire de connaissance (non seulement, par exemple, je sais ce qu'il sait, mais aussi je sais ce qu'il sait que je sais…). L'extension du raisonnement aux croyances concerne ici exclusivement les informations détenues, ou supposées détenues, par les uns sur les informations des autres. On songe à l'exemple classique en économie des asymétries d'information entre les assureurs et les assurés. Une telle connaissance se révèle toutefois insuffisante pour saisir comment fonctionne réellement l'intersubjectivité à l'occasion d'échanges interactifs. Toutes ces informations sont alors interprétées subjectivement par chacun, ce qui transforme leur contenu en états mentaux. Le dictionnaire imaginé par Aumann permet certes à chacun de comprendre cette transformation des états du monde en états mentaux pour autrui comme pour lui-même, sans, pour autant, fournir plus d'information sur ces états mentaux d'autrui.

L'autre moi-même et les limites cognitives d'une rationalité partagée

De quoi un individu dispose-t-il pour imaginer l'état mental d'un autre individu au moment où il réagit, c'est-à-dire lorsqu'il répond à une action de sa part, réelle, ou seulement supposée ? Une posture familière consiste, pour lui, à « se mettre *mentalement* à la place de l'autre ». Sa formulation est plus imagée en langue anglaise, puisqu'elle consiste à « se mettre dans les chaussures de l'autre ». Il existe du reste une petite différence dans ces deux formulations. Se mettre à la place de l'autre signifie, au sens propre, s'imaginer, soi-même, dans la situation où l'on sait (croit) que l'autre se trouve. Se mettre dans ses chaussures, c'est s'imaginer dans cette situation avec des moyens dont disposerait l'autre (ses chaussures) pour aborder cette situation, jusque dans ses implications motrices, afin de lui permettre d'en sortir en décidant un mouvement. Se mettre à la place d'un autre, selon la première acception, correspond à une opération mentale déjà complexe. Il faut d'abord pouvoir s'imaginer soi-même, comme extérieur (le « soi »), et considérer ce « soi » dans un contexte hypothétique différent. Mais se mettre dans « ses chaussures », au sens métaphorique où nous l'entendons ici, est, en réalité, impossible, parce que personne n'a directement accès au code par lequel un autre, placé dans cette situation, la transforme en états mentaux qui lui sont propres. Tout au plus pouvons-nous imaginer, de manière conjecturale, ce que pourrait être son appréhension de la situation. Adam Smith dans *La Théorie des sentiments moraux* fut sans doute l'un des premiers à prendre la mesure de cette difficulté et à en esquisser une solution sociale, à travers l'apprentissage réciproque d'une sorte d'intersubjectivité sociale, par l'intermédiaire d'un jeu de rôle entre l'acteur et le spectateur. Il existe, enfin, une troisième perspective selon laquelle un sujet cherche à se mettre à la place de l'autre lorsqu'il entend se représenter comment il est perçu par l'autre. On la trouve dans la formule familière « se regarder dans les yeux de l'autre » reprise dans plusieurs travaux de neurosciences et qui inspirent aujourd'hui la théorie de l'esprit. Là encore, Smith fut un pionnier qui, ayant identifié cette aptitude singulière des hommes, s'employa ensuite à en dégager quelques-unes de ses conséquences sociales (reconnaissance, adhésion…). L'autre moi-même constitue ainsi une hypothèse mentale à double usage, en quelque sorte réversible : elle me permet de me mettre à la place de l'autre, comme elle permet à l'autre de se mettre à ma place.

Ces différentes déclinaisons de l'autre moi-même montrent d'abord que l'interprétation de cette notion reste très dépendante de la perspective adoptée par l'agent. On y retrouve des raisons supplémentaires de l'attachement manifesté par une majorité des théoriciens des jeux à l'hypothèse de rationalité dans l'analyse des interactions. Sur la base de quelles informations un joueur peut-il, en effet, imaginer les états mentaux d'un autre joueur qu'il ne connaît pas ? Il s'agit, d'une part, de sa connaissance de la situation objective dans laquelle se trouve placé l'autre joueur. L'information dont il dispose sur cette situation peut être plus ou moins complète, mais cela est une autre question. Il utilise, d'autre part, une information d'ordre logique, que croit détenir le joueur et qu'il prête également à l'autre joueur, sous l'appellation de rationalité. Nous en avons dégagé les limites. C'est toutefois en articulant la connaissance subjective de cette information objective sur la situation de l'autre, avec sa connaissance logique, supposée partagée par l'autre, de cette rationalité, que le joueur est censé pouvoir imaginer mentalement la réaction de l'autre joueur à l'action (aux actions) qu'il

décide d'entreprendre. Point n'est besoin pour le joueur de savoir si la représentation subjective que l'autre joueur a de la situation objective où il se trouve coïncide avec la propre représentation subjective qu'il en a, dès lors qu'avec la rationalité, il détiendrait, en commun avec cet autre joueur, un code logique supposé transcender ces représentations subjectives. Derrière la rationalité, c'est donc la propriété d'invariance qui lui est communément associée, qui permet au joueur, selon ce schéma, de se mettre à la place de l'autre. Par invariance, il faut entendre ici que différentes représentations (subjectives) d'une même situation (objective) conduisent logiquement à un même choix rationnel. En d'autres termes, la rationalité d'un choix, ou d'une décision, serait logiquement indépendante de ses représentations. Grâce au principe d'invariance le plus souvent associé à la rationalité ainsi formulée, la difficulté inhérente à l'inaccessibilité des états mentaux d'autrui pourrait se trouver ainsi contournée.

Qu'en est-il vraiment ? L'invariance est une propriété logique qui s'applique seulement à la dimension de cohérence associée à la rationalité. Un même problème peut être formulé de façon différente, sans changer pour autant sa solution, c'est presque une évidence. À ce titre, et s'agissant d'un choix qualifié de rationnel, le principe d'invariance peut donner lieu à une interprétation normative, comme la plupart des règles logiques mobilisées par la référence à la rationalité. D'un point de vue psychologique cependant, il en va autrement, puisque aucun choix, fût-il rationnel, n'est réductible à une seule opération logique et encore moins à un calcul. Nous avons déjà vu que la justification logique de la rationalité n'excluait pas l'introduction de différents critères. Il faut maintenant dépasser ce constat et poursuivre l'analyse. Choisir une action implique, en effet, en dehors d'un calcul, une intentionnalité et un engagement du sujet qui sont indissociables de la représentation personnelle, et donc subjective, qu'il se fait de la situation de choix à laquelle il se trouve confronté, et à laquelle il répond par son choix. C'est là qu'interviennent les états mentaux. Plusieurs représentations subjectives différentes d'une même réalité objective sont donc, pour cette raison, légitimement compatibles avec des choix différents rationnellement justifiés. La rationalité, même partagée ou connaissance commune, ne peut pas, dès lors, garantir qu'il suffise, en la circonstance, de se mettre à la place de l'autre (première acception) pour entrer dans ses chaussures (seconde acception). Elle ne fournit pas même une sorte de dictionnaire qui permettrait à chaque joueur de comprendre l'état mental d'un autre joueur, en traduisant sa propre représentation du choix dans celle d'un autre.

On peut toutefois tirer de cette analogie une hypothèse d'inspiration différente. Sans pouvoir se mettre dans l'état mental de l'autre, le sujet peut néanmoins communiquer avec lui à travers un langage qui leur est commun, ce qui permet précisément aux joueurs de se comprendre, jusque et à partir d'un certain point. Le langage permet, du reste également, aux analystes de la théorie des jeux de comprendre les joueurs et de se faire comprendre par eux. Certaines recherches récentes de neurolinguistiques ont même identifié des bases neuronales à cette propriété d'intercognitivité du langage, en approfondissant ses fonctions de communication (Pinker et Bloom, 1990 ; Pinker et Jackendoff, 2005). C'est donc dans une direction différente de celle de la seule rationalité qu'il faut rechercher les voies par lesquelles un sujet appréhende un autre sujet en vue d'une interaction. Une

démarche qui rencontre, à ce stade, la perspective phénoménologique. Plus précisément, il est révélateur que Husserl ait consacré un cours entier à la phénoménologie de l'attention (Husserl, 1904-1905), à l'occasion duquel il analyse l'attention partagée comme une modalité de l'intersubjectivité[5]. On retrouve cette idée interprétée un peu différemment chez Merleau-Ponty (1946), comme nous l'avons montré dans un autre travail[6]. Pour autant, cette faculté des individus de pouvoir se projeter eux-mêmes dans la situation où ils pensent que se trouve autrui n'est pas sans relation avec la rationalité de leur choix, en cas d'interdépendance.

Reprenons la métaphore de se mettre à la place de l'autre entendue dans sa première acception. Afin de choisir rationnellement sa stratégie, le joueur imagine la stratégie adoptée par l'autre joueur, en se représentant comment il agirait s'il se trouvait lui-même dans la situation où se trouve cet autre joueur. Nous avons montré que rien ne lui garantit alors, qu'en procédant ainsi, son anticipation sur le choix de l'autre serait confirmée. Il peut toutefois légitimement penser que si l'autre choisit également sa stratégie rationnellement, il procédera de la même manière, c'est-à-dire qu'il se mettra aussi mentalement dans la situation où il se trouve lui-même. Ces deux « autres moi-même » n'ont évidemment aucune raison, *a priori*, de coïncider. Mais en interprétant la stratégie effectivement mise en œuvre par l'autre joueur, par référence à cette fiction de l'autre moi-même, chaque joueur pourra néanmoins avancer dans la connaissance de l'autre joueur véritable. En dynamique, un mécanisme mental de rapprochement mutuel par effet d'apprentissage et de correction des erreurs pourrait ainsi s'enclencher sur cette base, selon une procédure de type bayésien qui mobiliserait alors une autre forme de rationalité. On pourrait y reconnaître l'une des voies qui conduit à l'intersubjectivité.

Cette construction fait intervenir l'hypothèse d'une dynamique de l'apprentissage dans l'interaction qui, comme on l'a déjà dit au chapitre précédent, et comme on le verra plus en détail par la suite, s'est trouvée en grande partie validée expérimentalement. Mais ce qui rend possible son fonctionnement doit être recherché en amont, au niveau de l'organisation cérébrale des êtres humains, qui pourrait expliquer cette propriété fondamentale dont ils disposent, de se mettre à la place des autres et de savoir que les autres peuvent également se mettre à leur place. Cette aptitude de nos cerveaux à effectuer ce que l'on pourrait nommer « un apprentissage social » de l'interintentionalité entraîne de multiples implications, dont certaines seulement ont été recensées ici, à travers les différentes métaphores qui ont été rappelées. Ce sont elles qui sous-tendent de très nombreuses manifestations phénoménales que l'on retrouve au cœur des interactions sociales comme, par exemple, l'empathie, l'imitation, la simulation et, peut-être même, la coordination et la coopération. Ces deux dernières manifestations seront étudiées plus en détail dans la deuxième partie. Les informations fournies par les neurosciences, et en particulier par la branche cognitive des neurosciences sociales, permettent en tout cas d'avancer dans leur compréhension.

L

Les situations de jeu offrent un cadre général privilégié, simple et rigoureux, qui capte les traits essentiels des phénomènes d'interactions intentionnelles et raisonnées. Le jeu représente un « petit monde », à la fois extérieur pour chacun, mais commun à tous ceux qui y participent. Chaque joueur sait que, lorsqu'il choisit une action pour un objectif déterminé (intention), le résultat de l'action qu'il entreprend dépend également d'actions choisies par l'autre (l'autre) joueur (s) qui lui est (sont) extérieur(s). À la notion générale d'une rencontre des actions s'ajoute donc le phénomène plus spécifique d'une rencontre des intentions qui guident ces actions. D'un côté, cette interintentionnalité réduit l'incertitude de l'interaction, puisque toutes les actions retenues dépendent des intentions. D'un autre côté, elle est source d'ambiguïté pour chaque joueur, puisque les intentions des autres ne sont pas transparentes.

L'introduction de cette distinction entre la simple interaction et l'interintentionnalité a le mérite d'éclairer certaines contradictions qui sont apparues de manière répétée entre les résultats théoriques établis par la théorie des jeux et les comportements observés, à l'occasion de protocoles expérimentaux. Ainsi, les joueurs qui ne maximisent pas leur utilité individuelle, au risque de léser les autres, n'agissent pas de manière irrationnelle, ils sont seulement mus par d'autres intentions. Cette possibilité n'est pas retenue en théorie des jeux, où les intentions des joueurs se trouvent directement déduites des règles du jeu et où ces règles sont supposées connaissance commune entre les joueurs. Cette hypothèse forte concernant les règles du jeu ne se trouve pleinement vérifiée que dans le cas particulier des jeux de société, où l'intention commune prêtée aux joueurs et induite des règles est de gagner. Dans la réalité des interactions courantes de la vie économique et sociale, elle est, au contraire, loin d'être garantie.

Le concept de jeu s'est lui-même construit et développé à partir de celui de règles. On notera à ce sujet que le terme de jeu, choisi par les fondateurs de cette nouvelle discipline pour désigner son objet n'a pas seulement, pour eux, le sens métaphorique qu'on lui donne aujourd'hui. von Neumann, Nash, et avant eux Borel, élaborèrent plusieurs modèles de différents jeux de société et, en particulier, de jeu de Poker. Ce rappel est à rapprocher des observations développées au chapitre précédent sur la porosité entre le réel et le virtuel, qui caractérise les jeux d'enfants. Sous cet angle, on peut considérer que les jeux de société gardent cette propriété intéressante, qu'ils transmettent à la théorie des jeux stratégiques qui constituent aujourd'hui le cœur de la théorie. Tous les joueurs d'un jeu de société sont censés connaître ses règles. On distingue parmi ces jeux de société, à partir de leurs règles, ceux où l'interaction est aléatoire et ceux où elle est intentionnelle, avec toute une série de catégories intermédiaires. C'est, du reste, en partant de cette distinction qu'une des premières expériences de jeu combinée à l'imagerie cérébrale au début des années 2000 par des chercheurs en neurosciences, a mis en évidence une différence au niveau du fonctionnement cérébral entre la simple interaction et l'interaction dérivée d'une interintentionnalité (Gallagher *et al.*, 2002).

Cette expérience consiste à faire jouer plusieurs fois des sujets au jeu d'enfant « pierre », « feuille », « ciseaux », en les informant de ses règles, au demeurant élémentaires. Les expérimentateurs leur précisent qu'ils seront successivement opposés : 1) au pur hasard, représenté

par une variable aléatoire ; 2) à un ordinateur doté d'un programme, et 3) à un autre individu. En réalité, les expérimentateurs les confrontent chaque fois aux résultats d'une variable aléatoire. Compte tenu des règles de ce jeu, les modalités différentes annoncées dans son déroulement ne modifient pas sa structure logique, laissant sa résolution identique dans les trois cas. Les résultats recueillis n'ont, du reste, pas révélé de différences significatives dans les solutions observées entre chacun de ces trois cas. Durant cette expérience, l'activité cérébrale des sujets a été mesurée, au moyen d'un traitement par imagerie selon la technique de l'IRMf. Enfin, les commentaires des sujets sur les impressions qu'ils ont ressenties pendant l'expérience ont été recueillis après son déroulement.

Deux résultats importants ont ainsi été mis en évidence.

Tout d'abord, si un grand nombre de régions du cerveau se trouvent activées dans ces trois versions du jeu, une seule aire cérébrale spécifique, le cortex paracingulaire antérieur, n'a été activée que lorsque les sujets croyaient qu'ils jouaient contre un autre individu. L'activation de cette région cérébrale est associée à de multiples fonctions, notamment à la sélection d'une action. Il apparaît toutefois qu'elle joue ici un rôle spécifique dans la recherche et l'interprétation des informations concernant les postures intentionnelles des autres, une activité observée ici dans une situation où la relation à l'autre est compétitive, mais qui a été également mise en évidence dans des situations où cette relation à l'autre est, au contraire, coopérative.

Ce premier résultat a, depuis lors, fait l'objet de discussions de la part des spécialistes, en raison notamment des singularités de son protocole expérimental. Il semble néanmoins confirmer l'hypothèse avancée selon laquelle, indépendamment de la structure logique du problème posé par une situation d'interaction, l'agent appréhende ce problème de manière partiellement différente lorsqu'il sait (ou croit savoir) que cette situation résulte de sa confrontation avec un autre individu doté d'une intention. Il existe donc, au niveau neuronal, une différence entre l'appréhension de l'interaction et l'appréhension de l'interintentionnalité. C'est la raison pour laquelle ces chercheurs ont proposé de rattacher cette activation neuronale particulière à l'idée plus générale de détection d'une « posture intentionnelle ».

En second lieu, l'entretien de retour d'expérience a révélé que les sujets ont tous ressenti comme différentes leur confrontation à un ordinateur et leur interaction supposée avec un autre joueur. Ils gardent de leur affrontement avec l'ordinateur un souvenir plus tendu et intellectuellement beaucoup plus exigeant qu'avec un autre joueur, et cela en dépit des similitudes des actions qu'ils ont retenues dans les deux cas. Une relation précise a ainsi été mise en évidence entre, d'une part, un phénomène d'ordre physiologique, l'activation du cortex paracingulaire, qui relève de ce que l'on nomme communément le subpersonnel et, d'autre part, la sensation éprouvée par le sujet qui traduit un état mental personnel. En la circonstance, une telle relation est associée à la seule croyance du sujet qu'il est en face d'un autre sujet, puisque tous les autres paramètres de la situation sont identiques dans les trois cas.

On peut pousser plus loin cette analyse. La seule façon pour le sujet d'identifier la stratégie gagnante serait en définitive, dans ce jeu, de deviner la stratégie retenue par le joueur qui lui est opposé. Une telle stratégie n'est évidemment pas applicable dans le cas où il est confronté au hasard.

Elle apparaît difficile à mettre en œuvre dans le cas de l'ordinateur, plus difficile, en tout cas, que lorsque l'adversaire est une autre personne. Pourquoi ? Si, par exemple, cet autre joueur était visible, il lui serait peut-être possible de lire, dans son regard, le coup qu'il aurait choisi de jouer, ou d'inventer un piège pour qu'il se trahisse. Certes, durant le déroulement de l'expérience, il ne voit pas cet autre joueur auquel il est censé être opposé, mais son cerveau l'appréhende comme s'il pouvait le faire. Cela expliquerait ainsi le fait qu'il éprouve moins de peine à jouer contre un autre joueur que contre un ordinateur, même s'il ne connaît pas ce joueur et qu'il ne le voit pas au cours du déroulement de l'expérience.

L'interprétation esquissée ici rejoint curieusement une observation énoncée par René de Possel, qui, dans le prolongement des travaux de Borel sur les jeux de société, proposa une catégorisation des jeux de société. À côté des jeux de pur hasard, comme pile ou face, et de pure réflexion, comme les échecs, il introduisit une catégorie supplémentaire, celle des jeux de ruse, en définissant la ruse comme la possibilité donnée à un joueur de deviner les pensées des autres (Possel, 1936). Le jeu « pierre, feuille, ciseaux » en fournissait déjà un exemple. La ruse consiste ici à utiliser, à son avantage et à l'insu de (des) l'autre(s), la faculté d'appréhender mentalement l'intention de l'autre de jouer l'une de ses stratégies. On observera toutefois qu'une telle faculté n'est pas réellement utilisable dans le cadre de l'expérience qui a été rapportée, puisque le joueur ne dispose ici d'aucun moyen pour deviner ce que l'autre joueur aura choisi de jouer.

Cette observation est intéressante pour une autre raison. Elle met en évidence, au niveau neuronal, une différence entre la compétence et la performance. Le fait, pour un joueur, d'être confronté à un autre joueur supposé humain doté d'intentionnalité positionne d'une certaine manière son activité cérébrale, même si cette disposition ne débouche sur aucune performance. On objectera que la forme compétitive de ce jeu dicte ici l'intention de l'adversaire, ce qui ne permet pas, à ce niveau, de distinguer l'autre joueur humain de l'ordinateur. Pour autant, le seul fait, pour un joueur, de savoir qu'il est face à un autre joueur humain, plutôt que devant un ordinateur, active chez lui une zone du cerveau qui pourrait être utilisée pour « deviner ses pensées », selon la formule déjà retenue par Possel.

D

C'est à partir d'un type de jeu différent qu'une majorité de chercheurs s'est employée à dégager les principaux ressorts neuronaux du fonctionnement de cette interintentionnalité au cours d'interactions interprétables en termes de transactions économiques. Son archétype est le jeu dit de la confiance, parfois baptisé aussi jeu de l'investissement (Berg, Dickhaut et McCabe, 1995). Un premier joueur, nommé l'investisseur, peut choisir d'investir (ou de ne pas investir) une proportion I quelconque d'une somme X dont il dispose au départ. L'investissement consiste ici à confier ce capital à un second joueur, le mandataire (*trustee*), qui le fera fructifier de manière à en multiplier le

montant de I (1+r). C'est alors au mandataire de déterminer le partage des gains ainsi obtenus entre lui-même et l'investisseur soit, respectivement, Y pour lui-même et I (1+r)-Y pour l'investisseur. À son tour, l'investisseur peut, à nouveau, placer cette nouvelle somme correspondant à (X-I) +I (1-r) (ou tout au moins une fraction de cette somme) auprès du mandataire ou stopper le jeu ; et le jeu se déroule ainsi pendant un nombre de séquences déterminé à l'avance.

La structure de ce jeu comporte un certain nombre de caractéristiques intéressantes pour notre enquête. Elle permet, en premier lieu, de préciser les différences de traitement du risque, selon qu'il émane du monde extérieur anonyme, ou qu'il dépend de la décision d'un autre ; d'où l'introduction de la confiance (et de la défiance) dans le second cas. Dans ce jeu, en effet, toute l'incertitude réside pour chaque joueur dans l'action de l'autre, puisque le taux de rendement de l'investissement, r, est supposé exogène et connu d'avance par les deux joueurs. C'est la raison pour laquelle on l'appelle jeu de la confiance. Mais rien n'empêche de le faire jouer à des individus en substituant au joueur auquel il serait confronté, un ordinateur dont les actions seraient calculées sur la base d'un historique statistique des résultats antérieurement obtenus au cours de ce jeu. Ces réponses de l'ordinateur leur fournissent ainsi une mesure du risque auquel ils se trouvent exposés dans ce jeu. C'est ce qu'a réalisé une équipe de chercheurs qui a comparé de manière systématique les comportements des sujets dans ces deux types de situations dérivées du même jeu (Houser, Shunk et Winter, 2010).

Les résultats de ce travail montrent clairement que les comportements des joueurs se révèlent beaucoup plus tranchés (tout investir, ne rien investir ; tout garder, tout partager) lorsque le risque est perçu en termes de confiance en l'autre que lorsqu'il est appréhendé en termes de probabilités. Ces résultats sont à rapprocher de recherches antérieures visant à explorer les bases neuronales de ces comportements. Ainsi, a-t-on mis en évidence plusieurs différences relatives aux activations cérébrales concernant notamment des points précis de la région préfrontale du cerveau dans chacune des deux versions du jeu. Ces activations seraient beaucoup plus intenses dans le cas où le risque est réductible à une question de confiance (McCabe *et al.*, 2001). De même, l'effet de l'ocytocine, un neuropeptide connu pour ses propriétés anxiolytiques, n'a été observé que dans l'hypothèse d'une incertitude résultant d'un manque de confiance.

Cette expérience a mis en évidence une autre différence intéressante, liée, cette fois, aux situations respectives de l'investisseur et du mandataire. Tandis que l'ocytocine agit sur le comportement de l'investisseur, en réduisant son inhibition et en favorisant sa confiance, elle se révèle sans effet sur celui du mandataire. L'explication proposée pour en rendre compte consiste à introduire une distinction supplémentaire entre la pure confiance qui ne peut se fonder que sur un *a priori* de sociabilité, dans le cas de l'investisseur, et la confiance calculée, tirée de l'expérience, qui se manifeste dans le cas du mandataire (Kosfeld *et al.*, 2005). Une autre explication est également possible. Le risque auquel se trouve exposé l'investisseur n'est pas identique à celui que rencontre le mandataire. Le mandataire peut, en effet, garder pour lui la totalité du gain. Cette possibilité correspond, dans l'esprit de l'investisseur, à une perte par rapport au retour attendu de son placement. Il s'agit d'un risque irréductible, car il ne peut rien faire pour l'éviter. Si maintenant l'investisseur ne confie plus rien au mandataire à la séquence suivante du jeu, le mandataire ne le comptabilisera pas

forcément mentalement comme une perte. En outre, il contrôle en partie ce risque, puisqu'il sait que la décision de l'investisseur à la séquence suivante dépend de la part du gain qu'il aura antérieurement rétrocédée à l'investisseur. Il n'est donc pas surprenant que cette différence se trouve mise en évidence par l'effet de l'ocytocine. L'une des propriétés de ce neuropeptide est, en effet, d'inhiber l'activation de l'amygdale ; or l'amygdale est considérée comme le siège des émotions négatives engendrées par l'angoisse et la peur. Son rôle a notamment été mis en évidence dans le cas de la peur de pertes financières qui se trouve à l'origine d'une aversion au risque financier.

Revenons aux enseignements susceptibles d'être tirés du déroulement de ce jeu de la confiance pour comprendre les ressorts dynamiques de l'interintentionnalité. À des degrés différents selon sa position dans le jeu, chacun des joueurs sait que l'action qu'il choisit dépend de la confiance qu'il place dans l'action de l'autre, mais que l'action qu'il aura choisie, affectera, à son tour, la confiance que l'autre lui portera. La notion de confiance qui est ici déterminante dans le fonctionnement des interactions présente cependant une certaine ambiguïté : faut-il attribuer cette confiance directement à l'autre, ou seulement progressivement, au vu des actions accomplies par cet autre ? C'est la dynamique même enclenchée par le jeu qui semble, en définitive, résoudre ce dilemme. Chaque joueur peut, en effet, induire, de l'observation des actions effectuées par l'autre joueur au cours du déroulement du jeu, la confiance qu'il pourra lui accorder. Ce schéma bayésien permet ainsi d'associer l'autre aux actions qu'il a effectuées, en accord avec la seconde interprétation de cette confiance. Mais il ne permet pas d'expliquer comment la réciprocité de la confiance entre les joueurs démarre et par conséquent comment un tel schéma se met en place. Il ne peut pas, en effet, rendre compte de la confiance accordée au mandataire par l'investisseur lorsque celui-ci joue son premier coup. Si chacun, de son côté, peut, au terme de cette procédure construire sa confiance dans l'autre, il n'en résulte pas mécaniquement une interdépendance des intentions, surtout lorsque les rôles de chacun sont différents. Son explication doit donc être recherchée ailleurs.

Les expériences menées depuis quelques années par des équipes de chercheurs pluridisciplinaires en neurosciences fournissent aujourd'hui plusieurs informations susceptibles de nous mettre sur la voie d'une explication plus satisfaisante du fonctionnement de cette réciprocité dans la confiance. Il se confirme, tout d'abord, que ce sont des parties un peu différentes de la même région du cerveau qui se trouvent activées chez l'investisseur et chez le mandataire, lorsqu'ils se font respectivement confiance, soit la partie médiane du cortex cingulaire pour l'investisseur, et la partie antérieure du cortex cingulaire pour le mandataire. Cette différence en fonction des rôles, et surtout des séquences du jeu, pourrait expliquer pourquoi le cerveau de l'investisseur ne travaille pas selon les mêmes modalités que celui du mandataire. Elle renforce, en ce sens, l'observation précédente faite sur la sensibilité du comportement de l'investisseur à l'ocytocine. L'investisseur qui, au moins au départ, ne dispose d'aucune information sur le comportement du mandataire parie alors sur la confiance par une empathie sociale, dont on a montré qu'elle se trouve activée par ce neuropeptide. On observe, ensuite, une corrélation très significative entre l'activation du cortex cingulaire médian chez l'investisseur et l'activation du cortex cingulaire antérieur chez le mandataire. Plus précisément, l'imagerie cérébrale

révèle que la décision de coopérer du mandataire s'ajuste, d'abord au signal fourni par la somme remise en jeu par l'investisseur, puis, après quelques séquences du jeu, au signal anticipé de cette somme. Tout semble donc se passer comme si le mandataire faisait, au cours des premières séquences, l'apprentissage du modèle mental qui régit le mécanisme de confiance chez l'investisseur (King-Casas *et al.*, 2005). Mais l'interaction cérébrale qui régit le mécanisme de confiance réciproque entre les deux joueurs apparaît, en définitive, imputable à leur intersubjectivité. On a observé, en effet, qu'elle ne se manifeste plus lorsque l'on substitue aux véritables joueurs des informations visuelles sur le déroulement du jeu, pourtant identiques dans leur contenu, à celles obtenues par les joueurs dans le cadre du jeu de la confiance. En revanche, ce mécanisme paraît indépendant du résultat positif, négatif ou neutre, induit par la réciprocité, ainsi que des montants en jeu et n'est pas significativement sensible au nombre des séquences (Tomlin *et al.*, 2006).

Comment faut-il interpréter ces données ? Le paradigme de la confiance est intéressant pour notre enquête sur les modalités de fonctionnement de l'interintentionnalité, dans la mesure où faire confiance à un autre renvoie, chez cet autre, à une forme de réciprocité. Ainsi, peut-on penser qu'un signal de confiance perçu chez l'autre active directement l'intention d'y répondre, en donnant à son tour sa confiance à l'autre, selon un modèle où la réciprocité prendrait la place de l'imitation dans la dynamique d'un système du type de celui des neurones miroirs évoqué au chapitre précédent. Ressentir dans l'action de l'autre une manifestation de confiance pourrait ainsi déclencher automatiquement un système, ou plus probablement, une chaîne de systèmes, qui conduirait à faire confiance à cet autre. Ce mécanisme permettrait notamment d'expliquer pourquoi, à un certain moment du processus, la confiance devient un domaine commun aux deux joueurs, de telle sorte que la question qui se pose alors à eux est celle de la distinction entre moi « moi » et l'« autre (moi) », au niveau de sa manifestation. On notera à ce sujet que Iaccoboni et ses collègues ont émis sur ce point l'hypothèse que le mécanisme des neurones miroirs, mis en lumière à partir de la vision d'actes moteurs effectués par d'autres, pourrait être étendu aux intentions révélées par de tels actes (Iaccoboni *et al.* 2005). Cette conjecture a toutefois été discutée et contestée. Encore faudrait-il, pour déclencher ce mécanisme dans l'exemple du jeu de la confiance, qu'il existe un acte moteur initial.

Transposer le schéma des neurones miroirs au cas de la confiance exigerait, au préalable, de distinguer plusieurs niveaux dans le transfert de confiance par l'intermédiaire de la réciprocité : un premier niveau élémentaire, de caractère biologique, dont certaines traces neuronales ont pu être mises en évidence dans d'autres expériences ; un second niveau, plus élaboré, où l'observation d'une action de l'autre supposée inspirée par la confiance fait écho à l'intention de donner sa confiance ; un troisième niveau, enfin, où la vérification de cette réciprocité par l'expérience conduit à un modèle mental partagé de la confiance. Reste à comprendre les conditions de passage d'un niveau à l'autre. Nous proposons de reprendre ici les hypothèses de retranscription du codage d'un niveau à un autre formulées par Proust et Pacherie (2008). Ces processus de retranscription pourraient ainsi prendre leur place dans les phases d'apprentissage de la réciprocité qui ont été mises en évidence par l'imagerie cérébrale au cours du jeu. Le premier niveau correspond au découplage entre les « autres » et le « monde extérieur » ; le second niveau permet l'interprétation intentionnelle de l'information en

provenance de l'autre ; le troisième niveau conduit à la formalisation de la différenciation entre « moi » et l'« autre » dans un rapport de confiance réciproque.

L'éveil et la construction d'une confiance réciproque ainsi décrits permettent d'expliquer les résultats expérimentaux au terme desquels, on observe, dans une majorité de cas, que les joueurs s'engagent dans un processus de confiance mutuelle. Sont-ils pour autant irrationnels, comme le suggère la solution théorique proposée par la théorie des jeux ? Pas nécessairement. L'investisseur qui confie la somme au mandataire ne connaît rien de lui, au sens où il ne dispose d'aucune information sur son comportement. Selon cette perspective, il n'aurait donc pas plus de raison de lui faire confiance que de ne pas lui faire confiance. S'il se méfie, c'est en raison des règles de ce jeu, qui permettent au mandataire de ne rien lui rétrocéder des gains obtenus, au cours de cette première séquence. Par sa décision de lui confier cette somme au début du jeu, l'investisseur fournit toutefois au mandataire une information, en forme de signal, sur son intention de lui faire confiance, de telle sorte que le mandataire dispose, lui, au moment où il prend sa décision, d'un argument pour lui faire confiance. On retrouve ici un débat bien connu en théorie de jeux entre les mérites cognitifs respectifs des informations passées qui ont été expérimentées et qui sont projetées sur le futur (*foreward*), et des informations futures qui sont logiquement rétroprojetées (*backward*). Les recherches actuelles de neuroscience théorique semblent montrer que les réseaux neuronaux du cerveau travaillent plutôt sur un mode *foreward* (Dayan et Abbott, 2001), ce qui expliquerait le comportement du mandataire. C'est la raison pour laquelle les expériences de ce jeu réalisées en imagerie cérébrale ont privilégié, jusqu'à présent, l'étude des réactions du mandataire. Il reste à comprendre la confiance initiale de l'investisseur. Faute de repères mnésiques personnels (conscients ou non conscients), cette confiance initiale manifestée par l'investisseur ne peut s'expliquer de cette manière. D'autres ressorts du fonctionnement cérébral pourraient alors être concernés. Certaines caractéristiques biologiques subpersonnelles, propres à notre espèce, et peut-être à d'autres, ont été invoquées pour en rendre compte. De telles caractéristiques se développent au cours de la formation du cerveau, dès les premiers mois jusqu'aux premières années de la vie. Elles ont donné lieu à une interprétation en termes d'empathie sociale, dans la perspective de la mentalisation (*mentalizing*) développée par la philosophie de l'esprit (Frith C. et Frith U., 2003, 2006).

Les voies ainsi ouvertes par les neurosciences ne résolvent pas les difficiles problèmes posés à la théorie des jeux par une appréhension de l'autre dans un cadre logique. Elles permettent, néanmoins d'en mieux comprendre les contours et les racines mentales et d'éclairer, de cette manière, les voies qui pourraient conduire à leur résolution.

Commentaire Les subtilités de l’interintentionnalité

Christian Schmidt montre bien à la fois que la constitution d’une base de connaissance commune semble indispensable en théorie des jeux, et qu’elle pose problème dès qu’on prend au sérieux l’intersubjectivité des interactions, puisque alors on doit se demander comment des sujets différents peuvent parvenir à converger. Cependant nous avons montré que nos intentions et nos croyances personnelles se construisent dans ces mêmes interactions qui constituent les sujets les uns relativement aux autres, comme le suppose le concept d’interintentionnalité introduit par Christian Schmidt, et comme l’a suggéré le chapitre 1. Dès lors, parvenir à une solution du problème apparaît certes comme une tâche plus complexe, en raison de la complexité de ces coconstructions, mais aussi comme une tâche réalisable, même s’il faut sans doute renoncer à garantir l’unicité et l’optimalité d’une telle solution.

Nous allons donc revenir sur les conditions de cette coconstruction, d’une part en tentant de donner à ses bases sensorimotrices – du genre des neurones miroirs – leur juste place, d’autre part en analysant avec plus de détail les différentes positions possibles qui permettent que dans une interaction, autrui soit reconnu comme tel, enfin en montrant que l’établissement de la confiance pourrait bien nous faire sortir de la dualité moi-autrui pour esquisser une référence à des tiers. Et c’est d’ailleurs bien par cette introduction du rôle des tiers que la théorie des sentiments moraux d’Adam Smith s’était montrée si profonde.

B

La limite des processus du type « neurones miroirs » est qu’ils ne nous permettent pas de saisir autrui comme tel, alors même qu’ils nous permettent de nous constituer comme à la fois similaires et différents de lui. En s’activant à la fois quand nous accomplissons un mouvement orienté vers un but et quand nous voyons autrui accomplir un mouvement similaire, les neurones miroirs nous permettent

bien de saisir le mouvement d'autrui comme intentionnel, comme signifiant pour nous, mais non pas d'y voir une manifestation d'autrui comme autre. Il nous faut pour cela commencer par être capable d'inhiber notre propre mouvement qui tendrait à imiter le mouvement d'autrui – ce que ne fait pas à lui seul le neurone miroir. Et cette inhibition n'offre encore qu'une modalité de la constitution de soi comme différent d'autrui, mais pas de celle d'autrui comme différent de soi, puisqu'il demeure alors un autre, mais un autre moi-même. Pour appréhender autrui, et par exemple pour pouvoir imaginer qu'il peut avoir des croyances et des intentions différentes des nôtres, ce que Christian Schmidt relève comme problème pour la théorie des jeux, il faut d'abord couper court à notre tendance à l'imitation. Il faut ensuite pouvoir comprendre que la situation d'action pour autrui, même si elle offre des similitudes avec les activités dont nous sommes capables, n'est pas la nôtre. Mais en même temps, il nous faut continuer à utiliser nos propres dispositions d'action et d'expression pour interpréter la situation telle qu'elle se présente pour autrui, sinon il ne serait pas pour nous une autre personne. Il faut enfin pouvoir être sensibles, sur fond de cette similarité, à ses différences. Autrui se constitue donc dans un jeu complexe qui mêle le même à l'autre.

À la suite de Leslie, on a suggéré que cela demandait une capacité de découplage, puisqu'il fallait en quelque sorte activer dans la situation en cause à la fois notre propre point de vue et celui de l'autre, tout en distinguant les deux. Dire que nous activons deux esprits différents au sein du nôtre serait aller trop loin. Il nous suffit en effet d'activer nos propres processus cognitifs et affectifs, ce qui nous donne une base de comparaison, puis de tenir compte de quelques différences propres au point de vue d'autrui. Nous n'avons pas à constituer tout un esprit séparé, ce qui impliquerait d'ailleurs d'avoir un savoir de notre propre esprit dont nous ne sommes pas assurés.

Il faut de même rester prudent par rapport à une reconstitution du dualisme entre le physique et le mental, qui pointe même en neurosciences, et qui consiste à dissocier deux aspects des interactions avec autrui, celui des interactions physiques, en particulier motrices, lié au système des neurones miroirs, et celui des rapports mentaux avec autrui, qui serait lié à la « théorie de l'esprit », et au cortex préfrontal. Si les neurosciences se bornaient à reconstituer les préjugés de la philosophie classique (tout en critiquant, comme Damasio, un cartésianisme qui n'est pas celui de Descartes), elles n'auraient pas beaucoup d'intérêt pour le philosophe d'aujourd'hui. Certes bien des travaux insistent sur une différence entre l'empathie par contagion et imitation, et celle qui implique une activité de « mentalisation », ou d'inférence des états mentaux d'autrui (Fan, 2011, Zaki, 2012, Paulus, 2013, Corradini, 2013, par exemple). Cependant le cortex cingulaire médian ou encore l'insula antérieure sont activés aussi bien dans l'évaluation cognitive que dans l'empathie affective, alors qu'on relie la partie dorsale du premier à l'évaluation et la partie droite de la seconde à l'affectif. Cognitif et affectif ont donc des recouvrements. D'autre part il faut faire des différences plus graduelles, par exemple entre a) l'imitation d'un mouvement d'autrui vers une cible, mais avec inhibition de notre mouvement ; b) le partage d'une émotion parce que je vois réagir autrui et que je réagis au même trait de la situation[7] ; c) ma supposition implicite que la réaction émotionnelle que je ressens en face de telle situation, autrui la ressent lui aussi, d) l'attribution à autrui d'une émotion qu'il ne manifeste pas mais que j'éprouverais si j'étais à sa place, et e) l'assignation à autrui d'une émotion qu'il devrait

avoir – mais que je n'éprouve pas moi-même.

L'empathie qui est déclenchée au niveau de l'imitation pourrait passer pour primaire parce qu'elle passe par des activations liées au système moteur, que la source en soient les expressions – reliées à l'activation de muscles faciaux –, les postures, ou encore les mouvements (Paulus, 2013). Cependant, si expérimenter une conduite comme expressive ne demande pas de se projeter en autrui (Daly, 2014), cela demande bien de ressentir ses dynamiques corporelles comme des tendances affectives et des tendances à l'action (Braadbaat, 2013) en situation, ce qui va plus loin qu'une simple réaction primaire. L'empathie pour la souffrance d'autrui, bien qu'elle soit des plus spontanées, est d'ailleurs modulée par des informations plus élaborées, comme l'appartenance de celui qui souffre à mon groupe ou à un autre groupe (Eres, 2013), ou comme l'évaluation de son précédent comportement comme injuste (Engen et Singer, 2012). Une empathie qui active à la fois des régions cérébrales déjà activées pour des affects de contagion émotionnelle comme l'insula antérieure et des régions liées à la « mentalisation » ou identification des attitudes mentales d'autrui comme le cortex préfrontal médian est un bon prédicteur d'un comportement prosocial envers quelqu'un qui est victime d'exclusion (Masten, 2011).

La neuroéconomie, elle aussi, se permet bien des simplifications, mais elle nous oriente vers une autre piste que celle d'une opposition dualiste entre contagion affective et attribution purement cognitive d'états mentaux, quand elle tente d'articuler un système de la récompense (avec une base limbique) et un système de mise en balance de motivations opposées (lié au préfrontal). Car ce qui nourrit ce second système, ce sont des différences, des divergences entre les appréciations des conséquences d'une même décision. Or on peut penser que l'appréhension d'autrui en tant que tel se construit sur des bases similaires : sur la reconnaissance des divergences entre les buts d'autrui et les nôtres, entre la situation vue par autrui et vue par nous. Nous percevons d'abord des différences par rapport à nos propres expressions, actions et cibles – par exemple, si autrui et moi visons la même cible, il y a divergence pour savoir qui atteindra la cible. Comme c'est en ne pouvant d'abord m'empêcher de supposer autrui similaire à moi que je vais pouvoir remarquer des différences et ainsi reconnaître son altérité, la conjonction de réactions affectives de base et d'attributions d'états mentaux qui peuvent différer du mien (Lamm, 2011) est nécessaire pour cette reconnaissance.

Ces divergences, il ne suffit pas de les remarquer, il faut pouvoir les comprendre, et cet effort nous amène à enrichir notre gamme de scénarios variés d'actions et de réactions. Pour chacun de nous, la construction d'« autrui » constitue un répertoire de différences par rapport à nos attentes de base, qui nous amène à pouvoir nous décaler d'une attente sur une autre en fonction de variations de comportement d'autrui.

Peut-on alors penser que les similitudes des organisations neuronales des humains pourraient assurer la communication avec autrui mieux que ne le ferait une hypothèse trop forte de rationalité ? Les neurosciences suggèrent l'idée d'un réseau de représentations commun (*cf.* par exemple Decety et Sommerville, 2003) qui serait le point d'appui de nos raisonnements sur autrui. Or, si autrui est construit par apprentissage de différences, cette hypothèse n'est-elle pas mise en cause ? Mais il y a

bien une organisation qui nous est commune : c'est justement celle qui nous donne la capacité d'apprendre à partir de comportements d'autrui qui diffèrent de nos attentes initiales. Ce qui nous est commun, ce n'est pas que nous partageons des contenus mentaux et leurs corrélats neuronaux. C'est que nous disposons de manières similaires de déclencher des dynamiques interactives qui ont des trajectoires différentes, l'apprentissage étant l'effet mémorisé de ces interactions motrices, affectives, cognitives, quand nous sommes en relation les uns avec les autres. Autrui, c'est finalement un accélérateur et un démultiplicateur d'apprentissages.

On peut ainsi réinterpréter légèrement différemment les expériences qui montrent que nous n'activons pas les mêmes zones cérébrales quand nous pensons interagir avec un ordinateur ou avec une personne. L'ordinateur ne présente pas le même profil d'apprentissage interactif. Soit il donne des instructions à suivre à la lettre, soit, quand ses réponses deviennent non pertinentes, c'est à nous tout seuls de trouver comment nous y prendre. Au contraire, dans une interaction avec un humain, nous pouvons nous attendre à ce que sa conduite ait un sens qui soit relatif à la nôtre. Dès lors, soit l'interaction reste dans le cadre d'un scénario déjà appris (ce qui la rend plus aisée), soit, quand nous sommes perdus, nous pouvons la plupart du temps régler le problème par une nouvelle interaction.

L

Nous sommes donc amenés à différencier plusieurs manières pour un sujet de se mettre dans une perspective ou position qui lui permette un rapport à autrui.

Christian Schmidt en a déjà distingué trois : 1) « me mettre dans la situation d'autrui » cela consiste à identifier, selon mes propres critères, la situation dans laquelle se trouve l'autre, à imaginer comment je réagirais moi-même dans une telle situation, et à identifier ainsi les intentions d'autrui ; 2) « me mettre dans la perspective d'autrui » : cela consiste à tenter d'imaginer en quoi autrui a une perspective différente de la mienne et peut se représenter différemment la situation, ce qui me permet d'imaginer en quoi ses intentions ont des chances de différer de celles que j'aurais dans la situation telle que je la vois de ma propre perspective ; 3) « tenir compte de la manière dont autrui tient compte de mes propres attitudes » : cela ne consiste plus seulement à se représenter la situation visible pour tous deux et à moduler cette représentation différemment selon les différentes perspectives. Il faut d'abord que chacun tienne compte, pour déclencher telle ou telle conduite, des intentions de l'autre avant même qu'il agisse. Comme cette prise en compte par autrui de mes intentions supposées peut modifier les propres intentions d'autrui, il faut, par ricochet, que je tienne compte de la manière dont autrui interprète mes propres intentions.

Ces distinctions sont nécessaires pour mettre en place une réflexion sur les problèmes rencontrés par la théorie des jeux. Mais il peut être utile d'ajouter quelques paliers supplémentaires dans la succession de ces différentes étapes.

Il existe par exemple une étape intermédiaire entre les deux premières positions. Nous apprenons

certains rôles relationnels (moi et mes parents, moi et mon frère ou ma sœur, etc.), qui spécifient des scénarios typiques d'interintentionnalité. Avec ces structures relationnelles, nous disposons d'un stade intermédiaire entre « me mettre dans la situation d'autrui » avec mes propres modes de réaction, et « me mettre dans la perspective d'autrui » pris dans sa différence singulière. Dans ce stade médian, la perspective d'autrui est d'emblée intégrée à la situation relationnelle, que je perçois à partir de ma propre perspective, elle-même liée à un des rôles de la situation. C'est seulement en notant des différences de la conduite d'autrui avec ces rôles que nous pourrons ensuite nous interroger sur la spécificité d'autrui.

S'ajoutent à ce répertoire de rôles relationnels quelques éléments au moins de ce qu'on appelle la « théorie de l'esprit ». Ses partisans pensent évident que nous ayons des croyances et des désirs, des perceptions et des sensations, des intentions d'action et des émotions (certains voient dans les émotions des combinaisons de désirs et de sensations). Mais ils ne vont pas jusqu'à supposer que tous les humains disposent de ces concepts-là précisément et les maîtrisent. En revanche, tous ont appris, même s'ils ne le formulent pas exactement dans ces termes, qu'il est des désirs irréalisables, qu'il vaut mieux avoir des croyances conformes plutôt que contraires aux faits observés, que les mises en action des intentions peuvent amener leur ajustement, voire leur révision, etc. S'ils n'ont pas une maîtrise des concepts de croyances, etc., ils ont perçu les différences et conflits qui peuvent se produire, par exemple, entre croyances et désirs. Ils en ont une connaissance différentielle, un peu comme celle des différences de perspectives et de rôles.

La tâche de la fausse croyance, que nous avons évoquée au chapitre 1, implique ainsi une tension entre les croyances de l'enfant témoin – pour ce qu'il en est de la place actuelle des bonbons – et celles qu'il doit arriver à assigner à l'enfant qu'on a fait sortir. Rappelons que plusieurs expériences ont montré que cette tâche ne pose pas de problèmes quand il faut simplement la résoudre sur le plan pratique. Des enfants en bas âge savent aider l'enfant une fois de retour, et lui évitent de se diriger vers la boîte initiale[8]. De même, si l'enfant témoin avait lui-même caché les bonbons, il saurait que l'enfant qui est sorti désignerait leur place initiale, puisque c'est ce comportement qu'il aurait voulu induire. La situation de l'expérience classique est plus compliquée, parce que l'enfant témoin doit identifier dans la situation ce que vise l'expérimentateur qui fait sortir l'enfant et qui (lui ou un autre expérimentateur) pose la question à l'enfant témoin. Celui-ci doit chercher à identifier la « croyance » de l'autre enfant, et donc imaginer non pas ce que doit faire cet enfant, mais ce qu'il va « croire ». La situation n'est pas immédiatement claire pour l'enfant témoin : la bonne réponse attendue concerne-t-elle la place que devrait désigner l'enfant qui est sorti si celui-ci voulait donner une réponse conforme aux faits, ou celle déduite de sa croyance reconstruite ?

Le problème que l'enfant doit réellement résoudre n'est sans doute pas, dans ces protocoles, celui d'identifier les attentes pratiques d'autrui (là encore des expériences montrent que des enfants en bas âge comprennent qu'à partir d'informations fausses on a des croyances ou des attentes fausses). C'est plutôt celui d'avoir à construire des « croyances » isolées comme telles, c'est-à-dire de trouver la schématisation ou simplification pertinente de la situation qui sera en accord avec l'attente de l'expérimentateur concernant des « croyances » et avec une focalisation sur la perspective cognitive

de l'enfant qui est sorti de la pièce, sans plus tenir compte de la structure connue comme conforme aux faits.

Une fois cela fait, l'enfant témoin doit réussir à construire une catégorie de croyances qui sont des « vraies fausses croyances » : des croyances doublement en tension avec le rôle normal des croyances en théorie de l'esprit. Ces croyances particulières sont contraires aux faits, ce qui est déjà gênant, et de plus elles sont tenues pour vraies par leur porteur, ce qui aggrave encore la situation. Certes, l'enfant a bien déjà par ailleurs l'expérience de croyances fausses, puisque les *make-beliefs*, qui ne sont pas « pour de vrai », sont indispensables à ses jeux avec d'autres enfants. Mais elles sont tenues pour telles par leur porteur, alors même qu'il les utilise dans ses jeux. Et comme dans ces cas il ne croit pas que ces croyances soient la réalité, la tension est apaisée. On voit sur cet exemple que la théorie de l'esprit se ramène en fait à la maîtrise d'une série d'opérations qui puissent permettre de passer des croyances à la réalité ou des désirs à l'action et ainsi de suite pour les autres attitudes. Elle pourrait en fait être ancrée essentiellement sur l'identification de types d'opérations impossibles ou de sources de conflits entre ces attitudes, conflits qu'il faut régler.

La tâche de la fausse croyance est donc plus complexe que les expérimentateurs ne le pensent. Elle exige qu'à partir des différences de perspectives, l'enfant arrive à constituer un contenu et un type d'attitude qui aient la propriété remarquable de pouvoir se transférer, par une transformation bien définie des perspectives, d'un sujet à un autre : ce qu'on appelle une « croyance ». Et cette tâche présente donc bien des similitudes avec ce que le théoricien des jeux suppose ses joueurs capables d'avoir en tête !

Cependant la figure d'autrui est ici réduite à une perspective différente sur un contenu qui peut être réutilisé, *mutatis muntandis*, par le sujet. Or nous pouvons aussi vouloir repérer pour elle-même la singularité de la perspective d'autrui. Dès lors, comme nous ne pouvons nous mettre réellement « dans sa peau », nous devrons nous livrer à des interprétations. Et c'est parce que plusieurs interprétations différentes seront possibles, et que nous n'aurons pas la clef de ce qui nous permettrait de décider entre elles, que nous pourrons reconnaître (évidemment sans pouvoir la définir positivement) la singularité d'autrui.

C

Le troisième niveau distingué par Christian Schmidt, « tenir compte de la manière dont autrui tient compte de nos propres attitudes », peut aussi susciter des interprétations plus faibles ou plus fortes, et donner ainsi place à de nouvelles étapes intermédiaires. Il nous amène aussi à mieux spécifier comment peut se mettre en place l'interintentionnalité.

Quand nous transportons à deux un meuble lourd et encombrant dans un escalier, celui de nous deux qui est en haut apprend vite qu'il doit tenir compte de ce que certaines orientations qu'il donne au meuble rendent à son partenaire du bas la tâche plus difficile – par exemple parce que cela le

coince contre le mur ou contre la rampe. Le partenaire d'en bas préférera d'ailleurs lui-même imposer une orientation du meuble qui rende difficile au partenaire d'en haut de le coincer.

Nous sommes là en pleine interintentionnalité. Pour ce faire, cependant, les partenaires n'ont pas besoin d'identifier toutes les croyances et désirs de leur vis-à-vis, et de se situer à ce troisième niveau. Ils s'appuient essentiellement sur les contraintes que la situation leur impose. C'est en s'appuyant sur ces contraintes qu'ils peuvent interpréter des déplacements comme risquant de bloquer les manœuvres ultérieures, les lier à des intentions mal ajustées d'autrui, et procéder à des corrections de ces intentions.

Dans les jeux de la théorie des jeux, la définition des règles et des valeurs de paiement jouent le rôle de telles contraintes. Leur connaissance au moins partielle doit donc être supposée. Dans la vie sociale courante, les capacités spécifiques à ce troisième niveau ne sont mobilisées que si l'interaction envisagée ne tombe pas sous un scénario connu, qui porte avec lui sa structure de perspectives et sa structure cognitivo-pratique, et s'il faut se lancer dans une construction en commun de ces structures.

Le problème classique d'éviter de se télescoper quand on se croise sur un trottoir, par exemple, ne devient un problème de ce genre que s'il change de nature. Supposons que je me promène dans une ville renommée pour la fréquence des vols et attaques à main armée, et que je voie de loin que je vais croiser un piéton qui vient en face et qui n'a pas l'air débonnaire. Faut-il alors changer de trottoir, ce qui montre que je crains de sa part des intentions malveillantes, mais qui pourrait éveiller chez lui de telles intentions (ne serait-ce que des injures), ou bien continuer l'air de rien, mais en restant prêt à sauter de côté ? De ces deux scénarios, d'ailleurs, je n'ai aucune expérience préalable, si bien que je ne sais pas quelles contraintes effectives ils imposent dans l'interaction. Dans ce cas, on comprend que je tente d'imaginer quelle interprétation ce passant pourrait donner de mes propres intentions. Je serais bien heureux alors de disposer de règles qui fixent des limites à mon imagination. Mais précisément, nous n'avons besoin de nous situer à ce niveau que s'il n'existe pas encore de telles règles, ou bien si plusieurs règles différentes sont toutes disponibles au même degré.

À ce stade, il serait vain de croire que je peux me mettre à la place de l'autre tel qu'il pense et agit parce qu'il s'est lui-même mis à ma place. Nous sommes justement dans un cas où les contraintes de la situation ne sont pas suffisantes pour limiter l'indétermination. Celle-ci va se multiplier à chaque stade de réflexion et rendre le problème de la sélection d'une intention à assigner à autrui quasiment indécidable.

Nous sommes donc face à un dilemme : soit les contraintes de la situation d'interaction sont suffisantes pour ne laisser place qu'à peu d'interprétations différentes des intentions des partenaires, et nous n'avons pas alors vraiment besoin de passer au niveau des croyances sur les croyances d'autrui. Soit ces contraintes laissent encore une grande liberté d'interprétation – ce qui laisse le problème largement indéterminé, et n'encourage pas à nous lancer dans ces activités cognitives de niveau élevé, puisqu'elles ont peu de chances d'aboutir.

Les solutions envisagées dans diverses expériences confirment ce dilemme. Le recours à l'ocytocine revient à imposer une contrainte qui limite les interprétations, en nous faisant sélectionner les interprétations les moins anxiogènes de la situation. Le recours à une dynamique d'apprentissage

peut aussi se comprendre comme ce qui permet de trouver dans une histoire et une évolution des contraintes supplémentaires de cohérence entre ses étapes, qui éliminent certaines interprétations. Mais cela ne résout pas les problèmes qui se posent dans des circonstances où nous n'avons pas encore eu avec autrui d'interaction assez longue pour avoir construit un scénario autour de ses attitudes.

En fait dans ce genre de situation d'incertitude, nous avons deux attitudes génériques : l'une est d'activer malgré tout des scénarios connus, l'autre est de rester ouvert aux développements de l'incertain et de réviser nos attentes dès qu'un indice nous est donné. Ainsi, dans le jeu de la confiance (*Trust Game*), l'investisseur est mis en situation soit de faire confiance, comme dans un scénario d'investissement avec un partenaire déjà connu, soit de ne pas faire confiance, comme lorsqu'il est en encore en recherche d'indices. Le mandataire, lui, a plus d'éléments : il peut supposer que la transmission d'une forte somme vaut attente de réciprocité. Mais comme il interagit avec un inconnu, il peut malgré tout refuser ce scénario. Cela dépend, en fait, non seulement de ce qu'il imagine des intentions de son partenaire, mais de ce qu'il pense être ce que l'expérimentateur (qui occupe le rôle du tiers) tient pour la bonne réponse. Or ce dernier aspect de sa situation réintroduit de l'incertitude, si bien qu'il peut refuser la réciprocité en pensant qu'elle n'est pas forcément la réponse reconnue correcte par un expérimentateur économiste !

Nous retrouvons ici un trait de l'expérience de la fausse croyance. C'est la référence à un tiers qui expliquait la difficulté de cette tâche. Dans le jeu de la confiance, c'est inversement la référence implicite à des tiers (présents ou absents) qui donne des repères aux sujets. Les sujets commencent par rester dans un scénario connu, ce qui amène l'un à investir et l'autre à renvoyer l'ascenseur. Ils demeurent donc dans leur cadre social usuel, où les tiers vont juger défavorablement celui qui trahit les attentes d'autrui. Si on leur fait répéter longtemps ce jeu dans des conditions d'anonymat, comme ils sont engagés comme tout humain dans une dynamique d'apprentissage, ils vont chercher d'autres pistes. L'expérimentateur, en assurant l'anonymat, laisse envisager une autre attente que celle, classique dans les relations sociales, de coopération au moins partielle. On voit alors davantage de mandataires qui ne jouent pas le jeu de la réciprocité, et des investisseurs qui diminuent leur envoi. Mais le fait que ces transformations prennent du temps implique qu'il a fallu se déprendre d'un premier apprentissage qui renvoyait aussi à des tiers, mais qui se déroulait dans un autre cadre social que celui que semble laisser suspecter l'expérience. Le rôle des tiers, c'est celui de ces référents qui sont implicitement présents même quand ils sont absents, parce qu'ils font partie des scénarios d'interaction sociale que nous avons appris. Nos relations avec autrui ne sont pas complètes sans cette référence aux tiers (nous reviendrons plus loin sur ce point).

Quand nos relations avec autrui dépassent le cadre de nos proches, nous pouvons toujours nous référer implicitement à des tiers dont la présence est implicite, mais qui servent de référence commune aux deux protagonistes alors qu'ils ne participent pas à l'interaction – ils sont censés en quelque sorte observer les partenaires dans leur dos. Mais nous pouvons aussi être plus sensibles aux singularités d'autrui, ce qui exige dans un premier temps de ne plus pouvoir nous contenter de scénarios préconçus et donc de réintroduire de l'incertitude dans nos attentes.

1. Cette question des deux Adam Smith souvent présentée comme « The Adam Smith problem » a été longuement débattue par les historiens de la pensée économique, notamment dans la tradition germanique.

2. C'est dans cette seconde perspective que la tradition utilitariste, héritée de Bentham et de Sidgwick, a développé un mode de calcul économique qui intègre l'altruisme et préfigure, d'une certaine manière, l'économie du bien-être.

3. Concernant la question de l'influence des idées de Schütz sur la pensée d'Hayek, on consultera l'ouvrage de T. Aimar (2009), *The Economic of Ignorance and Coordination*, Cheltenham Glos (UK), Edward Elgar Publishing.

4. Cette question a été abondamment discutée dans le cadre de la théorie des jeux, sans qu'émerge une véritable réponse de la part des théoriciens. Tout au plus sont-ils tombés d'accord sur le fait qu'une entente préalable entre les joueurs, avant le commencement du jeu, n'est pas ici de nature à résoudre, ni même à éliminer, la question (Harsanyi et Selten [Postscript], 1988 ; Aumann, 1990).

5. Sur cette question on consultera Edmund Husserl, *Phénoménologie de l'attention*, introduction et traduction de Natalie Depraz, Paris, Vrin, 2009, et « De l'inter-attention à l'attention relationnelle. Le croisement de l'attention et de l'intersubjectivité à la lumière de l'attention conjointe », *Revue canadienne de philosophie continentale*, 2010, 14 (1), p. 104-118.

6. Merleau-Ponty conçoit cet « autre moi-même » au niveau de la perception d'autrui, comme le signe de ce qu'il nomme la croyance en un « être indivis » (Merleau-Ponty, 1946). Nous avons montré que cette croyance était indétachable d'une perception « active », c'est-à-dire, d'une certaine manière, intentionnelle (Schmidt, 2010).

7. Mais l'imitation affective n'est pas automatique, elle est affaiblie si on doit accomplir une tâche cognitive prenante (Rameson, 2011).

8. *Cf.* Baillargeon R., Scott R. M. et He Z. (2010), « False belief understanding in infants », *Trends in Cognitive Sciences*, 14, p. 110-118. Kovács A. M., Téglás E. et Endress A. D. (2010), « The social sense : Susceptibility to others' beliefs in human infants and adults », *Science*, 330 (6012), p. 1830-1834. Onishi K. H., Baillargeon R. (2005), « Do 15-month-old infants understand false beliefs ? », *Science*, 308 (5719), p. 255-258. Poulin-Dubois D., Sodian B., Metz U., Tilden J. et Schoeppner B. (2007), « Out of sight is not out of mind : Developmental changes in infants' understanding of visual perception during the second year », *Journal of Cognition and Development*, 8, p. 401-421.

CHAPITRE 3

Le cognitif interactionnel

Nous avons développé dans le chapitre 2 l'analyse des processus par lesquels un sujet devient peu à peu sensible à la fois aux similarités qu'autrui présente avec lui et à son altérité. Nous avions montré au chapitre 1 que le sujet est intersubjectif et qu'il se construit dans l'interaction avec son environnement et avec autrui. Pour être capable de se construire ainsi, il doit pouvoir *en lui-même* en arriver à disposer de registres différents qui soient eux-mêmes en interaction, ne serait-ce que pour pouvoir distinguer dans l'interaction ce qui relève des autres et ce qui relève du soi. On peut alors supposer que ces capacités à utiliser différents registres, qui sont utilisées dans son rapport à autrui, ont aussi une portée plus générale. Inversement, le développement des rapports à autrui doit avoir des effets en retour sur ces dispositions plus générales et leur apporter des potentialités supplémentaires. Dans ce chapitre, nous allons donc montrer en quoi la perspective interactionniste permet aussi de comprendre des capacités plus générales de cognition et de décision de l'individu, et indiquer quelques-uns de ces effets en retour.

Le problème soulevé par la perspective de construction interactive et intersubjective du sujet proposée au chapitre 1 est le suivant : dans ces interactions, des différences de registre se constituent (par exemple : le mouvement que je perçois, le mouvement que je commande). Mais ces registres doivent s'intégrer les uns aux autres, et cela de manière immanente, sans qu'on doive supposer le *deus ex machina* d'une unité organisatrice surplombante et déjà donnée. Cette intégration doit elle-même résulter ou émerger d'interactions. Un critère classique d'intégration est la cohérence d'un ensemble d'opérations qui s'enchaînent les unes aux autres dans le temps. Or la psychologie expérimentale montre des ruptures de cohérence dans plusieurs domaines.

Nous partirons de quatre types de phénomènes qui mettent en question cette cohérence. Ils ont donné lieu tous les quatre à des interprétations dualistes, qui opposent deux fonctionnements ou deux systèmes, et ils posent tous les quatre des problèmes à la théorie classique de la décision. Il s'agit des phénomènes de discount temporel ; des différences de rapidité de traitement cognitif et décisionnel

qui ont suscité l'hypothèse de deux systèmes de raisonnement et de décision (l'un dédié à des réponses rapides qui ne nécessitent pas d'inférences compliquées, l'autre à des raisonnements plus élaborés mais plus lents) ; de l'opposition entre émotions et cognition rationnelle ; de l'opposition entre catégories bien délimitées et catégories vagues. Nous tenterons de montrer que toutes ces dualités peuvent recouvrir des jeux d'interactions qui mêlent interactions à courte portée et interactions à longue portée, et que des variations dans l'importance donnée aux unes par rapport aux autres pourraient rendre compte de ces dualités apparentes. Il s'agit là non pas d'une théorie admise ni même pleinement constituée, mais plutôt d'une interprétation philosophique, dont le but est de montrer combien la variété des phénomènes auxquels une perspective interactionniste peut s'appliquer est grande, à partir de ces quatre types de phénomènes qui posent de nouveaux défis à une théorie rationnelle de la décision.

Q

Le « discount » temporel

Entre un gain qui va nous arriver très bientôt mais qui est moins important, et un gain un peu plus élevé qui nous est promis dans un futur plus lointain, nous préférons en général le premier. Cela nous amène à des renversements de préférence, puisque quand la date d'échéance du gain le plus lointain arrivera, nous préférerions avoir dans le passé opté pour ce gain alors futur. En revanche, entre deux gains lointains, même s'ils sont séparés par le même intervalle temporel que le gain proche et le premier gain lointain, nous préférons le plus important, sans trop nous soucier de l'éloignement temporel supérieur du plus lointain. La dévaluation des gains futurs diminue donc quand on passe d'une comparaison entre un futur proche et un futur lointain à une comparaison entre deux futurs tous les deux lointains. Ce « discount temporel » va au-delà de ce que peut expliquer notre méfiance envers des promesses lointaines, justifiée par l'incertitude du futur.

Ainslie, qui avait lancé dans les années 1970 des études sur ce thème, avait envisagé des situations où nous sommes attirés par un plaisir dans un futur proche alors même que nous savons que jouir de ce plaisir sera la cause d'un regret ultérieur, parce que les conséquences de cette jouissance nous auront privés d'un plaisir bien supérieur. Dans ces situations, le simple fait de passer, d'un temps t0 où les deux plaisirs sont lointains, d'abord au temps t1 proche du premier plaisir puis au temps t2 plus lointain où aurait pu intervenir le second plaisir suffit pour renverser deux fois nos préférences.

Supposons que je sache quelques jours à l'avance que la veille d'un concours important pour ma carrière, une soirée attractive aura lieu dans ma résidence. En t0, deux jours avant la soirée, je préfère ne pas participer à la soirée pour être suffisamment en forme pour le concours. Soit tds le début de la soirée. Peu avant tds, mes amis me pressent de venir à la soirée. Je décide de participer au début de la soirée, mais de la quitter assez tôt en tqs dans la nuit pour être encore en forme le lendemain, au matin

du concours en tmc. En t0, tqs et tmc étant tous les deux lointains, la distance temporelle entre tqs et tmc n'implique pas une dévaluation importante du plaisir d'être en forme pour le concours par rapport au plaisir que j'aurais à rester malgré tout à la soirée. Quand t0 sera proche de tqs, la distance temporelle entre tmc et tqs aura bien plus d'effet de dévaluation, si bien que l'anticipation du plaisir d'être en forme pour le concours pourra ne plus surpasser celle du plaisir de rester à la soirée. En t0 = tds, je peux donc penser que je vais bien quitter la soirée en tqs après avoir eu ma ration de plaisir, alors qu'en tqs, j'aurai tendance à rester à la soirée. En tmc, je changerai encore, rétrospectivement, de préférence. Mes décisions sont donc incohérentes dans le temps.

On a formalisé cela en supposant que le « discount », la dévaluation de l'appréciation des événements futurs est très lente entre deux événements du futur lointain, et qu'inversement l'évaluation croît très vite quand on se rapproche d'un événement positif dont on est proche (ce qui implique, réciproquement, qu'il suffit d'en être un peu moins proche, de se situer en en temps t0 qui ne soit pas en proximité immédiate avant l'événement pour que cette valeur décroisse rapidement). La différence de gains peut alors dominer la différence temporelle pour le futur lointain, mais l'accroissement de la proximité temporelle peut agir en sens inverse. Pour décrire ce peu de décroissance quand on se situe dans le futur lointain, conjugué à cette croissance très rapide quand on se rapproche très près, il faut utiliser une courbe hyperbolique. Or s'il est cohérent de conserver un facteur de discount constant, et de suivre une courbe exponentielle, qui donne moins d'importance aux biens du futur lointain que ne le fait la courbe hyperbolique, il n'est pas cohérent de changer le taux de dévaluation simplement parce qu'un même intervalle entre le moment du choix et celui de la réception du bien est déplacé vers le futur.

Les deux systèmes

Loewenstein a proposé de rapporter ces renversements de préférences, liés simplement au déplacement dans le temps, à la théorie des deux systèmes cognitifs (Lowenstein, 2000). Elle est fondée sur bien d'autres expériences, qui indiquent que, pour résoudre des problèmes, nous pouvons soit nous fier à notre « intuition », qui est souvent seulement basée sur quelques indices ou sur des analogies avec des problèmes pour lesquels nous disposons de routines menant à leur solution, ou bien explorer les éléments du problème de manière méthodique pour arriver à une solution que nous puissions garantir. Cette opposition a été radicalisée par la théorie des deux systèmes : le premier est celui des heuristiques frugales (selon l'expression de Gigerenzer), qui est fait pour donner rapidement aux problèmes que l'on rencontre des solutions qui ne sont pas forcément optimales mais qui ont fonctionné dans le passé de manière suffisamment satisfaisante pour que les individus qui les aient utilisées aient réussi à perpétuer notre espèce – ces solutions sont donc supposées fiables dans une perspective évolutionnaire. Le second système implique plus de contrôle et de réflexion, mais demande une élaboration plus lente et plus coûteuse en efforts cognitifs. Il est déclenché seulement si les solutions proposées par le premier système semblent ne pas devoir convenir. Préférer ce qui est

proche dans le temps serait propre au premier système, qui prend le pas sur le second quand la décision devient urgente, alors que lorsqu'on a le temps de réfléchir, on peut suivre une autre préférence et accorder plus d'importance à un futur lointain.

Le rôle des émotions dans nos décisions

Loewenstein a aussi proposé de relier le discount temporel et la théorie des deux systèmes au rôle des émotions dans nos évaluations et décisions. Il s'appuie sur des données d'imagerie cérébrale dont l'interprétation pourrait être que quand nous devons réfléchir sur des problèmes intertemporels, le cortex préfrontal est activé, et que pour des réactions bien plus rapides, ce sont nos régions limbiques – liées par ailleurs aux émotions – qui sont activées.

Certes les émotions dites « occurrentes » déclenchent des réactions rapides (la fuite devant un danger). Mais nous avons aussi des émotions en anticipant des joies ou des risques éloignés dans le futur ou encore en nous ressouvenant de moments signifiants dans nos expériences passées – pour ce que Kahneman (2011) appelle le *remembering self*, qu'il distingue de l'*experiencing self*, qui réagit en continu à ce qui lui arrive. De plus des petits détails dans la vie d'une journée peuvent induire une humeur positive ou maussade, sans même que nous sachions quels détails ont eus cette influence. Nos affects présentent donc des temporalités diverses.

De l'insensible aux transitions critiques : le vague

Nos décisions dépendent de la manière dont nous évaluons différentes options, et dont nous pouvons différencier les valeurs de ces options, leur donner un rang selon différents critères. Or nos capacités de discrimination sont limitées, si bien que nos distinctions entre une valeur de rang supérieur et une autre de rang inférieur sont sujettes aux problèmes que l'on dit « sorites », par analogie avec les tas de sable. Enlever un grain d'un tas de sable ne le fait pas changer de catégorie, c'est toujours un tas de sable. En enlever un deuxième non plus. D'une étape à l'autre, entre n grain enlevés et n+1 grains enlevés, il n'y a pas de différence. Pourtant, en répétant ces étapes, nous en arrivons à ne plus avoir de tas de sable. La catégorie « tas de sable » est donc vague. Des problèmes analogues se posent pour la distinction entre grand et petit, entre cher et pas cher, etc. Or nos décisions doivent trancher dans ce vague. Mais sur quelles justifications ?

Le problème du vague infeste d'ailleurs nos trois précédentes questions. Quand notre discount temporel passe-t-il du régime rapide au régime ralenti ? Y a-t-il transition ou bien rupture entre nos routines « intuitives » et nos raisonnements élaborés ? Nos émotions du moment dépendent-elles des émotions de nos anticipations et de nos souvenirs ?

Commençons par explorer ces interactions entre émotions.

L

On ne peut considérer les émotions seulement comme des « facteurs viscéraux » (l'expression est de Loewenstein) qui nous feraient agir impulsivement dans le présent. En fait nos émotions ne sont pas seulement des émotions « réactives », qui réagissent à un stimulus, ce sont aussi des émotions anticipatrices et comparatives, qui teintent nos anticipations du futur, et cela de plusieurs manières, et ce sont aussi des façons de récapituler nos expériences, là aussi de plusieurs manières.

D'une part nous anticipons les plaisirs ou les peines que peuvent nous causer des événements futurs, et ces évaluations se font par une comparaison en quelque sorte longitudinale – le long du temps – avec notre état présent. Si nous sommes déjà dans un état joyeux, en anticiper un autre de même nature n'est pas aussi attractif que si nous sommes dans un état neutre ou pénible. D'autre part nous éprouvons des émotions par une comparaison en quelque sorte latérale, celle de notre sort futur avec d'autres sorts possibles, et surtout avec celui d'autres personnes. Si notre sort futur est enviable mais que celui des autres soit encore meilleur, cela va gâcher notre plaisir anticipé. On a ainsi montré que notre joie de gagner ne dépend pas seulement de la valeur absolue du gain, mais aussi de la comparaison de ce gain avec le gain des autres.

De plus nous éprouvons des émotions en fonction de l'incertitude ou de la certitude de ces anticipations. Si un résultat souhaité est incertain, cela nous cause de l'angoisse, et inversement la forte probabilité d'un résultat que nous voulons éviter nous cause de la crainte. On voit sur cet exemple qu'il est parfois difficile de dissocier les probabilités et la coloration affective de l'évaluation des états futurs, évaluation liée à ce que des théoriciens de la décision comme Jeffrey ont appelé la désirabilité.

Or ce genre d'émotions liées à l'incertitude (être angoissé, craintif ou rassuré) a des incidences sur les émotions de comparaison. Supposons que le choix de telle action soit pour nous lié à l'anticipation d'un état futur bien plus attrayant, mais dont l'obtention est bien plus risquée, en comparaison d'un autre état qui peut être obtenu plus sûrement par une autre action et qu'une autre personne est supposée avoir choisi. Envisager que nous atteindrions cet état alors que l'autre personne n'obtiendrait que la satisfaction inférieure de l'état plus assuré va accroître notre jouissance. Inversement, si nous sommes plus assurés d'obtenir une satisfaction, même si elle reste bien inférieure à celle de cet état attrayant que nous pouvions espérer mais qui est plus risqué, nous allons nous rengorger de notre prudence quand nous envisagerons qu'un autre ait pris des risques et n'ait rien obtenu. Cela nous montre aussi que nous sommes sensibles à des comparaisons entre le résultat de l'action que nous avons choisie, et celui que nous, ou un autre, aurions obtenu si nous, ou cet autre, avions choisi l'autre action (Livet, 2010).

Une expérience menée par Coricelli (2007) l'a bien montré. Il a présenté à ses sujets des « roues de la fortune » qui figuraient d'une part les proportions entre les chances de gagner plus et les chances de gagner moins ou de ne pas gagner, et d'autre part les gains obtenus par le tirage de ces loteries. Si nous considérons une seule roue de la fortune, et qu'on nous indique par une flèche le résultat du tirage de cette loterie, nous serons déçus si la flèche se situe dans la région du gain moindre ou de

l'absence de gain. On peut aussi nous présenter deux roues de la fortune. Nous devons choisir l'une d'elles, et nous pouvons alors voir sur chacune des roues de la fortune le résultat du tirage. Quand le tirage de la loterie que nous n'avons pas choisi donne un gain bien supérieur à celui de notre loterie, nous n'éprouvons plus de la déception, mais bien du regret d'avoir choisi cette loterie. L'imagerie cérébrale confirme l'ancrage neuronal de ces distinctions entre regret et déception en montrant des activations différemment localisées pour ces deux émotions.

La théorie de la décision[1] considérait pourtant que, si nous pouvions envisager déception et regret avant notre choix (par exemple avec le critère du minimax du regret), ces émotions n'étaient pas justifiées *après* que nous avons envisagé toutes les issues des deux loteries et fait notre choix selon le principe du minimax, voire de minimax regret. Mais il n'est pas irrationnel, pour plus de réalisme, d'inclure dans notre fonction d'utilité ces émotions comparatives, y compris celles qui interviendront après coup, puisqu'elles résultent de comparaisons et d'anticipations qui n'ont rien d'irrationnel en elles-mêmes, et cela à condition que nous puissions toujours assurer une certaine cohérence entre nos décisions.

Par ailleurs, nos émotions récapitulent nos expériences passées. D'une part, les souvenirs qui activent en nous des émotions (et qui sont évoqués par des émotions similaires) portent sur le résultat, le bilan ou le point saillant des activités engagées dans ces souvenirs, et non pas sur les sentiments éprouvés en continu lors de ces expériences (Kahneman, 2011). Notre affect retient donc des sortes de leçons émotionnelles, qui nous servent de repères pour la suite. D'autre part, sans que nous en ayons forcément le souvenir, les dynamiques affectives à long terme de notre vie passée, les transitions ressenties au travers de périodes longues, entre espoir et déception ou entre angoisse et soulagement, modèlent nos attitudes envers notre environnement. Elles nous rendent enclins à la prudence, voire à l'anxiété, ou inversement à l'exploration et à la quête d'opportunités. Ces différences de dynamiques passées permettent de comprendre pourquoi les décisions de différentes personnes face à une même situation peuvent fortement diverger.

Une fois que l'on considère les émotions comme des combinaisons complexes en relation avec des récapitulations, des anticipations et des comparaisons, combinaisons qui sont le résultat d'une longue évolution et dont ont été extraits des scénarios simplifiés d'interaction, ou des types de situations interactives, on trouve quelque peu sommaires certaines interprétations de résultats expérimentaux.

Ainsi on nous apprend que des sujets qui ont dû d'abord accomplir publiquement certaines tâches et ont donc été soumis à une certaine pression voient leur cortisol s'accroître et leur capacité de mémorisation de l'association entre le nom d'une autre personne et l'image de son visage diminuer. Cela nous confirme simplement qu'une tâche où notre image de nous-même peut être mise en péril aux yeux des autres nous oblige à une certaine concentration qui nous fait manquer des informations données par ailleurs dans une interaction. Être soumis à des comparaisons avec les autres, étant facteur d'émotion, nous amène à privilégier cet aspect de la situation au détriment d'autres aspects.

On a aussi noté que des sujets qu'on a rendus plus disposés à la colère prennent moins en compte

les risques d'une action, et qu'inversement des sujets plus dépressifs ou plus anxieux les prennent davantage en considération. Cela montre, là encore, que nous traitons de manière interactive les risques ou incertitudes de telle anticipation. Nous prenons en compte des facteurs de comparaison interrelationnelle, qui nous conduisent à la colère contre ceux dont les interactions nous ont contrariés. Nous sommes aussi sensibles aux anticipations des perspectives offertes par les autres futurs possibles, qui sont toutes colorées négativement dans la dépression et l'anxiété. Nous ne traitons pas ces différents aspects de manière indépendante et isolée.

Rustichini a proposé un modèle de théorie des jeux (plus ou moins convaincant) pour rendre compte de ces liens entre émotions et comparaisons (Rustichini, 1999). Dans un jeu répété, il est utile de disposer de signaux sur les fréquences avec lesquelles tel joueur joue tel type de coup, et donc de disposer d'informations sur les séquences passées. Elles nous permettent de caler nos anticipations. Nous pouvons alors comparer à ces anticipations les résultats effectifs d'un coup donné, et faire des comparaisons de second degré, comparant ces différents rapports entre les diverses anticipations et les divers résultats des appariements de nos coups avec des coups différents d'autrui. C'est au prix de ces comparaisons de second degré, longitudinales et latérales, que nous pouvons espérer réduire les risques de regret. Les processus qui suscitent l'émotion de regret se révèlent donc très liés à ceux qui peuvent le minimiser. En fait ces mêmes mécanismes qui suscitent le regret à court terme peuvent permettre de le minimiser dans un apprentissage à long terme. L'émotion qui n'apparaît pas rationnelle sur une décision d'un instant peut se révéler sur une longue suite de décisions et d'apprentissages une incitation fiable, et qui nous permet de réagir en situation d'incertitude. Sans ces émotions qui ancrent les résultats de ces comparaisons dans notre esprit, nous apprendrions encore moins vite.

Quelques pistes pour une formalisation plus sensible aux interactions

Il se trouve que prendre en compte cette plus grande sensibilité aux interactions peut répondre à un autre problème rencontré par la théorie de la décision, celui relevé par de nombreuses expériences de Kahnemann et Tversky ainsi que leur école. Ils ont montré que nous accordions plus d'importance aux états de faible probabilité qu'aux états de probabilité moyenne, et que nous donnions bien plus d'importance à un état certain qu'aux états de forte probabilité. Ce traitement inhomogène des probabilités fait qu'elles cessent d'être additives, au sens où $f(a+b)$ devient différent de $f(a) + f(b)$, où le poids du composé ne se réduit pas à l'addition des poids de ses éléments.

Les interactions à longue portée

Or on peut relier cette non-additivité à la prise en compte d'interactions. On a remarqué en physique que les liens entre thermodynamique et statistique établis par Boltzmann et qui ont permis à

Gibbs de définir une notion d'entropie (quand on transforme la forme d'énergie, l'entropie croît) n'étaient valides que sous des conditions restrictives : il faut que l'importance des interactions à longue portée soit négligeable, et que l'on puisse ne faire fond que sur les interactions à courte portée. Or dans bien des phénomènes physiques, ce n'est pas le cas, en particulier pour les phénomènes fractals.

On sait qu'une manière de représenter formellement une indépendance entre deux éléments, c'est de penser que quand on les met ensemble, cela ne donne qu'une addition. Pour représenter le fait que cette mise ensemble produit des effets d'interaction, on introduit souvent une multiplication ou une exponentiation. Pour tenir compte des interactions à longue portée, on pouvait donc compléter une perspective additive par un produit. L'entropie de l'ensemble de deux sous-systèmes ne serait pas donnée par la somme des entropies de chacun d'entre eux, mais par l'addition à cette somme du *produit* de ces deux entropies. Si on pondère ce produit par un paramètre1 – q, quand la valeur de q est 1, 1 – q = 0, et l'élément qui représente l'interaction à longue portée est nul. Quand q est inférieur à 1, cet élément prend de l'importance[2].

On peut ainsi rendre compte de distributions statistiques qui n'obéissent pas à une loi normale – une distribution dont les extrêmes sont très peu fréquents – mais à une loi de Lévy – où les extrêmes sont mieux représentés. On sait que bon nombre de traders utilisaient une loi normale pour leurs estimations, alors que le monde de la finance est plus interactif et exige au moins une loi de Lévy (Mandelbrot et Hudson, 2004 ; Mandelbrot, 2005).

On a suggéré que notre perception conférait une importance de ce type à des interactions à longue portée, c'est-à-dire aussi, on le voit, qu'elle donnait un poids supplémentaire à des événements rares. Prenons un exemple dans le domaine de la perception. Supposons que sur un rocher gréseux fait d'un amas compacté de grains de sable de différentes couleurs, plusieurs grains de sable de même couleur se succèdent en ligne droite. C'est là à la fois une disposition rare et une interaction perceptive à longue distance – entre le premier et le dernier grain de la ligne, et pas simplement entre deux grains immédiatement voisins. Or notre perception va immédiatement se focaliser sur cette ligne. Notre perception est donc sensible à des dispositions rares quand elles présentent des régularités qui elles-mêmes assurent des interactions à longue distance – par exemple parce qu'elles nous permettent de distinguer les contours de différents objets. Cela nous permet, au lieu de devoir identifier les connexions entre chacun des grains, de pouvoir les supposer tous reliés par la ligne qui connecte le grain initial et le grain terminal, faisant ainsi une économie cognitive considérable.

Nous pouvons à partir de cet exemple définir une notion plus générale de saillance. Considérons les points de rebroussement, de changement de direction, vertex ou point de bifurcations, par exemple dans une courbe en V. Le point du bas du V sépare deux zones (les deux barres du V) où dans chacune les interactions de courte portée peuvent être résumées sans problème par une même interaction à distance, la barre en question. Par ailleurs, nos deux barres sont elles-mêmes reliées par des interactions à longue portée. Mais on peut toutes les ramener au lien entre d'une part le point de la base du V et d'autre part la distance virtuelle entre le haut des deux barres. *Via* ce point d'articulation, nos deux barres, nos deux premières interactions de longue portée, sont elles-mêmes reliées par des

interactions à longue portée. Ce point joue trois rôles : il est pris dans des interactions à courte portée, dans les deux interactions à longue portée des barres du V, et c'est la limite qui permet de résumer les interactions à longue portée entre ces deux barres. Nous dirons donc que, si un élément peut assurer une articulation entre plusieurs régimes d'interaction, à courte et à longue portée, il présente une saillance. Notre perception privilégie le saillant sur l'homogène, c'est-à-dire privilégie non pas le local, mais le local qui joue un rôle de pivot dans une articulation globale entre interactions, et *a fortiori* dans des articulations globales d'ordre supérieur.

Nous ne sommes d'ailleurs pas sensibles à la seule rareté de ces articulations multiéchelles entre interactions : il faut qu'elle se combine avec ce qui présente un intérêt pour nos motivations et pour nos buts. La saillance se combine avec la prégnance. L'interaction avec nos buts se rajoute à l'interaction assurée par ces patterns saillants. Cela implique, par exemple, que nous ne traitions pas les probabilités de manière homogène, et que nous soyons plus sensibles à des probabilités extrêmes (petites probabilités et forte proximité avec la certitude) qu'aux probabilités voisines de l'équiprobabilité. Si l'on prend en compte nos interactions avec notre environnement, dans lequel les irrégularités sont légion, et où quelques régularités multiéchelles sont significatives pour nos buts et donc prégnantes (parmi elles, il faut signaler les « singularités », ces inflexions brusques des courbes qui sont les signatures régulières de certaines formes), un tel comportement a des avantages, et on peut comprendre que l'évolution l'ait sélectionné.

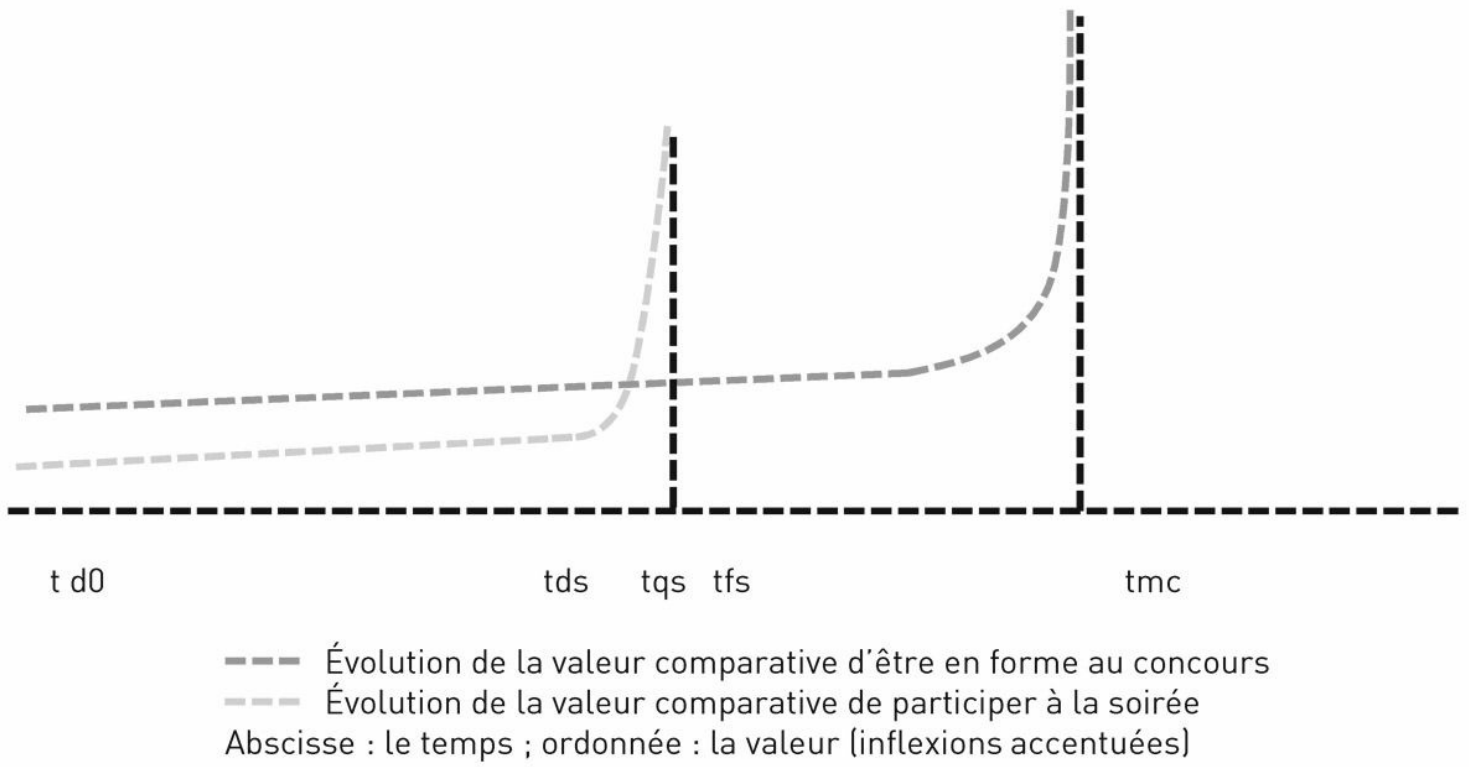
Montrons maintenant en quoi les phénomènes intrigants que nous avons évoqués peuvent aussi manifester cette importance pour notre cognition des interactions à plus longue portée, et plus précisément des interactions multiéchelles.

Le « discount » temporel revisité

À première vue, rendre compte du privilège donné aux gains futurs proches semble exiger d'aller en sens inverse de l'importance donnée aux interactions à longue portée, si l'on identifie futurs proches et interactions à courte portée, et futurs lointains et interactions à longue portée. En fait, il faut raisonner non pas sur les estimations de chaque gain pris isolément, mais sur les différentes comparaisons entre les gains. Ces comparaisons sont en cognition l'analogue de ce que sont en physique les interactions. Si nous étions seulement sensibles à des comparaisons locales, de proche en proche, nous n'observerions pas d'effet de renversement de nos préférences. Le futur lointain pourrait être dévalorisé mais notre taux de discount resterait le même, quelle que soit la position, lointaine ou proche dans le temps, d'un même intervalle temporel entre deux éléments de la comparaison. Mais nous sommes sensibles aussi à des comparaisons à plus longue distance temporelle, entre le futur proche et le futur lointain. Takahashi a donc pu utiliser la combinaison d'effets additifs et d'effets multiplicatifs indiquée plus haut pour rendre compte des effets de discount temporel.

Si ce discount est constant, il peut être représenté par une décroissance *exponentielle* de valeur : le même taux de discount se réapplique au résultat précédent de son application. Certes la courbe

exponentielle, qui monte modérément quand t est loin de l'événement gratifiant, se redresse quand t en est très proche, mais c'est simplement lié à l'effet boule de neige qui tient à ce que le même taux, la même fonction puissance, s'applique sur le résultat d'une précédente application. On peut donc dire que dans le cas d'une courbe exponentielle, il suffit de cumuler les effets des pas de courte portée, qui obéissent tous à la même progression, pour déterminer la valeur en un temps t lointain. Il n'y est pas nécessaire de faire des comparaisons à longue portée. Si nous voulons représenter géométriquement les effets de ces perceptions qui nécessitent dans la réalité de recourir à des interactions de longue portée, il nous faudra recourir à des courbes qui combinent une décroissance plus lente et une ascension plus rapide qu'une courbe exponentielle classique – c'est le cas des courbes hyperboliques, ou encore des courbes qui de manière un peu *ad hoc* combinent deux taux d'inflexion différents, ou de courbes exponentielles avec ajout d'un facteur q inférieur à 1 (Takahashi). Elles vont, par rapport à cette référence classique, donner lieu à plus de variations et donc à des inflexions qui, tout en étant moins tranchées que celle du point de rebroussement du V que nous avons pris pour exemple, vont cependant amener à supposer l'intervention d'interactions à longue portée.



Rappelons que, si notre discount temporel est représenté par des courbes hyperboliques (ou par les deux autres sortes de courbes mentionnées plus haut), nous pouvons changer d'ordre de préférences quand les deux courbes se croisent.

Nous avons suivi ici le mode de représentation d'Ainslie (2012) : quand le temps présent n'est

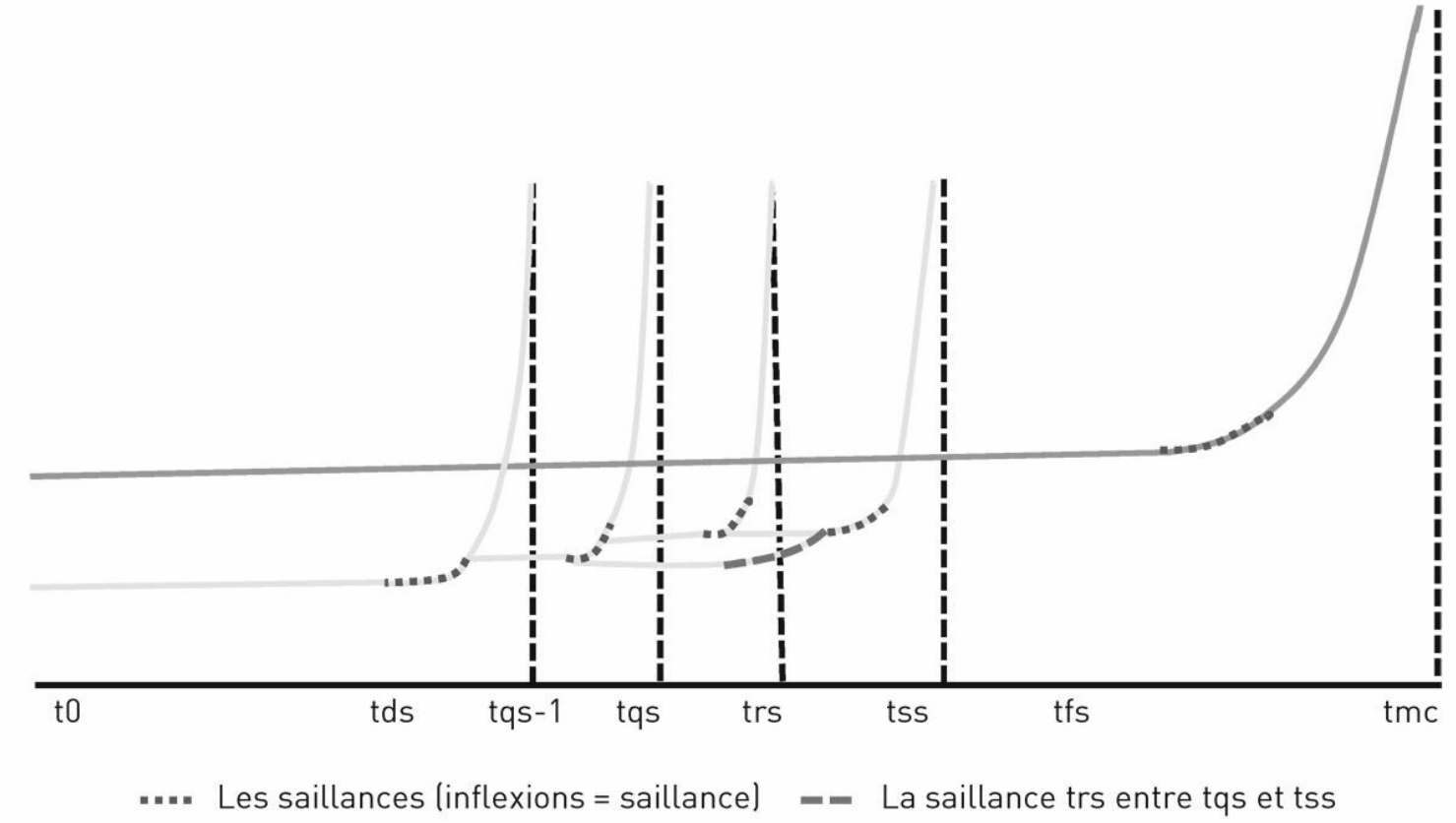
plus le temps de départ td0 et devient tds (le début de la soirée), la valeur anticipée de la soirée croît fortement, alors que celle d'être en forme au concours en tmc est toujours stagnante[3].

Mais les courbes ne sont ici que des modes de représentation. Les facteurs réels sont, selon nous, les relations entre interactions à courte portée et à longue portée, qui donnent lieu aux éléments saillants sur lesquels se repère notre perception. Or sur un quelconque des trois types de courbes mentionnés, nous ne verrons pratiquement pas de différences entre les points voisins de la partie de la courbe où la croissance est peu sensible, nous verrons quelques différences entre les points où la croissance est rapide, et nous verrons une inflexion importante dans la zone de transition entre ces deux parties. Cette zone d'inflexion correspond à la définition de la saillance que nous avons donnée en prenant l'exemple du V.

Revenons à l'exemple de la soirée, parce qu'il présente, on va le voir, des phénomènes d'intrication, dont nous avons vu qu'ils sont essentiels dans notre apprentissage émotionnel.

Vu du moment t0, le moment saillant est d'abord tds, le début de la soirée. Il sépare deux zones où l'on ne change pas de régime (donc des zones d'interactions à courte portée) : la zone homogène t0-tds et l'autre zone homogène tds-tfs-tmc. Maintenant, si de t0 on arrive à distinguer des périodes dans la soirée, on aura un second moment saillant, tqs, le moment raisonnable pour la quitter, qui sépare la soirée en deux zones, tds-tqs, et tqs – tss-tfs, tfs terminant la soirée et ouvrant la période qui va jusqu'à tmc, tss étant la suite de la soirée.

Or une fois venu à la soirée avec la ferme intention de la quitter en tqs, le moment trs, celui où nous serons sur le point d'abandonner l'idée de la quitter et où nous risquons de nous laisser aller à y rester, se révèle avoir un statut particulier. Quand nous envisageons ce moment à partir du début de la soirée (tds), il appartient au segment tqs – tfs-tmc, puisqu'il est au-delà de la saillance tqs. Il n'est donc pas saillant et il appartient à la même zone que tmc. Vu de t0, nous pensons donc pouvoir rester attentif dans cette zone à l'importance d'être en forme pour l'examen, et ne pas rester à la soirée.



Notre raison pour faire ensuite le choix contraire est la suivante : dans la zone de trs, il va y avoir d'autres saillances, donc d'autres inflexions séparant deux zones relativement homogènes. Il ne s'agit pas seulement des interactions entre le segment du voisinage temporel immédiat de tqs-1 et celui du voisinage immédiat de tqs – ou encore entre tqs et trs. Il s'agit d'inflexions et donc de saillances d'une portée plus longue : ainsi trs peut appartenir à la saillance liée à l'inflexion qui sépare tqs et tss. Dès lors, trs devient un moment saillant (en tirets plus foncés sur la figure).

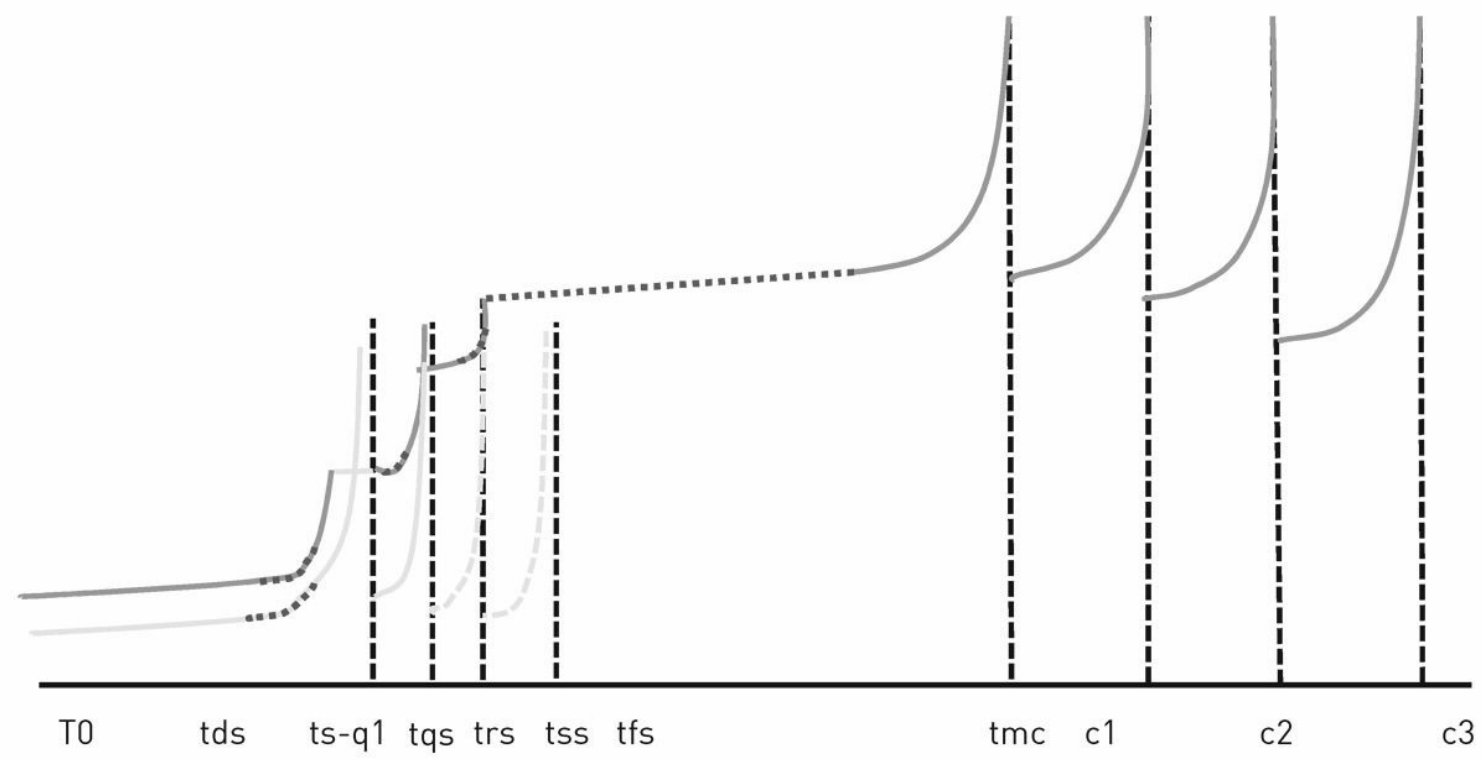
Or cette interaction à plus longue portée peut occulter l'interaction à encore plus longue portée qui relie le segment tds-tqs et le segment tqs – tmc : il en est comme de ces immeubles du voisinage non immédiat qui deviennent notre horizon en occultant des horizons plus lointains. En effet, alors que tqs semblait seul à présenter une saillance liée à une interaction lointaine, trs présente aussi une interaction de longue portée entre tqs et tss, voire entre tds et tfs, où tqs ne figure que comme un élément non saillant. Ainsi, quand une des deux interactions à longue portée fait de l'un des deux éléments trs ou tqs un élément saillant, l'autre est non saillant, et réciproquement. La saillance de tqs est donc brouillée. Il se peut donc que nous préférions rester à la soirée, parce que tqs n'a plus le privilège d'être la seule saillance multiéchelle, et que trs en est une autre.

Cela pourrait tenir à une limite de notre cognition et de nos affects. Elle privilégie les saillances, c'est-à-dire le local qui joue un rôle de pivot dans une ou plusieurs articulations globales entre interactions. Mais il n'y a pas qu'une manière de définir ces articulations[4], il y en a plusieurs, et elles peuvent se brouiller l'une l'autre.

Une fois que la saillance liée à l'interaction la plus lointaine est brouillée, et comme de nouvelles

saillances apparaissent à tout moment dans la suite de la soirée, nous sommes piégés par un effet fractal, par la découverte au sein des interactions supposées à courtes portées d'interactions à longue portée, qui vont occulter d'autres interactions à plus longue portée qui seraient plus importantes.

Comment échapper à ce piège ? Ainslie (2012) rappelle que si nous pouvons comparer non pas l'instant attractif de la soirée à un instant futur (le concours) mais une *série* de futures satisfactions reliées entre elles (le concours, différentes étapes de la carrière qu'il me permettra, notées c1, c2 ; c3 dans la figure suivante) et la *série* des instants plus proches reliés entre eux – ceux de la soirée – nous sommes moins facilement piégés[5]. Mais le problème est que, pris dans le piège, nous n'avons plus de moyen de dépasser l'horizon intermédiaire que nous offre la soirée, parce que c'est tout de même un horizon fourni par une interaction à longue portée. Et Ainslie ne nous dit pas clairement comment surmonter cette difficulté.



Quand les anticipations de la série des satisfactions à long terme sont appelées par les repères qu'on s'est donnés dans la soirée

Or, si le mal tient à ce qu'une interaction qui est à longue portée relativement à des interactions de plus courte portée nous brouille des interactions encore plus lointaines, le remède est évidemment d'utiliser le processus inverse, qui fait du local non pas un brouillage ou un écran, mais un relais du lointain. C'est comme si on mettait au sommet de l'immeuble qui nous masquait l'horizon un écran qui en donne une image. Plus abstraitement, c'est trouver dans les éléments de ces interactions plus

locales des repères qui puissent présenter des ancrages pour les interactions plus lointaines. Nous avions bien un repère : le moment tqs indique bien une articulation entre la soirée prise en bloc et le bloc plus important que constitue le concours, dont il ne faut pas oublier qu'il est lié à toute une série de satisfactions futures (*cf.* Ainslie). Mais n'avoir qu'un seul repère le rend sensible au brouillage, on l'a vu. Il faut donc disposer plusieurs repères successifs de ce type. Les atteindre les uns après les autres renforcera à chaque nouveau repère l'anxiété associée au risque de voir devenir inaccessible la série d'étapes importantes plus lointaines – anxiété qui renforcera la valeur accordée au concours – et mettra en place autant de contre-saillances supplémentaires.

Nous allons pouvoir utiliser ce type d'analyse, en termes d'effets des passages d'interactions locales à des interactions à longue portée, à propos des phénomènes de différence de rapidité et de modalité de raisonnement et de décision, qui ont donné lieu à la théorie de la dualité entre système frugal et système raisonneur, et nous pourrons aussi l'appliquer aux problèmes de décision à partir d'évaluations vagues.

R

Dualité et vague

Rustichini (2008) a proposé, en s'inspirant de Luce, une explication des effets de cette dualité dans des termes qui ont initialement été proposés pour rendre compte des effets de sorite et de vague. Il suppose simplement que dans l'ordre de nos préférences, on trouve des indifférences entre deux situations. Ces indifférences recouvrent en fait de petites différences, mais nous ne leur sommes sensibles que lorsqu'elles s'accumulent. Supposons que nous ne disposions que de sucre en poudre pour le verser dans notre café. Nous ne savons pas distinguer clairement une quantité x1 de sucre en poudre qui comporte 200 grains et une autre quantité x2 légèrement supérieure qui comporte 250 grains, ni cette quantité x2 comportant 250 grains d'une autre quantité x3 comportant 300 grains encore supérieure. Nous dirons qu'entre x1 et x2 nous sommes indifférents, de même qu'entre x2 et x3. Pourtant nous pouvons être capables de distinguer la quantité de sucre x1 de 200 grains de la quantité x3 de 300 grains et nous dirons donc que x1 est inférieur pour nous à x3.

Ce faisant, nous violons la transitivité de la relation d'indifférence, qui, de ce que x1 est indifférent à x2 et x2 à x3, nous obligerait à conclure que x1 est indifférent à x3. Notre ordre de préférences est un semi-ordre, disait Luce, parce que sa notion d'indifférence n'est pas transitive. Luce avait aussi montré qu'une évaluation qui procède selon ce semi-ordre donne des effets qui peuvent tenir à deux facteurs : 1) l'existence de petites différences, en dessous d'un seuil de discrimination consciente, et 2) un pouvoir de discrimination limité, soit très exactement les conditions des évaluations vagues.

Or nous pouvons interpréter ce changement dans la prise en compte des petites différences,

passant tout d'un coup de l'indifférence à la distinction, dans les termes de la problématique des interactions courtes ou longues. Les petites différences sont des interactions – des comparaisons – à courte portée. Les différences auxquelles nous sommes sensibles sont, du coup, des interactions – des comparaisons – à plus longue portée. Nous tenons compte des secondes quand l'accumulation des premières en vient à rendre non négligeables les interactions à longue portée.

L'intérêt de cette approche est de nous éviter d'avoir à considérer les « deux systèmes » comme strictement séparés. Notons d'ailleurs qu'en règle générale, pour toute étude d'imagerie cérébrale qui propose des combinaisons de zones d'activités, qui soient spécifiques pour chacun des systèmes, d'autres chercheurs montrent qu'en fait il y a des interactions et coactivations de certaines de ces zones. Comme le dit Kahneman (2011), les deux « systèmes » interfèrent. Car si le supposé système frugal se fixe immédiatement sur les cas saillants, nous avons montré que leur privilège tient à ce qu'au sein d'un cumul d'interactions à courte portée homogènes entre elles, ils assurent des interactions à longue portée, plus rares mais plus économiques, résumant ainsi un cumul de petites interactions en une seule interaction plus longue. Nos intuitions ne reposent donc pas simplement sur des interactions à courte portée. Inversement, nos démarches raisonneuses combinent aussi les interactions à longue et à courte portée. Le problème de nos raisonnements est d'assurer la transitivité, comme l'avait bien vu Descartes (*Règles pour la direction de l'esprit*). Il ne peut se satisfaire d'interactions à longue portée dont il n'aurait pas montré la cohérence avec les interactions à courte portée. Ainsi pour le problème du discount temporel, il amène à préférer la courbe exponentielle, parce qu'elle aligne sur un même régime les interactions à courte portée et celles à longue portée. Le système raisonneur ne renverse pas la préférence pour les interactions à longue portée et les formes saillantes de nos intuitions perceptives, mais il doit s'assurer que les interactions locales sont bien en cohérence avec l'interaction globale.

Vague et attitudes épistémiques

Revenons maintenant sur le problème du « vague », dont on voit, grâce à cette grille d'analyse, qu'il est au cœur de tous les autres.

Il faut noter que notre perception sait se tirer de la difficulté propre aux « sorites », du genre du problème du tas de sable. Nous arrivons bien à distinguer entre quelques grains de sable étalés sur une surface et un tas de sable. Notre catégorisation perceptive, malgré son faible pouvoir de discrimination, a donc trouvé des moyens d'échapper au paradoxe[6]. Mais il reste à donner au raisonnement les moyens d'y échapper. Le problème philosophique tient à ce que recourir à des repères stricts (tel nombre de grains, par exemple) ne résoudrait rien, puisqu'on pourrait encore discuter de la position de ce repère, et que l'identification de la bonne position resterait vague si nous sommes réellement dans un domaine vague.

Notre problème est de décider si nous rangeons un objet perçu dans une catégorie A ou dans une catégorie B. Supposons que la catégorie A qui nous intéresse soit celle de chamois, parce que nous

sommes chasseur de chamois – chasseur de photos, rassurez-vous. Il nous faut avoir des indices pour la distinguer d'une autre catégorie B. Il faut d'abord bien identifier les catégories à opposer, ce qui peut poser des problèmes si elles ne sont pas du même niveau de granularité : quand nous voulons déterminer si telle forme est celle d'un chamois, il serait peu pertinent de l'opposer à la catégorie « forêt de mélèzes ». Vient un moment où nous avons trouvé la bonne granularité : nous avons le choix, mettons, entre « chamois » et « souche fourchue ».

Nous sommes alors arrivés, par rapport à ce problème de distinction, au maximum de pertinence de chaque catégorie, si bien que les indices supplémentaires que nous recueillerons pourront nous faire basculer dans l'une ou dans l'autre. Si par exemple nous avons un indice du genre d'un mouvement, nous allons pencher pour le chamois.

Considérons alors comment cet indice a fait varier nos attitudes épistémiques, en l'occurrence nos rapports à un savoir. Quand nous en étions seulement au maximum de pertinence de l'opposition entre les deux catégories, nous savions qu'il s'agit de la bonne opposition, mais nous ne savions pas comment choisir : nous ne savions pas si c'est un chamois, et nous ne savions pas si c'est une souche. Nous avions donc deux non-savoirs symétriques. Nous avions aussi un savoir de second ordre, qui ne nous apporte pas grand-chose : nous savons… que nous ne savons ni l'un ni l'autre.

Quand nous disposons de l'indice du mouvement, nous ne pouvons pas encore garantir que c'est un chamois (le mouvement pourrait être un mouvement illusoire, lié à une saccade de notre œil plutôt qu'à un mouvement d'un animal), mais nos attitudes envers la catégorisation chamois et la catégorisation « souche » ne sont plus symétriques.

Notre attitude concernant la souche n'a pas changé : il reste vrai que nous savons que nous ne savons toujours pas que c'est une souche.

En revanche notre manque d'assurance concernant l'identification du chamois n'est pas le savoir d'un non-savoir concernant le chamois. C'est simplement l'absence d'une garantie, donc d'un savoir de second ordre, qui nous démontrerait que nous savons bien que c'est un chamois. Cette absence d'un savoir démonstratif est parfaitement compatible avec le savoir que c'est un chamois, même si elle ne garantit pas ce savoir. Paradoxalement, nous sommes passés d'un savoir de second ordre (mais qui portait sur une ignorance) à un non-savoir de second ordre. Mais ce non-savoir est maintenant compatible avec un savoir de premier degré concernant le chamois, ce qui rompt la symétrie entre nos deux hypothèses, souche et chamois.

Notre savoir reste donc vague (puisqu'il n'est pas totalement garanti) mais ce vague ne nous empêche pas d'avoir acquis des repères. Nous pouvons évidemment appliquer cette analyse au problème du sorite. Accumuler les grains de sable, c'est accumuler les indices, et cela finit par nous faire basculer du côté de la catégorie « tas », alors que nous étions partis d'une catégorie opposée, sans pour autant que nous ayons le savoir garanti – ce savoir qui assure notre savoir – que nous donnerait par exemple le savoir du nombre précis de grains qui serait nécessaire et suffisant pour en arriver à un tas. Nous n'avons donc pas besoin pour nous repérer dans le vague de disposer de repères qui échappent complètement au vague.

On notera alors, pour en revenir au rapport entre les interactions à longue et à courte portée, que

les petites différences qui sont d'abord négligées dans un problème du sorite le sont parce qu'elles se présentent comme des interactions à courte portée. L'accumulation de telles interactions peut donner lieu à la perception d'interactions à longue portée, à condition que les interactions courtes puissent s'organiser entre elles de manière cohérente. L'accumulation des indices en faveur de la catégorie chamois a ce genre d'effet. Ce qu'il ne faut pas oublier, c'est que l'accumulation d'indices ne peut pas faire oublier que ce ne sont que des indices, et que chacun d'eux restait insuffisant, ce qui grève toujours notre savoir (lié aux interactions à longue portée) d'une certaine fragilité.

Savoirs réflexifs

Bien évidemment, un savoir de second degré comme celui que nous venons d'analyser implique une forme de réflexion, puisqu'il porte sur un savoir ou un non-savoir de premier degré. L'important est de voir que nous sommes capables de comparaisons de second degré (dans notre exemple, entre un savoir de non-savoir, et un non-savoir de savoir) alors même que nous ne disposons toujours pas d'un pouvoir de discrimination suffisant pour nous garantir la différence de premier degré entre le chamois et la souche. D'ailleurs, quand devient-il raisonnable de nous demander si nous disposons ou non d'un savoir de second degré ? Quand il est sensé de nous interroger sur la fiabilité de nos informations, par exemple de nous demander si ce mouvement perçu n'est pas dû à un mouvement de notre tête ou à une saccade visuelle, donc quand nous passons à des questions non plus sur l'objet de notre perception, mais sur les garanties que celle-ci apporte.

Rustichini supposait qu'à un niveau inconscient nous puissions être sensibles à des différences que nous n'arrivons pas à faire à un niveau clairement conscient. Nous avons inversé sa position, puisque notre analyse en termes d'attitudes épistémiques est restée interne au domaine de la conscience. Nous avons distingué deux étapes. Dans la première, la symétrie entre les deux non-savoirs (chamois ou souche) nous indique (consciemment) que notre distinction entre les deux catégories est arrivée à son niveau de pertinence – sans que nous puissions encore déterminer quelle est la bonne catégorie. Dans une deuxième étape, nous avons suffisamment d'indices pour pouvoir passer au second degré de la connaissance. La faiblesse de notre pouvoir de discrimination n'est plus ce qui met immédiatement en danger la catégorisation de l'objet, mais ce qui induit des questions de second degré sur la fiabilité de notre accès à cette catégorisation. Mais ce déficit de second niveau n'exclut pas le savoir de premier niveau de cette catégorie, savoir conforme aux faits mais non garanti.

Renforçons les liens entre ce problème du vague, celui de l'articulation entre les deux « systèmes », et celui de notre sensibilité aux interactions à longue portée.

Raisonner, avons-nous dit, c'est vérifier que des interactions plus locales assurent bien la différence constatée au niveau d'interactions plus globales. C'est donc revenir de l'étape qui nous amène à classer notre objet dans la catégorie chamois à l'examen de la cohérence avec cette catégorisation de tous les indices relevés. Or cette enquête qui porte sur notre propre accès

épistémique peut cependant revenir sur une recherche d'indices, ceux qui peuvent donner à cet accès des garanties supérieures : nous bougerons la tête ou ferons d'autres saccades pour voir si le mouvement n'est pas illusoire, ce qui nous renseigne d'abord sur la fiabilité de notre perception, et seulement par ricochet sur le fait qu'il s'agit bien d'un chamois. Or bien que notre question soit de second degré, les indices qui permettront d'y répondre sont de premier degré puisque ce sont des perceptions, mais ce sont des perceptions plus affinées.

On peut tenter d'interpréter ce double mouvement comme un va-et-vient entre nos intuitions, qui décident sans garantie, et la recherche de garanties de nos raisonnements. Mais ce qui nous fait progresser, c'est de coupler les deux mouvements, celui dans le sens la montée épistémique et celui d'un nouveau recours à la base, à nos perceptions. Cela montre qu'il n'est pas possible de séparer un système « frugal » et un système plus réfléchi. Les manques du premier système peuvent donner des informations au second. Ainsi l'absence d'indices en faveur de la catégorie B va permettre de lancer une recherche de second degré sur la fiabilité des indices perceptifs en faveur de la catégorie A. Et, en sens inverse, le second processus a besoin, pour relancer la réflexion, de repartir de nouveaux indices que lui donne le premier mieux encadré. Les deux prétendus systèmes sont tellement intriqués qu'ils ne peuvent être isolés. On devrait plutôt étudier les entrelacements des dynamiques cognitives, chacune présentant des étapes différentes qui s'intercalent entre les étapes d'une autre dynamique. Notre système cognitif repose sur une véritable interactivité.

Ces considérations peuvent permettre d'esquisser une réponse à un problème bien connu : pourquoi allons-nous rarement plus loin que trois niveaux dans nos efforts réflexifs[7] ? Nous avons vu que nous ne pouvons jamais atteindre une absolue précision dans nos informations de base. Ce que nous permet une première enquête, c'est de faire une différence entre deux types de non-savoirs, un non-savoir de premier ordre, un non-savoir de second ordre.

Pour aller plus loin, il faut pouvoir relancer l'enquête sur les éléments qui permettent de définir une nouvelle distinction pertinente. Si nous considérons cela comme une étape supplémentaire dans la même lignée d'enquête, cela nous amène au troisième ordre (un non-savoir de troisième ordre qui reste cependant compatible avec un savoir de savoir). Le problème est alors que ces nouveaux éléments introduisent des oppositions de catégories supplémentaires : dans notre exemple du chamois et de la souche, « en mouvement/pas en mouvement », ou encore « deux cornes symétriques/une fourche asymétrique ». Mais ces catégorisations posent de nouveaux problèmes : la symétrie est-elle suffisante, le mouvement est-il suffisamment continu, etc. ? Or ces problèmes sont transversaux par rapport à la question initiale, et peuvent nous faire bifurquer dans notre enquête. Plus nous tentons de monter d'un niveau de savoir, de trouver des garanties supplémentaires, plus ce problème s'accentue. Cette ambiguïté entre montée d'un niveau et bifurcation dans l'enquête est sans doute à l'origine du fait que nous avons beaucoup de difficultés, sans le recours à un symbolisme logique, qui reste lui-même difficile à interpréter, à trouver des exemples de réflexivité d'ordre quatre.

U

La perspective des interactions nous a donc permis de revenir de notre analyse des étapes de l'appréhension d'autrui aux capacités plus générales d'une cognition interactionnelle.

Un retour sur le jeu des émotions va nous permettre de revenir de ces interactions, prises en un sens large, aux interactions sociales. Nos décisions, avons-nous rappelé, sont modulées par nos évaluations affectives, qui impliquent des comparaisons entre le présent et le futur anticipé, entre les différents futurs possibles anticipés, entre les récapitulations venues du passé et ces anticipations. Ces comparaisons par anticipation sont très sensibles à la possibilité qu'autrui parie sur d'autres futurs possibles. Ces récapitulations, ajouterons-nous, s'appuient sur les résultats d'expériences interactives avec autrui qui nous ont marqués.

Nous pouvons encore ajouter que nous pouvons distinguer davantage de niveaux de réflexivité si les savoirs ou non-savoirs qu'il s'agit de différencier sont ceux de plusieurs acteurs sur les informations dont ils pensent que les autres disposent ou non – pensez aux situations de vaudeville, que nous décryptons sans problème. Autrement dit, de même que les mouvements des autres nous fournissent un répertoire de mouvements typiques qui nous permet de mieux distinguer nos propres mouvements, de même les projets, anticipations et comparaisons supposées d'autrui – lisibles quand nous combinons la situation et les expressions émotionnelles d'autrui – nous servent de repères pour bâtir des réflexivités plus complexes que nos petits trois étages réflexifs.

Ce système intersubjectif de repérage exige de supposer des homogénéités entre des perspectives différentes, qui nous permettent des transpositions moyennant les transformations nécessaires. Comme ces différences indiquent une forme de distance, cela veut dire que nous établissons des interactions à distance, basées sur l'hypothèse que les autres sont sensibles aux mêmes saillances auxquelles nous serions sensibles dans leur position. Cela revient à bâtir des interactions à distance sur des capacités de sensibilité à des interactions à distance. Cela nous évite d'avoir à faire des hypothèses sur la manière dont est constitué l'esprit d'autrui : il nous suffit de repérer des saillances, et suivre les regards et mouvements des autres qui nous indiquent d'autres saillances, et de bâtir sur les différences entre ces interactions à longue portée d'autres interactions de même type.

On notera pourtant, en revenant au problème du discount temporel, que notre activation émotionnelle semble moindre, au niveau neurologique, quand on nous demande de faire des évaluations intertemporelles pour d'autres personnes, que lorsque nous sommes directement impliqués (Albrecht *et al.*, 2011) – notre investissement affectif est sans doute moindre – le discount temporel persiste, mais il est moins fort quand nous jugeons pour les autres.

Autrement dit, nous avons beau nous être construits subjectivement en utilisant des convergences entre nous et autrui sur des saillances, nous sommes davantage capables de donner la priorité aux interactions à plus longue portée temporelle quand il s'agit des autres. La raison semble pouvoir en être que, quand nous jugeons pour autrui, notre ancrage temporel personnel n'est pas en cause, et donc que nous pouvons désintriquer les interactions à longue distance des interactions à courte portée, voire échapper aux brouillages entre différentes interactions à longue distance que nous avons analysées

dans l'exemple de la soirée. En revanche, nous ne pouvons pas le faire quand nous sommes dépendants du déplacement dans le temps de notre propre ancrage temporel. Ce qui semble bien corroborer l'hypothèse d'une cognition sociale qui nous permet de nous appuyer sur les autres comme des repères plus stables que notre propre dynamique affective du moment.

Devons-nous alors, pour penser les interactions sociales au sens fort du terme, celles qui exigent une sensibilité non pas seulement à autrui, mais à des groupes, à des collectifs, à des normes sociales, ajouter à cet édifice des capacités supplémentaires ?

Certains ont avancé qu'aux prétendus deux systèmes, le système rapide et frugal, et le système plus lent et plus calculateur, il faudrait en ajouter un troisième, le système social, qui s'appuierait sur la rationalité des autres personnes.

Une première remarque s'impose : la cognition sociale exige de pouvoir supporter et traiter le vague. Dans un groupe social un peu important, nous ne connaissons pas forcément individuellement tous ceux qui font partie du groupe, et pourtant nous nous référons bien au groupe. Même si, dans un petit groupe, nous connaissons tous ses membres, nous ne savons pas ce que fait chacun d'eux à tout moment, et nous pouvons simplement supposer qu'ils agissent le plus souvent de manière à préserver la survie du groupe. Ces savoirs-là sont vagues, mais ce vague est nécessaire pour que le groupe puisse se maintenir et vivre. Exiger un contrôle de tous les instants paralyserait l'action du groupe. La relation entre ce vague de groupe[8] et la sensibilité à des interactions à longue portée est évidente.

Or, puisque notre cognition est interactionnelle, puisqu'elle sait traiter le vague et les interactions, elle dispose déjà des dispositions nécessaires pour assurer la formation de groupes et leur maintien. Encore faut-il qu'elle mette au point le double mouvement que nous avons esquissé plus haut, à propos des savoirs réflexifs, entre acceptation d'un vague irréductible dans les garanties de plus haut niveau de notre savoir, et recherche d'indices plus fins à des niveaux inférieurs. Quand il s'agit de la vie en groupe, ce double mouvement consiste à tenter des contrôles partiels, des mises à l'épreuve, mais sans pour autant bloquer les interactions et la confiance entre les membres du groupe, ce qui suppose d'accepter le vague de nos assurances dans la stabilité du groupe.

Une dynamique d'apprentissage est nécessaire pour arriver à connaître le bon dosage dans nos interactions sociales entre catégorisation vague et mise à l'épreuve d'autrui (et par autrui).

Peu à peu les interactions réussies, mais aussi les corrections que motivent des émotions du genre de la surprise, de la déception ou du regret, ainsi que les récapitulations émotionnelles qui les marquent, nous permettent de sélectionner par apprentissage les combinaisons les plus fiables d'acceptation du vague et de mise à l'épreuve. Or le repérage dans le vague implique la sensibilité aux interactions à longue portée. Et la mise à l'épreuve se fait par vérification de la cohérence entre ces interactions à longue portée et des indices locaux.

La dualité entre les deux systèmes supposés pourrait bien alors n'être qu'un effet de ces dynamiques d'apprentissage. Tant que nous n'avons pas trouvé les bons ajustements entre ces deux mouvements, nous devons passer par des circuits longs de contrôle, qui assurent des révisions successives d'un dispositif d'interaction par un autre. Une fois que nous avons trouvé des repères qui

permettent de court-circuiter ces circuits longs, nous devenons capables d'intégrations bien plus rapides, et qui, socialement, induisent dans les coordinations de groupe une confiance routinière. Mais les interactions à longue portée qui réalisent des courts-circuits entre ces repères présentent des biais : ces repères ne sont fiables que dans des contextes suffisamment similaires à ceux dans lesquels nous avons appris la viabilité de ces raccourcis.

Notre cognition a pu ainsi intégrer nos interactions avec notre environnement, avec autrui et avec nos groupes d'appartenance, mais la contrepartie est qu'elle reste partiellement dépendante de leurs contextes, ne pouvant s'en détacher qu'au prix de la relance de nouveaux apprentissages.

Des défis pour la théorie économique

La prise en compte de toutes ces interactions, celle des multiples capacités comparatives des émotions, celle des dynamiques d'évolution des évaluations au cours du temps, des modalités de repérage dans le vague, celle de l'existence de plusieurs couches sédimentées d'apprentissage qui nourrissent des réactions rapides ou plus lentes, ce sont là autant de défis pour la théorie économique, qu'il s'agisse de la théorie de la décision ou de la théorie des jeux.

La théorie de la décision définit une décision comme le choix d'une action, action qui est réduite à une fonction, partant des différents états du monde possibles et arrivant dans des conséquences. Les conséquences sont alors évaluées selon les préférences de l'agent, donnant lieu à des utilités, et l'agent décideur tient de plus compte des différentes probabilités des états possibles (ainsi traverser un gué est utile et aisé si on a affaire à un ruisseau et qu'il fait beau, mais un orage qui transforme le ruisseau en torrent peut donner pour conséquence à cette action une noyade).

Or une telle conception de l'action est assez éloignée de la dynamique effective de nos interactions. Nos préférences sont le résultat de nos apprentissages interactifs, cognitifs et émotionnels. Nos actions sont extraites d'interactions avec notre environnement et avec autrui, qui nous permettent de nous construire un répertoire de mouvements orientés vers des cibles ou d'expressions gestuelles. Nous évoluons dans un monde qui présente des incertitudes. Ces incertitudes ne sont pas aisées à probabiliser, parce que si nous pouvons assurément assigner des probabilités subjectives aux différents scénarios, nous ne savons pas comment en assigner à l'imprévu qui les bouleverse, auquel pourtant nous pouvons réagir émotionnellement. C'est d'ailleurs parce que nous vivons en situation d'incertitude que nous avons développé l'apprentissage émotionnel qui est le nôtre et ses comparaisons.

En situation d'incertitude, nos réactions prennent leurs repères sur nos interactions avec autrui. Ainsi ce sont des comparaisons avec les résultats d'autrui qui guident nos évaluations. Inversement, pour pouvoir être sensible à l'altérité d'autrui, il faut pouvoir être sensible à des différences de ce qu'il fait avec ce que nous attendions ; et c'est une capacité de sensibilité à l'imprévu, dans un monde en incertitude, qui nous permet d'être sensible à cette altérité.

Avant d'avoir des préférences entre des actions, nous avons des motivations pour telle ou telle

action. Une motivation est d'ordinaire supposée être le facteur qui pousse à la levée de l'inhibition d'une activation et lui permet de déclencher un comportement. Mais en fait elle oriente cette dynamique en cohérence avec la possibilité pour autrui de reconnaître l'orientation de notre comportement. Et ce sont ces motivations d'actions reconnaissables par autrui qui pourront être tenues pour nos intentions. De fait, c'est d'abord à autrui que nous assignons des intentions : nous n'avons pas besoin dans la plupart de nos actions de définir nos propres intentions, puisque cela ferait pour nous double emploi avec l'intention implicite dans notre action engagée. Mon intention est donc ce qui, si j'occupais le point de vue d'autrui, se traduirait par un comportement révélateur de telle intention pour un observateur semblable à moi.

Quand le théoricien de la décision parle de préférences révélées, il présuppose que mes choix puissent révéler mes orientations. En cela il ne fait que prolonger cette construction intersubjective des intentions. Mais si ces intentions sont sélectionnées parce qu'elles sont reconnaissables par autrui, et que cette sélection fasse partie du processus qui me construit comme sujet intersubjectif, alors mes préférences ne sont pas simplement révélées, elles sont bien réelles, mais comme des réalités interactives, et non plus comme des données internes au seul sujet.

Cet aller-retour entre nous et autrui nous permet d'entretenir en quelque sorte l'altérité en nous. Nous devenons capables d'entretenir en parallèle plusieurs attentes, correspondant à plusieurs scénarios. Certains d'entre eux ne correspondent pas à la réalité actuelle, si bien qu'ils sont dits « contrefactuels ». Cette capacité d'envisager des contrefactuels peut encore se complexifier, par un effet en retour de l'interaction avec autrui. La constitution de notre système sensorimoteur nous rend déjà capables d'inhiber le déclenchement de comportements sans pour autant inhiber leur préparation, et cela nous permet de « simuler » autrui. Nous devenons capables d'entretenir en parallèle sinon différentes préparations, du moins différentes mises en attentes de conduites possibles, quand nous faisons l'expérience des différences entre telle ou telle des autres personnes que nous côtoyons, et que nous apprenons à attendre des conduites différentes des uns et des autres. Quand les partenaires concernés ne sont pas présents, ces mises en attente vont évoquer des contrefactuels. La théorie des jeux, qui formalise nos réactions appropriées à tous les coups possibles d'autrui dans un jeu, se veut une généralisation de cette capacité[9].

Formaliser la dynamique de ces interactions, voire des interactions entre interactions à courte et à longue portée, cela semble bien être le défi proposé à la théorie économique, qui deviendrait alors véritablement une science sociale. Christian Schmidt va montrer que l'interaction entre neurosciences et économie nous engage bien dans cette voie.

Commentaire Pour une nouvelle approche économique de l'intertemporalité

Une multitude d'expériences, répétées depuis plus de vingt ans, a montré que les choix des individus se révélaient, d'une part, sensibles aux différentes échéances temporelles anticipées, et variaient, d'autre part, en fonction de leur position par rapport à ces échéances. Dans l'exemple décrit au début de ce chapitre, le choix, initialement arrêté quelques jours avant, de renoncer à une soirée sympathique en vue d'un concours professionnellement décisif, s'est trouvé progressivement remis en question au cours du déroulement du temps, au point de se trouver inversé au moment de cette échéance. Ces phénomènes sont appréhendés par l'analyse économique au moyen de deux concepts principaux. Il s'agit, d'abord, du taux d'actualisation, qui permet d'évaluer les conséquences des actions à des échéances différentes, plus ou moins lointaines, en les traduisant en termes de conséquences immédiates. Quant à la transformation de l'appréciation de leurs conséquences attendues, en rapport avec le rapprochement (ou l'éloignement) de leurs échéances, elle se trouve saisie à travers les modifications, voire les renversements, qu'elle entraîne dans les préférences des agents décideurs.

Ces deux indicateurs associent spontanément les phénomènes qui ont été rapportés à des anomalies économiques. Sans accident, en effet, un taux d'actualisation « normal » devrait pouvoir se traduire par un coefficient constant, dont la trajectoire prendrait la forme d'une courbe exponentielle. Or ce n'est pas ce que l'on observe le plus souvent dans les expériences, où le taux d'actualisation revêt plutôt la trajectoire d'une courbe hyperbolique ou quasi hyperbolique (Laibson, 1997). Quant aux préférences des sujets, dont on a montré au chapitre 2, tout à la fois l'importance et les limitations de leur traitement en économie, elles sont supposées traduire la cohérence des goûts des agents économiques et posséder, à ce titre, un certain nombre de propriétés logiques constantes (réflexivité, transitivité, antisymétrie, etc.). Là encore, ce n'est pas ce que l'on induit le plus souvent des comportements observés. Dès lors, ou bien les préférences qui inspirent les choix des agents présentent des incohérences, ou bien elles sont instables dans la durée. Dans les deux cas, ces

anomalies repérées sont imputables à différentes manifestations d'intertemporalité au cours des décisions. Comme le souligne Pierre Livet à la fin de ce chapitre, elles interpellent sérieusement la théorie économique de la décision.

U

Pour comprendre les véritables raisons de cette mise en difficulté de la théorie économique classique de la décision, il est nécessaire de remonter en amont, au niveau du découpage des domaines étudiés et de la construction des différentes catégories identifiées par la discipline économique. On notera que le concept de taux d'actualisation est principalement utilisé en macroéconomie, en relation avec les agrégats, représentant l'épargne, la consommation, l'investissement et, plus largement, les différentes affectations du capital. Les préférences concernent, en revanche, au premier chef, la microéconomie et renvoient, en premier lieu, à un agent individuel plus ou moins représentatif. Plus précisément, c'est à propos du capital que le problème de la temporalité a d'abord été posé en économie. Historiquement, la tradition autrichienne a même développé une théorie originale du capital, presque entièrement construite sur la base d'une analyse de la dimension temporelle dans laquelle s'inscrivent les actions économiques[10]. Aujourd'hui encore, les débats suscités par la définition et le choix d'un taux d'actualisation, même s'ils prennent leur source dans l'observation des comportements individuels, intéressent principalement la sphère de la macroéconomie financière et des travaux sur les cycles de vie du capital (Angeletos *et al.*, 2001). C'est, au contraire, à partir de l'analyse de la consommation, et dans la perspective du choix du consommateur individuel, qu'a été élaborée la théorie économique de la décision. C'est pourquoi, celle-ci peut s'entendre comme une manière de généralisation d'une théorie initialement destinée à comprendre et à éclairer le choix du consommateur.

Cette référence à la consommation permet d'expliquer, pour une part au moins, la formulation initialement statique de cette théorie. Son ancrage dans le temps présent est fourni par les préférences du consommateur, au moment de son choix. S'agissant d'un bien destiné à être consommé, on suppose que le délai temporel qui sépare le choix de sa conséquence, c'est-à-dire de la consommation du bien choisi, peut-être négligé. Lorsque je choisis entre des carottes et des tomates pour mon prochain repas, le délai qui sépare ce choix du déroulement de ce repas ne devrait pas normalement modifier les préférences qui ont orienté mon choix. La tentation est, dès lors, de traiter le choix comme une opération unique qui se déroule en un seul coup et où les préférences du consommateur sont supposées inchangées entre le moment où il effectue son choix et celui où il satisfait sa consommation. L'introduction d'une temporalité se trouve ainsi, en quelque sorte, greffée sur un schéma d'origine complètement statique.

Force est pourtant de reconnaître que les conséquences de ce choix, à partir desquelles se trouvent définies les préférences du décideur, appartiennent au futur. Elles sont, pour cette raison,

incertaines et requièrent une anticipation de sa part. Pour intégrer cette dimension d'incertitude dans la décision, les économistes ont introduit des probabilités subjectives résultant d'un traitement probabiliste des croyances de l'individu sur les conséquences de ses choix. Ces probabilités subjectives entrent alors en ligne de compte dans la définition des préférences du consommateur au moment où le décideur effectue son choix. L'instant du choix constitue toujours, pour cette raison, le point fixe auquel renvoie la décision. On peut dès lors interpréter ce recours aux probabilités subjectives comme un procédé permettant de transformer ces incertitudes en équivalents certains sur lesquels seraient définies les préférences. Afin d'éviter cette confusion simplificatrice, des logiciens comme Jeffrey, pourtant favorables à l'approche économique de cette logique de la décision, se sont efforcés de distinguer conceptuellement les probabilités subjectives, qui concernent l'occurrence de ces conséquences, et leur désirabilité pour le décideur (Jeffrey, 1983). En outre, la survenance d'événements imprévisibles, ou, tout au moins, imprévus par le décideur, peut remettre en cause la validité des probabilités subjectives associées aux conséquences d'un choix. Ainsi, un accident météorologique grave, subit et très rare, est le plus souvent absent du champ de conscience des décisions quotidiennes. Shackle est l'un des seuls économistes qui se soit attaché à critiquer cette faille des probabilités subjectives, ce qui l'a conduit à chercher à intégrer la surprise dans son traitement original de la décision en incertitude. On remarquera, à ce propos, et cela n'est pas fortuit, que presque tous les exemples choisis par Shackle pour illustrer son traitement de la surprise portent sur des choix d'investissement et ne concernent presque jamais la consommation.

Une extension dynamique de ce schéma statique s'est également imposée pour d'autres raisons ; lorsqu'il s'agit notamment de rendre compte des décisions, dont les procédures, et donc les conséquences, font intervenir plusieurs séquences, ou une période plus longue. On songe, par exemple, aux situations de négociations en plusieurs coups, où les choix séquentiels de chacun dépendent des choix observés et anticipés chez l'autre. Une dimension de temporalité se manifeste également au niveau individuel, lorsqu'il s'agit de prendre une décision dont les conséquences finales dépendent d'autres décisions à prendre à des dates ultérieures. Je décide, par exemple, aujourd'hui, de prendre une option sur un contrat à terme. Sans quitter le domaine de la consommation, un achat à crédit rechargeable, c'est-à-dire avec possibilité de renouvellement au moment de l'échéance, rentre dans cette catégorie.

Une version purement conséquentialiste de la théorie de la décision a été proposée pour répondre à son extension temporelle. Cette version soutient que la clé d'un choix rationnel réside dans le fait qu'il prend exclusivement en compte ses conséquences attendues. Rien n'empêche, dans ces conditions, que des modifications dans les préférences du décideur interviennent au cours du processus décisionnel. Il suffit de soumettre ces modifications à des conditions de cohérence dynamique, afin de préserver cette relation instantanée, « choix/conséquences » qui caractérise, dans cette optique, le moment de la décision. Ainsi, le sujet rationnel doit-il, lorsqu'il se trouve au moment d'un choix, prendre en compte, non seulement les conséquences attendues de ce choix, mais également les conséquences des choix intervenues dans les séquences antérieures (Hammond, 1976, 1988). Il en résulte plusieurs transformations dans la formalisation de la fonction qui représente, dans

ces nouvelles conditions, le choix intertemporel d'un consommateur rationnel[11]. Mais cette interprétation conséquentialiste de la décision ne change rien au fond. Le cadre statique, même rendu ainsi dynamique, sert toujours de référence à la théorie économique de la décision. Il reste, pour cette raison, incapable de rendre compte des véritables dynamiques d'interactions, bien plus complexes, comme l'a montré ce chapitre, entre le sujet et sa perception des différentes composantes de son environnement. Ces dynamiques commencent aujourd'hui à être explorées, à la lumière des connaissances naissantes qui s'accumulent, concernant les processus cérébraux intervenant au cours de ce que nous appelons décision.

Trois constats d'ores et déjà s'imposent. Une décision, quelle qu'elle soit, n'est jamais réductible à une opération mentale unique. Elle s'inscrit toujours dans une confrontation de mécanismes multiples. Sa dimension temporelle n'est pas intelligible en la rapportant à un seul point du temps (le moment présent). Elle fait intervenir, non seulement des souvenirs et des projections, mais aussi leurs déformations et leurs transformations, en fonction de la variation des distances qui séparent leurs conséquences les unes par rapport aux autres, et par rapport aux différents moments vécus par le décideur. Dans l'exemple choisi par Pierre Livet, la distance temporelle qui sépare la soirée programmée entre amis et l'examen n'est pas perçue de la même manière par le sujet, au moment de sa décision initiale (quelques jours avant), et à mesure que le temps s'écoule et que les échéances s'approchent. Le processus décisionnel, enfin, n'est pas un système fermé sur le sujet décideur, au moment de sa prise de décision. Il se déploie, au contraire, comme un système ouvert sur l'extérieur, c'est-à-dire en interaction permanente avec son environnement (matériel et humain) ; soit, directement, à travers les informations qu'il recueille, soit, indirectement, par l'intermédiaire de son imagination sensible. Il en résulte, en particulier, qu'associer la cohérence d'un agent et, partant, sa rationalité, à l'invariance intertemporelle de ses préférences, ou tout au moins, à quelques conditions strictes (transitivité…), comme le fait la théorie économique de la décision, n'est pas en accord avec le processus mental qui guide ces décisions. Ces constats, qui se sont progressivement imposés, justifient la place accordée à la discussion de la décision dans ce chapitre consacré à l'analyse de la cognition interactionnelle. Ils conduisent, comme on l'a montré, à une critique assez radicale de la théorie économique traditionnelle de la décision. Il nous faut maintenant examiner comment les différents matériaux recueillis, d'une part par l'économie et la psychologie expérimentales, et d'autre part par la neurobiologie et l'imagerie cérébrale, ouvrent la voie à l'élaboration d'une théorie alternative.

À la recherche des bases neuronales de l'actualisation temporelle

Commençons par réexaminer la question, classique en économie, du traitement de l'actualisation (« discount » temporel) par le décideur, au cours du processus de décision, avec les conséquences qu'il entraîne sur ses prises de décisions. Plusieurs travaux expérimentaux ont été réalisés dans le cadre de

protocoles différents, dont certains portent, précisément, sur des choix de consommation. La diversité des protocoles expérimentaux, qui correspondent à la variété des hypothèses que désirent tester les chercheurs, ne facilite ni l'élaboration d'une mesure fiable de l'actualisation subjective, ni la formulation d'une synthèse claire des résultats observés au cours de ces expériences. Un point semble toutefois acquis. Deux mécanismes cérébraux interviennent au cours des choix proposés entre plusieurs options, à échéances temporelles différentes. Ils correspondent respectivement à l'activation de régions cérébrales distinctes. Il s'agit, d'un côté, de l'activation de régions limbiques et paralimbiques, étroitement corrélées au système dopaminergique (striatum, NACC ou nucleus accumbens…) et, d'un autre côté, de certaines zones du cortex préfrontal et du cortex pariétal mobilisées de manière prioritaire dans les opérations de raisonnement. On observera que l'identification de ces deux mécanismes cérébraux distincts n'implique pas l'existence de deux systèmes séparés. Nous pensons plutôt que l'actualisation au cours du processus décisionnel est une opération mentale qui articule ces deux mécanismes, certes différents, et selon des modalités diverses, confirmant ainsi les ressources de plasticité du cerveau humain. C'est pourquoi les termes du débat qui oppose aujourd'hui, sur ce terrain, les partisans de deux systèmes à ceux du système unique nous paraissent mal posés (Rustichini, 2008).

C'est la publication du dernier ouvrage de Kahneman, intitulé précisément *Thinking Fast and Slow*[12] (2011) qui a élargi ce débat, en le faisant connaître à un public plus large que celui des seuls initiés. On peut certes regrouper un ensemble de fonctions cérébrales au moyen de cette partition. À ces fonctions peut également être rattachée l'activation de régions cérébrales de plus en plus précisément identifiées. Pour autant, nous avons montré, dans le premier chapitre, que l'activation et l'inhibition correspondaient à deux régimes opposés, mais dont le fonctionnement synchrone posait principalement un problème de coordination mis clairement en évidence dans les cas, parfois pathologiques, de dysfonctionnement. Nous verrons, en outre, que dans de nombreuses situations, si certains marqueurs du cerveau rapide (dopamine, striatum, NACC) se trouvent suractivés, cela n'empêche pas d'autres zones du cerveau lent, comme certaines parties du cortex pariétal, de travailler en arrière-fond. Enfin, l'opposition lent/rapide se révèle très insuffisante pour rendre compte de la dimension intertemporelle complexe dans laquelle évolue le fonctionnement de notre cerveau au cours du processus de prise de décision.

Sans surprise, l'activation de la première des régions cérébrales limbiques et paralimbiques apparaît prédominante lors des choix qui privilégient le très court terme, alors que c'est la seconde qui prévaut dans les choix qui optent pour le long terme. Mais, tandis que la première de ces deux régions cérébrales se trouve le plus généralement activée lorsqu'il s'agit d'échéances très courtes, il n'en va pas de même de la seconde zone qui est activée, à des degrés, certes variables, mais quelle que soit l'échéance temporelle. De plus, les modalités temporelles du déclenchement du mécanisme limbique à forte connotation émotionnelle semblent dépendantes des objets sur lesquels portent les choix et les délais correspondants. Des différences apparaissent, en effet, entre le traitement cérébral de gains financiers à diverses échéances, mesurées en semaines ou en mois, et celui de biens de consommation, comme un jus de fruit, à consommer immédiatement, 10 minutes plus tard, ou quelques minutes

encore un peu plus tard. Dans le cas de la consommation, en effet, l'activation du mécanisme limbique n'intervient que pour l'option d'une consommation immédiate. On ne la retrouve plus dans le cas d'une option juste un petit peu plus éloignée dans le temps (10 minutes), par rapport à une option encore plus éloignée (20 minutes) ; et cela, à l'inverse de ce que l'on observe dans le cas des choix portant sur des gains financiers pour des options éloignées, allant de l'immédiat à plusieurs semaines (McLure *et al.*, 2004, 2007). Cette différence montre d'abord que les échelles temporelles sur lesquelles le cerveau travaille sont très variables, en fonction des objets sur lesquels porte la décision. Elle conforte ainsi notre conviction qu'une théorie de la décision ne peut pas se construire sur la base d'une théorie générale élargie de la consommation, comme l'avait initialement énoncé la théorie économique. Elle révèle, ensuite, que les échelles de temps ne sont pas perçues par le cerveau comme linéaires et similaires dans ces différents cas. Si l'instant présent est toujours traité de manière à part, la distance qui sépare « tout de suite » et « dans 10 minutes » n'est pas du tout perçue de manière identique à celle qui sépare 10 minutes et 20 minutes après, dans le cas du jus de fruit (McLure *et al.*, 2007). Cette différence apparaît, en revanche, beaucoup moins nette dans l'expérience monétaire, lorsqu'il s'agit de comparer des gains à échéances de 2 et de 4 semaines (McLure *et al.*, 2004).

L'une des difficultés rencontrées lorsque l'on cherche à interpréter ces résultats expérimentaux tient au fait que ces expériences font intervenir simultanément deux variables différentes, l'importance des quantités attendues et la date des échéances. Une expérience plus récente s'est efforcée de distinguer leurs effets respectifs au niveau de l'activité cérébrale, en faisant varier de manière indépendante les montants des gains attendus et les dates où ils seront distribués dans le futur. Elle montre que l'élévation de l'activation de la région du nucleus accumbens (NACC), ainsi que d'autres régions du cortex cingulaire, se trouve directement corrélée à l'importance du montant attendu, tandis que d'autres régions, notamment corticales latérales, se désactivent en fonction du report temporel de l'échéance (Balard et Knutson, 2009). Le taux du discount temporel utilisé par les individus fait donc intervenir deux composantes neuronales distinctes, relatives aux quantités du gain attendu et aux délais de leur attente, mais dont la mobilisation apparaît étroitement imbriquée.

Il s'agit en définitive de mécanismes complexes, comportant chacun plusieurs composants. Lorsque les économistes ont cherché à les modéliser en les simplifiant, ils ont proposé de représenter la domination de l'un de ces mécanismes sur l'autre, au moyen de deux coefficients qui leur étaient familiers, β et δ, correspondant aux formes, hyperbolique, pour le premier, et exponentielle, pour le second, attribuées respectivement au taux d'actualisation. Ainsi, lorsque l'activation du système limbique l'emporte sur celle du système des régions du cortex préfrontal et pariétal qui ont été mentionnées, le sujet manifeste dans ses choix une forte préférence pour un résultat immédiat, fût-il moins intéressant en termes de capitalisation, ce qui correspond à un β très élevé. En sens inverse, lorsque l'on observe, chez le sujet, une beaucoup plus forte activation des régions du cortex préfrontal et pariétal, il ne manifeste plus de préférence pour le résultat immédiat et révèle un δ bien plus élevé. Tout se passe donc comme si la décision résultait d'un arbitrage, ou tout au moins d'une pondération, entre l'activation de chacun de ces deux mécanismes, eux-mêmes dépendants, comme on l'a vu, des

montants espérés. Cette hypothèse se trouve, d'une certaine manière, confortée par le constat d'une activation complémentaire au niveau de l'ACC, une zone du cerveau associée aux conflits, lorsque, par exemple, le mécanisme limbique prend le pas sur l'activation des régions préfrontales et pariétales concernées. On retrouve ainsi, formulé d'une autre manière, le dualisme précédemment évoqué.

Cette formulation schématique du discount temporel au moyen des coefficients β et δ a été récemment critiquée par Ainslie (2012) sur la base de son incompatibilité avec la notion de *self-control* de la volonté, qu'il privilégie dans son analyse des choix intertemporels, et dont Pierre Livet, a montré, dans ce chapitre, l'intérêt, mais aussi les limites. Ainslie préfère, pour cette raison, revenir à sa formulation hyperbolique initiale (Ainslie, 1975, 1991) de type DV = 1/1 + k.D, où DV représente le taux d'actualisation attendu rapporté à une échéance de temps objective, et k un coefficient hyperbolique traduisant l'impatience du sujet. Dans son dernier article Ainslie élargit toutefois l'interprétation de sa formule initiale, en introduisant un processus récursif de prédictions du sujet sur lui-même aux différentes échéances envisagées. Ce processus peut ainsi entraîner un renforcement ou, au contraire, un abandon des engagements pris par le sujet envers lui-même, au cours du déroulement du temps (2012). L'introduction de la volonté du sujet aux différentes échéances, reliée par Ainslie au traitement de la temporalité, permet ainsi de prendre en compte une véritable intertemporalité dynamique, de nature subjective. Ce traitement peut se traduire, dans les faits, par un ensemble de choix cohérents dans le temps, mais il peut également, comme le reconnaît Ainslie lui-même à la fin de cet article, donner lieu à une série de pseudo-rationalisations, de dénis, voire de mouvements impulsifs. Ces liens diachroniques entre les choix introduisent, certes, une manière de cohérence intertemporelle. Mais la mémoire qui les tisse n'est pas exempte, elle-même, de déformations et de biais affectifs. C'est pourquoi cette cohérence temporelle est bien éloignée de la rationalité économique. Pierre Livet a proposé dans ce chapitre de mettre l'accent sur la notion de perceptions successives, entendues en termes de saillances, qu'il interprète, dans la suite de son développement, comme les composantes d'une dynamique de catégorisations. Nous reviendrons sur cette question en des termes un peu différents.

Par-delà le problème posé par sa formalisation, qui fera l'objet d'une discussion séparée, l'hypothèse générale d'une articulation entre deux mécanismes a été confortée par d'autres résultats qui portent plus directement sur l'anticipation du décideur. Une dynamique cérébrale différente a été mise en évidence, lorsque la décision oriente les anticipations du décideur vers les conséquences immédiates de sa décision, comme dans le cas du choix du consommateur, et lorsque sa décision requiert de prendre en compte, différents termes, comme dans le cas de placements financiers (Tanaka *et al.*, 2004). L'intérêt de l'expérience de Tanaka réside dans une méthode originale qui inspire le protocole expérimental qu'il a choisi et son traitement statistique. Les sujets effectuent des choix en trois séquences. Leur option consiste à choisir, chaque fois, entre un cumul de petits gains à chacune des trois échéances différentes, ou l'acceptation de petites pertes compensées par un bonus plus conséquent à la dernière échéance. Cette étude est, du reste, l'une des rares à prendre en compte, en plus des attentes de gains, des anticipations de pertes, au moins temporaires. Ses résultats vérifient d'abord le rôle déterminant joué par les zones cérébrales qui régulent les émotions, et en particulier le

striatum et l'insula, dans les anticipations des attentes. Mais ils révèlent également, des différences plus fines dans les circuits des ganglions cortico-basals, respectivement activés dans les anticipations aux différentes échelles de temps. Se pose ainsi la délicate question des conditions neurophysiologiques de passage d'un système d'anticipation court-termiste à un système graduellement de plus en plus complexe, lorsque l'on élargit le spectre temporel.

D'autres chercheurs se sont efforcés de comparer le poids respectif donné par les sujets aux conséquences attendues de leurs décisions, dans l'immédiat et à un plus long terme. Ils ont d'abord vérifié l'hypothèse selon laquelle cette valorisation dépendait étroitement d'une activation conjointe du système dopaminergique (avec ses effets sur certaines parties du striatum) et de l'ACC. Fort de ce constat, ils se sont interrogés ensuite pour savoir si l'observation la plus fréquente d'un arbitrage en faveur du court terme, tenait davantage au fait que les conséquences immédiates se trouvaient survalorisées, ou, au contraire, si c'étaient plutôt les conséquences lointaines qui étaient dévalorisées. Ils ont alors distingué deux groupes d'individus dans leur échantillon : ceux qui privilégiaient les résultats immédiats et ceux qui privilégiaient les résultats à plus long terme. Or, tandis qu'une grande différence, entre ces deux groupes, était observée dans l'activation de ces zones cérébrales pour des résultats immédiats, aucune différence sensible ne se manifestait, au contraire, entre eux, concernant les résultats à plus long terme. Les auteurs en ont conclu que la préférence pour les conséquences immédiates devait plutôt être rapportée à une surévaluation de l'immédiat, déclenchée par cette suractivation émotionnelle, qu'à une sous-évaluation du futur plus lointain qui aurait été imputable à une sous-activation de la conscience réflexive (Cherniawsky et Holroyd, 2013).

Une confirmation indirecte du rôle prédominant joué par ce mécanisme dopaminergique et l'ACC dans la régulation des préférences intertemporelles, révélée par les choix des décideurs, peut être paradoxalement trouvée dans divers travaux de neurosciences interculturelles. Des études comparatives menées entre des sujets occidentaux et asiatiques ont montré que les régions du cortex préfrontal et pariétal, correspondant aux fonctions cognitives du raisonnement, étaient activées de manière semblable et dans des conditions analogues chez les Occidentaux et chez les Asiatiques. Mais une nette différence était, au contraire, observée concernant l'activation des régions cérébrales associées à la valorisation affective (système dopaminergique, striatum). Tandis que ces dernières se trouvaient le plus souvent suractivées, comme on l'a vu, chez les Occidentaux devant les options immédiates, ou à très court terme, elles s'activaient, au contraire, plutôt devant les options à plus long terme chez la plupart des Asiatiques. Ce constat établi par l'imagerie cérébrale correspond, comme attendu, au choix préférentiel du très court terme et du plus long terme, respectivement émis par des Occidentaux et par des Asiatiques au cours de l'expérience, en l'occurrence des Américains et des Coréens (Kim *et al.*, 2012). Des travaux expérimentaux antérieurs avaient déjà mis en évidence la même différence entre les taux d'actualisation qui orientaient les choix de sujets américains et de sujets japonais (Green et Myerson, 2002). En revanche, aucune différence notoire n'était apparue dans cette étude entre les choix des sujets américains et des sujets chinois.

Plusieurs facteurs peuvent expliquer cet écart, d'une étude à l'autre, entre les comportements

observés, chez les Coréens et les Japonais, d'une part, et chez les Chinois, d'autre part. Un élément doit être retenu. Dans cette expérience, les étudiants d'origine chinoise étaient depuis longtemps intégrés à la culture américaine, ce qui n'était pas le cas des étudiants japonais. Si elle était confirmée, cette explication pourrait renforcer notre hypothèse du rôle de variable de régulation joué par le système de valorisation affective dans les choix intertemporels. Une telle valorisation discriminante pourrait ainsi privilégier l'immédiat et le court terme, comme on l'observe le plus souvent chez les Occidentaux, mais également le plus long terme, comme cela semble s'observer chez de nombreux Asiatiques. Il reste alors à comprendre comment se construit cette différence, puisqu'il semblerait qu'elle soit imputable à un trait culturel, transmis, et donc appris par chaque individu. On peut sans doute rapprocher ce trait du phénomène plus général de reconnaissance sociale introduit dans les chapitres précédents.

Des informations supplémentaires, sur le rôle particulier de ce mécanisme dopaminergique dans la régulation des choix intertemporels, peuvent être tirées d'une étude portant sur la comparaison entre les résultats obtenus, lorsque la décision concerne le sujet lui-même (le « self »), ou un autre sujet que le décideur lui-même (Albrecht *et al.*, 2011). Au niveau expérimental, d'abord, les sujets les plus prompts à choisir pour eux l'option immédiate se sont montrés beaucoup plus circonspects lorsqu'il s'agissait de choisir pour un autre. Cette différence dans les comportements observés correspond également à certaines différences dans l'activation des zones cérébrales allouées à l'émotion. Les résultats de l'imagerie cérébrale s'avèrent cependant ici plus difficiles à interpréter, puisque, tout à la fois, les mêmes régions cérébrales se trouvent activées dans les deux cas, mais avec des intensités différentes et selon des modalités distinctes. Il semblerait cependant que le découpage temporel se révèle assez différent au niveau du ressenti affectif. Lorsque le sujet choisit pour lui-même, une première coupure se manifeste instantanément entre l'immédiat et le futur plus éloigné, quelles que soient, par ailleurs, les échéances de ce futur. On ne la retrouve plus lorsque le sujet effectue son choix pour une autre personne qu'il ne connaît pas. En d'autres termes, un lien émotionnel privilégié relierait chacun de nous à l'instant qui suit immédiatement le moment présent (celui-là même de la décision). Cette intimité singulière qui lie le « moi, maintenant », au « moi, immédiatement après », n'a pas d'équivalent lorsque le moi se trouve remplacé par un autre moi, même si le « je » imagine cet autre comme un autre lui-même, dans les termes que nous avons analysés au chapitre 2. D'autres différences dans la perception des échelles de temps ont également été mises en évidence dans une perspective différente, selon que les attentes du décideur le concernent lui-même ou s'appliquent à un autre (Takahashi, 2007). La perception de l'intertemporalité, qui intervient dans nos prises de décision, pourrait ainsi constituer un marqueur du découplage entre « moi » et l'(les) autre(s).

V

On peut déduire de tous ces travaux le schéma général suivant. Sur un arrière-fond d'activité cognitive du cerveau raisonneur (premier mécanisme), intervient une modulation d'ordre émotionnelle, chargée de pondérer les différentes échéances temporelles envisagées par le décideur (second mécanisme). Cette pondération valorise ainsi les attentes du décideur en différents points du temps. Si, au terme de cette pondération, un résultat immédiat, que l'on attend positif, se trouve souvent survalorisé, c'est en règle générale pour réduire la distance temporelle qui sépare, en la circonstance, l'instant de la prise de décision de l'instant de la réalisation de ses conséquences attendues. Cette pondération peut parfois valoriser, au contraire, des points plus éloignés du temps, en fonction notamment, comme on l'a vu, de références culturelles différentes. Mais ce phénomène serait alors plutôt la conséquence d'une absence de survalorisation de l'immédiat, lié à d'autres modes de pensée que celle qui domine l'appréhension du temps dans les cultures occidentales. De même, faudrait-il également introduire des différences dans ces pondérations des échéances temporelles, selon que les attentes du décideur portent sur ce qu'il interprète comme un gain (gain monétaire et matériel, mais aussi moral) ou, au contraire, comme une perte. Leur asymétrie, depuis longtemps mise en évidence sur une base expérimentale par Kahneman et Tversky, se répercute au niveau de la pondération de leurs échéances temporelles respectives. Mais on dispose, encore aujourd'hui, d'assez peu d'informations concernant les pertes. Ce schéma général repose moins, comme il peut paraître de prime abord, sur la survalorisation de telle ou telle échéance temporelle, que sur une modulation dans la valorisation des échéances, les unes par rapport aux autres, à travers des variations de pondération. Nous pensons qu'il s'agit là d'un mécanisme de régulation comparative des perceptions concernant les échéances temporelles de nos décisions à dominante émotionnelle.

Trois traits principaux semblent caractériser son fonctionnement :

- Il permet de sélectionner les points focaux de l'horizon temporel du décideur, qu'il va pondérer différemment, en lien avec ses attentes.
- Cette pondération traite d'une manière distincte (au moins pour le décideur) sa relation avec l'immédiat, conçu comme un prolongement direct du moment présent, et la comparaison entre plusieurs échéances différentes, mais éloignées dans le temps.
- Cette pondération varie, elle-même, au cours du temps. D'une part, le déroulement du temps fait changer de catégorie l'occurrence de certaines échéances. Ainsi, dans l'exemple de Pierre Livet, la soirée entre amis programmée avant l'examen, qui appartenait initialement aux échéances éloignées, rentre dans la catégorie des conséquences immédiates, dès l'instant où le décideur se trouve sur le point d'y participer. D'autre part, des informations internes et externes, recueillies par le décideur au cours du temps sur l'occurrence de ces différentes échéances, contribuent à transformer leurs pondérations qui gouvernent son appréhension de ces différentes échéances.

Il est clair que le schéma de ce processus de régulation n'est pas linéaire. Pour autant, sa traduction dans un modèle mathématique précis n'est pas encore assurée. La révélation de deux modes de pondération différents, lorsqu'ils portent sur l'immédiat, ou s'appliquent à des échéances plus éloignées, conduit à distinguer deux types de dynamique. L'une traduit une hypersensibilité à

l'immédiat, l'autre correspond aux appréhensions sensibles des échéances temporelles plus éloignées qui varient avec l'avancement du temps. Une première façon de traiter ces deux composantes distinctes, comme on l'a vu avec McLure, est inspirée de la présentation de Laibson. Elle consiste à associer à chacune d'elle un taux d'actualisation différent, nommé respectivement β, pour la première, qui prend une forme hyperbolique, et δ, pour la seconde, qui garde une forme asymptotique. On peut construire sur cette base un modèle dynamique d'ensemble qui revêt une forme quasi hyperbolique généralisée. Cette modélisation traduit la survalorisation des résultats immédiatement attendus, mais n'intègre pas les biais observés dans la comparaison des résultats à plus longue échéance. C'est la raison pour laquelle une autre modélisation mathématique de cette dynamique a été proposée. Elle prend en compte, maintenant, la tendance à une survalorisation de l'immédiat, mais également les biais de perception temporelle dont se trouve entachée l'appréhension des diverses échéances plus lointaines, du fait même de la dynamique temporelle dans laquelle elle s'inscrit (Takahashi, 2009).

Pour y parvenir Takahashi et ses collègues se sont inspirés d'une formulation proposée par Tsallis pour représenter des phénomènes physiques de dynamiques statistiques non additives et non extensives. Cette formulation permet, en effet, de représenter la déformation de notre perception temporelle aux différentes échéances de nos choix, par une fonction logarithmique associée à un taux d'actualisation exponentiel. Pierre Livet a montré que ce type de formulation permettait de traduire, à partir de ces non-linéarités, certaines différences de perception observées entre courte et longue portée.

Un détour par l'histoire des idées n'est pas sans intérêt pour comprendre les différentes implications de cette nouvelle formulation. En introduisant cette fonction logarithmique pour relier la durée objective du temps qui s'écoule à la perception sensible que nous en avons, Takahashi a retrouvé une ancienne approche développée par la neurophysique, au moyen de laquelle, un petit groupe de psychologues allemands de la seconde moitié du XIX^{e} siècle avait formulé la relation entre la mesure de l'intensité d'un stimulus extérieur et celle de la perception correspondante ressentie par le sujet. Cette forme de relation généralisée entre l'environnement physique et sa perception par les sujets est bien connue sous l'appellation de loi Weber-Fechner, du nom de ses découvreurs. On peut la représenter simplement par l'équation $\Upsilon = K (\log \beta + \log b)$, où β représente l'impulsion physique, Υ l'intensité de la sensation, et b une valeur seuil de β à partir de laquelle la sensation est perçue par le sujet (Fechner, 1860). Le principal intérêt de cette formulation logarithmique ici est qu'elle traduit le caractère comparatif des sensations ressenties par le sujet. Le programme initial de recherche de Fechner était celui d'une psychophysique externe, destinée à dégager les relations quantifiées entre les stimuli extérieurs de nature physique, des sensations ressenties par les sujets. Appliquée à l'origine par Weber à la comparaison des poids d'objets soulevés ou portés, elle avait été étendue par Fechner aux diverses perceptions, en particulier, visuelles et auditives. Mais à côté de cette psychophysique externe ainsi entendue, Fechner imaginait le développement d'un programme de recherche de psychophysique interne, permettant de relier, selon la même approche, les états mentaux aux processus de fonctionnement du cerveau. Plus précisément, Fechner imaginait le schéma selon lequel une impulsion physique, c'est-à-dire appartenant au monde extérieur, entraînait une activation

neuronale, qui, à son tour engendrait une réaction psychique selon les modalités de cette formulation mathématique. Il reconnaissait toutefois qu'il ne disposait pas, à l'époque, d'informations suffisantes sur le fonctionnement cérébral pour mener à bien cet autre programme. D'une certaine manière, par conséquent, la formulation de l'actualisation proposée aujourd'hui par Takahashi en s'appuyant sur les neurosciences se situe, comme il le reconnaît lui-même, dans le prolongement direct du programme interne exposé par Fechner (Takahashi, 2006).

Cette reformulation de l'actualisation intertemporelle, à partir des variations d'émotions engendrées par les attentes comparées d'un plaisir (ou, au contraire, d'un déplaisir, voire d'une peine) perçu à des échéances temporelles différentes, a le mérite de réintégrer la temporalité dans l'univers sensible de la perception des émotions. Le second avantage, qui en découle, est de relier cette perception à des valeurs seuils, en deçà desquelles les différences ne sont pas perceptibles, un peu comme dans le cas des perceptions visuelles et auditives. Cela a permis à Pierre Livet d'esquisser dans ce chapitre une interprétation intéressante de ces seuils en termes de saillances, sur la base desquelles s'organisent des catégorisations. Cette formulation en temps continu rend compte, en outre, de la dynamique de réactualisation qui s'opère dans l'esprit du décideur, en fonction des informations qui lui sont fournies par l'écoulement du temps. Du point de vue neuronal, cette modélisation s'appuie essentiellement, jusqu'à présent, sur des variations du mécanisme émotionnel (second mécanisme) observées par l'imagerie. Il reste à mieux comprendre, comment ces variations affectent, à leur tour, l'activation de certaines régions du cerveau raisonneur (premier mécanisme) qui a, jusqu'à présent, moins retenu l'attention des chercheurs en neurosciences.

R

La redécouverte de la loi de Weber-Fechner et la renaissance d'une psychophysique, à la faveur des emprunts faits à la modélisation physique de Tsallis, nous ramène, par un cheminement détourné, à la théorie économique de la décision. C'est en effet, en se fondant sur la loi de Fechner, et en s'inspirant des travaux contemporains de Wundt et de Helmholtz, que les utilitaristes anglais (et irlandais), comme Jevons, et surtout Edgeworth, entreprirent de fonder leur analyse des phénomènes économiques sur une « hédionométrie sociale ». Par hédionométrie, il faut entendre ici une théorie de la mesure du plaisir, traité comme la perception d'une sensation. Cette hédionométrie peut être qualifiée de sociale dans la mesure où, comme nous l'avons vu, ce plaisir pour Edgeworth et les utilitaristes hédonistes, n'est pas indépendant du cadre social dans lequel il se manifeste. La décision économique selon cette représentation ne consiste pas dans la maximisation des préférences du décideur, mais plutôt dans la maximisation du plaisir ressenti par le décideur, du fait des conséquences attendues de sa décision. C'est donc par une filiation en partie erronée que l'on fait traditionnellement remonter aux utilitaristes les sources de la théorie économique contemporaine de la décision au moins dans sa version dominante. Rappelons à ce sujet que Fechner, dont les travaux ont

influencé Edgeworth, cherchait à formuler, en termes mathématiques, la relation entre un stimulus, en provenance du monde physique extérieur, et les sensations qu'il provoque chez les individus. De même, Edgeworth se propose de mesurer la sensation de plaisir éprouvée par le décideur qui serait provoquée par les conséquences objectives de sa décision. Dans les deux cas, la formalisation mathématique adéquate est celle d'une fonction logarithmique sous-additive. Sur cette base mathématique commune, Edgeworth construit une fonction représentant les unités de plaisir ressenti qu'il nomme *sentients*, dont la dérivée première est positive et la dérivée seconde négative (1881). Il est intéressant de noter, qu'un siècle plus tard, Kahneman et Tversky associent les mêmes propriétés de concavité à la forme de la fonction qui décrit les variations des gains attendus, en relation avec leur probabilité de réalisation. On peut ainsi interpréter la *Prospect theory*, dans une optique edgeworthienne, comme une description des réactions de plaisir éprouvé aux stimulants extérieurs que représentent des gains futurs, plus ou moins certains (ou, au déplaisir éprouvé devant des pertes, elles aussi, plus ou moins certaines)[13].

On observera que si l'approche logarithmique adoptée par Fechner pour mesurer les sensations a été finalement délaissée par les économistes au profit de formulations linéaires, il n'en fut pas de même des psychologues. Des psychologues comme Stevens se sont attachés à la généraliser sous la forme de fonction de puissance (*Power Law*, Stevens, 1957). Plus près de nous, Luce en partant de l'hypothèse de Stevens a proposé une nouvelle estimation du quantum des sensations psychiques ressenties, en introduisant ce qu'il nomme des « traductions cohérentes des échelles de transformations » (Luce, 1990, 2002).

La forme logarithmique initialement choisie par Fechner, et reprise par Edgeworth, pour appréhender ce que ressentent les sujets, en réaction à un stimulus extérieur réel (Fechner), ou seulement imaginé et attendu (Edgeworth), présente plusieurs avantages dans notre perspective. Elle permet, d'abord, de mettre en évidence des valeurs seuils, en deçà desquelles, les différences d'intensité de ces stimuli ne sont pas perçues, ce qui ne signifie pas qu'elles n'existent pas. Elle rend, ensuite, possible, la distinction de plusieurs échelles différentes dans la perception des stimuli. C'est cette dernière propriété qui est utilisée par Takahashi pour distinguer les perceptions temporelles différentes du sujet, correspondant à différentes échéances, et qui varient, elles-mêmes, avec l'écoulement du temps. Elle permet, enfin, comme nous allons le montrer, et c'est peut-être là le plus important, de rapprocher le mécanisme mental qui semble régler cette perception de la temporalité, de celui qui guide notre appréhension subjective du risque, lui aussi sensible à ces distances temporelles. Nous verrons que l'un et l'autre peuvent être formalisés dans des modèles mathématiques équivalents, dérivés d'une même base logarithmique.

On peut dès lors esquisser dans cet esprit une théorie économique positive de la décision qui ne tomberait pas sous les critiques formulées au début de ce commentaire. Pour y parvenir, une première étape, qui s'inspire de l'approche Fechner-Edgeworth, consiste à traduire les conséquences attendues d'une décision dans le langage sensible des perceptions émotionnelles, ce à quoi nous invitent aujourd'hui les neurosciences. Ainsi ce n'est pas la somme d'argent que je pourrais gagner que prend en compte mon cerveau, mais le gain hédonique qu'il estime que je pourrais tirer de cette somme.

Cette transformation requiert d'introduire deux coordonnées. La première correspond à un point de référence, par rapport auquel les variations de ces perceptions émotionnelles hédoniques se trouvent étalonnées. Lorsque Kahneman et Tversky parlent de « gains », et de « pertes », dans le sens que nous proposons de donner à ces termes, ils se réfèrent explicitement à un point de référence ; encore faut-il préciser comment ce point de référence se trouve défini. Il s'agit à l'évidence d'une référence subjective. Par simplification, ce point de référence est traité dans la théorie de Kahneman et Tversky comme un point fixe qui correspond à la situation du sujet au moment où il prend sa décision. Il se trouve, par conséquent, associé au temps présent. Il apparaît pourtant, à la lumière des études qui ont été discutées, que ce point de référence n'est pas fixe, mais change avec l'écoulement du temps. Ainsi, dans l'exemple de Pierre Livet, le gain moral qui accompagne le fait de s'abstenir de participer à une partie entre amis avant un examen important est rapporté au moment correspondant au début du processus de la décision. Il va disparaître au moment où la partie entre amis commence, parce que, précisément, le point de référence temporel n'est plus le même. Il reste à comprendre comment le décideur, placé au point de référence to va se représenter ce qu'il ressentirait au point de référence to+n. La formalisation proposée par Pierre Livet en termes de saillances peut fournir ici une clé intéressante. Mais il faudrait connaître le processus mental, voire neuronal, par lequel s'opère cette transformation. Car il ne s'agit plus maintenant seulement de perception mais, cette fois, de cognition.

La seconde dimension renvoie à la perception des distances temporelles, en partant du point de référence. Ainsi, au temps to, correspondant au moment de la décision, une distance d'un jour, voire seulement de quelques heures, ou même de quelques instants peut être ressentie comme très longue, alors qu'elle devient presque indiscernable à l'horizon de deux ans. On retrouve donc, au niveau de la perception temporelle, l'existence de valeurs seuils qui jouent un rôle déterminant dans la psychophysique de la perception développée par Fechner. Ce phénomène, bien connu, semble étroitement corrélé au degré d'activation des zones cérébrales de l'émotion associée à l'impatience d'un résultat. On peut donc légitimement penser que leur activation régule cette perception des distances temporelles par une pondération qui valorise, et parfois même survalorise, le plus souvent, les distances proches du point de référence (actualisation hyperbolique). Cette observation est à rapprocher des saillances dans l'analyse de Pierre Livet.

Cette prise en compte de la perception des distances temporelles, telle que nous proposons de l'interpréter, rejoint également, sur ce point, la théorie du prospect. Kahneman et Tversky distinguent deux phases dans le processus mental qui conduit à la décision, le *framing*, parfois complété par l'*editing*, et le *weighting*, opération à laquelle ils consacrent l'essentiel de leur analyse. Ils introduisent, pour se faire, une fonction de pondération à laquelle ils ont donné d'abord la forme d'une fonction non linéaire de transformation des probabilités objectives associées aux options (Kahneman et Tversky, 1979), avant de proposer une formulation, directe et plus élégante, en termes de fonction cumulative de dépendance de rang (Tversky et Kahneman, 1992). Quelle que soit leur présentation, ces déformations sont, pour ces auteurs, imputables à la perception, par le décideur, du degré

d'incertitude qui affecte l'actualisation de ses attentes (risque subjectif). Les informations fournies par les probabilités objectives, dont disposent les sujets au cours des différentes expériences, représentent autant de signaux qui leur permettent d'évaluer subjectivement les distances mentales qui séparent les différentes options qui leur sont présentées de leur actualisation. On comprend mieux, dans ces conditions, les préférences majoritairement manifestées par les individus pour une option moins risquée, voire certaine, lorsqu'elle se traduit par l'attente d'un gain et plus incertaine, lorsque c'est une perte qui est attendue. Dès lors, l'hypothèse d'une similitude entre les évaluations subjectives de la temporalité, au cours d'un processus de décision intertemporelle, et l'appréciation du risque en termes de probabilités subjectives, en matière de choix risqué, s'explique naturellement.

Certes, la perception subjective de la distance temporelle qui sépare deux options et l'évaluation subjective de leur chance respective de se réaliser sont choses différentes. Il existe cependant, entre elles, plusieurs caractéristiques communes. L'une et l'autre s'organisent à partir du point de référence, dont on a vu qu'il renvoyait toujours à un moment du temps. La perception de l'incertitude n'est pas, en outre, indépendante de celle de la distance temporelle. La préférence pour le présent, observée dans le cas de l'attente d'un gain, traduit, à la fois, une forte pondération des distances (plus la distance qui sépare le résultat attendu de sa réalisation est longue, plus l'attente qui les sépare est perçue comme difficile à supporter, tout au moins dans la courte période), et une valorisation de la certitude, puisque cette distance accroît également le risque que ce gain ne soit jamais reçu (« un bon tiens vaut mieux que deux tu l'auras », selon le proverbe). On peut ainsi penser que le fait qu'un gain soit obtenu « juste maintenant » renforce son caractère certain (et *vice versa*). Il existe, en effet, une affinité mentale entre la certitude d'obtenir une récompense et la plus faible distance temporelle qui sépare de cette obtention. Pour autant, cette relation inverse entre le risque et la proximité temporelle peut recouvrir des modalités diverses et parfois même paradoxales, lorsque, par exemple, c'est le risque lui-même qui engendre, chez certains sujets, une forme d'excitation proche de l'impatience.

L'un des premiers chercheurs à avoir identifié, sur une base expérimentale, cette relation mentale entre le discount temporel subjectif et les probabilités subjectives est Rachlin (Rachlin, Rainneri et Cros, 1991). Rachlin a ainsi comparé les résultats des choix de deux groupes de sujets. Les sujets du premier groupe étaient confrontés à des options de gains monétaires certains, face à d'autres seulement probables (leurs probabilités variant de 5 % à 95 %). Les sujets du second groupe devaient choisir entre des gains monétaires immédiats et d'autres disponibles à différentes échéances (échelonnées de 1 mois à 50 ans). Il a mis en évidence de cette manière que la distribution statistique des choix dans les deux groupes prenait la même forme, et qu'une relation d'équivalence permettait de transformer la perception temporelle des échéances en probabilités. Il est intéressant de noter que Rachlin a montré, dans des travaux ultérieurs, que cette appréhension subjective de la temporalité par les individus se révélait sensible, dans les deux cas, aux interactions avec les autres. Cette dimension intersubjective de ces deux perceptions de la temporalité en situation de décision, et donc d'attente, a été mise en évidence dans le cadre de jeux expérimentaux (Rachlin et Jones, 2008, Jones et Rachlin, 2009).

Mais l'avancée décisive en la matière a été réalisée par Takahashi. Il a d'abord proposé un cadre

formel plus général pour rendre compte des deux phénomènes, en revisitant l'approche logarithmique de la perception temporelle développée par Fechner, à la lumière des formalisations hyperboliques ultérieures données au taux d'actualisation et au choix risqué. Il a, dans cet esprit, formulé ce qu'il appelle la q-exponentielle généralisation de l'actualisation temporelle, où les déformations par rapport à la forme exponentielle correspondent à des logarithmes de perception subjective du temps. Cette affinité mathématique entre les deux phénomènes lui a ainsi permis de dégager une interprétation de leur proximité en termes psychophysiques, rendue possible grâce au détour par un traitement logarithmique. Il s'agit, dans les deux cas, d'évaluer, aussi rigoureusement que possible, la relation entre le temps physique et le temps subjectivement perçu. Pour Takahashi, cet écart renvoie, dans le premier cas (actualisation), à l'attente elle-même, et, dans le second cas (risque), à sa pondération en termes de probabilités (Takahashi, Han et Nakumara, 2013). Il a enfin commencé à tester son hypothèse dans le cas de choix portant sur des gains et des pertes d'un montant différent, associés à des échéances temporelles différentes (Takahashi *et al.*, 2013).

Les perspectives ouvertes par cette analyse de Takahashi sont considérables. Son approche suggère qu'un choix réfléchi résulte nécessairement d'une réaction mentale à la réalité physique de la temporalité. On peut y retrouver une intuition forte déjà développée par Hayek dans *L'Ordre sensoriel*. Hayek y distingue l'« ordre physique » du monde externe avec ses stimuli, l'« ordre mental » ou « phénoménal », avec les sensations que nous ressentons, et un « ordre neural » qui transmet les impulsions du monde physique en les transformant en sensations au cours de processus dynamiques d'apprentissage dans lesquels s'inscrivent nos décisions et nos actions (Hayek, 1952, 2001)[14].

Pour autant, bien des points restent à éclaircir, avant de pouvoir tirer de cette approche et dégager de ces premières données une véritable théorie économique de la décision. En premier lieu, on connaît encore assez peu de chose sur les processus cérébraux qui aboutissent à ce schéma. On retrouve, certes, les activations des régions cérébrales déjà signalées. L'activation dopaminergique du striatum, observée dans le circuit de la récompense, et celle du cortex antérieur cingulaire (ACC), relevée dans la plupart des situations conflictuelles. Mais les modalités exactes de fonctionnement que pourrait traduire ce schéma sont encore mal connues. On commence seulement à associer certaines anomalies observées dans sa dynamique à des dysfonctionnements cérébraux répertoriés, comme c'est le cas, par exemple, dans les syndromes d'addiction. Nous ignorons encore largement comment ce schéma de perception du futur, principalement émotionnel et sensible, enclenche un processus cognitif plus élaboré. Nous avons montré qu'il fallait au décideur s'imaginer un temps présent, dans des avenirs séparés de son temps présent actuel par des distances temporelles différentes ; d'où l'introduction de degrés de conscience. De même, la perception de l'incertitude dans l'avenir implique d'abord la possibilité d'imaginer, en même temps, une situation et l'absence de cette situation, à l'horizon temporel donné ou choisi. Plusieurs exemples en ont été proposés par Pierre Livet dans ce chapitre 3. Cette faculté permet, ensuite, de pondérer les deux faces opposées de cette situation, ce qui aboutit à lui associer des probabilités subjectives. Mais cet exercice est plus complexe encore, car ces

probabilités ne sont pas rapportées à des états du monde, mais à des gains ou à des pertes, c'est-à-dire à des variations de plaisir ou d'agrément attendues par le décideur, par rapport à un point de référence. On a également montré que ce point de référence résultait d'une construction mentale, renvoyant elle-même à un moment du temps réel (le temps présent), imaginé (une échéance future), et même seulement imaginaire (le temps idéal). Or on sait encore peu de chose sur l'élaboration cérébrale de cette construction à étages multiples. On devine que l'ensemble des processus qui ont été identifiés ici s'insère dans une dynamique encore mal connue, qui inclut des apprentissages, des mémorisations et des corrections, c'est-à-dire une manière d'histoire.

Une théorie économique de la décision conçue sur ces nouvelles bases tourne-t-elle le dos aux hypothèses de rationalité sur lesquelles les théories économiques standard se sont construites ? Pas obligatoirement, et cela pour deux raisons. En premier lieu, ce que nous avons proposé ici ne concerne qu'une théorie positive ou descriptive de la décision, qui vise à expliquer comment les agents économiques prennent leur décision. Il laisse donc un champ ouvert à une théorie normative, qui, tout en s'inspirant de cette description, pourrait réintroduire la rationalité au niveau de ses critères. Cette question sera réexaminée dans la troisième partie de l'ouvrage. En second lieu, et surtout, les éléments que nous proposons pour l'élaboration d'une nouvelle théorie positive de la décision constituent un repérage préliminaire des différentes données à prendre en compte et des ressorts de leurs fonctionnements qui sont mobilisés dans l'acte de décider. Il faut, dès lors, prolonger cette démarche en précisant, par exemple, les degrés de contrôle dont peut disposer le décideur sur les mécanismes de sa décision, tels qu'ils ont été décrits. L'usage de ce contrôle par les agents, de façon à accroître, d'une manière ou d'une autre, leur satisfaction, permet de réintroduire ainsi la rationalité économique, mais à une place différente, et selon d'autres modalités. C'est dans cette direction que nous nous proposons, avec les contributions des neurosciences, d'orienter les recherches vers cette nouvelle théorie économique de la décision

---

1. Si on met à part la théorie du regret de Sugden et Loomes, qui cependant ne tient pas compte de toutes les comparaisons que nous venons d'évoquer. Loomes et Sugden (1982), « Regret theory : An alternative theory of rational choice under uncertainty », *Economical Journal*, 92, p. 805-824.

2. *Cf.* les travaux de Tsallis, que nous citons ici pour leur réutilisation par Takahashi.

3. Les économistes utilisent plutôt une représentation qui part de t0 = l'événement gratifiant (sommet de la courbe), et va vers le futur. Mais cela a l'inconvénient de présenter une baisse rapide de la valeur de la gratification quand on s'éloigne très peu vers le futur, alors que le phénomène important est que la valeur monte en flèche quand, partant d'un t0 situé plusieurs moments *avant* l'événement, on devient très proche du moment de l'événement.

4. Ainslie a bien noté qu'il n'a pas qu'une manière de relier des événements successifs pour constituer une série, mais il semble oublier que cela ne s'applique pas seulement aux événements lointains, mais aussi aux événements proches.

5. C'est une donnée dont ne rend pas compte l'hypothèse de Han et Takahashi (2012), qui réduisent l'effet de discount temporel au fait que nous avons une perception non linéaire du temps (les intervalles proches comptant plus que les lointains). En revanche cette perception non linéaire du temps peut s'expliquer par l'abondance des saillances dans le temps proche et leur raréfaction dans le temps lointain.

6. Les physiciens nous donnent un critère : si en ajoutant un grain de sable sur le côté du prétendu tas, on peut déclencher une

avalanche, c'est que c'est bien un tas. Sinon ce n'en est pas un.

7. Koechlin a pensé montrer que le cerveau humain ne pouvait accomplir plus de trois tâches simultanément, même si cette estimation dépend de la manière de regrouper ou séparer les tâches.

8. Nous préférons éviter l'incertitude que peut impliquer ce vague, et cette aversion à l'incertitude pourrait expliquer que notre empathie plus forte pour les membres de notre groupe aille jusqu'à des sentiments malveillants envers les membres d'autres groupes (Gonzalez, 2013) qui forment le côté « sombre » de l'empathie. Mais ces émotions peuvent être régulées si nous avons plus de temps pour contrôler nos réactions (Érès, 2013, citant Cunningham, 2004).

9. La théorie de jeux tente par exemple de tenir compte de ce formatage des motivations comme reconnaissables par autrui par une théorie du *signalling*, où l'on joue certains coups pour signaler une intention future.

10. On consultera sur ce point E. von Böhm-Bawerk, *The Positive Theory of Capital* (1889).

11. Plusieurs axiomatiques ont été proposées en ce sens, notamment par Bernheim (1986) et par Hammond (1994).

12. C'est du reste un retour à l'acception utilitariste du terme utilité que propose Kahneman en opposant ce qu'il nomme l'utilité expérimentée (*experienced utility*) à l'utilité décision, telle qu'elle est définie dans l'analyse économique standard (Kahneman, Wakker et Sarin, 1997).

13. Une intéressante filiation relie directement Hayek à Wundt et à Helmoltz, et maintenant Takahashi à Hayek.

14. L'analyse développée par Hayek dans cet ouvrage s'inspire directement des travaux du psychologue Stevens, qui, dans la tradition de Fechner et de sa psychophysique, propose une modélisation de la mesure de la relation entre l'impulsion physique et le ressenti mental.

PARTIE 2

COORDINATION ET COOPÉRATION

CHAPITRE 4

Les modes de coordination

Qu'il s'agisse de la construction mentale de l'intersubjectivité, ou de la mise en œuvre de l'interintentionnalité au cours des échanges entre les sujets, les chapitres précédents ont mis en lumière la pluralité des systèmes neuronaux qui se trouvent activés et inhibés, ainsi que la multiplicité des niveaux de codages et de rappels qui sont sollicités. Leur coordination se présente donc comme un facteur déterminant dans le fonctionnement des interactions. Beaucoup de dysfonctionnements, occasionnels ou pathologiques, peuvent ainsi être imputés à des défauts de coordination. Plusieurs chercheurs en neurosciences ont aujourd'hui recours aux métaphores économiques de la compétition et/ou de la régulation centrale, pour discuter de l'organisation des différents mécanismes neuronaux à l'œuvre dans la réalisation de cette coordination. En sens inverse, les modèles connexionnistes qui ont d'abord prévalu pour expliquer le fonctionnement en réseaux des neurones sont aujourd'hui utilisés par certains économistes pour rendre compte de mécanismes de coordination, lorsqu'il s'agit, par exemple, de transmission sociale de l'information. Mais ces emprunts métaphoriques, qui reposent souvent sur des approximations, sont parfois porteurs d'ambiguïtés et de méprises.

Un fait est certain. La coordination des actions humaines est multiforme et on la retrouve au cœur de phénomènes sociaux les plus divers. Que l'on songe à la circulation automobile, aux compétitions sportives, ou au fonctionnement/dysfonctionnement des places financières, les exemples ne manquent pas. L'analyse stratégique des interactions développée par la théorie des jeux devrait fournir le corpus théorique économique le mieux adapté pour en rendre compte. Il apparaît cependant, que la seule rationalité individuelle prêtée aux joueurs, qui constitue son socle logique, se révèle le plus souvent tout à fait insuffisante pour rendre compte de la coordination des mouvements des différents joueurs. Comme on l'a déjà montré au chapitre 2, ce problème de coordination se pose aux théoriciens des jeux, chaque fois que la situation décrite par le modèle admet plusieurs équilibres, entre lesquels la théorie ne dispose d'aucun critère logique suffisamment satisfaisant pour les départager et prédire celui qui prévaudra (ou devrait prévaloir). L'intervention de multicritères,

couramment utilisée en théorie de la décision individuelle, lorsque plusieurs solutions sont logiquement possibles, est de peu de secours en théorie des jeux, puisqu'elle ne fait alors que repousser la question centrale de l'existence d'une commune référence des joueurs à ces critères. De telles situations sont loin d'être des exceptions et l'on en trouve de très nombreuses illustrations dans la vie sociale la plus concrète.

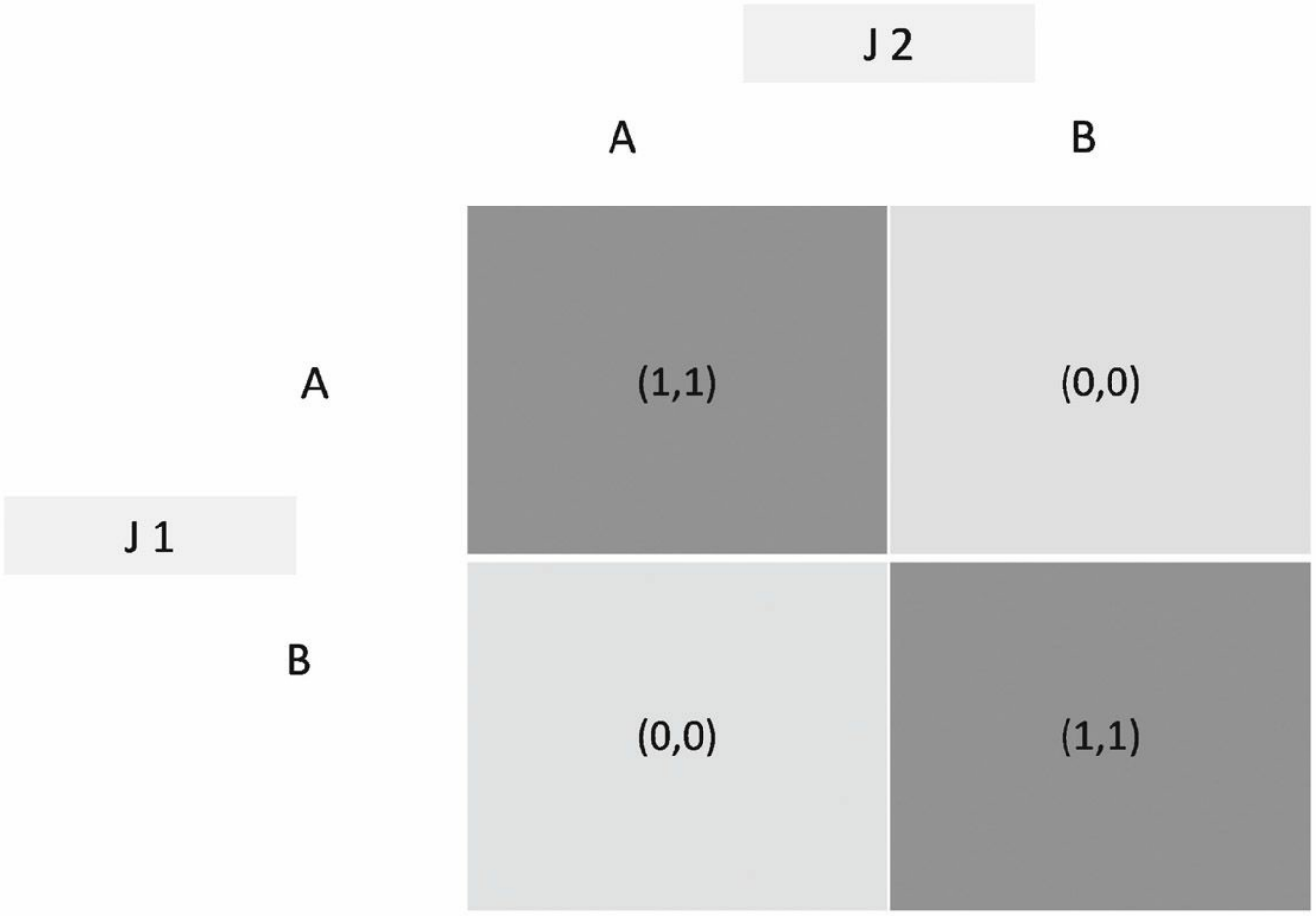
Pour chercher à résoudre les difficultés qui leur étaient ainsi posées par la coordination, les chercheurs en théorie des jeux ont exploré deux voies différentes et complémentaires. Au niveau strictement interindividuel, ils se sont concentrés sur les jeux les plus simples, où la coordination porte sur les décisions de deux joueurs seulement. Chaque joueur est un sujet confronté à un autre sujet, unique, comme lui-même. Au niveau d'une interface collective, les chercheurs ont étudié des jeux où interviennent un grand nombre de joueurs. Chaque joueur individuel interagit avec une (ou des) population(s) d'autres joueurs. Cette interaction comporte alors nécessairement une part d'anonymat, parce qu'il n'est pas possible, pour chaque joueur individuel, de ramener les autres joueurs à un seul individu. Ces deux types de travaux ont abouti à des solutions différentes. Mais, dans les deux cas, ils ont conduit à étendre, et même à transformer, le corpus conceptuel initial de la théorie des jeux. Dans le premier cas, la question de la coordination a ainsi été reformulée dans le cadre de jeux « psychologiques » ou « mentaux ». De tels jeux prennent explicitement en compte l'appréhension par chaque joueur de la représentation par l'autre joueur de la situation où ils se trouvent tous les deux confrontés. Dans le second cas, la coordination se trouve traitée dans le cadre de « jeux globaux », et plus récemment de « jeux de réseaux » », où la convergence vers un équilibre se trouve expliquée par la nature et les caractéristiques différentes des informations dont les joueurs peuvent disposer les uns sur les autres, et les uns par rapport aux autres.

La poursuite de notre enquête sur le fonctionnement des interactions nous conduit à nous demander si, et comment, les mécanismes de coordination des fonctions cérébrales, mis en évidence par les neurosciences, peuvent éclairer cette question de la coordination des actions sous ses multiples aspects. Pour ce faire, nous partirons d'une analyse des difficultés rencontrées ici par la théorie des jeux sur la base d'exemples schématiques très simples. On montrera ensuite les différentes voies explorées pour proposer des solutions aux problèmes ainsi posés par la coordination. On dégagera, enfin, de quelle manière et à quelles conditions les progrès dans notre connaissance des processus cérébraux enrichissent aujourd'hui les modèles, au moyen desquels nous appréhendons cette question.

Choix stratégiques et points focaux

Commençons par les jeux à deux joueurs. La spécificité de cette problématique de la coordination a été, pour la première fois, mise en évidence et discutée par Schelling à partir de petits jeux 2 x 2, qualifiés, pour cette raison, de « jeux de pure coordination » (Schelling, 1960). Voici l'une de ses versions les plus simples, que nous baptiserons « jeu du rendez-vous ».

Deux joueurs ont rendez-vous en un lieu déterminé sans avoir précisé le point de rencontre. Deux points de rencontre sont également possibles, A, la cafétéria, et B, l'office d'information. L'unique problème posé aux deux joueurs par cette situation est de porter, ensemble, leur choix sur A, ou sur B ; faute de quoi ils ne se rencontreront pas. Sa solution n'est pas fournie par la théorie des jeux, ou plus exactement, ce jeu possède deux solutions, puisque AA et BB sont deux issues parfaitement équivalentes, qui conduisent l'une et l'autre au résultat recherché par les joueurs. Traduites dans le langage technique de la théorie des jeux, elles désignent deux équilibres de Nash identiques.



Jeu du rendez-vous

Leur accessibilité par les deux joueurs dépend donc exclusivement de leur choix commun de l'un des deux points. Ils ne réussiront à se rencontrer que s'ils choisissent, tous les deux, le même point de rencontre. Mais les informations dont chacun dispose sur le jeu et les hypothèses classiques de rationalité qui sont associées à leurs comportements ne sont pas suffisantes pour y parvenir. Les hypothèses renforcées visant une connaissance commune des règles du jeu et de leur rationalité ne sont ici d'aucun secours. Considéré sous cet angle, le problème posé par la coordination des actions individuelles se révèle effectivement insoluble avec les outils logiques développés par la théorie des

jeux. Pourtant, les expériences menées sur ce type de jeu montrent que, dans une large majorité des cas, les joueurs parviennent à se coordonner sur l'un des points. C'est donc sous la forme d'un constat critique, qui met en lumière une limite importante de la théorie des jeux, que Schelling introduit, dès les années 1960, sa réflexion sur la coordination.

Que faudrait-il aux joueurs pour qu'ils puissent se retrouver ? Que chacun soit capable d'imaginer le point de rencontre qu'ils choisiront tous les deux. Le problème ainsi posé aux joueurs est donc celui d'une coordination mentale de leurs anticipations. C'est la raison pour laquelle Schelling considère cette convergence des anticipations des deux joueurs comme le résultat d'un phénomène d'ordre psychique, non réductible à la rationalité abstraite du jeu et de ses règles. Pour mieux le cerner, il convoque du reste des idées antérieurement développées par le psychologue gestaltiste Koffka. Koffka note, par exemple, qu'au cours d'un match de football, la position du gardien de but constitue une sorte d'attracteur pour les buteurs de l'équipe adverse, à partir duquel s'enclenchent leurs tirs au but, bien qu'ils sachent en même temps, pertinemment, que le gardien de but protège sa cage[1]. En partant de ce type d'observations, Schelling suggère une voie possible pour résoudre ce problème de coordination, tel qu'il se pose aux joueurs du jeu du rendez-vous. Elle consiste à introduire ce qu'il nomme des « points focaux », qui seraient conjointement perçus par les deux joueurs et leur fourniraient des attracteurs, capables de les guider dans le choix de leurs actions. Si, dans notre jeu du rendez-vous, le point de rencontre A possède, par exemple, une affiche peinte en rouge vif attirant l'attention, cela peut servir de signal aux deux joueurs pour coordonner leurs anticipations sur A, plutôt que sur B.

Cette notion de point focal reste cependant chez Schelling assez vague et, surtout, éminemment contextuelle. On peut la définir comme un trait qui s'impose (*salience*) dans l'organisation mentale des informations sur la situation et se distingue suffisamment de l'ensemble, pour pouvoir attirer, en même temps, l'attention des deux joueurs. Nous retrouvons ici la notion de « saillance », utilisée au chapitre précédent comme un marqueur cognitif. Pour Schelling la recherche de points focaux est une affaire purement empirique à examiner cas par cas. D'autres théoriciens des jeux, ainsi que des philosophes, ne se sont pas satisfaits de cette interprétation empirique. Ils lui ont recherché une acception plus théorique, afin de comprendre comment ces points focaux émergent dans l'esprit des joueurs et s'imposent à leur jugement. Ce faisant, ils ont été conduits à transformer le cadre analytique de la théorie des jeux.

Différentes formules ont été proposées pour y parvenir (Sugden, 1995 ; Bacharach, 1993). Si toutes s'inspirent de cette démarche, la majorité d'entre elles recherchent la solution dans une extension de l'une ou l'autre des procédures déjà imaginées par la théorie des jeux, ou seulement imaginables pour rendre compte des choix stratégiques des joueurs. Pour certains, la procédure retenue consiste à remonter la hiérarchie des croyances (je crois que ce point sera saillant ; je crois que l'autre joueur croit que ce point sera saillant ; je crois que l'autre joueur croit que je crois, etc.) (Stahl et Wilson, 1995 ; Camerer, Ho et Chong, 2004). Pour d'autres, la procédure privilégiée consiste à adopter le point de vue qui serait celui d'un agent collectif, représentatif du groupe, le « nous », à la

recherche d'un optimum collectif (Bacharach, 1999). Parmi eux, enfin, on trouve ceux qui considèrent que cet agent collectif se caractérise par ses intentions collectives, et ceux pour qui sa singularité réside dans un mode de raisonnement différent, mais cependant compatible, avec l'individualisme logique de la théorie des jeux (Gold et Sugden, 2007).

Ces propositions ont en commun d'accepter, sans vraiment la discuter, l'hypothèse selon laquelle la question de la coordination peut être appréhendée, par les joueurs, comme un choix rationnel dont il conviendrait seulement, pour y répondre, d'en modifier la formulation et/ou le mode opératoire. Avant de nous prononcer sur la validité de cette hypothèse, il importe de remonter en amont, au niveau de la représentation par les joueurs et de leur perception de la situation. Pour y parvenir, nous nous proposons d'appliquer au jeu du rendez-vous une procédure imaginée et développée par Bacharach dans différents travaux (1993, 1997, 2006).

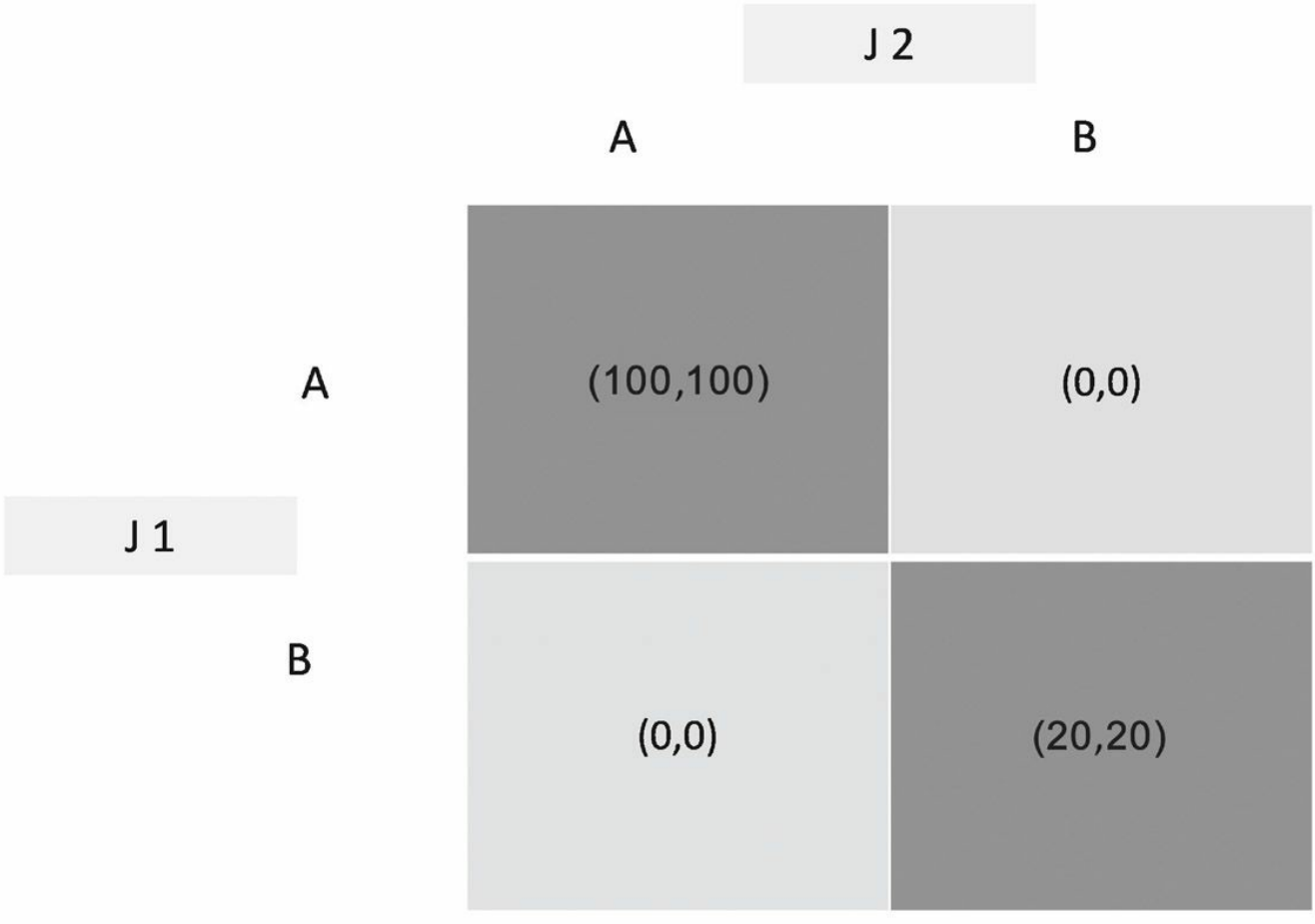
Sa première étape consiste à substituer à la description objective et statique du jeu qui reste la même tout au long de son déroulement (*cf.* Matrice 1), une description subjective et dynamique propre à chaque joueur, en fonction de sa situation, variant durant son déroulement. C'est la raison pour laquelle, Bacharach baptise « jeux à univers variable » cette nouvelle formulation (Bacharach, 1993). Une telle description des états du jeu dépend étroitement de l'action que les joueurs doivent entreprendre, et, par conséquent, de leurs intentions, ce que certains auteurs appellent des « descriptions-actions ». Dans l'exemple du jeu du rendez-vous, l'intention de chaque joueur est de rencontrer l'autre joueur en un point précis. Il convient donc d'ajouter aux états résultant des choix de l'une ou l'autre des deux stratégies possibles (se rendre en A, se rendre en B), les caractéristiques singulières de A et de B. S'agissant de représentations mentales, cela implique que les informations détenues par les joueurs sur ces points de rencontre s'organisent de manière significative au regard de l'objectif poursuivi. Les caractéristiques matérielles de chacun des points de rencontre A et de B (taille, forme, couleur…) deviennent alors des éléments de la description du jeu, susceptibles d'infléchir son déroulement dans la direction de l'objectif visé.

En retranscrivant le modèle de référence du jeu de pure coordination en modèles de jeux mentaux prêtés aux deux joueurs, on retrouve une idée chère aux psychologues cognitivistes, selon laquelle la manière dont les individus se représentent un choix dépend de son cadrage (*framing*). Ce cadrage résulte lui-même d'opérations mentales, au cours desquelles les informations pertinentes sont identifiées, sélectionnées et organisées. Dans leurs premiers travaux, Kahneman et Tversky avaient pris soin de distinguer, sous le terme d'*editing*, l'opération au cours de laquelle les informations sont sélectionnées et hiérarchisées, en la différenciant du *framing* lui-même (Kahneman et Tversky, 1979). Pour ce qui concerne la coordination, c'est au cours de cette opération d'*editing* que les points de saillance qui servent de support à la coordination peuvent s'imposer en même temps à l'esprit des deux joueurs. L'analyse des liens entre perception visée et attention, étudiés par Husserl dans la perspective d'une phénoménologie de l'attention fournit sur cette question un éclairage intéressant que prolongent aujourd'hui plusieurs recherches de neurosciences[2].

Une fois le modèle initial de jeu réécrit en termes de « descriptions-actions », les théoriciens des jeux traitent la sélection des points focaux retenus par les joueurs comme un choix « justifié », qui ne

serait guère différent, pour cette raison, d'un choix stratégique. Dans notre exemple du jeu du rendez-vous, chacun des joueurs choisira de se diriger vers le point A, en pensant que le rouge vif de sa peinture attire également l'attention de l'autre joueur. Ce résultat se trouve obtenu en élargissant, comme on l'a dit, l'ensemble des données retenues par les joueurs au moment de leur choix stratégique. Au lieu de considérer seulement les paiements associés aux deux stratégies pures à leur disposition, comme dans la théorie des jeux classiques, ils prennent également en compte dans leur calcul, la forme, la taille et la couleur des deux points de rencontre, au terme d'une opération d'*editing*. Les paiements n'étant pas ici discriminants, ils formulent le problème auquel ils se trouvent confrontés en ces termes : qu'y a-t-il de saillant dans ces données ? Après examen, ce n'est ni la taille ni la forme des deux points de rencontre, mais leur couleur, puisque le rouge vif de l'un d'entre eux attire l'attention. Chaque joueur en déduit que l'autre choisira pour cette même raison, comme lui-même, la stratégie A. Ce choix commun leur permettra ainsi de se retrouver au point de rencontre de couleur rouge vif. L'opération mentale qui s'avère déterminante pour y parvenir n'est pas ici le choix de la stratégie, mais la recherche et la sélection préalable d'indices coordinateurs. Une fois un tel indice identifié, le choix stratégique n'est rien d'autre qu'une forme de validation automatique. Mais rien ne permet, à ce stade de l'analyse, d'assimiler l'opération de recherche et de sélection d'un indice coordinateur à un véritable choix stratégique.

Bacharach, cependant, ne voit pas les choses ainsi. Le processus qu'il imagine consiste à combiner deux modes de raisonnements distincts, mais complémentaires, pour aboutir à un choix rationnel, en inversant seulement leur priorité par rapport à l'interprétation traditionnelle. Un optimum au sens économique possède aussi la propriété de pouvoir représenter un point focal, à partir duquel les joueurs peuvent coordonner leurs actions. Identifier et choisir cet optimum permettraient ainsi de garantir cette coordination. Il ne resterait plus alors qu'à interpréter un tel choix en termes de rationalité, d'où l'introduction de l'équipe (*team*) comme sujet véritable de ce choix (Bacharach, 2006). Ce schéma habile ne permet pas, cependant, d'expliquer comment nos deux joueurs du jeu du rendez-vous se coordonneront sur AA, puisque dans ce jeu, AA et BB, sont, non seulement, deux équilibres, mais correspondent également chacun à un optimum. De manière plus générale, la solution proposée par Bacharach pour résoudre le problème posé aux joueurs de la théorie des jeux par la coordination de leurs choix, n'est pas applicable à des jeux de pure coordination comme celui du rendez-vous.



Jeu haut/bas

Il est intéressant de noter, à ce sujet, que Bacharach, pour exposer sa solution, prend appui sur un exemple différent de jeu de coordination, qu'il appelle haut/bas (Hi, Lo) et utilise comme une sorte de paradigme de la coordination. En voici une illustration. Deux joueurs placés devant un carré ont à choisir entre deux options, le haut, noté A et le bas, noté B. Si tous les deux choisissent le haut, c'est-à-dire A, ils reçoivent chacun une somme de 100 euros. Si tous les deux choisissent le bas, c'est-à-dire B, ils reçoivent chacun 20 euros. Mais si l'un choisit A et l'autre B, ils ne gagnent rien, ni l'un ni l'autre. Traduit dans la terminologie de la théorie des jeux, AA et BB sont deux équilibres, au sens de Nash, mais seul AA est un optimum social et représente pour cette raison, selon Bacharach, le point focal qui permet aux joueurs de coordonner le choix de leur stratégie.

Le problème de coordination posé par ce jeu haut/bas n'est pas cependant le même que celui posé aux joueurs du jeu du rendez-vous. Dans le jeu haut/bas, les deux joueurs n'ont, en effet, aucune peine à choisir chacun A et à se coordonner sur AA. Dans le jeu du rendez-vous, la question de la coordination se présente aux joueurs de manière tout à fait indépendante d'une rationalité dérivée de l'optimalité de leurs choix. On constate que, tandis que la solution de la coordination dans le jeu haut/bas peut être trouvée en transformant seulement la représentation du jeu et les modalités de raisonnement des joueurs (équipe *versus* individus), celle de la coordination du jeu du rendez-vous

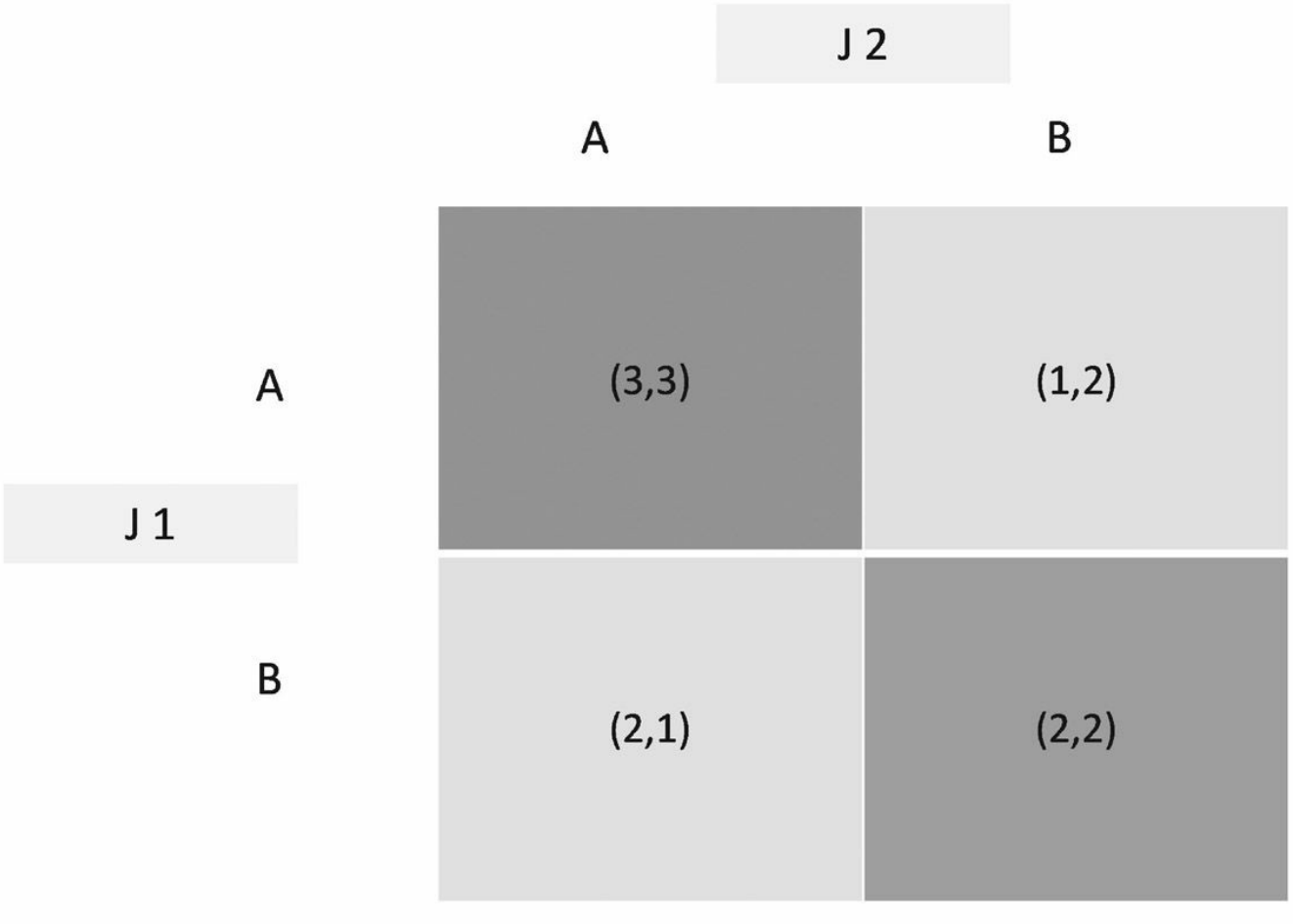
exige de rechercher et de traiter des informations qui sont extérieures au jeu lui-même. Dans notre exemple, il suffit même que chaque joueur choisisse sa stratégie dominante. Il en résulte que la notion de point focal n'a, ni exactement la même acception, ni surtout la même portée dans les deux cas.

Contrairement, par conséquent, à ce que pensent une majorité de théoriciens des jeux et d'économistes, il n'y a pas une coordination mais des coordinations qui posent des problèmes distincts aux joueurs qui y sont confrontés, selon les différentes situations où elles se présentent. On a vu, dans les deux exemples qui ont été rapportés, que l'hypothèse selon laquelle les problèmes posés par la coordination pouvaient toujours être ramenés à une opération de choix rationnel ne se trouve pas validée. L'idée selon laquelle la coordination, à travers ses diverses manifestations, correspondrait à une seule et même catégorie d'opération renvoyant aux mêmes processus mentaux chez les sujets est, comme nous le verrons, à l'origine de différentes méprises et, tout au moins, d'erreurs d'interprétation.

C

Pour éviter ces pièges, une première étape consiste à préciser la nature des différences entre ces divers problèmes de coordination. Revenons sur la formulation générale de la problématique de la coordination en théorie des jeux. Elle se manifeste, comme on l'a indiqué, lorsque l'interaction de joueurs rationnels conduit logiquement à plusieurs points d'équilibre entre lesquels la théorie ne permet pas d'arbitrer. Posée en ces termes la coordination relie en une même et seule question, la sélection de l'équilibre qui sera (ou devrait être) retenue, et les voies et moyens suivis par les joueurs pour y parvenir. Afin de montrer comment la relation entre ces deux composantes de la coordination est susceptible de varier, nous ajouterons aux deux exemples précédents, celui d'un troisième, déjà introduit au chapitre 2, celui de la chasse au cerf.

Les joueurs de ce jeu sont, cette fois, plus nombreux, puisqu'il s'agit de chasseurs.



Jeu de la chasse au cerf

C’est la raison pour laquelle cet exemple sert également de support à l’étude de la coordination dans le cas de jeux où nos deux joueurs emblématiques sont étendus à des populations de joueurs. On identifie alors deux catégories de joueurs par la stratégie qu’ils ont retenue, d’où la transformation du jeu en un jeu à deux joueurs. D’autres formulations, plus complexes, de nature dynamique, ont également été proposées, elles seront évoquées dans la dernière partie de l’ouvrage. Compte tenu des conditions supposées dans lesquelles opèrent les chasseurs, ils peuvent, soit chercher à attraper ensemble un cerf, ce qui correspond à la stratégie A ; soit capturer, chacun de leur côté, un lièvre, en adoptant la stratégie B. S’ils choisissent tous A, ils bénéficieront chacun d’un butin plus important, puisqu’un morceau de cerf a plus de valeur qu’un lièvre. Si chacun, au contraire, porte son choix sur B, tous s’assurent d’une prise minimale, certes plus modeste, mais indépendante des choix effectués par les autres chasseurs. Si, enfin, les uns choisissent A, tandis que d’autres optent pour B, ces derniers auront encore gagné un lièvre, contrairement aux autres qui reviendront bredouilles. Traduit dans le langage de la théorie des jeux, la situation décrite se trouve résumée par la matrice suivante :

Ici encore, AA et BB, comme dans les exemples précédents du jeu du rendez-vous et du jeu haut/bas sont deux équilibres ayant les mêmes propriétés formelles. La sélection de l’un d’entre eux y dépend, également, de la manière dont les joueurs choisiront de se coordonner, qui échappe, pour les

mêmes raisons, aux outils analytiques classiques de la théorie des jeux.

La question de la coordination se pose cependant de manière encore différente pour les chasseurs de la chasse au cerf et pour les joueurs des deux autres jeux précédemment analysés. En choisissant leur stratégie A, les chasseurs optent pour une coopération, ce qui n'est le cas, ni des joueurs du jeu du rendez-vous ni de ceux du jeu haut/bas. Or coopérer ne signifie pas seulement se coordonner. Le choix de coopérer est intentionnel. Il manifeste, de la part de chaque joueur, son intention de coopérer avec les autres dans une action commune. Il repose, de ce fait, sur une confiance mutuelle qui contrebalance le risque de défection. On comprend mal comment les joueurs des deux jeux précédents pourraient rationnellement préférer ne pas se coordonner, plutôt que de se coordonner. La situation n'est pas la même dans la chasse au cerf, où, faute de confiance et par aversion au risque, les joueurs peuvent préférer ne pas coopérer en toute rationalité ; d'où la référence, chez certains auteurs, à des métacritères qui ont été présentés au chapitre 2 (Harsanyi et Selten, 1988). D'autres auteurs voient plutôt une forme d'irrationalité dans le choix de ne pas coopérer. Selon eux, en effet, la rationalité étant supposée connaissance commune entre les joueurs, une telle défiance serait, de ce fait, assimilable à un défaut de rationalité (Aumann, 1990). Nous avons montré au chapitre 2 le caractère inefficace de cette hypothèse de connaissance commune pour comprendre comment les joueurs interagissent.

La comparaison de ces trois jeux met en lumière les différentes acceptions que prend la coordination dans les trois situations. Elle permet ainsi de préciser les opérations mentales requises de la part des joueurs dans chacun des cas. Se coordonner, c'est identifier un point d'accroche mental commun dans le jeu du rendez-vous, c'est choisir l'articulation des stratégies garantissant les meilleurs résultats pour tous dans le jeu haut/bas, c'est organiser une coopération commune, dans le jeu de la chasse au cerf. Il existe certes des relations entre ces différentes opérations. Sans une focalisation commune des esprits, il n'est pas de coordination possible. Un chasseur sceptique qui la refuserait exclurait, par exemple, toute coordination coopérative pour attraper le cerf. C'est la raison pour laquelle, la référence mentale à des points focaux supposés communs se présente comme une exigence incontournable pour le travail cérébral de tout individu en quête de coordination. Plusieurs modes d'accès sont également possibles dans la recherche de ces points focaux. Ils dépendent de la situation particulière dans laquelle se présente aux individus leur quête d'une coordination. Cette coordination peut, enfin, être poursuivie par les individus dans des buts différents : lever une incertitude, comme dans le jeu du rendez-vous, éliminer une possibilité fallacieuse, comme dans le jeu haut/bas, coopérer dans une action commune, comme dans le jeu de la chasse au cerf.

L'analyse comparative de ces trois exemples de coordination initialement tirés de la théorie des jeux n'est, certes, pas suffisante pour dégager une typologie complète des différentes formes de coordination. La mise en évidence de leur diversité rend, en tout cas, hypothétique, pour ne pas dire simplement utopique, la recherche de leur regroupement autour d'un principe unificateur. Des tentatives ont du reste été effectuées pour tester expérimentalement la validité de deux processus de raisonnement distincts, si ce n'est rivaux, le plus souvent avancés pour expliquer la coordination : une itération par niveaux cognitifs (je crois, je crois qu'il croit, je crois qu'il croit que je crois, etc.) ; un

raisonnement d'équipe (en tant que membre du groupe formé par l'ensemble des joueurs je m'identifie au « nous »). Elles ont jusqu'à présent invalidé l'hypothèse que l'une ou l'autre de ces modalités de raisonnement pouvait, toute seule, fournir ce principe unificateur (Bardsley *et al.*, 2008).

Si la théorie des jeux offre un cadre d'analyse à la fois simple et logique pour appréhender la coordination stratégique de plusieurs individus doués de raison, la simplification qu'elle introduit est également à l'origine de confusions génératrices d'erreurs méthodologiques. Une première confusion apparaît entre le choix d'une action et son but. Dans le cas de la pure coordination, comme le jeu du rendez-vous, le but visé est de faire émerger une règle. Dans le jeu haut/bas, cette règle se confond avec la simple maximisation d'un gain, d'où une ambiguïté. Dans le jeu de la chasse au cerf, les deux objectifs se trouvent distincts, mais étroitement dépendants, ce qui engendre une plus grande complexité. Une seconde confusion concerne, cette fois, l'intentionnalité qui guide les individus et leur mode de raisonnement (Gold et Sugden, 2007). Il est clair que, quel que soit le mode de raisonnement utilisé par les joueurs du jeu du rendez-vous, leur intention est seulement de pouvoir se coordonner. Telle n'est pas la situation des joueurs du jeu haut/bas, pour lesquels la solution logique de cette coordination passe par un raisonnement d'équipe, indépendant, cependant, de toute intention collective. Dans le jeu de la chasse au cerf, au contraire, c'est l'intention de coopérer qui conduira les joueurs à adopter le raisonnement d'équipe. Nous verrons, par la suite, que ces confusions continuent de fausser les hypothèses et de brouiller souvent l'interprétation des travaux de psychologie expérimentale et, plus récemment, d'imagerie cérébrale, en entraînant des comparaisons fallacieuses entre les différentes situations de coordination.

S'il faut néanmoins trouver une unité dans la diversité des types de coordination auxquels sont confrontés les individus, ce n'est, comme on l'a montré, ni au niveau des intentions, ni au niveau des modes de raisonnement, ni même au niveau des formes de représentation de la situation. Il semble, en revanche, que la coordination résulte, chaque fois, d'une série différente d'opérations, qui ont toutes pour origine une recherche préliminaire d'ancrages mentaux communs. Nous nous proposons de qualifier de niveau 0 de la coordination cette étape, puisque, c'est à partir d'elle que pourront s'enclencher des mécanismes cérébraux, plus complexes, destinés à coordonner les actions, en fonction des singularités du problème posé par chaque situation. C'est à ce stade qu'il convient d'introduire les points focaux de Schelling, dans un arrière-fond mental, qui conditionne, en définitive, le succès de tout phénomène de coordination. Dans certaines situations, comme dans le jeu haut/bas, l'identification de ces points focaux est presque automatique ; dans d'autres, elle requiert, au contraire, un travail mental conséquent encore mal connu. Les jeux de pure coordination, comme celui du rendez-vous, appartiennent à cette seconde catégorie. C'est la raison pour laquelle nous avons privilégié son étude, à titre d'introduction aux multiples formes que revêt le phénomène de coordination.

P

La recherche et la sélection de points focaux constituent, comme nous venons de l'expliquer, la première étape du travail mental des joueurs lorsqu'ils se trouvent dans une situation de pure coordination. On peut, en première approximation, considérer ces points focaux comme des moyens heuristiques destinés à aider chacun des joueurs à dégager (ou à découvrir) une règle implicite dont la mise en œuvre permettra leur coordination. Rappelons qu'en théorie des jeux, les règles sont supposées données et parfaitement connues des joueurs, à la manière des règles des jeux de société. Choisir une stratégie représente, pour cette raison, l'unique problème posé aux joueurs, qui se trouve traité par la théorie des jeux. La question qui leur est posée ici est cependant de nature différente et se présente en quelque sorte en amont de ce qu'étudie la théorie des jeux. Il faut pour les joueurs imaginer un moyen de trouver une règle de coordination, à partir de connaissances qu'ils croient pouvoir partager.

Des travaux récents de neurosciences permettent de jeter quelques lumières sur les modalités de l'activité cérébrale des sujets en quête de points focaux, dont une commune référence assurerait leur coordination. Une équipe de chercheurs a observé expérimentalement les performances, et examiné, par imagerie, le fonctionnement cérébral d'un groupe de sujets pendant qu'ils jouaient à deux versions différentes d'un jeu de pure coordination, cela pour éviter un éventuel biais de présentation. Dans le premier jeu, chaque joueur devait choisir un nombre entre 1 et 3 ; dans le second jeu, une boîte définie par une matrice à 3 dimensions. Chaque fois, l'objectif fixé était de choisir le même nombre dans le premier cas, la même boîte dans le second cas, que l'autre joueur. Ces deux jeux ont été transformés en jeux de coordination d'un type différent, en y adjoignant des règles supplémentaires permettant aux joueurs de résoudre le problème de leur coordination en suivant une procédure d'éliminations successives des stratégies dominées, au demeurant bien connue en théorie des jeux (Kuo *et al.*, 2009). Ce dispositif a ainsi permis de comparer les modalités du travail cérébral des sujets lorsque la coordination leur pose un problème soluble par une règle logique simple, et lorsqu'une telle règle logique n'existe pas.

Un premier résultat retient l'attention. Une large majorité des sujets a réussi, dans les deux types de jeux, à se coordonner sur une solution d'équilibre, c'est-à-dire à résoudre le problème différent qui leur était chaque fois posé ; soit dans une proportion de près de 80 % dans les jeux de coordination solubles par la théorie et près de 70 % dans les jeux de pure coordination. Des performances qui tranchent avec les résultats obtenus, lorsque ces jeux sont joués dans les mêmes conditions, sans règles déterminées. Tout porte donc à penser que la coordination émane, dans les deux cas, d'une opération mentale délibérée de la part des sujets. Plus surprenant peut-être, la proportion des joueurs ayant réussi à se coordonner dans les jeux de pure coordination est à peine inférieure à celle de ceux qui se sont coordonnés sur l'équilibre dans les autres jeux. On peut raisonnablement en induire que la pure coordination est le fruit d'un travail de l'esprit, tout comme l'est la coordination résultant de l'application d'une règle, en dépit du fait que, dans le premier cas, le travail effectué par le cerveau ne coïncide pas avec un traitement formel simple. Le fait avéré que les joueurs mettaient moins de temps et ressentaient moins d'effort pour se coordonner dans la version « pure coordination » de ces jeux que

dans leur version « coordination logique » suggère que le travail mental qu'il exige relève plutôt de la catégorie des heuristiques « rapides et frugales » que Gigerenzer associe à l'intuition (Gigerenzer, 2000 ; Gigerenzer et Gaissmaier, 2011). Mais cette qualification reste insuffisante pour caractériser le mécanisme mental qui guide cette forme de coordination.

Pour avancer dans cette direction, cette recherche a recueilli par la technique de l'imagerie fonctionnelle (IRMf) des informations sur le fonctionnement du cerveau des joueurs dans les situations différentes de chaque type de jeu. Comme on pouvait s'y attendre, l'IRMf révèle que ce sont des régions cérébrales traditionnellement dévolues au calcul, au raisonnement abstrait et à la mémoire immédiate, comme certaines parties du cortex frontal, et notamment le milieu du gyrus, ainsi que la région pariétale, qui sont prioritairement activées au cours du jeu, dont la solution dépend d'un traitement logico-mathématique. Ce sont, en revanche, d'autres aires cérébrales qui se trouvent principalement mobilisées dans les jeux de pure coordination. On observe, en particulier, une activation du cortex cingulaire antérieur (ACC), et d'une partie de l'insula. On sait que l'ACC réagit chaque fois que nous sommes confrontés à une situation conflictuelle génératrice d'incertitude cognitive (dilemme, impasse, devinette…) qui requiert attention et recherche d'informations. Quant à l'insula, elle se trouve notamment impliquée lors de variations, subies ou attendues, d'états physiologiques internes à l'origine, en particulier, d'impressions désagréables. Elle se manifeste aussi sous forme d'émotions empathiques à l'endroit de comportements observés ou imaginés chez les autres.

Les résultats d'une autre étude, interprétés pourtant par leurs auteurs comme contradictoires par rapport à ceux de ce premier travail, nous semblent, au contraire, les confirmer, *a contrario*. Elle émane d'une équipe de chercheurs qui a récemment conçu et réalisé une version expérimentale stylisée et séquentielle du jeu de la chasse au cerf, entre un individu et un ordinateur intelligent (Yoshida *et al.*, 2010). L'IRMf des sujets soumis à cette expérience a révélé que pour se coordonner sur la solution coopérative, les sujets activaient prioritairement des régions cérébrales préfrontales, comme le cortex paracingulaire et dorsomédian, des zones du cerveau qui sont précisément mobilisées dans le calcul stratégique à plusieurs niveaux. Ces chercheurs ont observé, en outre, que la partie ventrale du striatum de ces sujets se trouvait fortement activée au cours de cette expérience, une région cérébrale associée à l'anticipation des gains risqués. Ces chercheurs en ont déduit, un peu rapidement, que, d'une manière générale, le travail mental requis pour la coordination mobilisait davantage, chez les individus, des systèmes neuronaux prioritairement affectés aux opérations de calcul que des systèmes qui seraient principalement sollicités par des réactions émotives.

Ce jugement nous semble cependant reposer sur une interprétation erronée de la comparaison entre les deux groupes d'expériences. Un premier biais se trouve introduit par les caractéristiques différentes de ces expériences. Alors que la série des jeux de pure coordination de la première équipe sont des jeux joués en un coup, les autres sont des jeux séquentiels, au cours desquels les joueurs se trouvent amenés à traiter, par induction, les informations qui leur sont progressivement transmises. En outre, les jeux de coordination de la première équipe sont joués entre des joueurs humains, tandis que le jeu de la chasse au cerf de la seconde équipe est joué entre des individus et un ordinateur, ce dont

les joueurs sont informés. Cette différence peut également être à l'origine de biais. Mais l'élément le plus important concerne la nature de l'opération de coordination mise en évidence dans les deux catégories d'expériences. Comme nous l'avons déjà dit, dans un jeu de pure coordination, le problème posé aux joueurs est de discerner une règle implicite leur permettant de coordonner leurs actions, tandis que dans un jeu de coordination coopérative, comme celui de la chasse au cerf jouée de manière séquentielle, il leur faut anticiper correctement la stratégie de l'opposant, afin de choisir la meilleure réponse stratégique. Ainsi réinterprété, ce jeu de la chasse au cerf a plus d'affinités avec les jeux d'itérations hiérarchiques construits par la première équipe de chercheurs. Le fait que ses joueurs activent dans les deux cas des zones cérébrales localisées dans les cortex frontaux et préfrontaux confirme seulement, au niveau neuronal, la différence entre les opérations mentales requises pour choisir des réponses stratégiques coordonnées (jeux de la chasse au cerf, élimination itérative des stratégies dominées) et celles qui conduisent à l'identification de points focaux partagés (jeux de pure coordination).

Pour mieux cerner le fonctionnement cérébral au cours de cette seconde opération il convient d'examiner plus en détail la modalité des relations entre les deux régions du cerveau qui se sont révélées activées durant la recherche de points focaux partagés, de l'ACC et de l'insula. Différents travaux de neurophysiologie ont mis en évidence plusieurs connexions fonctionnelles entre ces deux régions cérébrales. De manière générale, ces connexions visent à relier la détection de saillances au choix de réactions appropriées. Dans le prolongement de cette hypothèse, certains chercheurs ont récemment dégagé deux systèmes distincts, en détaillant les localisations des différentes régions cérébrales ainsi connectées : un système large, reliant la totalité de l'insula à une partie médiane du cortex cingulaire, et un système plus restreint, reliant la partie antérieure de l'insula à une partie de l'ACC. Le petit système capterait les saillances ressenties au niveau des émotions subjectives internes, tandis que le grand système concernerait plus généralement les saillances perçues au niveau de l'environnement externe, en vue d'une réaction (Taylor *et al.*, 2008). S'il se confirme que ces deux systèmes participent à la recherche de points focaux en vue d'une pure coordination, il serait intéressant d'analyser comment ils s'articulent, puisque cette opération combine l'identification de saillances extérieures avec les réactions sensibles qu'elles suscitent. Or, par un mécanisme d'empathie, l'intensité des émotions subjectivement ressenties pourrait contribuer à renforcer, chez les sujets, leur croyance qu'elles pourraient être partagées, fournissant, de ce fait, un indice pour leur sélection en vue d'une coordination.

De tels éléments restent fragmentaires et partiels et font encore l'objet de controverses entre les chercheurs. Il serait donc naïf et prématuré de tirer de ces premières indications un modèle neuronal complet permettant d'expliquer le fonctionnement de la pure coordination. Elles conduisent toutefois à formuler quelques hypothèses qui ouvrent des pistes de recherches intéressantes. Ce serait ainsi au niveau des relations entre la perception des saillances de l'environnement et le ressenti subjectif, que l'on pourrait repérer le point de départ du processus mental qui aboutit à la coordination. Sur cet arrière-fond sensible s'élaborerait alors la trame d'un raisonnement, qui, selon des formes diverses,

intervient toujours dans la coordination. Des recherches ont, en effet, montré que des malades souffrant d'une maladie dégénérative du cortex préfrontal éprouvaient de grandes difficultés à coordonner leurs réponses dans une épreuve de pure coordination sémantique (choisir les mêmes prénoms…) (MacMillan *et al.*, 2011). Il est vrai que cette affection atteint également le cortex cingulaire antérieur (ACC), dont on a signalé le rôle majeur dans l'identification des points focaux. Une raison supplémentaire pour approfondir les recherches sur les relations fonctionnelles entre les différentes catégories de réseaux neuronaux qui sont activés au cours des différents types de coordination.

D

Considérons maintenant les situations de coordination, où chaque sujet individuel doit coordonner son (ses) action(s), non plus avec celles d'un autre sujet individuel, mais avec celles d'une multitude d'autres sujets qui forment une population d'individus. En changeant ainsi de dimension, l'explication de la coordination n'est plus ici réductible au seul niveau interindividuel. Il faut introduire dans l'analyse de nouveaux paramètres relatifs, notamment, à la taille de la population concernée (masse, petits groupes), à sa structuration (organisation en réseaux ouverts ou fermés, types de connexions) et aux modalités de cette structuration (choix des appartenances, formation au hasard). Du point de vue de la théorie des jeux, le problème posé aux théoriciens par la coordination reste néanmoins le même. Les joueurs se trouvent confrontés à des situations dans lesquelles il existe plusieurs équilibres, dont la solution, c'est-à-dire l'identification de l'équilibre qui s'imposera, n'est pas fournie par les outils d'analyse traditionnels de la théorie.

Les modèles conçus par les théoriciens des jeux pour le traiter sont toutefois un peu différents. La coordination ne peut plus être saisie maintenant dans le cadre schématique de petits modèles à deux joueurs, à un coup, comme nous l'avons montré au début de ce chapitre. Elle est supposée, cette fois, se dégager au cours d'une période de temps, plus ou moins longue. Les informations initiales dont disposent (ou ne disposent pas) les joueurs se trouvent maintenant complétées par des informations endogènes que les joueurs induisent du déroulement du jeu ; d'où la prise en compte d'effets d'apprentissage et de séquences d'essais et d'erreurs. Il est difficile dans ce contexte de retrouver les questions de pure coordination telles qu'elles ont été mises en évidence dans le jeu du rendez-vous. La coordination se trouve traitée ici comme un processus dynamique de réactions aux informations, dont il importe de distinguer l'impact en fonction de leurs différents types (informations publiques, informations privées, endogènes, exogènes…). Ces informations, enfin, ne peuvent être rapportées par chaque joueur à un individu déterminé, même, parfois, à un groupe d'individus qui seraient, chaque fois, clairement identifiés ; d'où la prise en compte d'une dimension statistique. Nous retrouvons ainsi les effets d'action de masse déjà anticipés par Nash dans sa thèse de doctorat (Nash, 1950) et évoquée au chapitre 2. La notion d'équilibre de coordination, enfin, n'a ni la

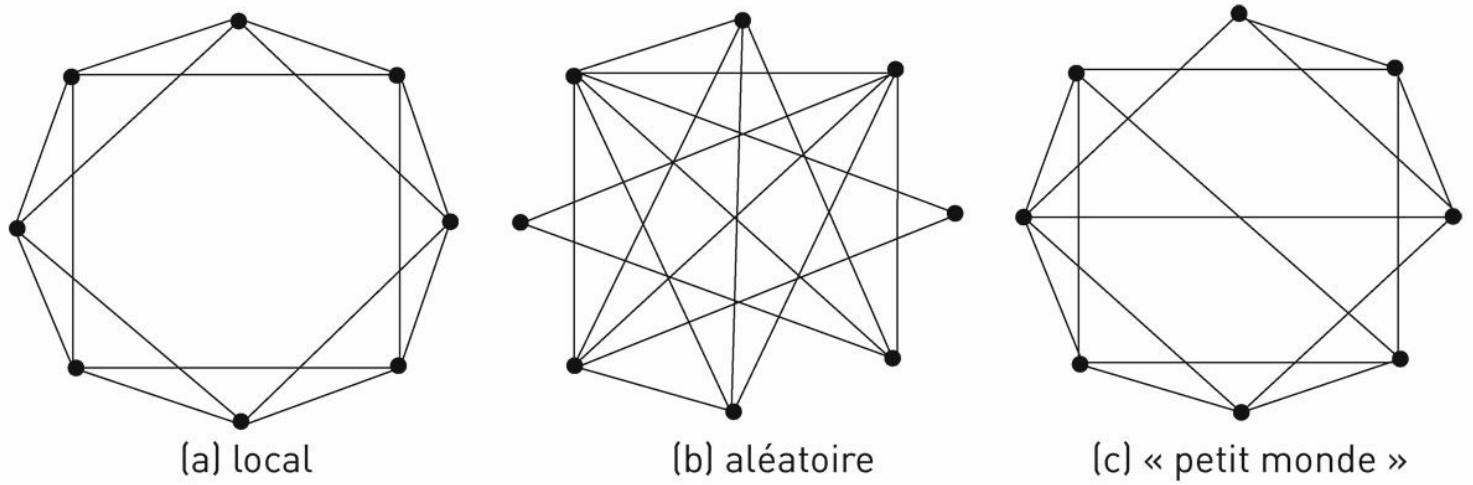
même portée ni les mêmes propriétés et, par conséquent, la même définition, lorsqu'elle se trouve appréhendée dans ce nouveau contexte. Dans un univers dynamique, l'équilibre sur lequel se coordonnent les actions des joueurs doit pouvoir être stratégiquement stable, afin d'éviter des mutations vers d'autres états du système, génératrices d'instabilités. Cela explique les liens étroits de ces modèles avec l'approche évolutionniste, qui a inspiré de nouveaux concepts d'équilibre, dérivés de la biologie et de la génétique, à partir de stratégies appréhendées comme évolutionnairement stables (ESS) (Maynard Smith, 1981 ; Weibull, 1985).

Ce nouveau cadre a étendu la problématique initiale des jeux de coordination. Les échecs qui menacent la coordination prennent la forme de deux phénomènes distincts, parfois regroupés sous un même terme dans les travaux qui leur sont consacrés. D'une part, les joueurs risquent de ne parvenir à se coordonner sur aucun équilibre, d'où une instabilité permanente. D'autre part, ils peuvent se coordonner sur un équilibre, mais cet équilibre est sous-optimal. Ainsi, par exemple, pour reprendre les termes utilisés au début de ce chapitre dans l'analyse du jeu de la chasse au cerf, les joueurs choisiront, sur la base d'arguments évolutionnistes tirés de l'apprentissage, de se coordonner sur l'équilibre *Risk dominant*, alors que c'est l'équilibre *Payoffs dominant* qui est optimal (L. Samuelson, 1997). C'est du reste cette seconde acception que la théorie des jeux retient le plus souvent dans ses analyses des échecs de la coordination, lorsqu'ils sont analysés dans une perspective dynamique. Enfin, les joueurs peuvent basculer d'un équilibre à un autre de manière parfois difficile à prévoir – un phénomène dont les ressorts dynamiques sont plus difficiles à cerner –, mais dont on observe fréquemment les effets, en particulier sur les marchés financiers. On notera que la durée et l'augmentation du nombre des informations accessibles aux joueurs ne semblent pas nécessairement favoriser leur coordination, comme on serait tenté intuitivement de le croire.

Plusieurs modèles théoriques ont été proposés pour rendre compte de ces situations, en partant, le plus souvent, du problème posé par la sélection d'un équilibre, dans l'hypothèse de populations de joueurs (Samuelson, 1997). Ils ont fait l'objet de simulations dynamiques de longue période sur la base de chaînes de Markov. Il s'avère, par exemple, difficile, pour cette raison, dans le cas d'un jeu du type de la chasse au cerf, de déterminer sur lequel des deux équilibres les joueurs finiront par se coordonner. Des recherches plus récentes ont complété ces travaux, en intégrant dans l'analyse le mode d'organisation des joueurs, amorçant ainsi une modélisation dynamique des réseaux, comme on le verra à la fin de ce chapitre. L'un des modèles les plus simples qui ait été imaginé vise à intégrer le coût que représente, pour chaque individu, la création d'un lien avec les autres individus, ce qui permet de supposer que le réseau s'organise sur la base d'un simple calcul coût-avantage, dont dépend le mode de coordination qui s'imposera. Selon ce raisonnement, on suppose que si les coûts d'entrée dans le réseau sont bas, les agents auront tendance à penser que leur connexion ne les protège pas contre le risque de comportements individualistes de la part des autres, de telle sorte qu'ils choisiront de se coordonner sur un équilibre qui minimise ce risque (*Risk dominant*). Si, au contraire, ces coûts sont suffisamment élevés, on suppose qu'ils choisiront, sur la base d'un raisonnement identique, d'entrer dans le réseau en se connectant aux autres, ce qui les conduit, alors, à se coordonner sur l'équilibre qui maximise leurs gains collectifs (*Payoffs dominant*) (Goyal et Vega-Redondo, 2005).

Ces travaux restent théoriques. Leurs résultats se fondent, comme on l'a montré, sur des hypothèses de comportements individuels rationnels. Leur exposition à des tests expérimentaux se heurte à l'obstacle du nombre des sujets testés, ce qui limite la représentativité des résultats obtenus dans ces expériences effectuées sur des échantillons relativement réduits. Indépendamment de protocoles expérimentaux différents, qui traduisent des phénomènes souvent distincts, la comparaison de leurs résultats rencontre deux difficultés supplémentaires. En premier lieu, ces expériences s'avèrent sensibles au nombre des joueurs et des séquences temporelles considérées. En second lieu, et peut-être surtout, la question se pose du degré et de la nature de la dépendance (ou de l'indépendance) de leurs résultats à la manière dont les joueurs sont organisés (ou s'organisent) en réseaux, au cours du jeu.

On dispose encore d'assez peu de travaux sur l'incidence des différents types d'organisation en réseaux sur la coordination des individus. Pour avancer sur cette question, il importe tout d'abord de distinguer les systèmes « clos » et « complets », où tous les individus se trouvent directement liés les uns aux autres, des systèmes « locaux » et « incomplets », où les individus ne sont directement connectés qu'à un sous-groupe de cette population. Dans cette seconde catégorie, on peut encore distinguer, selon que ce lien direct est déterminé (« réseau local ») sous la forme d'un voisinage avec quelques individus donnés, strictement aléatoire, ou encore, partiellement déterminé, du type « les amis de mes amis… », que l'on dénomme « petit monde », par référence à la formule bien connue « comme le monde est petit… ».



D'après Cassar (2006).

Les premiers travaux expérimentaux montrent cependant que les différents types de réseaux ont une incidence limitée sur l'équilibre à partir duquel les joueurs vont se coordonner. Dans les expériences de jeux de coordination conduites par Cassar, tous les joueurs se coordonnent finalement

sur l'équilibre *Payoffs dominant*, quel que soit le type de réseau dans lesquels ils évoluent, seule la vitesse avec laquelle ils parviennent à ce résultat varie d'un réseau à l'autre. Aux deux extrêmes, on trouve, d'un côté le « petit monde », où cette coordination est sensiblement plus rapide et, de l'autre côté, le « réseau local », où elle est plus lente. En revanche, c'est l'inverse que l'on observe, lorsque l'on cherche à tester l'aptitude des sujets à coopérer, à travers un réseau, dans un jeu du dilemme du prisonnier. Là encore, les joueurs parviennent finalement à coopérer, quoique plus difficilement, dans les différents types de réseaux considérés, mais, cette fois, c'est dans le réseau local qu'ils arrivent le mieux à s'entendre pour y parvenir, et dans le « petit monde » qu'ils y parviennent le plus difficilement (Cassar, 2007).

Ces résultats mettent en évidence la différence qui distingue la simple coordination de la coopération. Mais ils montrent surtout que les comportements observés dans ces situations sont peut-être moins sensibles aux structures des réseaux qu'on ne l'imaginait initialement. D'autres études, plus récentes, réalisées souvent sur des échantillons de populations beaucoup plus étendues, puisqu'elles portent, par exemple, sur plus d'un millier de sujets dans les expériences réalisées par Gracia-Lazaro (Grujic *et al*., 2010 ; Gracia-Lazaro *et al*., 2012) révèlent, de manière *a priori* assez surprenante, que les structures propres à chaque type de réseaux ont, en définitive, peu d'impact sur les comportements individuels des joueurs. On observera toutefois que la majorité de ces expériences concerne plutôt la coopération, à travers des jeux de dilemme du prisonnier répété.

Deux observations complémentaires peuvent être rapprochées de ce constat. Elles permettent d'avancer dans l'explication de ce phénomène. Il apparaît, tout d'abord, comme une singularité de l'espèce humaine, puisqu'on ne la retrouve pas dans la plupart des expériences réalisées chez des espèces animales, où l'organisation de ces réseaux de communication détermine assez largement leurs comportements. En second lieu, les expériences qui ont été citées ont été réalisées dans le cadre d'une organisation statique de ces réseaux. En clair, les sujets de ces expériences sont placés dans des réseaux dont la structure est donnée, et qu'ils ne peuvent donc pas modifier par leurs comportements observés au cours de ces expériences. Or d'autres travaux suggèrent que c'est précisément au travers d'une dynamique des structures d'organisation de ces réseaux, elle-même actionnée par les stratégies individuelles des joueurs, que pourrait se manifester, en retour, une influence de ces structures sur leurs comportements.

Les économistes ont tendance, comme on l'a déjà vu, à interpréter ces résultats comme une confirmation de leur hypothèse favorite, selon laquelle, même en réseaux, les choix des joueurs restent strictement individualistes. Telle n'est pas notre position. Si, en effet, les espèces animales réagissent de manière plus diversifiée que les hommes en fonction de l'organisation des réseaux dans lesquels ils opèrent, ce peut être, simplement, parce que, confrontées à ces environnements extérieurs variés, elles se trouvent dans la nécessité de s'y adapter. La situation serait, toutefois, un peu différente pour les hommes, qui, comme nous l'avons montré dans les chapitres précédents, n'appréhendent pas ces diverses organisations de leur environnement social comme leur étant tout à fait extérieures. Leur réaction à ces environnements ne consiste donc pas seulement à s'y adapter, mais également à chercher à les transformer, à les recréer, voire à en créer de nouveaux. Dans les

expériences qui ont été rapportées, ces réseaux sociaux revêtent la forme de structures statiques. Les sujets se trouvent alors dans l'impossibilité de les transformer, en les modulant au cours d'un processus dynamique. Or c'est à travers ce processus d'interactions dynamiques, entre, d'une part, les comportements, voire les préférences des sujets, et, d'autre part, l'organisation et le fonctionnement de ces réseaux, que se mettent en place les modes de coordination dans ces réseaux. C'est donc dans cette perspective qu'il faut chercher à les étudier. On verra dans les deux derniers chapitres que les schèmes d'organisation qui émanent de ces interactions prennent alors souvent la forme de conventions.

Une preuve indirecte de cette incidence de la formation des réseaux sur les comportements des acteurs sociaux face à la coordination de leurs actions est fournie par des situations inverses de celles que nous avons étudiées ici, où les équilibres dépendent, au contraire, d'une absence de coordination, comme c'est le cas, par exemple, d'une attaque surprise réussie dans certains types de jeux stratégiques. Des illustrations concrètes peuvent en être trouvées, en économie, dans le succès de raids de traders observés sur les marchés financiers. Dans de telles situations baptisées « anticoordination », il a été montré que la structuration en réseau et le nombre de ses participants transformaient les issues du jeu initial et pouvaient même, dans certaines configurations, aboutir à des équilibres de coordination, au moins en ce qui concerne certains groupes d'agents (Bramoullé, 2007). Ce genre d'effets se manifeste surtout lorsque le réseau est complet. Mais le plus intéressant réside dans la diversité de ses manifestations, selon les modes d'organisation que prendront les réseaux ainsi formés.

L'analyse de l'organisation dynamique de ces réseaux de communication est similaire à celle qui a conduit à la métaphore des réseaux neuronaux. Sous ce terme sont regroupés aujourd'hui un ensemble de modèles mathématiques assez divers, destinés à rendre compte de l'auto-organisation de systèmes en réseaux, afin d'en dégager les conséquences sur leur fonctionnement et leur évolution (Sayama *et al.*, 2013). La référence aux neurones renvoie ici aux modalités adaptatives de réseaux de connexion qui pourraient caractériser le fonctionnement d'un cerveau artificiel, ou tout au moins imaginaire. Les propriétés de tels modèles auraient ainsi la double fonction d'expliquer divers phénomènes de coordinations sociales (transports, circulation, etc.), tout en en éclairant la connaissance de l'organisation neuronale qui guide le fonctionnement du cerveau. La démarche qui inspire ces recherches repose sur l'hypothèse d'une similarité entre l'organisation en réseaux dynamiques qui détermine les règles de fonctionnement de ces phénomènes sociaux et l'organisation en réseaux dynamiques qui caractérise la neurophysiologie des cerveaux humains. Elle développe, sur cette base, une analyse des propriétés formelles de ces dynamiques de réseaux connectés, avant de chercher à les confronter ensuite, d'une part, au fonctionnement réel de nos cerveaux, et, d'autre part, à la dynamique observable de réseaux qui domine de plus en plus d'activités de notre vie sociale. Plusieurs travaux ont ainsi montré que différentes connexions neuronales étaient organisées sur le modèle de réseaux de type « petit monde » (*cf.* p. 135) (Watz et Strogatz, 1998). De même, des travaux d'économétrie appliquée se sont emparés de cette modélisation dite « des réseaux neuronaux », pour formaliser divers phénomènes macroéconomiques. Deux auteurs ont ainsi

récemment montré que ce type de modèle fournissait de meilleures prévisions sur l'évolution des taux de change entre les monnaies que les modèles économétriques standard (Jamal et Sundar, 2011).

Ce chantier ouvert dans la direction des réseaux neuronaux pose cependant de sérieux problèmes méthodologiques, voire épistémologiques. Ils n'échappent pas, du reste, à la conscience de quelques-uns des chercheurs engagés dans ce programme (Bullmore et Sporns, 2009). En premier lieu, l'hypothèse d'une correspondance rigoureuse entre cette modélisation mathématique des réseaux et l'organisation fonctionnelle des réseaux cérébraux reste, comme nous l'avons dit, une simple hypothèse qui repose, jusqu'à présent, sur quelques analogies assez naïves. Il faut donc encore la soumettre à des tests empiriques, souvent difficiles à construire, faute d'articulations fournies par des niveaux intermédiaires, comme nous l'avons signalé dans notre introduction (*cf.* Prologue : problèmes de méthode). En second lieu, même si cette étape préliminaire se trouve franchie, il n'est pas assuré que la mise en évidence d'une modélisation commune des propriétés des réseaux cérébraux et des réseaux sociaux résolve les questions posées par leur interaction dynamique. Or c'est de ces interactions que dépend, en définitive, le succès des coordinations qui ont été identifiées dans ce chapitre.

Une approche directe visant à une meilleure connaissance de la manière dont les cerveaux des individus, tout à la fois, contribuent à l'organisation de ces connexions sociales et sont modelés par elles, nous apparaît plus pertinente, au moins dans un premier temps, pour éclairer cette question. Elle conduit, en particulier, à réexaminer la relation entre le « je » et le « nous », à la faveur des transformations intervenues dans la communication sociale. Elle pourrait faire apparaître, en outre, de nouvelles distinctions entre les « autres » (membres de mon réseau) et les « autres, autres », étrangers à ce réseau, en fonction de l'organisation particulière de chacun de ces réseaux. Plusieurs questions très concrètes se présentent alors auxquelles on trouve déjà quelques éléments de réponse dans différents travaux de neuroscience sociale. Il semble, par exemple, que les individus, à travers les réseaux auxquels ils ont choisi d'appartenir, cherchent aujourd'hui à partager leurs informations (impressions, goûts, réactions, idées…) avec le « nous » qu'ils forment avec les autres membres du réseau. Il serait, intéressant de savoir si, et comment, le développement de ce mode de partage, d'origine virtuelle, transforme l'appréhension que le sujet peut avoir de ces différents « autres », et modifie, de ce fait, la perception qu'il peut avoir de lui-même. Une étude récente portant sur le partage d'un même bien commun a ainsi mis en évidence que le plaisir éprouvé par chaque participant était différent et plus intense, lorsque les autres appartenaient à des réseaux qui leur étaient communs, que lorsqu'ils étaient issus de groupes dont ils ne faisaient pas partie. Ce sentiment recueilli auprès des intéressés a été corroboré, au niveau cérébral, par une activation sensiblement plus forte de la région du striatum dans le premier cas (Fren *et al.*, 2012). Une autre étude a, par ailleurs, observé un impact direct du nombre et de la taille des réseaux sociaux auxquels appartenaient les individus sur le développement de leur insula, dont on connaît le rôle dans les manifestations d'empathie (Bickart *et al.*, 2012). Ces travaux restent cependant encore trop limités et parcellaires pour permettre d'en tirer des conclusions de portée plus générale. Mais c'est seulement en les multipliant que l'on pourra envisager, dans une seconde étape, d'élaborer une modélisation plus générale de ces effets de réseaux

sur les différentes formes de coordination étudiées dans ce chapitre.

Commentaire La coordination des perspectives du collectif et de l’individuel

Christian Schmidt a distingué trois modalités de coordination. 1) Je peux repérer une saillance notoire de l’environnement et compter sur la similarité entre les processus perceptifs humains pour penser qu’autrui va aussi la remarquer, si bien que nous convergerons vers le même point. C’est une coordination par point d’attraction commun, ou point focal. 2) Nous pouvons tous les deux repérer une différence saillante entre deux « équilibres ». Un équilibre se définit ici comme une situation telle que, si je m’y trouve, mon partenaire a intérêt à s’y trouver, et réciproquement. Mais nous pouvons tous les deux trouver une de ces situations beaucoup plus attractive qu’une autre. C’est une coordination par comparaison commune (le jeu haut/bas de Bacharach). 3) Nous pouvons repérer comme plus attractive une situation qui offre un avantage collectif, sur un critère que nous supposons dominant mais pas sur un autre que nous supposons moins décisif (jeu de la chasse au cerf). C’est une coordination par dominance collective.

C

Ces trois types de coordination peuvent chacun échouer, mais chacun pour des raisons différentes : le premier, parce que le point focal n’est pas suffisamment saillant, ou parce que l’attention du partenaire est parasitée par d’autres préoccupations. Ce problème a cependant peu de chances de se produire et, dans ces conditions, le fait même qu’il n’ait pas de remède le rend paradoxalement moins important : il ne sert à rien de nous lancer dans des spéculations compliquées pour tenter d’améliorer la situation, nous devons simplement nous fier à la fréquence de la réussite de ce mode de coordination.

Le deuxième peut échouer pour des raisons plus subtiles, parce que l’un des partenaires peut se demander si l’autre a bien privilégié de manière stable un équilibre, ou s’il n’a pas oscillé entre les

équilibres sans se décider. Cette possibilité d'oscillation tient justement à ce que l'autre situation est aussi un équilibre de coordination. À supposer que l'autre s'y soit fixé, j'aurais intérêt à la choisir, car préférer la situation qui est dans l'absolu la plus avantageuse ne me serait d'aucune utilité si je n'y rencontre pas mon partenaire. Il suffit que j'envisage que mon partenaire ne soit pas totalement certain que mon mode de raisonnement est strictement en phase avec les étapes du mien, ce qui est possible puisque on peut donner un avantage temporaire à une perspective sur l'autre, pour qu'il puisse en venir à osciller ! Ce problème tient à la structure « en huit » des inférences possibles : chacun peut avoir deux modes de raisonnement, l'un qui vise l'avantage commun, l'autre qui regarde si les conditions d'un équilibre sont réunies. Je me lance d'abord sur le premier mode, mais je ne peux conclure sans avoir envisagé que l'autre ait suivi le second mode, car il pourrait s'être lancé sur le premier mode, et supposer que j'aie suivi le second. Une fois qu'un parcours renvoie à l'autre, chacun doit emboîter dans son propre « huit » un parcours similaire d'autrui. Une fois que ces deux parcours s'emboîtent l'un dans l'autre nous risquons d'en rester à parcourir indéfiniment ce huit en passant d'une boucle et d'un parcours sur l'autre.

La difficulté tient à ce que, si chacun d'entre nous dispose d'au moins deux cheminements de raisonnement, la similarité avec autrui, qui devrait nous faire converger, peut aussi nous faire osciller : il suffit, sans même que nous nous situions réellement en opposition de phase, que chacun de nous doive envisager cette possibilité. Dans ce cas, la similarité entre les partenaires est source de divergence, parce que chacun d'eux présente dans le for intérieur de ses réflexions une dualité, donc une divergence possible, et doit, par similarité, intérioriser cette même dualité, mais vue par l'autre. On a parfois utilisé le terme de « spécularité », au sens de jeux de miroirs (« miroir » se dit *speculum* en latin) entre moi et autrui, pour désigner ce genre de situation.

Le troisième mode de coordination, celui de la chasse au cerf, renforce le risque d'échec, en ajoutant à cette spécularité une autre division, celle non plus seulement entre avantage collectif et équilibre interindividuel, mais entre risque collectif et risque individuel. En effet, la complexité de la coordination collective dans la chasse au cerf (dans une situation plus réaliste, pas dans une matrice de jeu ultrasimplifiée) introduit un risque – certains rabatteurs peuvent ne pas jouer correctement leur rôle, par exemple. Ce risque, le chasseur de lièvre solitaire ne le prend pas. Il en prend un autre – les lièvres peuvent ne pas passer à l'endroit de ses collets –, mais ce type de risque individuel existe aussi pour chaque chasseur de cerf. Le risque d'une coordination collective imparfaite s'ajoute donc dans la chasse au cerf au risque d'une réussite individuelle imparfaite.

S

Comme dans le commentaire du chapitre 2, il nous faut, quand nous revenons de la théorie des jeux aux pratiques effectives, introduire quelques processus intermédiaires, qui apparaissent quand on tient compte des dynamiques d'apprentissage. Considérons la première cause d'échec, le cas où le

point focal n'est pas très saillant. Malgré tout, si nous avons réussi à converger une première fois, cette première fois va renforcer la saillance – exactement comme l'activation d'une connexion neuronale renforce cette connexion. Dans la dynamique d'apprentissage de l'interintentionnalité, nous coconstruisons des « affordances[3] » communes, qu'on peut considérer comme des « coaffordances ».

Ces apprentissages en commun permettent de réduire les altérités entre les partenaires. On pourrait donc penser que toute coordination va procéder par réduction des altérités et par renforcement de similarités. Pourtant la difficulté rencontrée dans le deuxième mode de coordination nous a montré que la similarité elle-même peut faire naître des problèmes.

Là encore, cependant, recourir à une dynamique de coconstruction nous offre une échappatoire. Il s'agit alors de coconstruire le *point de vue* à partir duquel le critère de l'avantage commun sera dominant. Et c'est ce qu'a visé Bacharach, semble-t-il, en parlant de *team reasoning*, de raisonnement d'équipe. Encore faut-il que moi et mon partenaire, nous envisagions ce que serait le point de vue de notre équipe. Mais là aussi, une fois que nous l'avons fait, alors ce point de vue se renforce. En effet, du point de vue de l'équipe, la convergence vers l'équilibre communément avantageux domine l'oscillation de chaque partenaire livré à ses suppositions spéculaires. En revanche, si nous en restons au point de vue individuel, nous ne savons pas comment décider. Nous pouvons alors associer le raisonnement d'équipe à un équilibre de second ordre. De même que dans un équilibre ordinaire, une fois que nous y sommes, aucun de nous n'a intérêt à en dévier, de même, une fois que nous avons adopté le point de vue du raisonnement d'équipe au lieu d'un raisonnement strictement individuel, aucun de nous n'a intérêt à en dévier – alors que ce n'est pas le cas du point de vue de l'« équipe » de degré zéro, pour ainsi dire, que constitue chaque individu de notre duo raisonnant spéculairement chacun de son côté.

Il semble donc que la meilleure manière de nous coordonner, ce n'est pas de se lancer dans des raisonnements compliqués, mais bien de choisir un point de vue qui implique coordination, et de voir si les autres nous suivent, ce qui permet alors de renforcer ce point de vue. Cela pourrait expliquer pourquoi, dans l'expérience de Kuo *et al*. les sujets ressentent moins d'effort dans la situation de « pure coordination », celle où aucune procédure de raisonnement par étapes n'est proposée. C'est qu'en fait, raisonner par étapes pour résoudre la plupart des problèmes de coordination nous entraînerait dans les affres de la spécularité, alors qu'il vaut mieux faire en quelque sorte un saut hors de ces jeux spéculaires pour nous situer d'emblée dans ce que nous venons d'appeler un équilibre de second ordre.

R

Cependant, dans la chasse au cerf, ce saut quasi quantique – puisqu'il y a bien là une discontinuité irréductible – n'assure pas la stabilité du point de vue collectif. La différence avec la matrice du jeu de Bacharach est que dans ce dernier, si chacun choisissait une option différente de

l'autre, les deux joueurs n'obtenaient rien. Dans une hypothèse similaire concernant les joueurs de la chasse au cerf, le joueur qui aurait visé la chasse au cerf ne gagnerait rien, mais celui qui aurait fait le choix du solo obtiendrait tout de même son lièvre. La solution du raisonnement par équipe revenait à réduire l'altérité des partenaires en faisant ce saut au niveau de l'équipe. Mais la solution du chasseur solitaire nous suggère de réduire cette altérité d'une manière inverse : en nous rabattant sur le choix dont le résultat ne dépend pas du choix d'autrui.

Ce rabattement n'implique pas de refuser tout raisonnement par aller-retour entre un joueur et l'autre. Certains auteurs (comme Yoshida) pensent que l'adoption de la solution « chasse au lièvre » ne doit déclencher aucune montée réflexive, alors la solution « chasse au cerf » est liée à une telle remontée (je dois croire que l'autre chasseur croit que je vais chasser le cerf, etc.). Pourtant il n'est nullement exclu que je puisse, et même que je doive, pour décider de chasser le lièvre, monter d'abord en réflexivité, en envisageant les croyances des autres sur mes croyances sur les croyances des autres, etc., pour en conclure finalement que ces étages supérieurs de croyances mutuelles étant fragiles, je dois revenir à la chasse au lièvre.

Plaçons-nous cependant dans une situation un peu plus réaliste que celle de la matrice du jeu. Il n'est pas nécessaire que tous les membres du groupe participent à la chasse au cerf. Il est même souhaitable pour le groupe que certains continuent à chasser le lièvre, puisque cela peut assurer de la nourriture en cas d'échec de la chasse au cerf. Si l'on passait à des jeux évolutionnaires, où une stratégie est évaluée en fonction de la capacité de survie et de reproduction qu'elle assure, une stratégie mixte, qui joue tantôt la chasse au lièvre, tantôt celle au cerf, pourrait être l'équilibre optimal pour le groupe. Mais la stabilité d'un tel équilibre sur le long terme est sujet à caution.

Il faut ajouter que, dans le groupe de chasse au cerf, la réussite de chacun n'est pas nécessaire. Nous pouvons donc inverser notre argument précédent, celui du risque collectif, qui tenait à ce qu'il peut toujours y avoir des imperfections dans une coordination collective, et partir de collectifs dont le fonctionnement invalide cet argument. Un collectif « robuste », donc un collectif qui a des chances sérieuses de survie et de reproduction, c'est un collectif qui peut se permettre des imperfections de coordination et pourtant parvenir à ses fins. Notre critère d'un équilibre de groupe, d'un équilibre collectif robuste doit alors être double : 1) le point de vue collectif doit, une fois adopté, présenter des avantages pour l'*individu en tant que membre du collectif* par rapport au point de vue individuel ; 2) les imperfections de coordination du collectif doivent être possibles et admissibles, elles ne doivent pas être telles qu'elles obèrent sérieusement ces avantages. Un collectif de ce genre est robuste par rapport à la défaillance d'un de ses membres, alors que, par définition, la réussite d'un individu ne peut être robuste par rapport à sa propre défaillance.

Certes, la chasse au lièvre a aussi sa robustesse, et peut réussir malgré des imperfections. Cette robustesse est d'origine différente : chaque chasseur a un contrôle direct sur la manière dont il mène sa chasse, ce qui assure un certain niveau d'adaptabilité aux circonstances imprévues. En revanche, dans l'organisation collective d'une chasse au cerf, le contrôle est le plus souvent indirect : on a posté des tireurs et on fait avancer le front des rabatteurs, mais les rabatteurs n'ont pas un contact direct avec les tireurs et ne peuvent leur dire précisément d'où va venir le cerf. Le groupe reste plus robuste

face à la défaillance de membres isolés, puisqu'un membre peut remédier à la défaillance d'un autre, mais son mode de contrôle indirect n'a pas le type de robustesse du contrôle individuel, parce qu'il est moins aisé à adapter à des imprévus locaux et moins précis dans ses adaptations. Les activités individuelles présentent elles aussi un double aspect : elles sont moins robustes parce que si elles font défaillance, rien ne permet d'y remédier au niveau de l'individu, mais elles sont plus robustes dans la mesure où le contrôle direct permet des adaptations plus rapides et plus précises.

Il ne faut pas oublier, cependant, que le contrôle collectif n'est rien sans le contrôle direct de chacun de ses membres. Inversement, un collectif dont le contrôle indirect fonctionne suffisamment bien donne plus de chances de survie et même de reproduction aux individus. Nous pouvons alors concevoir un équilibre de troisième ordre. Il ne met pas en rapport la meilleure réponse d'un individu à la meilleure réponse d'un autre individu. Il met en rapport, pour ainsi dire, la meilleure réponse du point de vue collectif aux meilleures réponses du point de vue individuel. On commence donc par définir les équilibres individuels. Puis on adopte le point de vue du collectif. Il faut évidemment pour cela que ce point de vue puisse se définir : une situation où les deux meilleures possibilités communes sont la première en faveur d'un partenaire, la seconde de l'autre, de manière symétrique (du genre « bataille des sexes », l'un préférant aller au match de boxe, l'autre à un ballet, tous deux ayant une préférence de second rang pour être ensemble), ne permet pas cette définition. On évalue alors dans quelle mesure les stratégies des équilibres individuels permettent de maintenir les coordinations utiles que permet l'adoption de cette perspective, et on élimine les stratégies qui détruisent ces coordinations (dans le jeu de la chasse au cerf, on élimine les stratégies de chasse au lièvre, sauf en cas de disette). Il faut ensuite s'assurer que les stratégies individuelles sélectionnées, qui sont les meilleures possibles pour assurer les coordinations collectives du genre chasse au cerf, vont bien contribuer à l'amélioration des sorts individuels : elles doivent leur assurer des gains préférables à ceux d'un équilibre du point de vue individuel. C'est le cas dans la chasse au cerf. Dans un jeu de dilemme du prisonnier (DP), les stratégies de l'équilibre individuel sont des défections mutuelles (DD). Elles ruinent la coopération mutuelle (CC). On considère alors ce qui se passe si on les écarte. Il faut ensuite s'assurer que les stratégies qui restent (CC) assurent un meilleur gain aux individus (ou aussi bon) que celles de l'équilibre individuel (DD). C'est le cas. On en reste donc à la coopération mutuelle – du moins dans un raisonnement à long terme, puisque si chacun est capable de ce raisonnement, celui qui envisagerait de revenir à une stratégie de défection plus avantageuse pour lui (parce qu'il ferait défection en profitant de la coopération des autres) s'apercevrait qu'il détruit ainsi l'équilibre collectif, et se désavantage pour de futures interactions.

Smerilli (2012), revenant sur les pistes de Bacharach, a proposé de définir d'une part un équilibre collectif – le collectif n'a pas intérêt à dévier unilatéralement de la stratégie de cet équilibre – et d'autre part un équilibre individuel – l'individu n'a pas intérêt à dévier unilatéralement de la stratégie de cet équilibre. Il y a des cas où les deux équilibres correspondent pour l'individu à la même stratégie, d'autres où les deux stratégies diffèrent, et d'autres, comme le DP, où on a un intérêt à dévier de l'individuel vers le collectif, mais aussi du collectif vers l'individuel : on oscille sans se

stabiliser. Mais ce que nous envisageons est plus exigeant : nous évaluons les équilibres individuels du point de vue de leur capacité à maintenir ou restaurer un équilibre collectif, et les équilibres collectifs du point de vue de leur capacité à améliorer la situation des individus par rapport aux équilibres individuels. Dans le DP, l'équilibre individuel est la défection mutuelle (DD), qui ruine l'équilibre collectif. On envisage donc de l'abandonner. L'équilibre collectif est la coopération mutuelle (CC), il améliore le résultat individuel par rapport à DD, il doit donc être choisi. Nous entrecroisons les perspectives d'évaluation au lieu de les maintenir séparées.

Il est évidemment possible de généraliser ce genre d'équilibre en ne le limitant pas aux relations entre un collectif et ses membres individuels, mais en l'étendant aux relations entre des collectifs différents qui seraient membres d'un collectif plus englobant.

Nous n'avons fait là que privilégier le point de vue de l'interaction par rapport aux points de vue des éléments (nous reviendrons sur les effets de réseaux dans le chapitre 5). Comme les individus, en tant que sujets intersubjectifs, sont coconstitués dans leurs interactions, on comprend que cette perspective d'interaction entre point de vue collectif et point de vue individuel leur soit naturelle. Et c'est aussi cette perspective qui en fait des êtres sociaux.

---

1. Schelling cite en note cette observation tirée du psychologue gestaltiste Koffka (1955). Koffka prend soin de préciser que le buteur, dès qu'il a pris conscience de cette attraction exercée sur lui par la perception du gardien de but de l'équipe adverse, réorganise le centre de gravité de son espace mental, autour d'un autre attracteur. Cette observation est importante, car elle suppose un mécanisme d'apprentissage au niveau de cette perception que l'on peut qualifier ici de stratégique.

2. Husserl, comme on l'a vu au chapitre 3, consacra une année entière de cours à Göttingen, durant l'hiver 1904-1905, à l'étude d'une phénoménologie de l'attention. Pour Husserl, l'attention joue le rôle d'un modulateur de la perception. Cette approche husserlienne du rôle de l'attention dans les mécanismes de perception, avec ses implications en termes de saillances émotionnelles, s'est trouvée validée aujourd'hui, au plan neurochimique, par de nombreux travaux de chercheurs en neurosciences (Yu et Dayan, 2005 ; Roelfsema *et al.*, 2010 ; Todd et Andersson, 2013).

3. C'est à Gibson qu'on doit ce terme, difficile à traduire. Un élément dans notre champ visuel présente une affordance si sa forme indique sa disponibilité par rapport à l'une de nos activités. Par exemple, une poignée de porte présente une affordance, celle de la tourner et de la tirer pour ouvrir la porte.

CHAPITRE 5

De la coordination à la coopération

Le chapitre précédent a montré, que, sous le terme de coordination, se trouvent regroupés des phénomènes différents. Un tel regroupement peut, comme on l'a vu, porter à confusion, lorsque l'on cherche à identifier les opérations mentales qui interviennent dans cette coordination chez les sujets. On a observé, notamment, un fréquent glissement sémantique de la simple coordination à une véritable coopération.

Rappelons d'abord que la question de la coordination se pose chez un même sujet. Presque toutes les activités motrices intentionnelles (marcher, se lever, saisir un objet…) exigent une coordination des mouvements. Beaucoup d'activités mentales font également intervenir une coordination de plusieurs des réseaux neuronaux récemment mis en évidence par les techniques d'imagerie. Cette coordination dans l'interaction des systèmes neuronaux pourrait intervenir au moyen d'une modulation correspondant à un mode de synchronisation des rythmes d'activation des zones cérébrales concernées (Womesfeld *et al*., 2007). On est dès lors en droit de s'interroger pour savoir si, et de quelle manière, la coordination, lorsqu'elle concerne des actions émanant de sujets différents, présente des affinités, mais aussi des différences, avec la coordination que requièrent ces activités chez un même sujet.

Plusieurs travaux de neurosciences ont récemment mis en évidence des mécanismes de coordination de mouvements observés dans un groupe d'individus qui aboutissaient à des rythmiques collectives. Il serait sans doute préférable de parler plutôt à leur sujet de synchronisations (Oullier, Kirman et Kelso, 2008 ; Oullier *et al.*, 2010). De tels mécanismes se déclenchent toutefois indépendamment de la conscience que peuvent en avoir les sujets individuels entre lesquels ils se manifestent. L'enjeu, du point de vue de l'analyse des interactions, est de détecter comment le sujet appréhende l'autre (les autres), lorsqu'il cherche à coordonner son action avec celle de l'autre (des autres). L'autre est-il alors conçu comme une manière d'extension du « soi-même », ou conserve-t-il l'essentiel de son altérité ? La première hypothèse rapprocherait la coordination entre différents sujets

de la coordination chez un même sujet. Si cette hypothèse se confirmait, elle entraînerait une seconde question. Cette extension de soi-même doit-elle s'entendre comme un simple prolongement du moi (l'autre moi-même *stricto sensu*), ou faut-il, plutôt, l'interpréter comme la manifestation d'une appartenance commune à un « nous » ? Considéré dans une perspective cognitive, le « nous » selon cette seconde interprétation correspondrait à ce que Bacharach nomme l'équipe, faisant ainsi de l'équipe le véritable sujet du *team reasoning* et, partant, le centre de la volonté consciente de se coordonner. Nous nous rapprochons alors de la coopération.

Une seconde source de difficultés rencontrée dans l'analyse de la coordination provient de la diversité de ses acceptions, selon les problèmes particuliers qu'elle pose aux individus dans les différents contextes où elle se présente à eux. Les situations de jeux qui servent de références aux protocoles expérimentaux discutées au chapitre précédent en fournissent quelques illustrations. Ainsi la question de la coordination, telle qu'elle se pose aux deux joueurs du jeu du rendez-vous, n'est pas exactement la même que celle que doivent résoudre ceux du jeu haut/bas. Dans le premier cas, il leur faut, rappelons-le, imaginer un point focal qui leur serait commun ; dans le second cas, il suffit à chacun de penser que l'autre choisira, comme lui-même, l'action dont les conséquences sont les plus avantageuses. Avec le jeu de la chasse au cerf, la question de la coordination se pose aux joueurs de manière encore plus différente. Se coordonner équivaut, pour les chasseurs, à coopérer dans une entreprise commune (la chasse au cerf). Contrairement au jeu du rendez-vous, un échec de la coordination n'entraîne pas ici l'absence de solution, puisque chaque chasseur peut toujours attraper un lièvre de son côté. L'existence de ces solutions alternatives permet d'interpréter la coordination comme le résultat d'un choix raisonné. Un tel choix se révèle précisément problématique dans le jeu du rendez-vous et trivial dans le jeu haut/bas. En outre, et peut-être surtout, la coordination est ici indissociable de la coopération, puisque chasser le cerf implique la mise en œuvre et l'exécution d'un objectif commun aux chasseurs, celle de la prise du cerf. Tel n'est pas le cas dans les deux autres jeux, où la coordination des mouvements des joueurs garantit seulement la rencontre de leurs intentions.

Si des liens étroits existent entre la coordination des agents et leur coopération, ne serait-ce que parce que coopérer nécessite, au préalable, de pouvoir coordonner ses actions avec celles des autres, il n'est pas, pour autant, possible d'induire leur coopération d'une simple coordination entre les actions qu'ils entreprennent. La théorie des jeux 2 x 2 fourmille d'exemples de ce type, où il existe plusieurs équilibres et donc de points de coordination, mais dont aucun ne traduit une coopération entre les joueurs, bien au contraire. En outre, dans certains jeux de combat, le choix victorieux de stratégies, pour le moins non coopératives, exige cependant une coordination des mouvements entre les adversaires (Schmidt, 1993)[1]. Preuve, s'il en fallait, que la coordination n'a aucune raison de conduire nécessairement à la coopération. La coopération fait donc intervenir une relation particulière à l'autre (aux autres) qui se concrétise dans l'intention, consciente ou non, de réaliser une action qui puisse être communément profitable.

Cette définition liminaire de la coopération est toutefois insuffisante pour capter les multiples formes que peut également prendre la coopération dans différents contextes et saisir ensuite ce qui, tout à la fois, les relie et les sépare. Ici encore, comme avec la coordination, de premiers

enseignements peuvent être tirés des expérimentations de plusieurs jeux bien connus, à condition d'interpréter leurs résultats dans une perspective inverse de celle qui a dominé jusqu'à présent. Au lieu de rechercher les points communs de ses expériences, il nous apparaît plus fécond de mettre l'accent sur leurs différences, afin de contribuer ainsi à dégager les diverses composantes du concept de coopération. Parmi les jeux emblématiques à deux joueurs le plus souvent utilisés pour identifier les principaux traits caractérisant les comportements « coopératifs », les chercheurs en économie expérimentale et, ultérieurement, en neurosciences, ont privilégié, bien sûr, le dilemme du prisonnier dans ces différentes versions (à un coup et répété), mais aussi et de manière moins évidente, le jeu de l'ultimatum et ses variantes (jeu du dictateur), et le jeu de la confiance (ou de l'investissement). On y joindra ici, pour les raisons expliquées au début de ce chapitre, le jeu de la chasse au cerf à plusieurs joueurs.

L

Les jeux qui ont été énumérés présentent une particularité commune. Ils appréhendent tous la coopération par référence à une non-coopération, entendue plus ou moins explicitement comme le résultat d'un comportement normal des joueurs. Plus précisément, en théorie des jeux classique, les joueurs de ces jeux, en choisissant chacun une stratégie qui maximise leur gain personnel, accèdent à un équilibre de Nash, sans se soucier de ce qu'il advient à l'autre. Un tel équilibre constitue, pour cette raison, une référence privilégiée pour définir la non-coopération, ou plus précisément, une manière d'a-coopération. Or les comportements observés au cours des expérimentations répétées de ces différents jeux ont révélé que les joueurs de ces jeux ne procédaient généralement pas de cette manière. La remise en cause de cette hypothèse implicite d'« a-coopération » a constitué, dans bien des cas, le point de départ des recherches comportementales qui se sont développées sur la coopération. Cette définition indirecte de la coopération par défaut de non-coopération peut surprendre. Elle s'explique cependant lorsqu'on la replace dans le cadre logique des jeux non coopératifs qui domine aujourd'hui la théorie des jeux, au moins en économie. Elle porte, en effet, l'empreinte d'une grille de lecture particulière de la théorie des jeux qui s'est progressivement imposée.

Dès son origine, ou presque, la théorie des jeux s'est scindée en deux branches distinctes, correspondant à deux catégories différentes de situations sociales, traitées respectivement par des jeux coopératifs et non coopératifs. Dans les jeux coopératifs, les premiers étudiés dans l'ouvrage fondateur de von Neumann et Morgenstern (1943), les joueurs sont supposés pouvoir prendre entre eux des engagements qui sont exécutoires (*binding agreements*). Par cette référence à ces engagements exécutoires, les différentes solutions proposées des jeux coopératifs renvoient toutes aux alliances et aux coalitions que les joueurs peuvent former au cours du jeu. Alliances et coalitions y occupent ainsi le rôle de marqueurs de la coopération. Coopérer dans cette perspective c'est s'allier

avec d'autres joueurs, et, éventuellement, contre d'autres. À l'opposé, dans les jeux non coopératifs de tels engagements sont exclus, puisque les choix stratégiques des joueurs reposent exclusivement sur les hypothèses de rationalité strictement individuelle qui leur sont associées. La coopération y semble donc exclue, ou tout au moins, absente au point de départ.

En développant son approche non coopérative des jeux, Nash introduisit cependant, d'une manière différente, on pourrait même dire détournée, la notion de coopération. Partant de l'hypothèse que les agents ne peuvent pas prendre entre eux d'engagements exécutoires au cours du jeu, faute d'institutions capables de les faire respecter, il démontra dans deux articles célèbres consacrés à la négociation que la coopération peut résulter d'une simple entente entre deux ou plusieurs joueurs strictement rationnels dans son acception individualiste, indépendamment de tout engagement réciproque de leur part (Nash, 1950, 1953). Ainsi entendue comme l'issue d'une procédure de négociation explicite (ou seulement implicite) entre deux ou plusieurs agents, la coopération prend alors également sa place à l'intérieur des jeux non coopératifs, avec, toutefois, une différence majeure. Tandis que les divers concepts de solution imaginés pour résoudre les jeux coopératifs reposent tous sur la coopération des joueurs interprétée dans sa première acception (formation réfléchie d'alliances et de coalitions), l'équilibre de Nash, qui représente la principale solution des jeux non coopératifs, n'implique aucune forme de projet commun, c'est-à-dire de coopération au sens fort entre les joueurs. Grâce cependant à l'idée d'entente négociée, l'équilibre de Nash se révèle toutefois compatible avec la notion de coopération, entendue sous un mode différent. Mais, dans la plupart des situations décrites par ces jeux, leurs solutions restent non coopératives, au sens où seuls comptent pour les joueurs les avantages individuels qu'ils tirent du choix de leur stratégie, sachant que les autres procèdent de même. C'est la raison pour laquelle coopérer traduit souvent, de la part des joueurs, un comportement en rupture avec la rationalité non coopérative qui prévaut dans la théorie des jeux (non coopératifs) ; d'où cet *a priori* que la coopération serait le résultat d'un comportement en quelque sorte déviant par rapport à une norme supposée non coopérative. On notera toutefois que la théorie des jeux coopératifs n'a pas pour autant été abandonnée. Elle continue notamment de se développer dans certains pays, en particulier en Russie, où elle représente une majorité des travaux de recherche. Mais ces développements n'ont donné lieu, jusqu'à présent, qu'à peu d'expérimentations récentes et de travaux en neurosciences.[2]

D

Revenons aux jeux non coopératifs à partir desquels ont pu être identifiés expérimentalement des comportements coopératifs. Une analyse plus détaillée de leurs règles fait d'abord apparaître des particularités qui permettent de les distinguer les uns des autres.

Dans les jeux de l'ultimatum et de la confiance, comme, du reste dans celui de la chasse au cerf, l'équilibre de Nash non coopératif n'est pas le seul équilibre. Le jeu de l'ultimatum et le jeu de la

confiance portent l'un et l'autre sur la détermination de règles de partage. Dans les deux cas, presque toutes les règles de partage, qui renvoient à autant d'équilibres, peuvent être considérées comme coopératives, au sens où elles émanent d'une entente (certes un peu forcée). C'est selon une acception différente de la coopération, plus proche de l'équité, que seules certaines règles de partage seront tenues pour révélatrices d'une attitude coopérative.

La situation de la chasse au cerf est différente. L'enjeu ne porte plus sur un partage, mais sur un choix entre deux solutions correspondant, comme l'a vu, à deux équilibres de Nash. L'un (la capture de lièvres) est strictement non coopératif, dans la mesure où il n'implique pas même une entente de chaque chasseur avec les autres. L'autre équilibre, au contraire, peut être qualifié de coopératif, mais dans un sens beaucoup plus fort que dans les deux autres jeux. Il exige, en effet, des chasseurs qu'ils organisent une collaboration qui leur permette d'atteindre un objectif commun (la prise du cerf). Cette solution coopérative présente en outre l'avantage (la prise du cerf) d'être bénéficiaire pour tous et de correspondre ainsi à un optimum social. Si coopérer en ce sens signifie la participation volontaire à un projet commun, sa mise en œuvre, en revanche, ne requiert ni véritable altruisme, ni même souci d'équité de la part des joueurs.

Le jeu du dilemme du prisonnier n'est pas de même nature que les trois jeux précédents. Considéré du point de vue économique, ses joueurs n'ont guère de choix, puisque ce jeu ne possède qu'un seul équilibre de Nash. Or cet équilibre est à la fois non coopératif et sous-optimal. Coopérer signifie donc ici enfreindre la règle du jeu non coopératif pour échapper au piège tendu aux joueurs par cette situation, en construisant une issue alternative fondée sur une confiance réciproque. Cette confiance implique une croyance en des références partagées, qui se manifestent le plus souvent par la réciprocité. Mais de telles références peuvent être d'ordre très divers (altruisme, équité, ou simple communauté d'intérêt ou d'appartenance).

Derrière, par conséquent, un rejet commun de solutions non coopératives, les comportements coopératifs prêtés aux joueurs dans ces différents jeux se différencient, tant par leurs motivations que par les modalités qu'ils empruntent.

La logique des situations permet de mieux saisir certaines constantes, tout en approfondissant la diversité des motivations qui peuvent inspirer ces comportements coopératifs[3]. Une première distinction s'impose à ce niveau entre les jeux où les joueurs occupent une position symétrique, et ceux où ils sont en position asymétrique en termes d'information.

Les jeux de l'ultimatum et de la confiance appartiennent à cette seconde catégorie. Dans le jeu de l'ultimatum, l'asymétrie dont bénéficie le joueur qui fixe les règles de partage et jouer en premier lui offrent la possibilité d'augmenter son gain au détriment de l'autre joueur. On interprète parfois le refus d'une maximisation de son gain, révélée dans ces conditions par l'offre du joueur en premier, comme la manifestation d'un comportement coopératif. Il peut s'expliquer, comme on le verra dans la dernière partie, par un souci d'équité. Mais il peut également se comprendre comme le résultat d'une anticipation lui faisant redouter un refus de l'autre joueur. Les choses se présentent de manière assez voisine dans le jeu de la confiance lorsqu'il n'est pas répété. Dans sa version économique, qui porte sur un investissement, le mandataire qui joue en second dispose également d'une asymétrie en sa

faveur. Il peut ainsi, en toute rigueur économique, maximiser son gain personnel, en réduisant la partie qu'il reverse à l'investisseur. Les motivations qui le pousseraient à ne pas utiliser cet avantage de position à son profit peuvent être considérées comme identiques à celles prêtées à celui qui fixe les règles de partage dans le jeu de l'ultimatum. Il existe cependant entre ces deux jeux une différence importante, lorsque le jeu de la confiance est répété. L'autre joueur, c'est-à-dire l'investisseur, peut en effet, à la séquence suivante, confier, ou ne pas confier, au mandataire, au vu de son comportement, une autre somme à placer. Le fait, pour le mandataire, de ne pas maximiser à son avantage le partage des gains obtenus au détriment de l'investisseur, dès la première séquence du jeu, peut alors être cette fois davantage motivé par une anticipation intéressée que par un sens de l'équité, ou un penchant altruiste. De manière plus générale, du reste, le fait qu'un jeu s'étende sur plusieurs séquences modifie la perspective des joueurs et, par conséquent, leurs intentions de coopérer. Non seulement l'horizon de leur anticipation se trouve étendu, mais ils disposent surtout d'informations supplémentaires sur le comportement des autres joueurs, de nature à renforcer, ou, au contraire, à modifier leurs intentions premières. Ces informations se révèlent décisives pour la motivation de réciprocité susceptible d'inspirer les comportements coopératifs.

De telles asymétries de position n'existent ni dans le jeu de la chasse au cerf, ni dans celui du dilemme du prisonnier, où aucun des joueurs ne dispose d'un avantage initial. Dans le jeu de la chasse au cerf, l'équilibre coopératif est optimal et assure par conséquent un gain maximal à chaque joueur. Les motivations de ceux qui choisissent une stratégie coopérative ne sont donc pas à rechercher du côté de l'altruisme et de l'équité, mais plutôt dans l'intérêt individuel bien compris, avec toutefois une limite à ce pur individualisme. Pour réussir cette coopération, il leur faut manifester, en plus, une confiance dans le comportement également coopératif des autres joueurs. Une telle confiance qui contrebalance une forme de risque peut, là encore, prendre sa source dans des considérations très diverses : rationalité partagée, réciprocité bien comprise. La simple conscience d'appartenir à une même communauté culturelle ou autre peut même ici suffire.

Les joueurs du dilemme du prisonnier se trouvent également placés dans une position symétrique de départ, l'un par rapport à l'autre. Mais, contrairement au jeu de la chasse au cerf, le choix individuel d'une posture coopérative expose celui qui l'adopte au risque beaucoup plus dommageable d'une exploitation de la nouvelle situation par l'autre joueur à son avantage, au détriment du coopérateur. Un tel risque provient de ce que le candidat à la coopération crée de ce fait une asymétrie dont bénéficie celui qui refuse de coopérer. Il n'en demeure pas moins qu'une coopération partagée assurerait à chacun des deux joueurs un bénéfice supérieur à celui qu'ils peuvent tirer de la solution non coopérative du jeu. Adopter dans de telles conditions un comportement coopératif n'implique pas, non plus, nécessairement, une motivation altruiste, ou empreinte d'équité. Il s'apparente davantage à un pari sur la conscience partagée d'un calcul du risque commun. Le choix d'un tel pari repose sur les croyances de chaque joueur, sur les intentions de l'autre joueur, avec les différents niveaux d'itérations propres à ces croyances. On retrouve ici encore le ressort de la confiance qui se trouve confirmée et enrichie par l'expérience, lorsque le jeu se trouve répété.

Ce survol des situations dans lesquelles se manifestent des comportements coopératifs qui ont été le plus souvent observés expérimentalement permet d'identifier deux facteurs de différenciation des motifs de coopérer liés aux règles du jeu : les positions respectives occupées par les joueurs (symétrie, asymétrie) et les conditions de déroulement du jeu (jeu à un coup, jeu séquentiel ou répété).

L'approche alternative des neurosciences

Les différentes hypothèses qui ont été développées, à partir de la théorie des jeux et des protocoles expérimentaux qu'elle a inspirés, ont montré la variété et la complexité des ressorts de la coopération. Pour comprendre leur fonctionnement il est nécessaire de pénétrer plus avant les mécanismes mentaux qui conduisent les individus à coopérer ou, au contraire, à refuser de coopérer dans les diverses situations qui ont été examinées. Les neurosciences, dans leur branche comportementale et sociale, se sont penchées sur cette question depuis une dizaine d'années. Pour la majorité d'entre eux, ces travaux ont pris appui, comme on l'a vu, sur une transposition expérimentale des jeux dont les conditions ont été rapidement analysées (jeu de l'ultimatum, jeu de la confiance, dilemme du prisonnier et même jeu de la chasse au cerf), en les confrontant aux données fournies par l'imagerie cérébrale. D'autres études ont également cherché à déceler les effets de la sécrétion de certaines hormones, et notamment de l'ocytocine, sur les comportements coopératifs ou non coopératifs des sujets au cours de ces jeux (Kosfeld *et al.*, 2005 ; Baumgartner *et al.*, 2008).

En dépit de leurs références aux jeux, les contributions des neurosciences restent difficiles à articuler avec les distinctions qui ont été mises en évidence sur la base d'une analyse logique des situations. Dans une première phase, qui remonte au début des années 2000, les recherches de neurosciences sur la coopération, sous l'impulsion, en particulier, de Rilling et de Sanfey, se sont essentiellement concentrées sur l'étude du fonctionnement cérébral des sujets participant à deux jeux, le jeu de l'ultimatum à un coup, et le jeu du dilemme du prisonnier répété. Leur objectif était assez précisément circonscrit. Ces travaux s'efforçaient, d'une part, de distinguer le fonctionnement cérébral des sujets lorsqu'ils sont confrontés dans des situations identiques soit à d'autres sujets humains, soit à des machines. Ils cherchaient, ensuite, à valider, à travers l'interaction mentale impliquée par la coopération, les fondements neuronaux de la théorie de l'esprit, à l'époque en plein essor (*Theory of Mind*, *TOM*) (Callaguer et Frith, 2003). Leurs investigations des processus mentaux reliés à la coopération portaient principalement, si ce n'est exclusivement, sur des manifestations de réactivité entre les sujets, où la coopération se manifeste essentiellement à travers une réciprocité. La première étape de ces travaux concerne l'interprétation des résultats recueillis au cours de ces expériences.

Dans le jeu de l'ultimatum, c'est le fonctionnement cérébral du sujet qui se trouve conduit à refuser le partage proposé par le joueur jouant en premier, qui a été principalement étudié. Refuser la posture consistant à accepter le partage et par conséquent rejeter la coopération, a été interprété

comme la manifestation d'une aversion à l'iniquité (Fehr et Schmidt, 1999, 2010), ou plus précisément comme la manifestation d'une sanction du non-respect d'un principe de réciprocité (Sanfey, Rilling *et al.*, 2003). Le délicat problème soulevé par l'interprétation des résultats expérimentaux en termes de comportements des joueurs lorsqu'ils semblent renvoyer à des normes sera analysé plus en détail au chapitre 7 et dans son commentaire.

Dans le jeu du dilemme du prisonnier répété, l'analyse a porté sur les zones cérébrales activées en fonction des réponses données par chaque joueur aux actions coopératives (ou non coopératives) de l'autre joueur à la séquence précédente (Rilling *et al.*, 2002). Coopérer s'entend ici comme une réponse positive donnée à une action coopérative, ou plus précisément à une action perçue comme émanant d'une intention coopérative de l'autre joueur. Le moteur de la coopération serait, là encore, recherché dans un mécanisme de réciprocité[4].

Les mêmes équipes de chercheurs ont ensuite mis en évidence l'existence d'une certaine similitude des régions du cerveau activées dans les deux types de jeux. Plus précisément, il s'agit du cortex paracingulaire antérieur et du sulcus temporel supérieur, deux des régions auxquelles sont justement attribuées les bases neuronales de la théorie de l'esprit (Rilling, Sanfey *et al.*, 2004). Ce rapprochement semble à première vue contredire les interprétations que nous avons dégagées des résultats expérimentaux qui soulignaient les différences qui caractérisent la coopération dans ces deux types de jeux. On peut toutefois penser, bien que la coopération requière des opérations mentales différentes dans les contextes respectifs de ces deux jeux, qu'il existe néanmoins un certain recouvrement dans les zones du cerveau qui sont chaque fois sollicitées. La question que nous avons soulevée se traduirait alors dans les conditions différentes de leur fonctionnement dans les deux cas. Mais d'autres travaux sont nécessaires pour creuser cette hypothèse.

La portée véritable des premiers résultats qui ont été rapportés reste cependant encore limitée, en raison des hypothèses initiales sur la coopération qui sous-tendent ces expériences et du détail des protocoles utilisés. Conformément à la théorie de l'esprit qui inspire une majorité de ces chercheurs, ils entendent par coopération un échange réciproque positif entretenu par l'attente et la révélation des intentions de l'autre. Ce que l'on recherche, par conséquent, c'est donc moins l'intention initiale de coopérer que la réaction à une action de l'autre que l'on suppose inspirée par une intention coopérative, (ou, au contraire, non coopérative). Dans le jeu de l'ultimatum, c'est la réaction du cerveau du second joueur à la proposition de partage du premier joueur qui est retenue. Dans le jeu de la confiance (ou de l'investissement), c'est celle de l'investisseur au partage effectué par le mandataire qui est tenue pour révélatrice de la coopération (King-Casas *et al.*, 2005). Dans le jeu du dilemme du prisonnier répété, l'enclenchement du processus coopératif est traité comme une donnée, et la recherche porte sur les réponses successives des joueurs aux intentions des autres qu'ils induisent de l'observation de leur comportement (Rilling *et al.*, 2002). Ces travaux ne nous apprennent donc rien sur le fonctionnement du cerveau du premier joueur du jeu de l'ultimatum, lorsqu'il énonce sa proposition de partage. On peut certes imaginer, en prolongeant la perspective de la théorie de l'esprit, qu'il anticipe l'accueil de l'autre joueur à sa proposition, mais cette hypothèse reste conjecturale. De même, aucune explication n'est donnée du rejet conjoint d'un comportement non coopératif par

chacun des deux joueurs du dilemme du prisonnier, lorsqu'ils choisissent ensemble de coopérer à la première séquence du jeu.

Une réflexion plus approfondie fait apparaître que le cadre théorique dans lequel ces recherches neurophysiologiques ont été menées est, d'une certaine manière, l'inverse de celui dans lequel la théorie des jeux, et derrière elle, l'analyse économique, aborde la coopération. Pour les chercheurs en neurosciences, une réciprocité altruiste constitue, en quelque sorte, un penchant naturel des cerveaux humains. Ils en veulent pour preuve la fréquence avec laquelle les joueurs du dilemme du prisonnier adoptent une stratégie coopérative, dès la première séquence du jeu, lorsqu'il est répété. Son explication serait alors plutôt à rechercher au niveau génétique dans la perspective d'une anthropologie évolutionniste. Dès lors, ce qui pose problème, ce sont moins les comportements coopératifs que les comportements non coopératifs. Comme la coopération fonctionne principalement sur le mode d'un échange dont le moteur serait la réciprocité, ces chercheurs ont été conduits à s'intéresser aux réactions cérébrales des sujets lorsqu'ils étaient confrontés à des comportements de non-réciprocité (Rilling *et al.*, 2007). Ils ont ainsi observé, en reprenant le jeu du dilemme du prisonnier répété, qu'une rupture de la réciprocité au cours du jeu engendrait, chez le joueur qui en était victime, une activation renforcée de deux régions cérébrales spécifiques, l'insula antérieure et une partie du cortex orbitofrontal. Ces deux régions sont traditionnellement associées, la première, l'insula, à la peine et aux pertes, la seconde, le cortex orbitofrontal, à la relation entre le ressenti affectif et le cognitif. Ces réactions à la non-réciprocité du partenaire s'accompagnent, en outre, d'une activité accrue de l'hippocampe gauche prioritairement alloué à la mémoire épisodique, suggérant une mémorisation instantanée plus forte des comportements non coopératifs que des comportements coopératifs de l'autre joueur.

Cette première élaboration d'une théorie neuronale de la coopération s'oppose, de prime abord, à la construction proposée par la théorie des jeux non coopératifs pour rendre compte du phénomène de la coopération (et de la non-coopération) entre les individus. L'interrogation centrale des théoriciens des jeux à son sujet peut être formulée en ces termes : pourquoi les joueurs préfèrent-ils adopter un comportement coopératif, dès lors qu'ils pourraient soit obtenir davantage (cas du joueur en premier dans le jeu de l'ultimatum), soit sécuriser leur gain (cas des deux joueurs dans le jeu du dilemme du prisonnier), en choisissant de ne pas coopérer ? Les réponses à cette question sont évidemment variées, puisque la motivation des joueurs à coopérer dépend étroitement des situations particulières dans lesquelles ils se trouvent placés qui, comme on l'a montré, sont différentes d'un type de jeu à l'autre.

Pour les chercheurs en neurosciences la question soulevée par la coopération se pose en sens inverse : pourquoi les joueurs se trouvent-ils conduits à refuser de coopérer, alors qu'une coopération réciproque pourrait leur assurer un gain, fût-il modeste (cas du joueur en second dans le jeu de l'ultimatum), ou leur offrir la possibilité d'un gain supérieur (cas des deux joueurs du jeu du dilemme du prisonnier) ? Leurs réponses à cette question sont cette fois identiques. Quelles que soient, en effet, les spécificités des jeux où ces comportements sont observés, on retrouve, chaque fois, le sentiment

que la réciprocité n'a pas été respectée. Cette manière différente d'aborder la coopération oriente les recherches sur ce sujet dans des voies distinctes, si ce n'est parfois même opposées. Les théoriciens des jeux se trouvent conduits à privilégier sa dimension cognitive. La coopération représente pour eux une solution conçue par les joueurs pour résoudre un problème complexe qui leur est posé par telle ou telle situation particulière d'interaction. Les neurobiologistes, de leur côté, s'attachent plutôt, à approfondir la dimension émotionnelle de la coopération. La coopération traduit pour eux une disposition des cerveaux humains, dont il s'agit de découvrir les racines, en deçà des intentions raisonnées des individus, avant d'identifier les obstacles qui s'opposent à sa manifestation.

V

En dépit de ces divergences, plusieurs travaux récents portant sur la coopération, considérée du point de vue de la théorie des jeux, et sous l'angle du fonctionnement cérébral des joueurs tendent néanmoins à se rapprocher. Cette convergence repose sur une conviction partagée que la coopération est un phénomène dynamique qui fonctionne sur la base d'une réciprocité autoentretenue par les comportements des joueurs au cours du déroulement des jeux. Deux facteurs ont favorisé ce rapprochement.

Du côté de la théorie des jeux, plusieurs chercheurs ont, au vu des résultats expérimentaux, transformé leurs hypothèses concernant les préférences des joueurs. Au lieu de définir exclusivement ces préférences sur la base des conséquences attendues des actions (*cf.* matrices de paiements), ils ont intégré dans la définition de ces préférences les intentions prêtées par chaque joueur aux autres joueurs, transformant ainsi les jeux classiques en jeux psychologiques (Geanakoplos *et al.*, 1989 ; Rabin, 1993). Cette transformation a permis de prendre en compte, dans le mécanisme de réciprocité qui caractérise la coopération, l'évaluation subjective par chaque individu de l'intention bienveillante, ou malveillante, de l'individu avec lequel il interagit (McCabe *et al.*, 2003), et d'en tirer les conséquences sur l'évolution de la coopération en longue période (Levine et Pesendorfer, 2007). Ces travaux ont abouti, le plus souvent, à opposer cette motivation dérivée de la perception des intentions des autres à la motivation, traditionnelle en théorie des jeux, des gains attendus. Ils ont, d'autre part, invoqué un mécanisme d'imitation, pour expliquer la dynamique de cette coopération sur le long terme.

Dans un modèle à visée plus ambitieuse, Falk et Fischbacher ont d'abord montré comment ces deux motivations des joueurs pouvaient se concilier. Ils ont, ensuite, substitué à la simple imitation un mécanisme plus complexe de réciprocité fonctionnant sur la base de stratégies conditionnelles mises en œuvre par les joueurs. Ces stratégies conditionnelles sont définies non seulement en fonction des gains attendus, mais également en relation avec les informations sur les intentions des autres joueurs, induites des coups précédents. Plus concrètement, les joueurs, au moment d'effectuer leur choix entre différentes stratégies alternatives, tiennent compte du caractère coopératif ou non coopératif des

autres joueurs, au vu des stratégies qu'ils ont déjà mises en œuvre (Falk et Fischbacher, 2006).

De leur côté, plusieurs chercheurs en neurosciences ont pris quelques distances avec une version naïve de la théorie de l'esprit selon laquelle nous disposerions de la faculté mentale de lire directement dans l'esprit des autres (*mind reading*). Comprendre l'autre implique, en effet, d'abord de pouvoir s'identifier soi-même comme moi (le « self »), et de pouvoir ainsi traiter l'autre comme un autre moi. Cette représentation de l'autre, tout à la fois facilite et complique la prise en compte des intentions d'autrui, dans un contexte d'interactions stratégiques. Se pose ainsi au cerveau le délicat problème des niveaux de croyances emboîtées et mutuellement dépendantes (l'intention de l'autre à mon égard dépend de l'intention qu'il me prête à son égard, et ainsi de suite…), un problème bien connu des théoriciens des jeux. Pour le résoudre, ces chercheurs se sont d'abord attachés à élaborer un modèle calculable, inspiré de la théorie des jeux dynamiques, et baptisé, pour cette raison, jeu neuronal. Ils l'ont ensuite testé sur un jeu de coopération réciproque, en l'occurrence celui de la chasse au cerf, et ont recueilli, par l'imagerie, les zones du cerveau activées durant le déroulement de ces opérations (Yoshida *et al*., 2008 ; 2010). Il est ensuite apparu à d'autres chercheurs, que la notion clé de réciprocité, telle qu'elle était testée dans la plupart des protocoles expérimentaux visant à détecter la coopération, se trouvait entachée d'une certaine ambiguïté, puisque, dans les jeux répétés à deux joueurs, le prédécesseur et le successeur de chaque joueur sont la même personne. Ils se sont ensuite attachés à distinguer les différentes composantes des comportements coopératifs et non coopératifs construits sur la base de la réciprocité. Ce constat les a conduits à chercher à séparer expérimentalement les éléments tirés de l'information directe, ou indirecte, dont pouvait disposer un joueur sur l'autre joueur au moment où il se trouve en interaction avec lui, sur son intention de le gratifier ou, au contraire, de le sanctionner à la séquence suivante (Suzuki *et al*., 2011). Ce découplage a permis de dégager les substrats neuronaux des bases cognitives qui guident cette coopération conditionnelle.

Ces percées permettent aujourd'hui de concilier les contributions respectives de la théorie des jeux et des neurosciences à la compréhension de la coopération dans une perspective évolutionniste. Le modèle formel de réciprocité élaboré par Falk et Fischbacher conduit à la définition d'un équilibre de réciprocité, ou plus exactement, puisqu'il s'agit d'un modèle dynamique, d'un équilibre de *reciprocation* (pour reprendre le terme anglais). Ce modèle général trouve autant d'applications particulières dans les situations spécifiques qui correspondent aux principaux jeux utilisés pour analyser les différentes manifestations de la coopération. En effet, l'évaluation par un joueur des intentions coopératives ou non coopératives de l'autre joueur avec lequel il se trouve en interaction se fait en fonction des contraintes stratégiques et des asymétries d'information qui caractérisent les règles de chacun de ces jeux. Évaluer l'intention de coopérer (ou de ne pas coopérer) d'un joueur, à partir de l'observation de la stratégie qu'il a mise en œuvre, nécessite en effet qu'il ait disposé, au moment de son choix, d'autres stratégies alternatives. Le problème posé par le jeu de l'ultimatum est pour cette raison plus délicat. Il existe certes une bande d'équilibre de réciprocité, mais il revient à celui qui propose la règle de partage de l'estimer sans disposer au préalable d'information sur les intentions coopératives de l'autre joueur. Dans le jeu du dilemme du prisonnier répété, en revanche,

un équilibre de réciprocité peut être atteint par un processus d'essais et erreurs, qu'il appartient, cette fois, aux deux joueurs, de mettre en œuvre en procédant à des stratégies de coopération conditionnelles. Un tel équilibre tend naturellement à se renforcer, sous l'effet de la répétition. Il peut néanmoins se trouver brutalement interrompu par une action de l'un des joueurs perçue par l'autre (les autres) comme non réciproque. Ce risque s'accroît avec l'augmentation du nombre des joueurs qui rend plus incertaine l'accessibilité de cet équilibre de coopération.

Cette nouvelle approche de la coopération développée par la théorie des jeux, à partir d'une dynamique de réciprocité conditionnelle, recoupe précisément plusieurs travaux récents de neurosciences. Une équipe japonaise a mis en évidence les bases neuronales de cette notion de coopération conditionnelle, en distinguant expérimentalement, dans un jeu du dilemme du prisonnier répété, la réaction des sujets aux comportements coopératifs (et non coopératifs) observés ou prêtés aux autres joueurs, de l'intention de récompenser ceux qui pratiquaient la réciprocité et de punir ceux qui, au contraire, rompaient cette réciprocité (Suzuki *et al.*, 2011). Pour y parvenir, le protocole classique du jeu du dilemme du prisonnier a été transformé. Chacun des joueurs disposait d'informations, directes ou indirectes, sur le joueur qui l'avait précédé, mais ignorait celui qui jouerait après lui, puisque ce dernier était tiré au hasard dans un ensemble de joueurs potentiels. Cette expérience a permis de dégager l'esquisse d'un modèle neuronal de cette coopération conditionnelle fondée sur la *reciprocation*. Pour qu'un tel modèle ait une suffisante généralité pour rendre compte du phénomène de la coopération, il est nécessaire qu'il puisse également s'appliquer à d'autres situations de jeu. Or il semble aujourd'hui, qu'il puisse également rendre compte des comportements de coopération dans différents jeux de partage.

Le ressort neuronal de ce modèle reposerait sur le fonctionnement de deux systèmes distincts. D'un côté, un système de stimulation autoréférentiel, activé par la partie médiane du cortex orbitofrontal (OMPFC) et la partie postérieure du cortex cingulaire (PCC), entretiendrait le processus de coopération. D'un autre côté, un système d'inhibition, activé par la région droite dorsolatérale du cortex préfrontal (DLPFC), bloquerait ce processus en réaction à un comportement non coopératif observé, ou présupposé tel, de la part d'un autre. Le premier système constituerait le support neuronal de cet équilibre dynamique de coopération reposant sur la réciprocité ; le second, le support neuronal de l'activité mentale qui détecte la non-réciprocité et ajuste, en conséquence, le comportement du sujet.

Ces recherches ne sont encore qu'à leur début. Il convient de confirmer ces deux mécanismes et de préciser la nature de l'antagonisme entre le rôle incitateur, joué par l'OMPFC, et le rôle inhibiteur, joué par la partie droite du DLPFC, dans cette dynamique de coopération. Différentes expériences ont montré que lorsqu'on réduisait artificiellement l'activité de cette partie du DLPLC, il en résultait, chez les mêmes sujets, une propension plus grande à coopérer, même en cas de manquements à la réciprocité (Knoch *et al.*, 2006).

Il faut toutefois interpréter ces résultats avec prudence, puisqu'ils ont été mis en évidence dans le cadre différent du jeu de l'ultimatum. Plus informatives pour ce sujet, car établies à partir d'un jeu du

dilemme du prisonnier répété, sont les données transmises par des travaux concernant l'incidence de pathologies psychotiques sur le mécanisme d'équilibration entre l'activation de ces deux zones cérébrales. Sous l'effet de l'OMPFC, il apparaît que les sujets normaux auraient tendance à enclencher spontanément la coopération avec autrui, tout en plaçant sous le contrôle du DLPLC le mécanisme ainsi initié. Tel ne serait plus le cas chez les sujets psychotiques, qui faute d'une activation suffisante de l'OMPFC, se trouveraient contraints de recourir au DLPLC, qui fonctionnerait, cette fois, dans un sens inverse, pour stimuler leur intention de coopérer (Rilling, Gleen *et al.*, 2007).

Il reste encore à comprendre comment s'articulent les activations des régions de l'OMPFC et du PCC au cours du processus de coopération conditionnelle. Ces aires cérébrales ont, comme on le sait, toutes les deux un rôle capital dans le traitement des informations relatives à l'appréhension de l'autre en référence à soi (*cf.* chapitre 2). Des travaux récents suggèrent l'existence d'une certaine complémentarité entre elles dans cette fonction, en distinguant ce qui relève des apparences de l'autre, de ce qui appartient au jugement porté sur l'autre. Ainsi les informations détenues par un coopérateur potentiel concernant un candidat à cette coopération du type « il a l'air coopératif » se trouvent traitées par le PCC, tandis que la formulation d'un jugement qu'il porterait sur lui, comme, par exemple, « il est coopératif », mobilise l'OMPFC (Moran *et al.*, 2011).

Se pose, enfin, la question de la coordination entre le système qui entretient la coopération conditionnelle (OMPFC et PCC) et le système qui permet de l'interrompre (DLPFC). Sur ce point, des recherches concernant les troubles mentaux occasionnés par des lésions observées dans ces différentes régions cérébrales sur la manifestation (ou la non-manifestation) des comportements coopératifs fournissent de premiers matériaux. D'autres études portant sur la plasticité comparée de l'OMPFC et du DLPFC chez des patients atteints de lésions dans l'une ou l'autre de ces régions apportent des compléments d'information (Forbes *et al.*, 2012).

Est-il possible, à partir de ces résultats, de reconstituer une manière de circuit cérébral de la réciprocité qui constituerait la base neuronale de cette coopération conditionnelle sur laquelle repose, en définitive, l'essentiel de la dynamique coopérative ? Une analyse plus détaillée de l'incidence de deux pepsines chimiques, principalement l'ocytocine, mais également la vasopressine, sur les comportements susceptibles d'être associés à cette coopération conditionnelle, semble pouvoir fournir quelques informations dans cette direction. On sait que l'ocytocine et subsidiairement la vasopressine sont deux produits qui, comme la dopamine, ont des effets ciblés sur l'activation de certaines régions identifiées du cerveau. L'idée est ainsi venue à des chercheurs de comparer les comportements observés chez des joueurs ayant inhalé une dose de ces produits avec ceux des joueurs auxquels on avait donné un placebo, au cours de jeux dont on soupçonnait que la dynamique reposait sur un mode de coopération conditionnelle. Ces expériences ont d'abord été réalisées dans le cadre d'un jeu de l'investissement (Kosfeld *et al.*, 2005), puis, dans celui d'un jeu du dilemme du prisonnier répété, complétées, dans le second cas, par une investigation en neuro-imagerie (Rilling et Sanfey, 2011).

Dans le cas du jeu de l'investissement, où l'effet seul de l'ocytocine a été testé, il ne s'est manifesté que lorsque les joueurs étaient confrontés à d'autres joueurs. Il n'a pas été observé, en revanche, lorsque les joueurs savaient qu'ils jouaient avec un ordinateur programmé par un

algorithme. Plus significatif, son incidence n'a été sensible que sur le comportement de l'investisseur. Les joueurs dans ce rôle qui avaient absorbé de l'ocytocine ont proposé de plus larges sommes aux mandataires que ceux qui avaient reçu des placebos. En revanche l'ocytocine s'est révélée sans effet sur les comportements des mandataires, dont les propositions de partage ont été similaires dans les deux cas. Or nous savons que dans ce type de jeu la position des deux joueurs n'est pas symétrique. Dans le jeu de l'investissement, en effet, seul l'investisseur, prend sa décision sans connaître comment se comportera l'autre joueur. C'est donc sur lui que repose l'engagement du jeu dans la voie d'une coopération conditionnelle qui s'organisera sur la base d'une réciprocité. Il n'en va pas de même pour le mandataire qui, jouant en second, peut déjà interpréter la décision de l'investisseur de lui confier une somme d'argent comme la manifestation d'une volonté de coopération conditionnelle. La prise de risque au sens strict repose donc initialement sur l'investisseur. Mais lorsque l'investisseur est devant un autre joueur humain – et non pas devant une machine – ce risque dépend entièrement du comportement qu'adoptera le mandataire. Pour l'investisseur, par conséquent, ce risque varie en raison inverse du degré de confiance qu'il accorde *a priori* au mandataire. L'effet de l'ocytocine observé seulement chez l'investisseur lorsqu'il se trouve confronté au mandataire semble donc vérifier l'hypothèse du rôle initial joué par un mouvement d'empathie confiante, qui correspond précisément aux activités des zones du cerveau dynamisées par l'ocytocine.

Dans le jeu du dilemme du prisonnier répété, les chercheurs ont testé à la fois les effets de l'ocytocine et ceux de la vasopressine. On retrouve dans leurs résultats la différence observée lorsque les joueurs jouent face à d'autres joueurs ou devant un ordinateur, confirmant ainsi que l'effet de l'ocytocine ne favorise la coopération que lorsque cette coopération s'organise entre des êtres humains. Cependant, s'agissant d'un jeu symétrique, où les mêmes joueurs occupent alternativement le rôle du joueur en premier et celui du joueur en second, aucune différence n'a été observée entre l'un et l'autre de ces deux rôles et des modifications de comportement qui seraient imputables à l'absorption de ces produits. Les comportements de ceux qui les ont absorbés se révèlent à peu près identiques, dans les deux cas, à ceux auxquels on a donné des placebos. Des différences apparaissent assez nettement, en revanche, au niveau des manifestations de la réciprocité. Ainsi l'ocytocine accroît nettement les manifestations de réciprocité positive (« je vais coopérer, parce que tu as coopéré au coup précédent »), mais ne semble pas modifier très sensiblement les réactions de réciprocité négative (« je ne coopérerai pas, parce que tu as fait défaut au coup précédent »). Elle aurait même plutôt tendance à les atténuer. Le mécanisme de réciprocité, sur lequel se construit la confiance, au terme d'un échange de coopérations conditionnelles, semble donc ici confirmé par le renforcement opéré sélectivement par l'ocytocine.

Cela fait déjà assez longtemps que de nombreux travaux ont mis en évidence le rôle de l'ocytocine dans la stimulation des attitudes et des comportements coopératifs envers autrui, et plus généralement des manifestations d'empathie et de plaisir partagé avec autrui. L'intérêt supplémentaire de ces dernières recherches est de mettre en évidence leurs effets sur les comportements des sujets dans deux types de situations interactives distinctes, où l'émergence et le

renforcement de la confiance jouent un rôle déterminant (jeu de l'investissement, jeu du dilemme du prisonnier répété). La seconde étude révèle en outre le rôle spécifique, en la matière, d'un circuit différent, mais encore mal connu, régulé par un autre neurotransmetteur, la vasopressine. La vasopressine agirait notamment sélectivement sur la connexion entre l'amygdale et une partie de l'insula, souvent associée à une sensation désagréable. Son activation, paradoxale de prime abord, pourrait cependant s'entendre ici comme le signal d'un risque de désagrément social qu'un comportement interactif prosocial pourrait permettre d'éviter. Sans entrer dans les détails neurophysiologiques fournis par l'imagerie, les expériences de cette dernière étude permettent d'identifier avec plus de précisions les différentes zones du cerveau qui sont activées et connectées et celles qui sont désactivées, au cours de cette dynamique de *reciprocation* qui caractérise le fonctionnement de la coopération conditionnelle. Cette réciprocité conditionnelle répétée qui, selon nous, fournirait la clé de la coopération entre les individus se trouverait ainsi modulée par des mécanismes cérébraux complexes à l'origine de ce que l'on nomme communément la confiance.

Deux observations importantes doivent être soulignées en conclusion à leur sujet. En premier lieu, cette stimulation de la confiance, sous l'effet de ces neurotransmetteurs, intervient en réponse au comportement d'un autre (*reciprocation*). Mais ce comportement de l'autre peut être observé, comme dans le cas du dilemme du prisonnier répété, ou anticipé ou même seulement imaginé, comme dans celui de l'investisseur, dans le jeu de l'investissement. La confiance fait donc intervenir l'interprétation, par le sujet, du comportement attendu de l'autre, avec tout ce qui contribue à cette interprétation (apparence, réputation, mémoire, désir…). En second lieu, ce renforcement de la confiance par l'action des neurotransmetteurs, ainsi circonscrite, s'accompagne d'une réduction dans l'appréhension émotionnelle du risque qui accompagne, chez un sujet, l'action d'un autre qu'il ne contrôle pas, mais dont dépend sa situation. Il s'agit, toutefois d'un risque particulier, puisque sa matérialisation dépend entièrement ici du seul comportement de cette autre personne. Nous verrons, par la suite, que l'introduction de cette dichotomie entre les risques, selon qu'ils sont imputables à l'environnement, ou à une personne particulière, se retrouve au niveau des types de confiance qui seront distingués.

D

Plusieurs faits concernant la coopération entre les individus semblent aujourd'hui établis. En premier lieu, il se confirme, d'expériences en expériences, qu'une large majorité de sujets se trouve disposée à engager une coopération avec les autres, lorsque cette occasion leur est offerte. Ainsi dans un très grand nombre de cas, le joueur qui joue en premier dans ces différents jeux expérimentaux opte pour la coopération, même lorsqu'il s'agit de jeux asymétriques à un coup. Un tel comportement ne peut donc pas s'expliquer ici par la dynamique de réciprocité précédemment esquissée, sauf comme nous l'avons suggéré, dans la perspective imaginaire d'une amorce de ce processus. Son

origine serait donc peut-être à rechercher en amont, au niveau de la génétique. Des travaux comparatifs avec des comportements observés chez certaines espèces animales, et en particulier les chimpanzés, sont à cet égard instructifs, tant par les similitudes que par les différences qu'elles mettent en évidence (Boesch et Tomasello, 1998 ; Cosmides et Tooby, 2005 ; Langergraber et Boesch, 2011).

En second lieu, la coopération elle-même est un phénomène qui n'est intelligible que dans une perspective dynamique. C'est l'une des raisons pour lesquelles le cadre initial structurellement statique, dans lequel a été initialement élaborée la théorie des jeux, s'est d'abord révélé inadapté pour en rendre compte. Les choses ont changé, comme on l'a montré, avec le nouveau formalisme des jeux dynamiques. Plusieurs raisons peuvent être avancées pour expliquer la fécondité de ce nouveau cadre théorique pour recueillir, regrouper et formaliser les informations d'origine psychologique expérimentale, neuronale, voire génétiques, concernant la coopération. Il permet de dégager, par-delà les règles initiales d'un jeu considéré, des règles implicites que les joueurs élaborent au cours de son déroulement. Les règles de réciprocité qui s'instaurent par l'intermédiaire de ce que nous avons appelé la *reciprocation* en fournissent une illustration. Elles permettent aux joueurs, et, par extension, à tous les individus placés dans les mêmes situations ou dans des situations analogues, d'entretenir une pratique coopérative qui, sauf accident, se renforce avec le temps.

D'autres règles ont été proposées qui seront étudiées et discutées dans les chapitres suivants. Le concept central d'équilibre en théorie des jeux s'est trouvé modifié dans cette perspective dynamique. Il se construit au cours du jeu, au gré d'un mécanisme complexe d'actions/réactions qui fait intervenir, à chaque séquence, la mémoire des coups précédents et l'anticipation que les joueurs en tirent pour les coups suivants, sur la base de cette observation. En outre, l'évaluation des joueurs qui guident leurs anticipations ne dépend pas seulement de leurs gains attendus du fait de la coopération, mais, également, des intentions coopératives (ou non coopératives) des autres joueurs susceptibles d'être induites de leurs actions observées. La coïncidence positive de ces deux dimensions a même permis à certains auteurs de proposer une nouvelle définition de l'équilibre, entendu comme un « état mental » perçu comme satisfaisant et qui serait partagé par les joueurs (Bhatt et Camerer, 2005). Les nombreux travaux menés par les chercheurs en neurosciences pour saisir les bases neuronales de la coopération ont ainsi trouvé un support dans cette nouvelle définition d'un équilibre. Il est, en effet, possible de mieux connaître maintenant le fonctionnement des cerveaux correspondant à cet état mental associé à une forme d'équilibre coopératif.

Le rôle particulier joué par la temporalité dans l'émergence et le développement de la coopération doit ici être précisé. Si les expériences directes induites de l'observation des actions des autres au cours des séquences précédentes représentent, pour le sujet, autant d'indices des intentions des autres de coopérer (ou de ne pas coopérer), ils ne sont pas les seuls. Faute d'expériences réelles, les apparences des autres peuvent fournir les éléments d'une sorte de mémoire imaginaire, construite par rapprochements avec d'autres expériences passées. Ainsi les souvenirs accumulés à partir d'expériences antérieures, avec d'autres personnes, contribuent à alimenter l'intention de coopérer. En outre, le rôle joué par la répétition dans le processus de mémorisation est aujourd'hui connu. Il

intervient ici dans la stabilisation du mécanisme de *reciprocation* que l'on trouve à la base des interactions coopératives. On sait, par exemple, qu'au bout d'un certain nombre de répétitions et après un laps de temps suffisant, le cerveau anticipe le comportement d'autrui, sans attendre la confirmation apportée par la dernière séquence observée. Cette mémorisation antérieure pèse fortement dans la projection mentale qui conduit les sujets à coopérer. On l'a notamment vérifié lorsque la volonté de coopérer au cours de séquences répétées se manifeste chez des sujets, qui, forts de leurs expériences antérieures avec d'autres joueurs, sont prêts à coopérer avec la nouvelle personne en face d'eux qu'ils ne connaissent pas. Enfin, transformer les informations recueillies sur les comportements des autres en un jugement porté sur eux, même temporaire, fait nécessairement intervenir la durée. C'est ainsi que se forment notamment les réputations qui, en engendrant la confiance, facilitent la coopération. Quant aux stratégies visant une coopération, même conditionnelle, elles font intervenir des anticipations sur un horizon temporel de longue période, qui exige un travail cognitif important et mobilisent les régions cérébrales correspondantes. De telles anticipations, qui portent sur un système d'interactions, sont d'autant plus fines qu'elles nécessitent différents niveaux d'itérations (je pense que l'autre coopérera, je pense que l'autre pense que je pense qu'il coopérera, etc.). Ce mode de raisonnement se complique encore, lorsque le nombre des joueurs augmente. On retrouve ici la question de la confiance et ses liens avec les réseaux. Loin par conséquent d'être primaire et spontané, le raisonnement qui conduit à l'adoption d'une stratégie coopérative peut au contraire se révéler beaucoup plus élaboré.

Le troisième acquis concerne la diversité des opérations mentales qui concourent, ou peuvent concourir, à la coopération, selon les formes variées sous lesquelles se présentent aux individus les occasions de coopérer. Après avoir d'abord cru pouvoir regrouper dans un même moule les diverses informations fournies par l'imagerie cérébrale au cours des différents jeux expérimentaux qui ont été rapportés, les chercheurs en neurosciences ont maintenant une autre représentation de l'activité cérébrale qui guide la coopération. En étudiant en détail les corrélats neuronaux correspondant aux spécificités de chaque jeu, les différentes motivations de l'investisseur et du mandataire dans le jeu de l'investissement (Van den Bos *et al.*, 2009), la formation et les corrections des anticipations des joueurs, au cours du jeu du dilemme du prisonnier répété (Suzuki, 2011), le modèle spéculaire de raisonnement des chasseurs dans la chasse au cerf (Yoshida *et al.*, 2010), ils ont réussi à mettre en évidence une grande variété de mécanismes, le plus souvent complexes. Cette découverte traduit d'abord les propriétés de plasticité et de modularité qui permettent aux cerveaux humains d'adapter le traitement de la coopération aux variétés des situations sociales dans lesquelles elle se manifeste. L'évolution récente des recherches sur la coopération en neurosciences les rapproche, comme nous l'avons montré, des travaux d'approfondissement des différentes logiques de situations qui caractérisent ces jeux. La propriété d'adaptabilité des réseaux neuronaux sollicités permet d'expliquer, notamment, comment se construisent, par l'intermédiaire de l'interaction des cerveaux, des règles tacites entre les joueurs, chaque fois différentes, destinées à favoriser et à renforcer la coopération. Elles ne sont pas exactement les mêmes lorsque la coopération porte sur un problème de

partage (jeu de l'ultimatum et de l'investissement), ou sur la question d'actions coordonnées visant l'acquisition d'un avantage commun (jeu du dilemme du prisonnier et de la chasse au cerf, jeux des biens publics). Des sous-distinctions seraient à introduire dans cette seconde catégorie, selon que le non-coopérateur pourrait (dilemme du prisonnier) ou ne pourrait pas (chasse au cerf) exploiter à son profit les actions à visée coopérative des autres. Quant aux règles tacites de coopération élaborées par les joueurs au cours du déroulement de ces différentes situations de jeux, elles peuvent s'entendre comme des ferments de sociabilité qui facilitent le développement social de la coopération.

Deux ressorts principaux de la coopération s'imposent toutefois assez généralement, par-delà la diversité des circuits cérébraux activés et désactivés en chaque circonstance. Il s'agit, d'une part, d'une prédisposition à coopérer, attribuable à l'appartenance des sujets à un même groupe, ou peut-être, plus largement, à une même espèce. On observera que l'attractivité de cette prédisposition varie entre les groupes (degrés de proximité, système culturel…), mais également d'un individu à l'autre. Des facteurs génétiques y jouent sans doute un rôle qui commence à être identifié, mais ils ne suffisent pas à en rendre complètement compte. Des éléments culturels, mis en évidence à l'occasion d'autres travaux d'ordre socioanthropologique, portant notamment sur la confiance, peuvent également contribuer à expliquer des comportements qui traduisent, de la part de ceux chez qui ils sont observés, une confiance aux autres que l'on pourrait qualifier d'« inconditionnelle », par opposition à la confiance « conditionnelle » qui s'instaure peu à peu, au fil des interactions interpersonnelles (Krueger *et al.*, 2007).

On retrouve, d'autre part, dans presque toutes les manifestations de coopération qui ont été étudiées, la référence à un mécanisme de *reciprocation*, qui fait intervenir à la fois, comme on l'a vu, des composantes affectives et des composantes réflexives. Ces différentes composantes se trouvent activées selon des modalités diverses. L'étude de leur système de fonctionnement a débuté en partant de l'hypothèse que cette *reciprocation* pourrait reposer sur des stratégies conditionnelles mutuellement pratiquées par les coopérateurs. Cette hypothèse d'une interaction de stratégies conditionnelles réciproques a fait l'objet d'un modèle sous forme de jeu (Falk et Fischbacher, 2006). Elle a, d'une certaine manière, été validée au niveau neuronal, dans le cas d'un jeu du dilemme du prisonnier répété (Suzuki *et al.*, 2011). Elle peut également rendre compte d'autres situations, où s'instaure une manière d'équilibre coopératif. Il reste cependant à vérifier si cette apparente similitude correspond effectivement aux mêmes supports neuronaux, dont les modalités de fonctionnement seraient alors seulement modifiées pour s'adapter aux caractéristiques particulières de ces différentes situations.

La transposition de ces résultats, principalement établis à partir d'échantillons limités sur des relations directes de personne à personne, à des populations plus larges d'individus n'est pas immédiate. Or c'est à cette échelle que les conséquences économiques et sociales des phénomènes de coopération sont les plus significatives et socialement les plus déterminantes. Nous avons montré dans le chapitre précédent que des travaux récents, concernant l'organisation de la communication en réseaux sociaux et le fonctionnement de tels réseaux, ouvrent des perspectives sur le passage du niveau strictement interindividuel de la coopération à un niveau plus proche du collectif. La

coopération représente un facteur majeur de transmission de la sociabilité, qui transite de plus en plus aujourd'hui par l'organisation de réseaux. C'est pourquoi il devient nécessaire de relier plus étroitement les recherches en cours sur les déterminants neuronaux de la coopération aux travaux menés sur la communication des hommes à travers ces réseaux sociaux.

Commentaire Le « nous » de la coopération et les modes de confiance

Ce chapitre a abordé une étape supplémentaire des interactions. Au « raisonnement par équipe », et même au jeu de la chasse au cerf, correspondait déjà une prise en compte du collectif. Mais choisir la coopération dans le dilemme du prisonnier et plus généralement dans le dilemme des biens publics exige de se rapporter au collectif en tant que différent des individus. Le raisonnement par équipe requiert seulement de se placer du point de vue du groupe. La coordination collective de la chasse au cerf exige de passer d'un contrôle individuel direct à un contrôle collectif interactif, voire indirect. Mais il reste clair, dans chacune de ces deux situations, qu'une fois parvenu au point de vue collectif, chaque individu a intérêt à y rester. En revanche, dans le dilemme des biens publics, une fois que nous avons choisi la coopération, il devient au contraire intéressant, individuellement, de la trahir. Poursuivre la coopération dans ce cadre, c'est donc faire le choix du collectif contre les intérêts individuels. Nos sociétés humaines sont habitées par cette tension entre le choix du collectif et la possibilité pour des individus de tirer profit de la coopération des autres.

Nous avions rappelé que passer au point de vue du collectif impliquait un saut. Mais ce saut, l'évolution avait pu le faire pour nous. Il suffisait que la coexistence quelque peu compétitive entre les espèces se combine avec des mutations des individus pour amener des lignées dotées de caractères plus sensibles aux relations de groupe à survivre et à se reproduire avec un peu plus de succès que les autres, pour que, dans une population, des traits qui renforcent les possibilités de coordinations collectives deviennent dominants. L'espèce humaine étant une de ces espèces « sociales », le mode par défaut pour chacun de ses membres est d'adopter le point de vue du groupe. Le mode individualiste est alors un mode moins fréquent.

On notera que les mécanismes de l'évolution finissent par obtenir les mêmes effets que l'équilibre entre point de vue collectif et point de vue individuel que nous avons esquissé dans le chapitre précédent. Une fois qu'à l'échelle collective d'une lignée, des traits coopératifs sont apparus, ils avantagent bien les individus membres de cette lignée, du point de vue de leur groupe – par des

phénomènes d'entraide ; et les tendances des individus sont bien alors celles qui améliorent les chances de cette lignée de se reproduire, voire de s'étendre.

Cette coordination – voire coopération – comme mode par défaut semble bien apparaître dans les jeux du dictateur (un joueur choisit la répartition que l'autre ne peut refuser) et de l'ultimatum (le partenaire peut ou non accepter cette répartition). La tendance dominante des joueurs est de choisir une répartition qui ne soit pas trop inégalitaire. Dans les dilemmes du prisonnier et des biens publics, cette tendance apparaît aussi, mais la tentation individualiste peut s'auréoler des lauriers de la prudence de la rationalité stratégique, puisqu'elle combine stratégie dominante (préférable y compris dans le cas où autrui ne coopérerait pas) et équilibre de Nash. Dès lors, la dominance de la tendance atavique à la coopération n'est plus si aisée à interpréter ni à conserver. Mais du coup, pour ne pas céder à la tentation de l'exploitation des coopérateurs, il faut accorder une valeur au collectif supérieure à celle des intérêts individuels ; et, pour ne pas jouer la prudence en faisant défection, il faut que notre confiance ou notre espoir dans le collectif surpasse notre crainte des exploiteurs individuels. De plus, dans le dilemme des biens publics (mais pas dans celui du prisonnier), quand nous coopérons, c'est en ne sachant pas qui parmi les membres du collectif va coopérer et qui pourrait faire défection. Cela reste dans le vague. Or c'est une caractéristique des sociétés humaines, dès que la population est suffisamment nombreuse, que l'on n'ait plus de moyen individuel de s'assurer des conduites de chacun des autres. La confiance devient alors non plus une confiance en autrui, mais une confiance dans le collectif, un collectif dont, pour chacun des membres, une partie des membres reste anonyme.

C

On voit que le terme de confiance peut recouvrir des processus bien différents. Notre tendance atavique à coopérer ne semble pas mériter le nom de confiance, puisque nous ne faisons alors que suivre une tendance héritée de l'évolution. Cette tendance n'implique pas l'incertitude sur autrui qui est un composant du concept de confiance, celle-ci consistant à surmonter celle-là. Les coordinations collectives du type de la chasse au cerf peuvent exiger de la confiance, puisque nous devons faire le pari que les défaillances des partenaires ne seront pas trop importantes. Mais cette incertitude ne remet pas en cause la dominance du point de vue du collectif, une fois que nous sommes parvenus à considérer les choses de ce point de vue et que nous avons atteint l'équilibre entre points de vue collectif et individuel. Atteindre cet équilibre, c'est ce à quoi nous amène l'éducation interactive que nous avons tous reçue. Comme on l'a dit, nous sommes les sujets d'apprentissages qui se poursuivent tout au long de notre vie. Qui dit apprentissage dit à la fois cadres présupposés pour cet apprentissage – il nous faut des repères – et sensibilité aux différences par rapport à nos attentes. On peut se demander quel est le degré de normativité de ces cadres. Une première théorie des économistes pour expliquer la coopération dans le dictateur et l'ultimatum était que nous avions une aversion à

l'iniquité – au sens de distribution inéquitable, voire inégale, des gains. Mais les expériences ont montré que nous ne jugeons pas simplement selon les résultats de la répartition, selon les conséquences des choix des joueurs. Nos évaluations changent si nous prêtons attention aux intentions des partenaires. Notre apprentissage reste donc interactif. Il nous amène seulement à adopter une attitude coopérative au sein de notre groupe, surtout quand nous en connaissons tous les membres, tout en restant sensible à l'incertitude des coordinations collectives. Notre éducation nous fournissant les cadres de notre vie sociale, cadres que nos conduites ne cesseront de présupposer, nous pouvons ici parler d'une confiance-cadre.

En revanche, dans le dilemme du prisonnier ou des biens publics, on parlera plutôt d'une confiance-pari, parce qu'il nous faut surmonter notre incertitude pour parvenir à la confiance, au lieu que cette confiance se présente à nous d'emblée comme l'attitude sociale de base entre membres connus d'un même groupe. On peut se demander si dans le *Trust Game*, c'est la confiance-cadre ou la confiance-pari qui est en jeu. La sensibilité aux intentions des partenaires est saillante dans le *Trust Game* (Falk, Fehr, Fischbacher 2008), et il faut que l'investisseur y parie sur la tendance du mandataire à lui renvoyer une partie de ses gains. Si dans un jeu en un coup, on peut penser demeurer dans la confiance-cadre, la répétition du jeu implique des ajustements en fonction des précédentes réactions, et semble plutôt relever de la confiance-pari.

Les deux modes de confiance semblent nécessaires au fonctionnement des sociétés humaines. La confiance-cadre est nécessaire pour permettre aux apprentissages sociaux de disposer d'un cadre stable de référence – ainsi dans le milieu familial, ou dans le groupe des connaissances. La confiance-pari est aussi nécessaire, justement pour permettre des extensions à des réseaux de relation plus larges que celles de groupes où nous connaissons tout le monde.

R

La neuroéconomie nous conduirait-elle à oublier cette différence entre modes de confiance ? On pourrait le croire car, dans le flot d'études sur l'effet d'hormones comme l'ocytocine (*oxytocin* en anglais), on a souvent du mal à retrouver ces différences entre confiance-pari et confiance-cadre, entre confiance entre membres quelconques d'une société qui peuvent ne pas se connaître et confiance réglée sur des stéréotypes de relations à l'intérieur d'un groupe connu. Des études comme celle de Baumgartner (2008) ne font pas cette différence et suggèrent simplement que l'ocytocine réduit l'appréhension, ce qui vaut évidemment aussi bien pour la confiance-cadre que pour la confiance-pari. Revenons après Christian Schmidt sur quelques autres particularités des effets de cette substance.

Une revue d'un grand nombre de ces travaux (K. et T. MacDonald, 2010) montre que l'ocytocine est liée aux interactions sociales. Une interprétation moins spécifique serait que l'ocytocine est essentiellement un anxiolytique (Evans *et al.*, 2013), ce qui ferait supposer quelque anxiété dans nos relations sociales. Cependant, si on compare des réactions à des images d'humains qui sont saillantes

émotionnellement et des réactions des images de scènes porteuses d'émotions mais sans interaction sociale (situations dangereuses), l'injection d'ocytocine réduit la portée des saillances émotionnelles qui pourraient donner lieu à interaction sociale, mais pas celle des images de scènes non interactives. De même, si des sujets soumis à ocytocine montrent plus de confiance dans la tendance de partenaires à leur renvoyer l'ascenseur (dans le *Trust Game*), cela ne vaut que si le risque est pris dans un contexte d'interaction, pas s'il s'agit d'une loterie. Une hypothèse évolutionniste (par exemple Churchland, 2011) est que l'ocytocine joue d'abord un rôle dans l'attachement de la mère à ses enfants, et que son fonctionnement a ensuite été « recruté » pour des fonctions de socialisation plus larges.

Mais de quelle socialité s'agit-il ? De Dreu (2012) se basent sur leurs propres études précédentes pour soutenir que l'ocytocine accroît la confiance – ou diminue la méfiance – surtout vis-à-vis des membres de notre propre groupe, et bien moins pour les membres d'autres groupes. Elle pourrait même diminuer la préférence pour une répartition équitable dans des jeux avec des partenaires anonymes (Pfeiffer, 2013). L'ocytocine peut d'ailleurs aussi bien accroître l'agressivité vis-à-vis des intrus que renforcer la bienveillance pour les proches ou les membres du groupe. Lambert *et al*. (2014) vont jusqu'à penser que l'ocytocine n'accroît pas la capacité à détecter les personnes dignes de confiance, mais plutôt la capacité à détecter celles qui ne le sont pas, ce qui en ferait un facteur de ségrégation entre mon groupe et les autres groupes. Certes, Shamay-Tsoory *et al*. (2013) ont trouvé que chez des sujets soit israéliens soit palestiniens à qui on demande d'évaluer la souffrance manifestée par des personnes prises en photo, et leur propre empathie avec cette souffrance, l'ocytocine accroît la sensibilité à la souffrance des membres du groupe adverse, alors qu'elle n'a pas d'effet pour la sensibilité aux membres du groupe d'appartenance. Mais on peut penser que c'est là un effet de la présentation de souffrances et de questions portant sur le degré d'empathie.

Notre hypothèse pour tenter de rendre compte de ces expériences dont les résultats semblent aller dans des directions assez variées est que l'ocytocine a surtout des effets dans des situations où on pourrait osciller entre deux options : traiter une personne comme nos proches ou nos relations, ou bien la traiter comme n'étant pas des nôtres. Le cas de la souffrance est particulier, parce que normalement notre empathie en ce domaine va au-delà de notre groupe. Dans d'autres cas, la différence entre les deux groupes est déjà fixée. Or ce ne sont donc pas dans les situations déjà tranchées que l'ocytocine peut avoir un effet. Ainsi dans l'expérience citée, l'ocytocine n'accentue pas l'empathie pour le groupe d'appartenance. Ou encore, l'ocytocine n'atténue pas la différence que nous faisons entre des visages dont l'orientation implique une saillance sociale forte (ils sont tournés vers nous) et ceux qui ne présentent pas cette saillance.

Il faut d'ailleurs distinguer ici entre la présence endogène d'ocytocine et les effets de son injection exogène. Dans le *Trust Game*, de hauts niveaux *endogènes* d'ocytocine sont bien en corrélation, chez le receveur ou mandataire, avec un retour de sommes plus importantes à l'investisseur. Autrement dit, la présence endogène d'ocytocine est un indicateur que le sujet se sent dans une relation de proximité avec ses partenaires. Zak et Fakhar, (voir Zak, 2006) ; Zak a multiplié les publications qui mettent l'ocytocine à la base des interactions sociales, avec sans doute un enthousiasme un peu excessif ; certains ont critiqué sa tendance à surinterpréter ses résultats (Conlisk,

2011) et ont d'ailleurs noté (en comparant un jeu de *Trust Game* avec un jeu qui donnait des gains similaires, mais qui était en fait une loterie) que le taux d'ocytocine endogène augmente chez le receveur de l'investissement seulement une fois que l'investisseur a donné un signal de sa confiance en envoyant une somme importante.

Mais ce qui nous intéresse ici, ce sont les influences d'une injection exogène (injection intranasale, sans qu'on sache clairement comment l'ocytocine peut passer la barrière qui protège le cerveau de substances d'origine externe). Or celle-ci a un effet dans des situations où le sujet peut se poser des questions sur sa relation avec un nouveau partenaire. Ainsi dans le jeu de l'ultimatum et le jeu de la confiance (*Trust Game*), le proposant est dans l'incertitude : le partenaire va-t-il accepter la répartition proposée, le receveur de l'investissement va-t-il renvoyer l'ascenseur ? L'injection d'ocytocine a donc un effet sur celui qui est dans l'incertitude, sur l'investisseur, comme l'a rappelé Christian Schmidt. Elle n'en a pas sur les choix du receveur de la somme investie, le mandataire, qui n'est pas dans cette incertitude. Dans le jeu du dictateur, où le partenaire receveur ne peut influer sur celui qui décide de la répartition, l'ocytocine n'a pas d'effet sur le décideur. Celui qui répond à l'interaction sans avoir à se soucier d'un retour futur n'est pas influencé par une injection d'ocytocine.

L'ocytocine est-elle plutôt liée à la confiance-cadre ou à la confiance-pari ? Notre hypothèse est que l'ocytocine endogène est liée à l'installation d'une confiance-cadre, alors que l'injection exogène d'ocytocine a un effet significatif dans les conditions où il faut décider de se lancer ou non dans une confiance-pari. Ainsi, les premiers contacts entre mère et progéniture déclenchent une montée endogène d'ocytocine, qui est donc liée au démarrage de la confiance-cadre la plus basique. Inversement, une étude montre que l'injection d'ocytocine n'a pas d'effet sur le degré d'empathie de sujets vis-à-vis de la souffrance de leurs parents, et ces relations entre proches ressortent de la confiance-cadre. Mais ensuite, l'ocytocine intervient surtout dans des situations où il s'agit de démarrer une relation de confiance. Or les relations sociales qui dépassent le cadre des proches ou du groupe d'appartenance nous exposent à des situations de confiance-pari – dont certaines vont ensuite donner lieu à une confiance-cadre.

T

L'inspiration évolutionniste que nous signalions chez Churchland à propos de l'ocytocine, nous la retrouvons, liée à des calculs économiques, dans des modèles proposés par Gintis, Bowles, ou Zak. La coopération émerge comme un résultat collectif, à l'échelle d'une population. Les modèles proposés font alors interagir des acteurs qui sont supposés chacun appartenir à un type stylisé fixé : des coopérateurs qui coopèrent inconditionnellement – nommés « altruistes » ; des individus qui coopèrent si les autres coopèrent, voire qui ne le font que si de plus les non-coopérateurs encourent eux-mêmes des sanctions. Ces derniers peuvent se diviser en ceux qui recherchent leur propre intérêt mais s'adaptent aux tendances de ceux qui les entourent, les « machiavéliens », un sous-groupe des

acteurs centrés sur leur propre intérêt, et en ceux qui ont tendance à coopérer si leurs partenaires coopèrent, qu'on peut appeler des altruistes « si réprocité » ou « conditionnels ». Parmi ces derniers, on peut trouver ceux qui présentent une « réciprocité forte ». Ils vont plus loin que les autres altruistes « si réciprocité ». Ceux-ci coopèrent si les autres coopèrent, ou s'ils peuvent en espérer des bénéfices futurs, par exemple par des effets de réputation, et font défection dans les cas contraires. Les individus à forte réciprocité non seulement coopèrent, mais punissent les non-coopérateurs même s'ils ne peuvent espérer de cette conduite de punition aucun gain présent ni futur.

Ces divisions entre des types d'acteurs semblent exiger des séparations plus rigides entre les conduites des individus que celles qu'on peut observer dans des interactions sociales. Ainsi, l'altruiste pur est censé coopérer non seulement avec ceux qui lui ont laissé des expériences positives de coopération mais aussi avec des acteurs anonymes. Or quand on observe dans le cadre restreint d'une expérience un sujet réel qui présente cette conduite, on ne peut exclure qu'il agit ainsi parce qu'il projette sur ces acteurs anonymes les leçons qu'il a tirées d'interactions coopératives avec des acteurs avec qui il était en échange régulier. Inversement, quand on observe une conduite de méfiance, on pourrait l'attribuer aussi bien à un altruiste « conditionnel », qui a pu être marqué par des expériences de défection, qu'à un individu seulement guidé par son propre intérêt. Les acteurs sociaux ont eu une histoire d'interactions sociales avant de participer aux expériences. L'expérience est d'ailleurs pour eux une interaction sociale d'un type nouveau, qu'ils interprètent le plus souvent comme leur proposant un défi : celui de deviner quelle est la « bonne » conduite, celle qui ferait que le sujet aurait le comportement le plus socialement adapté à ce qui est attendu de lui – par d'éventuels partenaires, voire par les expérimentateurs. Leur simple participation à l'expérience montre déjà que les sujets ont une certaine tendance à la coopération, tendance qui peut être soit diminuée par une compétition implicite pour répondre au défi en question (par exemple quand on leur dit que leurs gains dépendent de leur réponse), soit accrue si les sujets pensent que le défi à relever est celui de la coopération.

L'économie expérimentale devrait donc tenter de dépasser sa tendance – encore plus marquée qu'en psychologie expérimentale – à raisonner sans prendre en compte l'histoire des apprentissages sociaux de ses sujets. De même, on peut se demander si les modèles qui exigent de constituer des populations composées d'acteurs stylisés, lesquels s'en tiennent à agir rigoureusement selon un des types répertoriés, pour pouvoir faire tourner leurs simulations, peuvent avoir des résultats qui soient similaires à ceux qu'on pourrait observer dans une population réelle. Dans la vie sociale, chaque acteur peut présenter en coexistence les caractéristiques de ces divers types que sont les différentes propensions à la coopération, à la défection, à la punition des non-coopérateurs, même si c'est pour chacune à différents degrés. La question est importante pour une épistémologie qui procède par simulations, en recourant aux SMA (simulations multiagents) pour obtenir des résultats qu'on pense pertinents pour analyser les sociétés humaines. Mais elle reste ouverte. Même si on est tenté de répondre positivement pour ce qui est de la pertinence des simulations, celle-ci peut rester indirecte, et ne pas nous permettre de transposer sans autres conditions les résultats obtenus. Il faut noter cependant un avantage des simulations : elles permettent de voir évoluer des populations, et donc d'étudier des dynamiques de toutes formes, alors qu'un raisonnement ne pourra porter que sur un petit

nombre d'agents différents. L'économie, en recourant à ces simulations, peut donc échapper à sa tendance à se focaliser sur les situations d'équilibre. En effet la plupart des dynamiques des simulations ne vont pas converger vers un équilibre, ce qui nous permet un peu plus de réalisme, puisqu'il est difficile de rencontrer de telles convergences et stationnarités dans les interactions sociales.

Il resterait cependant à doter ces agents stylisés non seulement d'une dynamique collective, mais d'histoires personnelles et interpersonnelles. Nous ne sommes pas les mêmes acteurs à chaque étape de notre histoire. Il est probable que, si nous n'avions pas eu un cadre protecteur pendant notre petite enfance, nous ne présenterions pas les mêmes tendances à la coopération une fois sortis de ce cadre familial. À titre de contre-épreuve, des enfants qui ont subi des brimades répétées peuvent avoir des comportements systématiques de défiance et de refus de coopération. Certes, la coopération avec de nouveaux partenaires n'est pas du même type que celle du cadre familial. Mais l'incertitude de sa réussite est mieux supportée si l'on dispose d'une base qui reste sûre avant de se lancer dans cette aventure. Une fois des coordinations avec des partenaires extérieurs réussies s'installe progressivement un cadre d'attentes mutuelles, dont on a appris à pouvoir supposer de manière fiable la robustesse. Ce processus de lancement d'une nouvelle extension à partir d'une base plus assurée peut se reproduire ensuite. Cette construction progressive d'attentes ne peut être très rapide. Mais sa lenteur a un versant positif : quand notre confiance est trahie par un partenaire, nous n'abandonnons pas pour autant toutes nos attentes de coopération avec les autres partenaires, nous considérons plutôt sa défection comme une exception – à moins de vivre dans un milieu extrêmement compétitif et où la recherche de l'intérêt personnel est érigée en vertu.

On peut avancer sans trop de risque que les situations que nous avons rencontrées dans nos apprentissages sociaux ont présenté soit des interactions fiables et coopératives, soit des interactions où la coopération était plus risquée et plus sujette à exploitation, ou encore où il valait mieux y être l'exploiteur que l'exploité. Les tendances à la coopération observées dans les jeux expérimentaux comme le *Trust Game* ou le jeu du dilemme du prisonnier ou celui des biens publics ne seraient pas possibles d'une part si les situations d'interactions fiables n'avaient pas pu constituer une base de confiance-cadre, d'autre part si elles ne nous avaient pas offert aussi des occasions de nous risquer hors de nos cadres usuels et de tenter des interactions nouvelles et plus risquées, mais qui nous offraient une extension attrayante de notre réseau d'interactions sociales et pouvaient de surcroît nous donner quelque aura sociale supplémentaire auprès de notre entourage habituel.

Si ces hypothèses sur nos apprentissages sont correctes, alors nous y avons fait l'expérience de basculements de la confiance-cadre à la confiance-pari – quand nous nous risquions sur un terrain social moins connu – comme aussi du basculement inverse, une fois que les risques pris ont été payants et que notre nouveau réseau d'interactions est devenu fiable. Nous avons aussi fait l'expérience du passage de la confiance-pari à la méfiance et au repli sur notre cadre usuel plus fiable, quand nos paris ont été perdus (il vaut donc mieux considérer que notre cerveau ne réagit pas simplement à une récompense ou à une peine, mais entretient un certain nombre d'attentes, une

distribution de satisfactions et d'insatisfactions) et qu'il réagit à des variations importantes par rapport à ces attentes – *cf.* Xiang, 2013). Il est heureusement bien plus rare, mais malheureusement possible que notre cadre habituel de confiance s'écroule lui aussi, et cela mène certaines personnes parfois jusqu'au suicide.

On peut alors se demander si l'on peut réduire nos processus cognitifs, dans ces situations où nous sommes amenés à prendre le risque de coopération, à un calcul coûts-bénéfices entre les avantages de la coopération et ceux de la défection – calcul qui peut amener à envisager de punir les non-coopérateurs pour les inciter à coopérer. Car ces situations semblent mettre en œuvre une dynamique de basculement d'une perspective d'attentes sociales à une autre, de la confiance-cadre à la confiance-pari ou inversement. Or cette dynamique sociale est complètement passée sous silence, quand les auteurs des expériences de neuroéconomie interprètent les données d'imagerie cérébrale corrélées aux réactions de punition à des ruptures de coopération, en y voyant l'activité d'un calcul d'optimalité (opéré dans le préfrontal) entre les avantages d'avoir affaire à un partenaire coopératif et le coût de punir celui qui ne coopère pas. Pourtant l'imagerie cérébrale montre souvent dans ces expériences l'activation de zones liées à nos représentations des perspectives des autres, ceux dont nous pouvons espérer la coopération ou craindre la défection. Ce sont ces espérances et craintes qui nous engagent dans une dynamique prospective et modifient les cadres de nos calculs, ce qui pourrait correspondre aux processus qui nous amènent à basculer de la confiance-cadre, qui demande moins d'analyses, à la confiance-pari.

L

Élargissons encore la perspective en revenant au rôle des tiers dans la mise en place des interactions sociales. Nous avons été formés socialement dans des interactions sociales, et ce qui en témoigne est l'importance fondamentale dans la mise en place de ces interactions du rôle des *tiers*, ceux qui nous observent et nous jugent. Dès lors, nos comportements de coopération ne sont pas simplement des apports à la construction du collectif des coopérateurs. Si nous allons coopérer entre « nous », c'est aussi par référence à la position de ce « nous » *par rapport aux tiers* (notons que l'économiste peut en un sens prétendre intégrer une représentation formelle des points de vue des partenaires et des tiers quand le théoricien des jeux injecte dans ses acteurs la connaissance qu'a ce théoricien sur le jeu en question).

C'est assez récemment que ce rôle des tiers a commencé à être analysé dans les expériences de neuroéconomie – en introduisant dans ces expériences des acteurs qui sont en position d'observation par rapport aux interactions entre les partenaires (ainsi dans Chavez et Bicchieri, 2013, où ce sont les tiers qui peuvent compenser des échanges inégaux ou punir les profiteurs). La présence de ces acteurs-observateurs peut s'y réduire à des images de visages présentées sur l'écran, ou encore se manifester par la mise à notre disposition d'évaluations externes des comportements des partenaires. La tendance

actuelle de la neuroéconomie est donc de ne plus se borner à voir dans les activités cérébrales simplement des évaluations des plaisirs et des peines et des bilans de coûts et de bénéfices immédiats, mais à tenir compte de l'influence sur ces évaluations de notre sensibilité aux variations d'intention des partenaires des interactions, mais aussi aux évaluations manifestées ou supposées des tiers observateurs. C'est cette prise en compte qui a conduit Falk, Fischbacher et Fehr (2008) à abandonner la théorie initiale de Fehr (dans les années 1990) qui postulait une aversion à l'iniquité, partant comme seules données des répartitions finales des gains, et à intégrer dans la description des situations les intentions des partenaires et l'influence des tiers, ces constituants qui tiennent davantage compte de la complexité des interactions sociales telles que nous les avons analysées (voir aussi plus loin la référence à Sanfey, 2011).

Les données d'imagerie cérébrale dont nous disposons aujourd'hui sont tout à fait compatibles avec cette perspective plus intégrative. Krueger *et al.* (2007) ont osé sortir du paradigme de la coopération avec des partenaires anonymes pour revenir à des situations de coopération mieux articulées sur nos interactions ordinaires. Leurs sujets sont engagés dans une suite de *Trust Games* où ils connaissent leur partenaire et peuvent échanger leurs rôles. On repère alors les zones cérébrales qui présentent des différences d'activités avec celles des conditions contrôles qui offrent les mêmes possibilités de gains, mais sans interaction. L'intention de faire confiance se manifeste par l'importance de l'investissement, puis, quand on souhaite maintenir la confiance d'autrui, par l'importance de la somme renvoyée.

Lors du choix de la confiance par l'investisseur, on observe un différentiel d'activation dans le cortex paracingulaire, qu'on pense lié à nos représentations des intentions d'autrui. Les investisseurs qui font d'emblée défection ont une activation moins élevée dans cette zone. Mais chez les coopérateurs l'activation de cette aire du cortex décroît au fil des interactions, alors qu'elle croît chez ceux qui font plus souvent défection. Par comparaison avec ceux qui font défection, les coopérateurs ont une activation plus élevée dans l'aire septale, que ce soit dans la phase de décision en faveur de la confiance ou dans la phase de maintien de la confiance (en renvoyant une somme importante). L'aire septale est un élément du circuit de la récompense et du renforcement qu'elle peut produire. Elle semble aussi liée à la production d'ocytocine. Quand un de ceux qui font le plus souvent défection décide cependant de renvoyer une somme importante, et donc de maintenir la confiance, cela est lié chez lui à une activation supérieure dans l'aire tegmentaire ventrale, qui est aussi impliquée dans le circuit de la récompense, associée à la production de dopamine, et liée à la motivation. Cette activation est encore supérieure chez ce groupe de sujets quand ils décident en tant qu'investisseurs de faire confiance.

Autrement dit, si vous faites défection, vous commencez par ne pas trop faire attention aux intentions d'autrui, mais vous devez de plus en plus en tenir compte quand vous poursuivez ce style d'interaction, alors que si vous faites confiance, c'est le contraire. Dans la première posture, le fait que vous refusiez la confiance-pari ne vous en laisse pas moins dans l'incertitude, qui, si elle se poursuit, exige des efforts cognitifs plus importants. Dans la seconde, vous passez peu à peu de la confiance-pari à la confiance-cadre, et vous avez donc moins d'efforts cognitifs à faire pour deviner

les intentions d'autrui. Si vous entrez dans des relations de confiance ou que vous les maintenez, vous pouvez renforcer votre satisfaction – vous passez de la confiance-pari à la confiance cadre. Si vous choisissez de faire confiance alors que ce n'est pas votre attitude la plus fréquente, il vous faut évidemment accroître votre motivation.

C

Cette étude un peu plus poussée de la dynamique des interactions dans des conditions où la coopération est possible amène sans doute un accroissement de complexité. Cependant notre cognition sociale n'est pas condamnée à toujours impliquer davantage de complexité, et les modèles de neuroéconomie n'ont pas forcément à suivre une telle inflation. Notre cognition, y compris sous ses aspects sociaux, tend à économiser nos efforts, que ce soit dans la perception, où nous préférons nous focaliser sur les formes saillantes, ou dans nos raisonnements et nos décisions, quand nous préférons suivre des scénarios routiniers, ou encore dans nos interactions avec autrui, quand nous commençons par lui supposer des intentions similaires à celles que nous aurions dans sa situation.

On a d'ailleurs pu voir dans la confiance une forme d'économie cognitive (Declerck *et al.*, 2013). Si nous étions toujours méfiants, nous devrions consacrer beaucoup d'efforts à imaginer les différentes possibilités machiavéliques ouvertes à nos partenaires, pour finir d'ailleurs par ne plus savoir sur laquelle tabler, sinon sur la pire pour nos propres desseins. La confiance, comprise alors comme une capacité de transition entre confiance-cadre et confiance-pari – selon les situations, on passe de la première à la seconde ou inversement – nous permet, d'une part, de nous épargner cet effort calculatoire qui est souvent vain et, d'autre part, quand nous sommes dans un contexte d'interaction qui active plutôt la confiance-pari, de rester sensibles à des signaux qui pourraient éveiller notre méfiance.

La dualité entre la confiance-cadre et la confiance-pari pourrait donc être un effet de cette tendance à rechercher des solutions qui économisent l'effort cognitif tout en nous conservant une capacité de veille par rapport aux fluctuations et à la complexité d'un environnement de rapports sociaux qui comporte toujours une part d'incertitude.

On peut interpréter dans cette perspective une observation (du genre de celles que Christian Schmidt a analysées dans le chapitre 4) faite sur les réseaux d'interactions sociales, quand on y distingue les réseaux de proximité (où tous les acteurs sont reliés par des liens directs ou qui ne comptent qu'un très petit nombre d'étapes) et les réseaux dits « petit monde ». Leur nom vient d'une expérience du sociologue Milgram, montrant qu'en confiant une lettre à une connaissance, qui la transmettait à une de ses propres connaissances, on pouvait atteindre n'importe quel destinataire en un nombre d'étapes (5 ou 7) qui reste peu élevé, ce qui nous fait dire que le monde est petit. Or si dans les réseaux de proximité la connectique est dense, puisque chacun est relié à un autre, quel qu'il soit, par une ou deux étapes, dans les réseaux qui assurent cette transmission rapide à des acteurs éloignés,

les structures de connexion sont assez différenciées, des groupes assez compacts, donc où chaque nœud ou membre du groupe entretient plusieurs liens avec les autres, étant reliés entre eux seulement par peu de liens passant par peu de nœuds. Le principe de Granovetter, « la force des liens faibles », fait en plus intervenir l'intensité ou la fréquence d'utilisation des liens. Les liens intragroupes étant plus fréquemment utilisés que ces liens intergroupes, ces derniers peuvent donc être des liens plus faibles, mais pourtant nécessaires pour pouvoir créer des réseaux complexes à circulation rapide. Ces nœuds qui jouent le rôle de « hubs » ou de nœuds de transferts entre groupes plus compacts donnent à ces réseaux des propriétés intéressantes – un changement d'échelle conservera une variation linéaire de la connectivité (au sens où la non-linéarité de la connectivité (loi en puissance) est réductible à une variation linéaire quand on passe à une représentation logarithmique) ; on parle d'invariance d'échelle. Cela implique un certain type de robustesse : si l'on supprime des nœuds au hasard, comme on a peu de chances de tomber sur un hub, la probabilité que le réseau soit séparé en deux est faible, plus faible que dans un réseau où tous les nœuds sont reliés avec des probabilités similaires. Mais inversement le réseau est très fragile face à une attaque qui ciblerait les hubs.

On a alors observé qu'il y avait moins d'interactions coopératives sur les réseaux petit monde que sur les réseaux de proximité, alors même que les premiers permettent des coordinations à plus longues distances (Andras, 2011). Notre sensibilité aux effets en retour qu'ont les succès ou échecs de nos interactions sociales sur nos attentes pourrait en être la cause. Les réseaux de proximité permettent bien plus de face-à-face ou de relations en triangles. L'expérience de Gracia-Lazaro (2012), qui montre surtout que la coopération est plus faible dans les réseaux petit monde que les modèles de réciprocité ne le supposaient, présente l'effet coutumier de décroissance de la coopération par répétition du même jeu, ce qui doit nous rappeler que nos coopérations effectives – dans la réalité de la vie sociale – se font en variant les circonstances. Dans une relation triangulaire, chacun peut exercer un contrôle direct sur les deux autres, ce qui est nécessaire quand les circonstances varient. Les membres de ces réseaux s'attendent donc à devoir avoir à rendre compte aux autres de leurs comportements, et cela selon un temps de retour assez bref. Ils doivent donc veiller à assurer rapidement la réciprocité des coopérations (ce serait cependant une illusion que de croire que dans les réseaux de proximité on contrôle tout : par exemple, dans des groupes relativement compacts de pêcheurs, on peut observer des surexploitations des ressources, qui sont pourtant nocives pour leur collectif ; il s'agit alors d'effets cumulés qui ne se font sentir qu'au niveau collectif et qui ne sont pas perçus dans les interactions locales). Dans un réseau petit monde, on voit aussi revenir de l'information. Mais d'une part cela prend un peu plus de temps – même si le cheminement de cette information est bien plus court que les chemins les plus longs possibles du réseau – et surtout c'est moins fréquent, si bien que la pression de réciprocité est moins forte. D'autre part, comme on ne peut pas assurer un contrôle direct de l'interaction (sur 5 nœuds, on n'en contrôle qu'un ou deux directement) la coordination se fait *via* des intermédiaires qu'on peut suspecter : on peut toujours soupçonner la personne qui occupe un nœud hub de manipuler les choses, et les autres sous-groupes d'avoir des intérêts divergents, ce qui justifie d'éventuelles défections.

Inversement, on a pu montrer (Van Segbroeck *et al.*, 2011) que pour que la coopération puisse

subsister dans un réseau de type petit monde, où elle est mise en danger par des exploiteurs et par le genre de soupçon que l'on vient d'évoquer, il suffit que les types de liens sociaux puissent être assez variés, tout comme les vitesses auxquelles on peut changer de type de lien. Les coopérateurs pourront coopérer avec différentes intensités ou fuir les non-coopérateurs. La coopération n'est d'ailleurs un enjeu que dans ces réseaux où tout le monde n'est pas relié à tout le monde, si bien qu'il y a une certaine incertitude sur ce qui est ou non partagé par leurs membres. La conjonction entre coopération et structure dite « petit monde » (dont les liens sont plus inégalement répartis que dans les réseaux de proximité directe) exige alors de disposer de moyens de transférer les informations, comme les rumeurs ou les commérages (Andras, 2011). Plus noblement, une société propose à ses membres des représentations communes, qui font fonction de nœuds reliés à tous les acteurs. Mais chaque acteur peut avoir une appréhension d'une telle représentation qui est un peu différente de celle des autres, si bien que l'on retrouve à un moindre degré l'incertitude et le vague des commérages. Quand la coopération sociale se poursuit et que chacun s'attend tout de même à pouvoir se coordonner avec la plupart des membres de la société, c'est que ces représentations restent supposées communes malgré le vague qu'introduisent ces variations. Le fait que les acteurs sociaux deviennent capables de s'imposer la contrainte de surmonter ce vague amène alors à soupçonner qu'ont émergé dans cette société des attentes normatives, qui sont les bases des normes sociales. Nous étudierons cette émergence au chapitre suivant.

---

1. Un exemple bien connu en théorie des jeux est le jeu dit de « la poule mouillée » (*Chicken Game*), où la coordination sur un équilibre implique que l'un des deux joueurs lance son bolide, alors que l'autre lui dégage la voie. Plusieurs interprétations militaires en ont été tirées sous la forme, par exemple de modèles d'attaques éclairs (Brams, 1990, 2003). On peut également interpréter de cette manière la dissuasion nucléaire bilatérale qui a prévalu durant toute la période de la guerre froide (Schmidt, 2001).

2. Il est intéressant de noter toutefois que les premières expériences de théories des jeux, auxquelles participèrent notamment Nash, portaient sur des jeux coopératifs. Il est vrai que le corps entier de la théorie, conçu par von Neumann, relevait alors des jeux coopératifs (Kalisch *et al.*, 1954).

3. Cette idée a été développée par Greenberg, qui s'est efforcé de construire, sous forme d'alternative à la théorie des jeux, ce qu'il a appelé une « théorie des situations sociales ». Cette théorie s'efforce d'intégrer, dans le cadre élargi du formalisme des jeux, la description des environnements sociaux qui contribuent à moduler l'interaction des joueurs (Greenberg, 1990). Mais cette tentative est restée isolée dans le monde des théoriciens des jeux.

4. Indépendamment des travaux expérimentaux et des recherches neuroéconomiques, il est intéressant de noter que les théoriciens ont de leur côté conçu des modèles logiques de réciprocité séquentielle qui ont abouti à identifier de nouveaux concepts de solution, comme par exemple, l'équilibre séquentiel de réciprocité (Dufwenberg et Kirchsteiger, 2004).

PARTIE 3

RÈGLES ET NORMES

CHAPITRE 6

Règles et conventions

Nous avons déjà vu que la référence aux jeux de société dans la théorie des jeux n'est pas le résultat d'une simple métaphore. Ses pères fondateurs, Borel, von Neumann et Nash ont tous élaboré sur ses bases plusieurs modèles de jeux de société, notamment de poker[1]. Or ce qui caractérise un jeu de société ce sont ses règles, au moyen desquelles chaque jeu se trouve défini. Que faut-il entendre précisément par règles du jeu ? Il s'agit d'un ensemble cohérent et ordonné d'instructions auxquelles les joueurs doivent se conformer au cours du déroulement du jeu. On peut ainsi considérer chaque jeu comme un « petit monde », au sens logique, qui se trouve régi par un code composé de cet ensemble de règles. Les joueurs constituent la population de ce « petit monde ». Pour pouvoir jouer dans le petit monde, les joueurs connaissent préalablement son code, mais ce n'est pas tout. Ils savent, comme nous l'avons déjà dit au chapitre 2, que les autres joueurs le connaissent également et que les autres savent qu'ils le connaissent. Le code qui rassemble les règles d'un jeu représente, pour les joueurs, ce que Lewis, avant les théoriciens des jeux, a appelé une « connaissance commune » qui sert de fondement à sa théorie des conventions (Lewis, 1969).

Nous avons eu l'occasion de montrer qu'en stricte logique, une connaissance n'était commune qu'au terme d'une opération d'itérations à l'infini (je sais que tu sais, que tu sais que je sais, etc.), dont la réalisation (« implementation ») dépasse les capacités cognitives des cerveaux humains. Si l'on considère dans cette perspective le raisonnement stratégique des joueurs individuels, il est clair que celui qui est capable de pousser plus loin que les autres ce processus d'itération dispose d'un avantage sur eux, par l'effet d'une asymétrie d'information dont il dispose alors. Ce n'est cependant pas ainsi que se présente la propriété de connaissance commune lorsqu'elle s'applique à un code social, comme celui formé par les règles d'un jeu. Considérons, à titre d'exemple, les règles du code de la route. Certaines répondent à des normes d'organisation de la circulation (définition des priorités…), d'autres sont de simples conventions (rouler à droite ou rouler à gauche…). Mais tous les automobilistes sont censés les connaître et savoir que les autres automobilistes les connaissent

également et qu'ils savent qu'ils les connaissent. Il ne servirait à rien ici de pousser plus loin la profondeur du raisonnement de chacun des automobilistes, puisque, par destination, la connaissance de ce code a vocation à être un savoir partagé, ce que les économistes nomment, dans leur jargon, un « bien public pur ». Cet exemple permet de mesurer la portée de cette distinction des niveaux de connaissance, selon que cette connaissance concerne une information dont l'intérêt exclusif réside dans le fait d'être partagé, comme les règles d'un jeu et de n'importe quel code social, ou une information dont la détention présente un intérêt privé.

P

Les choses sont claires et bien tranchées dans les jeux de société. Les règles y représentent un ensemble d'informations publiques explicites, dont la connaissance est commune à tous les joueurs. Chaque joueur peut évidemment utiliser les règles du jeu pour construire un raisonnement stratégique privé, sur la base d'informations qu'il tire d'itérations de niveaux plus ou moins poussés. Mais la transposition n'est pas immédiate des jeux de société aux situations sociales, interprétées comme des jeux, par analogie avec les systèmes d'interactions qui en constituent les trames. S'il est relativement aisé de traiter un jeu de société comme un système d'interactions sociales, l'inverse n'est pas vrai. Lorsqu'il s'agit de situations sociales courantes qui ne sont pas codifiées, la référence aux règles doit plutôt s'entendre comme une métaphore. Les règles y font figure d'instructions partielles et lacunaires dont le statut de connaissance commune relève davantage de l'interprétation implicite que de l'énonciation formelle.

Considérons, à titre d'exemple, une situation de négociation. Traduite en termes de jeu, on peut la représenter sous la forme d'un échange alterné de propositions et de contre-propositions entre deux ou plusieurs parties. Quelles sont précisément les règles d'un tel jeu ? La formulation séquentielle alternée de propositions, au terme de laquelle chaque joueur énonce à son tour une proposition et répond par une acceptation, ou un refus, à la proposition énoncée par l'autre, fournit des informations suffisantes pour décrire un jeu de négociation. Mais elles ne constituent pas, à elles seules, les règles de ce jeu. Ainsi, par exemple, qu'est-ce qui détermine la fin de ce jeu si le nombre de ses séquences n'est pas fixé à l'avance ? Un observateur naïf serait tenté de répondre que le jeu s'arrête dès que les différentes parties s'estiment satisfaites. Mais la satisfaction de chaque joueur n'appartient pas à la catégorie des informations publiques, connaissance commune. C'est la raison pour laquelle la satisfaction respective des joueurs ne peut être considérée comme une règle du jeu de négociation qu'au terme d'un raisonnement analogique du type « le jeu de négociation se joue comme si la satisfaction conjointe des joueurs faisait partie de ses règles ». À l'évidence cependant, cette satisfaction est une information privée détenue par chaque joueur qui peut difficilement être assimilée à un bien public. Il faut, dès lors, déterminer comment ces interprétations subjectives de la situation formulées par chaque joueur peuvent être transformées en une règle objective de caractère public et

connaissable par tous. La réponse proposée par la théorie des jeux consiste à considérer que le concept de solution associé au jeu permet cette transformation. À l'équilibre, dans notre exemple, les satisfactions respectives éprouvées par tous les négociateurs sont rendues compatibles. Le jeu des propositions et contre-propositions n'a donc plus alors de raison de se prolonger. L'équilibre correspond ici à la solution du jeu. Peut-on, pour autant, en conclure que la solution ainsi définie fait partie des règles du jeu de la négociation au sens où nous l'entendons, au même titre que l'ordre séquentiel des propositions ? Rien n'est moins sûr.

Plusieurs objections peuvent être avancées à l'encontre de cette présentation. En premier lieu, le concept d'équilibre est un outil d'analyse construit par les théoriciens qui permet d'étudier le fonctionnement idéal d'un jeu de négociation. Sa connaissance peut certes aider les joueurs, mais elle n'est pas nécessaire au déroulement du jeu. En outre, la détention de sa connaissance par l'un ou l'autre des joueurs, voire par tous, ne garantit nullement qu'il s'agisse d'une connaissance qui leur est commune. En second lieu, l'équilibre constitue ici un concept de solution possible, mais d'autres concepts de solution peuvent également être associés au jeu de négociation, et les théoriciens des jeux ne se sont pas privés d'en proposer différents, même d'en inventer de nouveaux[2]. Dès lors, si la solution d'un jeu prenait sa place dans l'énoncé de ses règles, ou bien un même jeu pourrait avoir plusieurs règles, ou bien il existerait autant de jeux de négociation que de concepts de solution susceptibles de leur être associés. Comment les négociateurs pourraient-ils alors savoir, comme les joueurs d'un jeu de société, dans quel jeu ils jouent ? Le système de règles ainsi redéfini se révèle, pour ces raisons, difficilement praticable dans les deux cas. On ne peut donc pas incorporer la solution, entendue au sens technique de la théorie des jeux, dans l'ensemble des règles d'un jeu de négociation. D'autres difficultés se manifestent lorsque la situation de jeu contient plusieurs équilibres et donc plusieurs solutions, ou lorsque les informations à la disposition des joueurs sont insuffisantes pour leur permettre d'y accéder. La nature des problèmes posés aux joueurs dans de telles situations a déjà été présentée et développée au chapitre 4, consacrée à la coordination.

D'un autre côté, cependant, de très nombreuses expériences ont révélé que les individus, lorsqu'ils étaient placés dans des situations d'interactions sociales construites en forme de jeux, complétaient les systèmes lacunaires de règles explicites associées à ces situations, par d'autres règles, plus ou moins implicites. Il est apparu, au cours de ces expériences, que certaines de ces règles tacites ou implicites pouvaient même contredire les règles qui auraient été déduites de leurs solutions théoriques en suivant la démarche précédemment décrite. De telles règles, au lieu d'être données *a priori* comme les règles des jeux de société, sont mises en œuvre par les sujets, au cours du déroulement de leurs interactions. Il reste à expliquer comment elles sont élaborées

L'objet de ce chapitre est de comprendre à quoi correspondent exactement de telles règles pour les personnes qui se trouvent en situation d'interactions. Pour y parvenir, nous disposons, d'abord, d'informations sur les comportements qui ont pu être observés au cours de l'exécution de différents protocoles expérimentaux d'interactions, plus ou moins directement dérivés d'exemples tirés de la théorie des jeux. Des informations complémentaires sur le fonctionnement cérébral des sujets sont fournies par les résultats d'investigations menées, selon diverses techniques d'imagerie, aujourd'hui

associées à plusieurs de ces expériences. D'autres expériences ont été plus spécifiquement conçues en vue de tester, au regard des circuits neuronaux activés et désactivés, la pertinence des hypothèses avancées pour expliquer le contenu de ces règles. Mais le traitement de ces données très disparates soulève de sérieux problèmes méthodologiques, laissant encore souvent le champ à des interprétations contradictoires. C'est à démêler cet écheveau complexe, pour dégager les lignes de force qui commencent à s'imposer, que s'attache ce chapitre. Il débouche sur la mise en évidence de processus dynamiques qui engendrent et renforcent des standards de comportement sociaux.

D

Partis de l'hypothèse que beaucoup des règles des jeux sociaux pouvaient s'entendre comme des conventions partiellement implicites partagées par ses acteurs, la première démarche des chercheurs dans ce domaine a consisté à s'efforcer de relier les comportements observés dans ces situations à des règles supposées correspondre à ces conventions. Pour y parvenir, ils ont souvent utilisé une approche familière aux économistes de la décision et imaginée depuis longtemps par Paul Samuelson, sous l'appellation de « théorie des préférences révélées » (Samuelson, 1947, 1948). Comme son nom l'indique, cette approche repose sur l'hypothèse que, par leurs choix, les consommateurs et, plus généralement, les agents, économiques révèlent leurs préférences[3]. De même que les préférences individuelles, les règles implicites, qui sont supposées guider les acteurs sociaux dans ces situations d'interactions, ne sont ni directement observables, ni même facilement repérables empiriquement. Il convient donc de trouver un stratagème pour permettre leur révélation. Pour y parvenir, l'un des moyens est de supposer que ces règles peuvent être révélées par les sanctions appliquées par les joueurs eux-mêmes aux autres joueurs, lorsque ces règles hypothétiques ne sont pas respectées. Les règles se caractérisent par leur portée normative. À toute règle, doit, en effet, être associée, une obligation, une interdiction, ou une permission de faire. Le non-respect de ces injonctions expose donc à des sanctions. C'est dans cette perspective qu'on peut interpréter le concept, à première vue déroutant, de « punition altruiste » imaginé par plusieurs chercheurs en neurosciences, en vue, notamment, d'induire des comportements individuels observés, les règles sociales implicites qui les inspirent.

Pour illustrer cette démarche le plus simple est de partir du jeu de l'ultimatum. Le jeu de l'ultimatum représente un exemple de jeu où les règles procédurales sont incomplètes, même en y incluant les spécifications normatives de l'équilibre de Nash, comme sont tentés de le faire les théoriciens des jeux. L'enjeu y consiste précisément à trouver une règle de partage acceptée par les deux joueurs. Dans la logique de la rationalité nashienne, toutes les règles de partage proposées par le joueur qui joue en premier sont acceptables par celui qui joue en second, dès lors qu'elles lui assurent un gain, fût-il minime. Il est peu probable, en revanche, qu'elles le satisfassent toutes. Pour s'entendre sur un partage, les deux joueurs du jeu de l'ultimatum doivent donc se référer à une norme commune,

qui ne figure pas dans ses règles du jeu. Une difficulté se présente cependant ici, du fait précisément des règles de ce jeu. Le joueur qui joue en premier dispose, en effet, du privilège de proposer les termes de ce partage, dans le cadre d'une négociation asymétrique qui prend la forme d'un ultimatum. S'il refuse la règle de partage proposée par le premier joueur, le second joueur, en le sanctionnant, s'inflige à lui-même une punition, puisque la somme initialement mise à la disposition des deux joueurs leur sera retirée, au terme des règles formelles de ce jeu. On peut l'interpréter, dès lors, comme un coût économique de cette punition qui serait supporté par le punisseur. Si l'on entend par « altruiste » en un sens large, toute règle de partage qui prend équitablement en compte l'intérêt de l'autre, le rejet d'une proposition de partage inéquitable peut se comprendre comme l'expression d'une punition altruiste. Ce rejet punirait, en effet, l'auteur de cette proposition, au nom d'un principe altruiste que ce dernier a violé par son comportement, et il s'accompagnerait d'un coût matériel pour celui qui le décide. Il punirait donc, en définitive, les deux joueurs dans un dessein altruiste de portée sociale. Contrairement, par conséquent, à l'interprétation étroite la plus généralement retenue par les théoriciens des jeux, l'absence d'un partage trop inégalitaire, le plus souvent observé au cours du déroulement de ce jeu, ne contredit pas la recommandation de l'équilibre de Nash ; il révèle d'abord son insuffisance. Il suggère, ensuite, l'existence d'une référence à une règle supplémentaire qui serait implicitement acceptée par eux deux. Le fait que, dans une majorité des cas, on observe que la proposition du premier joueur ne s'éloigne guère de cette règle corrobore, d'une certaine manière, cette hypothèse. C'est parce qu'il anticipe un refus de la part de l'autre joueur dans le cas où sa proposition, pourtant légitime en référence à la rationalité nashienne, serait jugée par lui inéquitable par rapport à cette norme implicite d'équité, qu'il propose un partage qu'il pense acceptable pour les deux[4].

L

Ce n'est pas, cependant, sur la base de la version classique du jeu de l'ultimatum que le concept de punition altruiste a été introduit expérimentalement dans le corpus des neurosciences. Pour en préciser la signature neuronale, l'équipe de De Quervain a conçu un protocole de jeu différent, un peu plus compliqué (De Quervain *et al.*, 2004). Il s'agit d'un jeu séquentiel à deux joueurs, proche du jeu de l'investissement (*cf.* chapitre 5). Son déroulement dépend de la confiance respective que s'accorde chaque joueur. Les deux joueurs reçoivent chacun, au départ, une somme de 10 unités. Celui qui joue en premier peut soit garder pour lui cette somme, soit la confier au joueur qui jouera en second. Ce dernier verra alors quadrupler son montant. Il disposera donc de 40 u + 10 u. Il peut, à son tour, dans cette seconde séquence, soit garder la totalité pour lui, soit la partager, en parts égales, avec le premier joueur. Trois issues du jeu sont donc possibles : 1) le premier joueur garde ses 10 u, les deux joueurs gagnent alors chacun 10 u. 2) le premier joueur confie ses 10 u au second joueur qui garde le tout pour lui, il gagne alors 40 u + 10 u = 50 u et le second joueur perd 10 u. 3) le premier joueur confie ses 10 u

au second joueur qui partage les gains de manière égale, chacun des joueurs gagne alors 25 u. Contrairement au jeu de l'ultimatum, les règles du jeu sont ici complètes, puisque la formule de partage se trouve incorporée dans ses règles. Ce jeu ne possède qu'une seule solution, qui correspond à son unique équilibre de Nash. L'issue la plus avantageuse pour le joueur qui joue en second est évidemment celle où le premier joueur lui a confié sa dotation initiale, ce qui lui permet de réaliser un gain maximal. Elle est aussi la plus désavantageuse pour celui qui joue en premier, qui perd alors le montant de cette dotation initiale. Anticipant cette issue, le joueur qui joue en premier gardera donc pour lui la somme mise à sa disposition au début du jeu. On se trouve ainsi ramené à la situation bien connue du dilemme du prisonnier, à l'origine de nombreux paradoxes résultant de l'opération logique dite d'induction à rebours (*Backward induction*), couramment utilisée en théorie des jeux[5].

Dans l'expérience conduite par De Quervain et son équipe, un seul des sujets placés dans le rôle du joueur en premier a gardé pour lui la mise, et mis ainsi fin au jeu, dès sa première séquence, au cours des 7 essais. Presque personne n'a donc suivi la procédure mentale d'induction à rebours préconisée par la théorie des jeux. Quant aux sujets placés dans le rôle du joueur en second, seuls 4 sur 7 ont, en toute logique, gardé pour eux la somme dont ils disposaient. Rappelons qu'aucune règle formelle, dans ce jeu, ne sanctionne les joueurs en second qui adoptent un comportement égoïste au détriment des autres. Le fait qu'une très large majorité des sujets placés dans la situation du joueur en premier aient cependant choisi de confier à l'autre joueur la somme mise à leur disposition au départ laisse à penser qu'ils croyaient plutôt que ce dernier choisirait de partager les gains ainsi obtenus au terme de la procédure du jeu. On peut dès lors se demander sur quoi peut s'appuyer cette croyance, si ce n'est sur la conviction de partager, avec les autres joueurs, une référence commune à une règle informelle concernant une allocation équitable des gains réalisables. Une telle règle, non seulement, n'appartient pas aux règles formelles de ce jeu, mais elle s'oppose même, ici, comme on l'a vu, aux recommandations tirées des implications logiques de sa solution. Il n'est guère surprenant, dans ces conditions, que les joueurs, lorsqu'ils ont été victimes de leur confiance dans cette règle de partage équitable, désirent infliger une punition aux autres joueurs qui l'ont violée.

L'objet de cette expérience n'est pas de mettre en évidence une règle implicite de partage qui serait révélée par la sanction des joueurs à son manquement, puisque cette règle fait partie des options offertes aux joueurs et s'inscrit, de ce fait, dans les règles de ce jeu. Les auteurs de cette expérience ont d'abord voulu savoir, si les joueurs, placés dans cette situation, préféraient gagner davantage, en faisant confiance aux autres, ou s'assurer un gain personnel moindre sans prendre de risque. Mais ils ont surtout cherché à étudier les ressorts du mécanisme mental, qui poussaient les joueurs à punir les autres joueurs qui, abusant de leur confiance, s'étaient approprié la totalité des gains, même si une telle punition devait s'accompagner, pour eux, d'un coût.

C'est la raison pour laquelle ce sont les sujets qui avaient choisi de confier leur dotation initiale au second joueur qui ont ainsi été retenus pour la suite de l'expérience. Il n'est pas surprenant, non plus, ici, qu'ils se soient tous prononcés en faveur d'une punition sanctionnant les sujets qui, dans le rôle du joueur en second, avaient gardé pour eux la totalité du gain ainsi obtenu. Il serait toutefois intéressant de savoir quel aurait été leur propre comportement s'ils avaient eux-mêmes été placés dans

la position de jouer en second.

Mais le but principal de cette expérience étant d'explorer les bases neuronales qui poussent, en la circonstance, les sujets à punir, trois niveaux de punition ont été introduits dans ce jeu : a) une sanction symbolique, sans coût ; b) une sanction pénalisant seulement l'auteur du comportement jugé répréhensible, sans coût pour celui qui punit ; c) une sanction pénalisant l'auteur du comportement jugé répréhensible, avec un coût pour celui qui punit. À ces trois niveaux de punition a été ajoutée une catégorie supplémentaire : d) correspondant à la sanction pénalisant le comportement jugé répréhensible d'un sujet tiré, cette fois, au hasard. Seule la sanction de niveau c) traduit, selon cette typologie, une punition altruiste, au sens de la définition particulière qu'en ont donnée les auteurs de cette expérience, puisque, d'une certaine manière, toutes ces punitions peuvent être considérées comme altruistes.

C'est en recourant à la technique de l'imagerie cérébrale qu'ils ont mené, ensuite, leur investigation du fonctionnement cérébral qui guide ces comportements de punition et, en particulier, celui qui correspond à la punition altruiste selon leur définition. Quelle que soit la punition retenue, ils ont observé qu'elle active toujours, chez les sujets, des régions cérébrales du noyau caudé associées à la récompense (striatum dorsal, notamment). Cette activation est cependant plus nette, lorsque la punition est effective et pas seulement symbolique (b), et elle s'accroît encore lorsque la punition s'accompagne d'un coût pour celui qui punit (c). Dans ce dernier cas, différentes zones supplémentaires du cortex préfrontal et du cortex orbitofrontal, respectivement allouées au raisonnement et aux relations entre l'émotion et la cognition, se trouvent également activées. On retrouve encore des activations liées au réseau de la récompense lorsque la punition vise le comportement d'un sujet tiré au hasard (d), mais ces activations sont alors très atténuées. Les sujets sont, du reste, beaucoup moins nombreux à vouloir le punir.

Pour De Quervain et son équipe, la décision de punir, dans ces différentes variantes, serait d'abord mue par l'attente d'un plaisir, celui éprouvé par le punisseur lorsque les personnes visées se trouvent sanctionnées. À travers la personne visée, c'est la sanction de son intention supposée qui semblerait à l'origine de ce plaisir, puisque le mécanisme neuronal qui active le circuit de la récompense est beaucoup moins marqué dans le cas de la punition symbolique. En revanche, l'attente d'une récompense, génératrice de cette jouissance chez le sujet, grandit, paradoxalement, lorsque la punition a un coût pour le punisseur. On sait, en effet, que le circuit de la récompense se traduit par une stimulation de la jouissance provoquée par un plaisir attendu. Cette jouissance serait plus intense dans cette hypothèse, parce qu'elle doit, alors, vaincre des résistances opposées par le raisonnement (cortex préfrontal ventromédian), et par d'autres émotions qui lui sont associées (cortex orbitofrontal). De Quervain en tire la conclusion que cette punition, lorsqu'elle s'accompagne d'un coût pour le punisseur serait « altruiste », en un sens biologique (accepter de perdre personnellement un peu pour renforcer une norme sociale profitable à tous), différent de son acception psychologique courante (De Quervain *et al.*, 2004). D'autres interprétations ont, du reste, été proposées pour en rendre compte.

L'un des mérites de cette expérience est d'avoir mis en évidence, à travers le cas singulier pour l'économiste de cette « punition altruiste », un système neurobiologique visant la protection et le renforcement de règles sociales implicites, dont la portée pourrait être bien plus large. La décision de punir y résulte, comme on l'a vu, d'un arbitrage de type coût-avantage. Le choix de la punition entraîne le déclenchement du circuit de la récompense qui se trouve, en la circonstance, suractivé par le sacrifice individuel qu'elle implique. Le plaisir ainsi attendu qui serait tiré de cette punition l'emporterait, dans cette hypothèse, sur l'intérêt individuel bien compris.

Une interprétation neurobiologique, voire génétique de cette « punition altruiste » a été, comme on l'a dit, proposée sur la base de ce constat. La jouissance individuelle éprouvée, en la circonstance, par le punisseur pourrait s'entendre comme une sorte de ruse de la génétique, qui, en assurant de cette manière le respect de règles sociales informelles, permettrait à l'espèce humaine de se maintenir et même de prospérer. Sans trancher sur la pertinence de cette interprétation, on observera qu'elle laisse plusieurs questions en suspens. Cette suractivation du circuit de la récompense, observée en la circonstance, est-elle provoquée par la seule souffrance qui accompagne la punition, ou par la valorisation qu'elle imprime à sa portée ? Existe-t-il un lien direct entre ces deux facteurs d'explication et, si oui, comment se manifeste-t-il ?

L'exemple de la punition altruiste a permis de mettre en évidence les racines neuronales d'un mécanisme de socialisation, par l'intermédiaire de cette punition qui sanctionne le non-respect de règles comportementales implicites. Le rôle joué par ce mécanisme dans des situations diverses et ses nombreuses implications l'ont transformé aujourd'hui en une sorte de nouveau paradigme pour les neurosciences sociales. Pour autant, le contexte expérimental spécifique dans lequel il a été initialement repéré et l'interprétation particulière qui a été proposée de ses résultats rendent les comparaisons difficiles avec les résultats obtenus dans d'autres expériences ultérieures qui visaient souvent un but différent, à partir de protocoles expérimentaux distincts. On observera, tout de même, à ce sujet, que presque toutes ces expériences ont été plus ou moins directement dérivées du schéma d'un jeu de partage, dont le jeu l'ultimatum et le jeu du dictateur sont les versions les plus connues.

Si l'on entend maintenant la punition altruiste dans une acception plus large, qui recouvre tout comportement visant à sanctionner, aux frais éventuels du punisseur, la violation de règles implicites se référant à des normes sociales, trois composantes différentes doivent être distinguées : la punition, la règle sanctionnée, et la (les) personne(s) visée(s) par la sanction. Pour chacune d'entre elles, il existe plusieurs variantes. La punition elle-même est mise en œuvre par l'un des joueurs. Elle sanctionne un comportement qui ne respecterait pas une règle implicite ; mais cette punition peut être symbolique, ou effective, et entraîner un coût matériel, ou seulement moral, pour le punisseur. La règle implicite violée, qui fait l'objet de la sanction, porte, dans une majorité de ces expériences, sur un partage ; mais cette règle implicite de partage peut renvoyer à une norme égalitaire abstraite, ou correspondre, plus concrètement, à l'attente d'une réciprocité dans les comportements. Quant à son non-respect, il peut se manifester sous la forme d'une proposition jugée inéquitable, comme celle du joueur en premier dans le jeu de l'ultimatum, ou, plus directement, par le refus de partager, comme

dans l'expérience de De Quervain. La personne visée par la sanction est celle qui a violé cette règle implicite. Ce peut être un autre joueur dont le comportement lèse personnellement le punisseur, comme le joueur en premier du jeu de l'ultimatum lorsqu'il propose une offre de partage inégalitaire. Ce peut être, également, un joueur tiré au hasard et donc anonyme pour le punisseur, au même titre qu'un ordinateur. Ce peut être, enfin, le joueur d'un autre jeu social, dont le punisseur observerait directement le comportement inéquitable, ou même, seulement, en serait informé. D'autres extensions, plus éloignées du paradigme de la punition altruiste, ont également été explorées en isolant, pour l'étudier, un seul de ces paramètres, comme la décision de punir ou la norme d'équité.

C'est à partir de cette catégorisation des divers paramètres qui interviennent dans la punition altruiste, et en tenant compte des différentes interprétations retenues par les auteurs de ces expériences, que l'on se propose maintenant de formuler quelques hypothèses sur le statut de ces règles sociales implicites auxquelles semblent renvoyer les comportements de punition qui ont été observés dans des situations expérimentales, nécessairement singulières. Pour y parvenir, nous distinguerons successivement les informations recueillies sur les normes auxquelles peuvent se rattacher ces règles, puis sur les personnes visées par les punitions qui sanctionnent le non-respect de ces règles.

R

Les différents jeux de partage, dont l'archétype est le jeu de l'ultimatum, ont, le plus souvent, servi de cadre pour mettre en évidence ces règles sociales implicites. Or ces jeux font intervenir deux considérations différentes, mais étroitement imbriquées, dans les décisions qui ont été observées chez les joueurs : d'une part, une référence abstraite à une norme de partage égalitaire, ou tout au moins équitable, d'autre part, une attente concrète de la part du sujet d'une certaine réciprocité dans le comportement de l'autre avec lequel il interagit dans le cadre du jeu. On forme intuitivement l'hypothèse que cette norme est l'objet d'une croyance des joueurs que chacun pense partager avec les autres joueurs ; équité et réciprocité se trouvent alors intimement liées, d'où la difficulté de distinguer entre ces deux ancrages. Il en résulte notamment que, si le comportement d'un joueur conduit à un partage inéquitable, le joueur, qui en est victime, se trouve, en quelque sorte, doublement lésé. Il souffre, au premier degré, de la perte, ou plutôt du manque à gagner, qui en résulte pour lui, et il endure, au second degré, la peine de s'être trompé sur cette norme qu'il croyait partager avec cet autre joueur. Cette dualité complique l'interprétation de sa réaction qui se traduit par un rejet, voire par une punition, très largement observé chez les joueurs qui sont victimes d'un partage inégal. D'un côté, en effet, leur décision sanctionne directement ou indirectement le manquement à une règle dérivée de cette norme, mais, d'un autre côté, elle vise l'auteur de ce manquement sur la base d'un principe simple de réciprocité, à l'œuvre, par exemple, dans la loi du talion. Dans le cas de la première motivation, la punition ne s'accompagne pas nécessairement d'animosité particulière à l'encontre de

ceux qui, par leur comportement, ont enfreint cette norme sociale de partage égalitaire ; dans le second cas il pourrait s'agir d'une simple vengeance, partiellement indépendante de cette norme. Les chercheurs se sont donc efforcés de disjoindre ces deux possibles motivations.

Il apparaît cependant très difficile, en restant à un niveau strictement comportemental, de désenclaver ces deux composantes de la motivation à punir, dans les diverses situations de partage inégal fournies par les jeux expérimentaux. Si elles interviennent le plus souvent en symbiose en se confortant mutuellement, il reste probable que chacune puisse être associée à une activation cérébrale, au moins partiellement, différente. C'est l'hypothèse formulée par une majorité de chercheurs en neurosciences qui a conduit récemment plusieurs équipes à s'efforcer d'identifier ces différences au niveau neuronal. Cette démarche paraît ici particulièrement féconde, dans la mesure où, à l'exception de témoignages tirés de l'introspection, les matériaux fournis par les sciences du cerveau sont pratiquement les seuls dont on peut disposer pour éclairer les questions posées par la distinction entre ces deux composantes des réactions observables face à des partages inéquitables.

Pour éliminer tout facteur de vengeance et la référence à un simple principe de réciprocité, afin de déterminer si l'acception d'un coût personnel pouvait s'expliquer par la simple volonté de faire prévaloir une règle de partage équitable, une équipe de chercheurs a procédé à une expérience différente. Plusieurs distributions de revenu entre les joueurs d'un groupe leur étaient proposées au hasard et il leur était ensuite demandé s'ils étaient disposés à payer pour rétablir une distribution plus équitable. Non seulement une large majorité des joueurs s'est révélée prête à accepter de payer pour obtenir un partage plus équitable, mais ce sont ceux qui se montrèrent le plus sensibles à cette inégalité, qui dans un autre jeu, proche de celui de la punition altruiste, ont également accepté de punir, à leurs frais, ceux qui ne respectaient pas ce principe d'équité. D'une certaine manière du reste l'idée d'un sacrifice altruiste, au nom d'un principe ou d'une règle, même si elle ne se traduisait pas ici par la décision d'une sanction, était cependant présente dans cette expérience (Dawes *et al.*, 2012).

Le fait que cette distribution, déterminée par l'ordinateur, était, dans cette expérience, indépendante de toute volonté humaine a permis ainsi d'isoler la référence à une norme d'équité dans les choix observés en faveur d'une redistribution plus égalitaire, même lorsque de tels choix s'accompagnaient d'un coût personnel pour les intéressés. Conformément aux expériences antérieures, plusieurs régions du cerveau ont été activées à cette occasion, en particulier celles du cortex préfrontal dorsolatéral et ventromédian d'une part, et celle du cortex orbitofrontal, d'autre part. En outre, la partie antérieure du cortex insulaire s'est également révélée activée. Le rôle joué par cette région cérébrale dans les relations des individus avec autrui, en particulier en termes d'équité sociale semble déterminant à l'issue de cette expérience. Cette partie du cortex insulaire est, en effet, la seule à avoir été activée aux différentes séquences de l'expérience, lors de la transformation de la distribution initiale en une distribution plus égalitaire (Dawes *et al.*, 2012). Elle est également la seule à avoir été activée dans toutes les expériences connues correspondant aux différentes versions de partage inégal, tirées du jeu de l'ultimatum et du jeu du dictateur. Cette zone du cerveau est aujourd'hui considérée au centre d'un système qui régulerait les différentes manifestations attribuées à une sorte d'empathie sociale. Cette norme égalitaire d'équité pourrait donc trouver son origine dans un mécanisme cérébral

d'émotion empathique à l'endroit d'autrui de portée très générale. Elle constituerait alors l'un des facteurs déterminants de la décision de punir les comportements qui s'éloignent de cette norme, même et peut-être surtout, lorsqu'elle s'accompagne d'un coût économique pour le punisseur.

Cette hypothèse, fondée sur l'observation de régions cérébrales sélectivement activées, peut être rapprochée d'autres résultats obtenus dans des travaux différents mais complémentaires, portant sur les relations entre les réactions des sujets lorsqu'ils sont confrontés à des partages inégaux, et le niveau de leur taux de sérotonine. La sérotonine, comme on l'a déjà signalé, est un neurotransmetteur chimique qui agit sur la régulation de certaines émotions et des comportements qu'elles entraînent chez les sujets, en raison, précisément, du rôle de modulateur qu'elle exerce sur les aires cérébrales qui ont été mentionnées. L'action de la sérotonine se manifesterait ainsi déterminante, à l'intersection entre l'aversion (ressentie) et l'inhibition (déclenchée). Certains chercheurs soupçonnent l'existence d'un système de régulation sérotoninergique, comme il existe un système dopaminergique, qui déclenche le circuit de la récompense, aujourd'hui bien connu. Le fonctionnement d'un tel système de modulation par la sérotonine pourrait éclairer nos rapports sociaux et plus particulièrement nos réactions face aux inégalités et à l'iniquité (Crockett, 2012). La difficulté rencontrée pour identifier son fonctionnement tient aux relations apparemment contradictoires qui ont été observées entre les niveaux de sérotonine et les comportements des sujets, devant des situations équitables et inéquitables résultant notamment d'un partage. On observe, en effet, d'un côté, une baisse du niveau de la sérotonine chez les sujets qui refusent un partage inégal dans le jeu de l'ultimatum à un coup, et plus encore chez les punisseurs lorsque leur punition s'accompagne d'un coût personnel. D'un autre côté, il apparaît que toutes les propositions de partage équitable s'accompagnent, au contraire, d'une augmentation du taux de sérotonine (Crockett *et al.*, 2010). Crockett propose une explication assez convaincante du rôle pro-social de la sérotonine dans les deux cas. L'activation de la sérotonine stimule une relation émotionnelle d'empathie positive vis-à-vis d'autrui, visant notamment à éviter tout ce qui pourrait entacher cette relation et, en particulier les comportements inéquitables. Dans le cas de la punition altruiste, la désactivation de la sérotonine permet d'inhiber cette émotion immédiate, en vue précisément de permettre de faire prévaloir une règle sociale d'équité garantissant la coopération entre les individus. Cette hypothèse se trouve indirectement validée par la corrélation inverse observée entre la variation du taux de sérotonine et la réciprocité, suivant qu'elle résulte d'une « réciprocation positive » ou d'une « réciprocation négative », comme le sont les propositions de partage inéquitables (Siegel et Crockett, 2013). La sérotonine serait donc un modulateur des comportements sociaux en fonction des différents environnements. Principes de réciprocité et normes égalitaires pourraient, en fin de compte, dépendre d'un même système, dont l'origine est peut-être d'ordre génétique. Mais les conditions précises de son fonctionnement sont encore mal connues.

On sait, en outre, que ce système sérotoninergique n'est pas sans lien avec le système dopaminergique. Or on a vu que l'activation du circuit de la récompense était également l'un des ressorts de la punition altruiste. Il n'est pas surprenant, dans ces conditions, que les sujets les plus prompts à sanctionner à leurs frais ceux qui ne respectent pas la norme d'équité soient précisément

ceux qui sont personnellement les plus attachés à cette norme, quelle que soit l'interprétation associée à cet attachement. On peut ainsi conjecturer que la punition altruiste fait intervenir simultanément une intensification du système dopaminergique et une réduction du système sérotoninergique.

La manière dont fonctionne cette référence à une norme égalitaire implicite n'est pas, comme le suggèrent les travaux mentionnés, indépendante des circonstances dans lesquelles elle se manifeste. C'est ainsi que des effets de contexte ont été mis en évidence, en partant de réponses observées dans un jeu de l'ultimatum, où les procédures ont été modifiées, de manière à permettre aux joueurs des comparaisons entre l'offre qui leur était proposée et des offres alternatives (Radke *et al.*, 2012). Par ailleurs, la référence à cette norme n'exclut pas, bien au contraire, l'anticipation par le sujet du comportement de l'autre avec lequel il se trouve en interaction. Plusieurs expériences ont ainsi montré que la réaction à une règle de partage proposée par le joueur en premier, dans le jeu de l'ultimatum, variait en fonction des données mises à la disposition du joueur en second, lui permettant, de ce fait, de formuler une anticipation sur cette offre. Ainsi une offre antérieure très basse lui fera accepter plus facilement une offre, pourtant encore bien inégalitaire (Sanfey, 2009). On ne peut pas invoquer pour l'expliquer une forme de réciprocité, puisque le joueur ajuste son exigence à ce qu'il connaît du comportement de son partenaire, alors même que ce dernier s'est montré antérieurement peu soucieux d'équité. Cette modulation de la norme sociale de référence, en fonction de l'anticipation circonstancielle du sujet, s'accompagne de l'activation d'un réseau neuronal clairement identifié, celui organisé autour du cortex cingulaire antérieur (ACC). Ce réseau est, comme on l'a vu, associé au traitement de situations incertaines, voire conflictuelles, en raison de références contradictoires. Or cette région cérébrale est précisément la seule à se trouver activée, lorsque la perception d'une violation de la norme sociale du partage égalitaire se juxtapose à des informations capables de nourrir cette anticipation (Chang et Sanfey, 2011). Tout porte donc à penser qu'il s'agit d'un mécanisme d'adaptation de cette norme sociale aux données factuelles dont dispose le sujet pour la mettre en œuvre dans la situation spécifique d'interaction où il se trouve. On peut, pour cette raison, le rapprocher d'autres mécanismes de régulation bayésienne qui ont été identifiés à l'occasion de comportements faisant intervenir d'autres fonctions cérébrales. Un tel mécanisme contribue ici à renforcer les règles dérivées de cette norme sociale d'équité. Elle pourrait même avoir facilité leur émergence.

P

S'il semble établi que la punition altruiste sanctionne le manquement à une règle fondée sur une norme implicite, cette punition constitue également une action intentionnelle prise par le décideur à l'encontre d'une autre personne. Dans les exemples précédemment analysés, cette décision de punir, lorsqu'elle se manifestait, émanait d'un joueur et visait un autre joueur, dans le cadre d'une interaction régie par les règles formelles d'un jeu. Elle était qualifiée d'altruiste dès qu'elle entraînait

une pénalisation, directe ou indirecte, pour le punisseur. Une règle, toutefois, et plus encore la norme qui l'inspire, se doit d'avoir une portée générale. Cet attribut des règles devrait donc conduire les individus qui s'y réfèrent à sanctionner toutes les personnes qui les violent. C'est la raison pour laquelle le champ d'application de la punition ne se limite pas aux seules personnes dont le comportement a directement lésé ceux qui l'actionnent. Afin de savoir si tel est le cas pour les punitions sanctionnant cette norme d'équité du partage, même lorsqu'elles s'accompagnent d'un coût pour le punisseur (cas de la définition étroite de la punition altruiste), les chercheurs ont conçu différents dispositifs expérimentaux, grâce auxquels ils ont observé si, et dans quelles conditions, cette décision de punir ceux qui ne respectent pas les règles implicites qui ont été mises en évidence était également prise à l'encontre de tierces personnes, extérieures à l'interaction directe qui caractérise le déroulement du jeu.

Une équipe autour de Strobel a ainsi fait réagir des sujets en les confrontant successivement à des partages inégaux dans deux situations différentes. Dans la première situation, le sujet est partie prenante, et donc personnellement victime de cette inégalité. Dans la seconde situation, il observe ce partage inégal entre deux acteurs qu'il ne connaît pas. La victime de ce partage lui est donc, cette fois, extérieure (Strobel *et al.*, 2011). Les résultats expérimentaux ont montré que l'échelle des punitions décidées dans les deux situations était peu différente, à l'exception, contre-intuitive, d'un partage 18/2, pour lequel la punition décidée est apparue sensiblement plus sévère, dans la seconde situation.

Les informations sur le fonctionnement cérébral qui ont ainsi été recueillies par IRM sont plus précieuses. Les mêmes régions ont été activées dans les deux situations, avec, seulement, une intensité un peu plus faible, lorsqu'il s'agissait de sanctionner des partages inéquitables concernant une personne étrangère, dans la seconde situation. Plus précisément, et comme attendu, le circuit de la récompense, associé à la punition altruiste, est apparu plus fortement activé dans le cas d'une punition sanctionnant un partage inégal dont le sujet était personnellement victime. En revanche, les régions sollicitées du cortex préfrontal droit et de l'ACC ont été activées avec une intensité tout à fait comparable dans les deux situations. Cette dernière constatation est d'autant plus intéressante que ce sont ces zones du cerveau qui avaient été activées avec une plus forte intensité dans l'hypothèse de la punition altruiste, telle qu'elle avait été mise en évidence dans l'expérience originelle de De Quervain. La jouissance immédiate, tirée de la punition altruiste, est donc peut-être un peu moins intense dans le cas où elle vise une personne étrangère, mais ses ressorts neuronaux apparaissent identiques dans les deux cas.

Si ces résultats sont confirmés, on pourrait être en présence d'un système neuronal de sanction de la norme d'équité qui présenterait une certaine analogie avec celui des neurones miroirs. En effet, les résultats obtenus par Strobel et son équipe suggèrent que la vision d'une inégalité dont serait victime un autre déclencherait, chez les sujets, les réactions d'un même système neuronal que si le sujet en était lui-même la victime. Mais contrairement aux exemples des neurones miroirs, il ne s'agit pas cette fois d'un système moteur, mais d'un système de ressenti émotionnel identique, qui conduirait, ici, dans les deux cas, à une même intention, celle de punir.

L'introduction du point de vue du tiers dans des protocoles expérimentaux a permis de mieux

identifier ce que représente une règle implicite qui se réfère à une norme sociale, en transformant la nature de l'implication du sujet dans sa mise en œuvre. C'est la raison pour laquelle cette mise en perspective a été souvent utilisée dans des contextes très différents, afin de mieux cerner comment les sujets appréhendent des règles sociales implicites. Une extension de cet usage concerne la norme de partage elle-même, qui constitue l'exemple qui a été encore privilégié. Sans référence à une punition, cette fois, une équipe de chercheurs a cherché à identifier les réactions des sujets confrontés à différentes règles de partage. Ces partages leur étaient présentés de manière aléatoire, c'est-à-dire indépendamment de toute intentionnalité qui pourrait être rapportée à un sujet. On distinguait alors : a) le cas où les sujets étaient eux-mêmes directement concernés par le partage ; b) le cas où le partage concernait les membres d'un groupe, auquel ils n'appartenaient pas. Comme dans le jeu de l'ultimatum, les sujets dont le comportement était étudié se trouvaient placés dans le rôle du joueur en second et pouvaient, de ce fait, accepter où refuser la règle de partage qui leur était proposée. Mais cette fois, et cela contrairement aux règles du jeu de l'ultimatum, leur réaction n'intervenait pas en réaction à la décision d'un joueur en premier, puisque les propositions leur étaient transmises de manière aléatoire (Civai *et al.*, 2012).

Ce protocole expérimental complexe, en raison des différents types de partage retenus dans les deux situations, rend difficile l'interprétation des résultats comportementaux qui ont été obtenus. Sans surprise, une majorité des sujets rejetèrent les partages inégaux dans la situation b. Dans la situation a, en revanche, les résultats sont moins clairs. Si une majorité rejeta les offres de partage qui leur étaient défavorables, plusieurs acceptèrent cependant des offres inégalitaires, dès qu'elles n'étaient pas désavantageuses pour eux. Cette différence observée dans les comportements, selon que le partage les concernait directement ou concernait des tiers s'est révélée étroitement corrélée aux différences d'activation de certaines régions cérébrales, dans les deux situations. Ainsi des aires correspondant à des tâches cognitives, comme certaines zones du cortex cingulaire antérieur, ont été sensiblement plus activées dans la situation a, que dans la situation b. Ce constat n'est guère surprenant. C'est dans cette situation, en effet, que le dilemme entre l'intérêt personnel et la norme d'équité est évidemment le plus aigu. Plus intéressant, peut-être, certaines zones cérébrales, en particulier dans la région de l'insula, se sont montrées activées lorsque les sujets sont confrontés à une offre de partage inéquitable, dans la situation a, comme dans la situation b, et cela quelle que soit la réponse donnée à cette offre. Une référence commune de dégoût par rapport à l'inégalité d'un partage se manifesterait donc au niveau neuronal, indépendamment de l'implication personnelle et directe du sujet.

L'interprétation de ces résultats et leur confrontation avec ceux mis en évidence dans l'expérience de Strobel ne sont pas simples. En apparence, au moins, ils semblent même se contredire partiellement. Certes, dans les deux expériences, les dispositifs ont permis d'identifier des régions du cerveau qui se trouvent activées lorsque la norme sociale d'un partage équitable se trouve violée, que cette violation pénalise le sujet lui-même, ou seulement d'autres sujets. Mais ces régions cérébrales ne sont pas les mêmes dans chacune des expériences. La zone du cortex préfrontal droit, identifiée par Strobel comme commune aux deux situations, ne se trouve activée que dans la situation où le sujet est

personnellement concerné par le partage, dans l'expérience de Civai.

L'expérience de Strobel porte sur la décision de punir l'auteur d'un partage jugé inégal, dont la victime est, dans le premier cas, le sujet lui-même, et, dans le second cas, un autre, que le sujet observe. Dans l'expérience de Civai, il est seulement question, pour le sujet, d'accepter, ou de rejeter, une offre de partage inégal qui, en a le concerne personnellement, ou qui, en b, concerne d'autres. Certes, si le sujet décide de rejeter l'offre de partage, il en résulte une sanction pour ceux qui sont les parties prenantes de ce partage, puisque, alors, personne ne reçoit rien. On peut même parler, à cette occasion, pour cette raison, de punition altruiste dans la situation a. Mais, dans l'expérience de Civai, ce n'est pas le responsable du partage qui se trouve ainsi sanctionné. En outre, dans la situation b, la victime du partage inégal est, dans l'expérience de Strobel, directement vue par le punisseur, tandis qu'elle reste abstraite dans l'expérience de Civai. On peut penser, dès lors, que le sujet interrogé sur l'acceptabilité (ou le refus) du partage se trouve conduit à s'imaginer à sa place dans l'expérience de Strobel, alors qu'il adopte un point de vue extérieur dans celle de Civai. Civai invoque, du reste, la figure du spectateur impartial dans un article plus récent sur le même sujet (Civai *et al.*, 2013).

Leur point commun porte, en définitive, sur l'étendue du champ d'application de cette norme de partage équitable, comme référence implicite des comportements observés. S'agissant, dans les deux expériences, d'actes différents, il n'est pas étonnant que cette référence opère selon des modalités propres qui correspondent à l'activation de régions cérébrales distinctes. Considérés sous cet angle, les résultats recueillis dans ces deux types d'expériences devraient donc plutôt être considérés comme complémentaires. Sur la base d'une réaction première commune de dégoût, provoquée par une formule de partage inégal, se construirait, ensuite, une réaction comportementale adaptée à chacune des situations.

Le point de vue du tiers a également été retenu dans une perspective un peu différente, s'agissant, cette fois, de définir la punition à infliger à un individu coupable d'un comportement délictueux, au sens juridique, c'est-à-dire en référence à une norme explicite. L'épreuve consistait à demander aux sujets de l'expérience de choisir, sur une échelle, une peine à infliger à différents individus qu'ils ne connaissaient pas, mais qui leur étaient présentés en relation à des actes délictueux, au terme d'une cinquantaine de scénarios brièvement exposés (Buckholtz *et al.*, 2008). La référence à une norme de justice n'est plus ici portée par un partage et il n'existe plus de lien entre la violation de cette norme et un dommage que cette violation pourrait concrètement causer au punisseur. Dans une telle expérience, les sujets de cette expérience endossent en quelque sorte les habits du juge, ou, si l'on préfère, adoptent le regard du « spectateur impartial » d'Adam Smith.

En dépit de ces différences avec les expériences précédemment discutées, certains résultats fournis par l'exploration du cerveau par IRMf dans cet autre cadre présentent un intérêt pour notre recherche. Ainsi retrouve-t-on activée, dans l'expérience de Buckholtz, la même région du cortex préfrontal droit, dont l'activation avait été observée dans l'expérience de Strobel, lorsque les sujets décidaient de punir ceux qui ne respectaient pas la norme d'équité, qu'ils en soient ou non personnellement victimes. Ce rapprochement n'a pas échappé à Buckholtz, qui y a reconnu, derrière la décision de punir, l'activation d'un même substrat neuronal, qu'il s'agisse de sanctionner les

comportements de ceux avec lesquels le punisseur est en interaction directe, comme dans les situations étudiées par la théorie de jeux, ou ceux de tiers par rapport au punisseur, comme dans les situations judiciaires (Buckholtz, Marois, 2012).

C

Une comparaison directe des résultats mis en évidence par ces différentes expériences n'est pas possible. Les rapprochements trop hâtifs de ces résultats peuvent même aboutir, comme on l'a signalé, à des contradictions. Il est clair en tout cas qu'il n'existe, aujourd'hui encore, aucune théorie neuronale unifiée permettant de rendre compte des comportements observés regroupés autour des différentes formes de punitions altruistes.

On peut néanmoins tirer de la confrontation des données recueillies par l'imagerie dans ces situations plusieurs éléments intéressants pour notre enquête. Ainsi trois mécanismes neuronaux différents semblent pouvoir être distingués, qui interviendraient successivement lors de la révélation de ces règles sociales implicites, par l'intermédiaire de la punition. En premier lieu, la détection, par le sujet, d'une situation qui violerait une norme de justice (partage inégal, acte délictueux), dont la signature neuronale serait l'activation de la région de l'insula. À cette phase préliminaire correspondrait une baisse du niveau de la sérotonine, associée à l'émotion négative qu'elle suscite. En second lieu, la décision prise par le sujet, de punir celui qui est responsable de cette violation, serait principalement repérable, au niveau neuronal, par une activation du cortex préfrontal ventromédian, et surtout du cortex préfrontal dorsolatéral droit. La transformation de cette norme en règle se trouverait ainsi confirmée et entretenue par la punition. Enfin, la mise en œuvre de cette punition serait stimulée, chez le punisseur, par un mécanisme de jouissance de type du circuit de la récompense, qui se manifesterait par une augmentation de la dopamine et une activation de certaines aires cérébrales, comme le striatum. Une telle jouissance serait renforcée lorsqu'elle s'accompagnerait d'un coût pour le punisseur. Le ressort de cette jouissance paradoxale pourrait être recherché au niveau génétique. Il s'agirait alors, comme le pensent certains neuroanthropologues d'une modalité biologique de la sociabilité aux fins de préservation de l'espèce.

Ces résultats rapportés par les neurosciences semblent conforter notre hypothèse initiale, selon laquelle des règles sociales implicites, renvoyant à l'égalité et à la justice, se trouvent révélées par la punition infligée à ceux qui les violent. L'argument déterminant en sa faveur est fourni ici par la mise en évidence d'un ensemble de mécanismes cérébraux que l'on trouve, selon des modalités parfois différentes, associé à tous les modes de punition qui ont été distingués. Leur mise en œuvre se retrouve, à quelques variantes près, dans les divers types de situations qui ont été étudiés, quelle que soit la position occupée, dans ces différentes situations, par le responsable et par la victime, par rapport au punisseur. L'introduction, sous des formes différentes, d'un point de vue de tiers dans les expériences récentes qui ont été mentionnées, a permis de montrer que les supports neuronaux de ces

punitions sont presque identiques, lorsque ces punitions visent une personne, extérieure, par définition, au jeu interactif, et lorsqu'elles sont destinées à d'autres joueurs, comme dans les jeux où elles avaient initialement été étudiées. Ce constat alimente notre présomption qu'un même ressort mental anime ces punitions chez les punisseurs : conforter, affermir et transmettre certaines règles sociales implicites.

L'existence de règles sociales implicites, qui contribuent à orienter les comportements des individus dans leurs interactions, fait donc partie intégrante de l'intersubjectivité sous ses différentes dimensions qui ont été repérées dans la première partie. Elles s'inscrivent dans une interaction directe avec l'autre acteur (deuxième personne) dans un même jeu. Mais elles ont également leur place dans l'interaction entre deux, ou plusieurs « autres », lorsqu'ils sont considérés, du point de vue d'un tiers (troisième personne). Elle concerne, enfin, le « moi », en tant qu'il se trouve, lui-même, acteur dans cette interaction avec l'autre (première personne) et qu'il représente, à ce titre, un autre pour l'autre et donc, aussi, pour lui-même, dans cette situation. Cette tripartition, à laquelle renvoient ces règles sociales, trouve ainsi une illustration particulière aux différents niveaux où s'exerce leur sanction par la punition. Elle permet, en particulier, d'éclairer le sens de la « punition altruiste », que le sujet décide de s'infliger (première personne) en sanctionnant l'autre (deuxième personne) avec lequel il se trouve en interaction directe, ou un autre extérieur à cette interaction (troisième personne). On observera que dans ce dernier cas le sujet s'imagine lui-même dans la situation d'une troisième personne, au sens du spectateur impartial.

Il reste à comprendre comment s'enchaînent les différents mécanismes neuronaux qui ont été repérés, chaque fois, au niveau des cerveaux individuels. C'est dans cette perspective dynamique que la dimension sociale du phénomène pourrait être élucidée. Pour mieux le saisir, on commencera par rappeler que la connaissance d'une norme destinée à fonder une règle doit être, au moins, partagée avec les autres, et même, peut-être, commune entre eux. Nous proposons d'introduire ici, pour cette raison, une phase intermédiaire entre l'émotion négative première, assimilable à un dégoût, suscitée par la perception d'une situation inéquitable ou injuste, et la décision de punir ceux qui en seraient responsables. Ce dégoût s'accompagne, pour celui qui l'éprouve, d'une surprise provoquée par la découverte que la norme d'équité, au moyen de laquelle il évalue un partage, n'est pas partagée par l'autre et/ou les autres.

Nous retrouvons ici le rôle joué par l'anticipation dans cette référence à des normes sociales implicites, mise en évidence par Sanfey. Mais Sanfey cherchait à montrer que des informations antérieures, dont peut disposer le sujet, modifient ses anticipations et, partant, son comportement au regard d'une même norme sociale (Chang et Sanfey, 2011). Cette modification qui, dans l'expérience de Sanfey, est assimilable à un effet de « cadrage », n'est pas cependant ici l'élément central qui caractérise l'anticipation du sujet. Ce dernier s'attendait, en effet, à ce que sa norme fût partagée. Le comportement de l'autre lui révèle brutalement qu'il n'en est rien. Il se trouve, par conséquent, confronté à une incertitude d'un type particulier, que l'on peut qualifier d'incertitude « inattendue », par opposition à l'incertitude attendue qui caractérise les situations ordinaires, où l'incertitude porte sur l'occurrence d'un état parmi des états prévisibles, et donc, attendus. D'autres recherches en

neurosciences ont découvert une base neuronale à cette distinction. Le travail cérébral se trouverait régulé par des neuromodulateurs différents dans les deux cas, l'acétylcholine, lorsqu'il s'agit d'une incertitude attendue et la norépinéphrine lorsqu'il s'agit d'une incertitude inattendue (Yu et Dayan, 2003). Or ces deux neuromodulateurs ne fonctionnent pas de la même manière, de telle sorte que les procédures mentales qui leur sont associées sont également différentes. Ainsi, la norépinéphrine, impliquée dans l'hypothèse d'une incertitude inattendue, fournit au cerveau un signal d'alarme, qui interrompt le modèle familier de traitement des informations, selon une procédure *top down* guidée par l'acétylcholine (Dayan et Yu , 2006). C'est précisément ce qui semble se produire, lorsqu'un sujet, qui s'attendait à un partage équitable, au regard d'une norme implicite qu'il croyait partager avec son interlocuteur, réagit à sa proposition de partage inégalitaire. Cette proposition entraîne pour lui une rupture dans son modèle d'anticipation. Cette rupture pourrait expliquer la transition entre l'émotion négative première, que lui inspire cette manifestation d'injustice, et sa décision de réagir en décidant de punir, fût-ce en se pénalisant lui-même. La décision de punir traduirait alors une réponse stratégique adaptée à cette incertitude qu'entraîne la perte de cette référence à la norme sociale d'équité.

Une telle analyse rejoint, sur ce point, ce que De Souza, et d'autres philosophes, nomment l'émotion épistémique (*epistemic feelings*). De Souza a lui-même relié la surprise qui se trouve à l'origine de cette émotion épistémique à la prédominance de la norépinéphrine, qui caractérise l'incertitude inattendue selon Yu et Dayan (De Souza, 2009). L'introduction de cette dimension de surprise, à l'origine de la décision de punir, permet également d'expliquer les interprétations psychologiques divergentes qui ont été proposées pour rendre compte de sa relation avec la baisse constatée du taux de sérotonine chez le punisseur. Cette décision de punir marque, en effet, une rupture brutale par rapport à l'appréhension de l'incertitude, souvent associée à une forme d'emportement. Mais elle révèle également un choix stratégique réfléchi dans le nouveau contexte. Il reste, enfin, à comprendre comment s'opère ce choix stratégique. Il s'agit, pour le punisseur, d'un arbitrage entre un gain personnel, modeste, mais immédiat, et un manque à gagner, de même importance, mais ouvrant la voie à un futur retour au partage équitable. Considéré dans cette perspective le rejet d'un partage inéquitable dans le jeu de l'ultimatum et, plus généralement, les différentes variantes de la punition altruiste, correspondent à un pari social sur l'avenir. Il n'est donc pas étonnant, dans ces conditions, que la décision de punir active également un circuit de la récompense associé au risque qui accompagne l'effet, cette fois attendu, de cette sanction.

Retour sur les jeux : normes sociales et standards de comportement

Les travaux qui ont été analysés ont permis de dégager l'arrière-plan neuronal des comportements de rejet et de la volonté de punir, observés chez des sujets lorsqu'ils se trouvent confrontés, du fait d'autrui, à des situations qu'ils estiment inéquitables ou injustes. Cette approche a

permis, de cette manière, d'identifier certaines normes implicites qui règlent nos comportements d'interaction sociale. C'est en adoptant une autre perspective sur les jeux que von Neumann et Morgenstern se proposaient, à travers la théorie des jeux qu'ils esquissaient, de rendre compte des règles sociales susceptibles d'émerger des interactions. Selon eux, tout système d'interaction formalisé par un jeu peut donner lieu à différentes solutions, qui varient en fonction du concept de solution retenu.

Sous cet angle, la solution d'un jeu sert à définir les conditions de cohérence internes associées, par cette interaction, à l'obtention d'une richesse collective et à son allocation entre les joueurs. Il existe, de ce fait, pour von Neumann et Morgenstern, plusieurs concepts de solution différents, qui renvoient, chaque fois, à ce qu'ils nomment des « standards de comportement acceptés », qui équivalent, pour ces auteurs, à « des ordres sociaux établis » et se traduisent par des organisations institutionnelles (von Neumann et Morgenstern, 1946)[6]. Les ordres sociaux qui correspondent à ces standards de comportement, tout à la fois, reposent sur certains principes implicitement acceptés par les joueurs (normes sociales), et modèlent l'organisation sociale de leurs interactions (institutions). À partir de cette grille d'interprétation, le jeu de De Quervain, par exemple, peut donner lieu à deux solutions différentes, selon le concept de solution retenu. Au terme de sa solution classique, comme on l'a rappelé, le fruit de leur interaction n'augmentera pas leur dotation initiale, soit 10u + 10u, correspondant à la somme allouée à chacun au départ. Ce résultat dépend directement du standard de comportement incorporé dans le concept de solution adopté, qui implique ici que chacun des joueurs maximise son intérêt individuel. Une autre solution est également possible, où l'interaction des joueurs aboutit, cette fois, à un gain, partagé entre eux, en parts égales, soit 25u + 25u. Une formalisation en ce sens a été proposée par Berge. Cette solution satisfait les mêmes exigences de logique interne que la précédente, mais elle se fonde sur un autre concept de solution, l'équilibre de Berge, qui renvoie à un standard de comportement différent. Ce standard de comportement répond, dans la formulation de l'équilibre de Berge, à la norme sociale d'un partage égalitaire.

La théorie des jeux ainsi esquissée par von Neumann et Morgenstern dans leur ouvrage fondateur postule l'existence de standards de comportements acceptés qui éclairent la solution du jeu. Selon eux, ces standards de comportements, adoptés par les joueurs, ne sont pas réductibles à leur seule rationalité individuelle. Ils considèrent seulement que les recherches sur l'origine de ces standards de comportements sortent du champ d'investigation de la théorie qu'ils se proposent élaborer.

Depuis lors, la majorité des progrès réalisés par la théorie des jeux, sous l'influence des économistes, a été accomplie en adoptant une perspective différente sur les jeux initiée par Nash. Afin de réduire le caractère arbitraire du choix des normes sociales qui inspirent ces standards de comportement, on peut tout simplement postuler que les joueurs choisissent de manière rationnelle leur stratégie et que ce choix rationnel consiste seulement à adopter une stratégie qui maximise leur gain personnel. En assimilant ainsi le standard de comportement à la seule rationalité individuelle de joueurs égoïstes ainsi définie, on aboutit à une théorie plus dépouillée et plus parcimonieuse dans ses hypothèses. Formulé dans le jargon de la théorie des jeux, c'est le paradigme des jeux non coopératifs qui, pour cette raison, s'est progressivement imposé à une majorité de chercheurs, au moins en

économie. Ces avantages ont cependant leurs contreparties. L'économie expérimentale, puis, plus récemment, la neuroéconomie ont montré qu'une décision réfléchie n'est jamais réductible à une décision rationnelle. Elle fait intervenir un ensemble beaucoup plus riche d'éléments. Certaines de ces composantes relèvent de la cognitivité, elle-même distincte de la simple logique ; d'autres dépendent de la sensibilité et des émotions éprouvées par le sujet.

Ces nouvelles données ont non seulement conduit à réexaminer un ensemble de modèles de jeux, dérivés du standard des choix individuels rationnels, mais elles ont, surtout, contribué à réouvrir la question des standards de comportement, initialement posée par Neumann et Morgenstern, avec cependant un changement épistémologique important. Au lieu de les tenir pour des facteurs, certes importants, mais extérieurs à la théorie des jeux elle-même, il s'agit désormais de les endogénéiser, de manière à pouvoir en rendre compte, au moins partiellement, en utilisant des développements plus récents de la théorie des jeux.

Pour y parvenir, les chercheurs ont tout d'abord essayé d'expliquer ce qui pousse les joueurs à conformer leur propre comportement à ce qu'ils croient être des standards de comportements socialement acceptés par les autres. Ils ont ainsi dégagé une dynamique du conformisme, où prime, chez chacun, l'idée de ce que pensent les autres de leur comportement. Cette dynamique se trouve entretenue par le fait que les comportements conformes à ces standards se trouvent récompensés, tandis que les comportements déviants sont sanctionnés, voire pénalisés (Jones, 1994 ; Bernheim, 1994). Bernheim montre sur des exemples que ces récompenses et ces sanctions peuvent être matérielles, mais qu'elles peuvent également être purement morales (l'estime des autres, la réputation…). Les économistes seront tentés d'invoquer alors le rôle de « métapréférences », dont il reviendrait aux neurosciences de dégager les ressorts cérébraux. On peut y voir un argument supplémentaire en faveur du rôle joué par la punition altruiste, entendue, au sens large, dans la transmission des standards de comportements, et donc des règles auxquelles ils renvoient.

Une branche particulière de la théorie des jeux se propose d'aller plus loin dans la recherche des déterminants sociaux de ces standards de comportements. Elle cherche ainsi à rendre compte des évolutions dans les comportements sociaux, en étudiant directement la dynamique des interactions sur une période suffisamment longue, non pas seulement entre deux ou quelques joueurs individuels, mais également et surtout entre des populations d'un grand nombre de joueurs. Ces travaux ont ainsi mis en évidence l'émergence et l'évolution de normes sociales qui, indépendamment des intentions des individus, orientent leurs actions.

Ici encore le jeu de l'ultimatum peut servir d'illustration. Dans sa formulation statique, toutes les propositions de partage émanant de celui qui joue en premier débouchent sur une forme d'équilibre faible, puisque celui qui joue en second a toujours intérêt à l'accepter – un refus entraînant, pour lui, un manque à gagner. Le standard de comportement associé à ce concept de solution, tel que nous l'avons interprété, recommande donc au joueur en premier de proposer la formule de partage la plus inégale, à son avantage, et, au joueur en second, de toujours l'accepter. Imaginons maintenant que le jeu soit répété un très grand nombre de fois. Les comportements que prescrit la solution statique

deviennent alors caducs. Le joueur en second peut avoir intérêt à refuser certaines des offres de partage inégales qui lui sont faites, et celui qui joue en premier à modifier ces offres, pour tenir compte de ces refus. De tels ajustements interviennent à la lumière des comportements observés chez l'autre et des anticipations que suscitent ces observations chez chacun. De ce fait, la pluralité des équilibres possibles réapparaît et un nouveau problème surgit. Il concerne la sélection de l'un de ces équilibres. Sa solution dépend, cette fois, de la dynamique du système et, en particulier, de la durée du déroulement du jeu. Le concept de solution pertinent repose alors sur la stabilité du système et renvoie à d'autres standards de comportements. Quelques auteurs ont ainsi calculé, qu'en partant d'une offre de partage initiale très inégalitaire refusée par le partenaire, un jeu de l'ultimatum parvient à se stabiliser autour d'un équilibre, dès que l'offre de partage s'élevait au-dessus de 20 % du total (Gale, Binmore et Samuelson, 1995). Selon ces auteurs, les standards de comportements destinés à guider les joueurs découleraient, en réalité, de cet équilibre stable que les joueurs n'appréhendent que progressivement, par tâtonnements au cours de l'évolution du jeu.

En multipliant le nombre des joueurs et en raisonnant sur une longue période, les conséquences de ces mécanismes se précisent sur l'émergence, le renforcement et même l'évolution de ces normes sociales. Qu'il s'agisse du jeu de l'ultimatum, du jeu de la confiance (ou de l'investissement), ou du jeu de la punition altruiste de De Quervain, tous ces jeux de partage qui servent de supports à la mise en évidence de règles de comportements implicites reposent sur une asymétrie initiale d'information dont bénéficie l'un des deux joueurs au détriment de l'autre. Cet avantage tend à s'émousser lorsque la dynamique du jeu met en relation, sur plusieurs périodes, non plus deux joueurs, mais deux catégories de joueurs. Il va même pratiquement disparaître si un même individu sait à l'avance qu'il pourra appartenir à l'une et à l'autre de ces deux catégories au cours de l'évolution du jeu. On comprend mieux, dans ces conditions, que les standards de comportements qui émergent de ces processus tendent naturellement vers des normes de partage plus égalitaires (Young, 2007).

La mise en relation des données, tirées de l'exploration cérébrale de sujets individuels au cours d'expériences nécessairement brèves et singulières avec les informations déduites de modèles évolutionnistes faisant intervenir des populations d'individus sur de longues périodes, n'est pas aisée. Elle représente l'un des principaux défis pour la recherche sur la question de l'origine et du fonctionnement de règles implicites dans les interactions humaines. Deux hypothèses peuvent être avancées pour faciliter leur articulation. La première est d'ordre génétique. Elle pourrait prendre la forme de la transmission génétique d'une mémoire sociale La seconde, qui n'est pas exclusive de la première, viserait l'activation ou la réactivation de cette mémoire par l'intermédiaire de comportements individuels en situation, dont les neurosciences commencent à fournir une description plus complète. Le rejet d'un partage inéquitable et, peut-être, plus encore, la punition infligée à celui qui s'en rend coupable au prix d'un coût personnel, constitueraient ainsi l'une des modalités repérables destinées à favoriser, voire à accélérer l'enclenchement de ces processus d'adaptation sociale décrits par ces modèles. La décision de punir pourrait alors s'interpréter comme une tentative du sujet individuel pour corriger un changement risquant de faire tendre vers le déséquilibre, ou vers un équilibre pervers, le système général dans lequel il évolue. Ce changement serait d'abord perçu, au

niveau du sujet, comme une rupture entraînant une perte de référence qui serait à l'origine de ce que nous avons appelé, à la suite de De Souza, une « émotion épistémique ».

Commentaire Règles, altruisme et reconnaissance sociale

Quand avons-nous besoin de règles dans nos interactions sociales (ces règles pouvant être implicites) ? Quand dans une situation plusieurs modes d'interaction sont possibles, et qu'il est avantageux collectivement de privilégier l'un d'eux. Pourquoi cette référence au collectif ? Certes un sujet isolé peut sans doute se fixer à lui-même des règles, mais il peut aussi changer de règle à discrétion[7]. S'il en arrive à se sentir coupable de ne pas avoir suivi la règle qu'il s'est fixée, c'est par référence à une opinion qu'il pense généralisable, et il est ainsi entré dans le champ d'un collectif au moins virtuel. En revanche, on n'a pas besoin d'une référence à un collectif pour se donner des objectifs et un plan d'utilisation des moyens pour y parvenir. Le plan dépend alors d'une rationalité instrumentale, qui est censée fournir les meilleurs moyens possibles, et ceux-ci dépendent du choix de l'objectif, qui reste totalement dépendant, par hypothèse, du choix individuel. Au contraire, si nous suivons une règle, c'est non seulement parce que d'autres règles sont possibles pour nous, mais aussi parce que d'autres partenaires pourraient en choisir d'autres, et que, comme nous voulons vivre ensemble, selon ce que Wittgenstein appelle une forme de vie, il faut que nos conduites s'accordent. Les règles, dans la vie en commun, supposent donc une référence à un point de vue collectif.

Cela implique que toute conduite sociale renvoie à des règles, mais pas forcément que ces règles aient été choisies de manière délibérée par les partenaires. Il se pourrait qu'en fait personne dans le premier groupe n'ait choisi volontairement ces manières de se conduire, et que pourtant on puisse en observant les conduites d'un groupe penser que ses membres obéissent à des règles, parce qu'on voit qu'un autre groupe en suit de différentes. Qu'est-ce qui pourrait alors justifier l'usage du concept de règle ? Ce serait que des déviations par rapport à ces types de conduite donnent lieu dans le groupe à des évaluations négatives, voire à des sanctions – et quand ces sanctions peuvent elles-mêmes être considérées comme régies par des règles, on parlera de normes. Nous pourrons alors soutenir que, du point de vue du groupe, ceux qui dévient non seulement ont un comportement qui ne se *conforme* pas à la règle, mais ne *suivent* pas la règle. Nous pourrons donc faire la différence – même si ce n'est que par une approche négative – entre se conformer à une règle – sans forcément s'y référer – et suivre la

règle, ce qui suppose au moins de pouvoir s'y référer.

L'économie expérimentale et la neuroéconomie ont été sensibles à ce critère d'évaluation négative et de sanctions, et l'ont nommé « punition ». Les sociologues auraient parlé de « sanctions ». Nous avons déjà vu que l'on parlait de « réciprocité forte » (Gintis) quand les coopérateurs punissaient les non-coopérateurs même s'ils ne pouvaient espérer que cette conduite de punition leur rapporterait quelque chose. Christian Schmidt a évoqué dans ce chapitre le concept de « punition altruiste ». Comme celui de règle, il peut sembler lié à une conduite délibérée. Pourtant le terme d'altruisme est ici utilisé dans un sens très voisin de celui où il peut s'appliquer en théorie de l'évolution, pour des interactions entre animaux de toutes sortes, à la limite des bactéries, pour lesquels le concept de délibération ne semble pas approprié. Nous allons donc revenir sur ce qu'est l'altruisme en théorie de l'évolution.

Nous pourrons ensuite revenir sur l'usage que font les neuroéconomistes de la notion de punition. Ils l'utilisent dans un but quelque peu paradoxal. D'une part ils observent des activités de punition qui sont coûteuses pour celui qui les exercent, conduites qui semblent irrationnelles pour un agent économique. D'autre part ils ont tendance à conclure, de l'existence de ces coûts, et de l'hypothèse de rationalité économique des agents, qu'il doit bien y avoir pour un avantage qui compense ces coûts, et qu'il faut donc supposer un « plaisir de punir ». Nous montrerons, revenant de plus près sur les interprétations que De Quervain *et al*. donnent de leur expérience, que la recherche de reconnaissance sociale serait un candidat plus plausible pour expliquer ces conduites. Mais celle-ci exige, là encore, une référence à une évaluation possible par des tiers, ce qui nous confirme la complexité de la structure des interactions sociales. Elle implique, outre des relations entre moi et autrui, des relations triangulaires entre ce duo et des tiers, ainsi que des références à un ou à des points de vue collectifs.

L'

Ce terme est utilisé en plusieurs sens[8]. Ce n'est pas le sens usuel – être motivé à aider les autres –, ni le sens économique théorique – préférer les intérêts des autres – que nous étudierons. Ce sont d'une part le sens évolutionniste, lié à la génétique et à la reproduction, d'autre part le sens usité en économie expérimentale, où l'altruisme est révélé par des comportements qui sont coûteux pour la personne et avantageux pour les autres.

Dans la perspective évolutionniste, l'« altruisme » est un concept qui combine le point de vue du biologiste et celui de l'économiste. Il désigne, comme pour l'économiste, un comportement d'un agent qui n'est pas dans son intérêt personnel. Comment de tels comportements ont-ils pu être sélectionnés par l'évolution, si l'on suppose que la sélection a plus de chances de favoriser les comportements qu'on pourrait dire optimaux (qui seraient rationnels, pour un agent supposé rationnel) ? Parce que, s'ils n'ont pas de bénéfices directs pour l'individu, ils ont des bénéfices indirects. De quels bénéfices s'agit-il ? De ceux apportés à d'autres individus : ceux qui sont

apparentés génétiquement aux premiers ; ceux dont le phénotype (les propriétés manifestes de l'individu) est similaire à celui de leurs parents génétiques ; ceux qui jouent un rôle utile pour la maintenance d'un groupe ou collectif en réseau, quand le maintien de ce réseau offre un intérêt pour les deux premiers types d'agents.

Dans le premier cas (parenté génétique) on observe bien l'existence de groupes, mais ce n'est pas par référence au groupe en tant que tel que le « bénéfice direct » peut être défini, c'est seulement par rapport à un capital génétique et à sa reproduction. Dans le deuxième cas, c'est toujours le capital génétique qui est la référence. Cependant, de manière parasitaire par rapport à cette référence, des interactions qui n'ont pas pour fondement effectif le partage d'un même capital génétique sont favorisées. La raison en est que les individus ne disposant pas, avant une période très récente dans l'évolution, de tests ADN, seules des ressemblances de surface, des ressemblances phénotypiques, peuvent leur donner des indices de parenté – mis à part, évidemment, les relations de filiation directe ou de fratrie. Des agents qui favorisent ceux qui leur sont similaires ont cependant des chances de favoriser aussi des agents qui leur sont apparentés génétiquement, si bien que leur capital génétique se reproduit et que l'évolution lui permet de se maintenir, voire de se diffuser.

Dans le troisième cas, celui d'un réseau d'individus, il nous faut faire une distinction entre les réseaux qui forment des organismes (nous-mêmes, faits de cellules, de bactéries intestinales, etc.) et les collectifs sociaux qui ne sont pas des organismes. Johannes Martens, dans son excellente thèse (2012), dont nous nous inspirons ici, pense que le mode de cohésion des organismes tient à ce que les intérêts de ses composants convergent entre eux au niveau du collectif qu'est l'organisme et que de plus ce mode de coopération y est robuste par rapport à l'introduction de mutants – avec cette limite que, lorsque les mutants l'emportent, l'organisme meurt. Un collectif social n'implique pas une telle convergence. Nous avons déjà montré la différence entre un groupe aux intérêts convergents et un collectif social : dans le second, la possibilité d'exploiter la coopération collective pour satisfaire des intérêts individuels persiste malgré tout. La survie du collectif pourrait alors tenir à la présence de processus compensatoires – dont les « punitions » peuvent faire partie.

Ce qui importe au théoricien de l'évolution, c'est que, même dans ce dernier cas, l'altruisme a encore des bénéfices indirects : les coordinations et coopérations au sein du réseau du collectif augmentent les chances, pour un capital génétique donné, de survivre et de se reproduire. La différence avec les deux autres justifications des avantages de l'altruisme au sens évolutionniste est qu'un collectif social augmente ses chances pour plusieurs lignées génétiques qui coexistent, voire qui peuvent collaborer.

Quand les économistes parlent de « punition altruiste », ils continuent à supposer que le comportement de punition ne sert pas les intérêts directs de celui qui le présente, et que ce comportement présente en première analyse un coût non compensé pour cet individu. Mais ils ne font plus référence à un capital génétique et aux bénéfices indirects qui lui sont liés. Du coup, ils sont amenés à imaginer un bénéfice direct mais masqué, qui compense ce coût pour l'individu. Et ils en arrivent à parler du plaisir de punir. Les biologistes et théoriciens de l'évolution n'ont nul besoin de ce concept d'apparence sadique. Quand la punition a pour effet de maintenir le réseau de coopération,

elle offre des avantages quant à la survie et diffusion d'un capital génétique. Cependant, les humains ne se livrent pas à n'importe quel type de punition. Ils semblent sélectionner celles qui ont des chances de maintenir le réseau social, sous la forme culturelle que préfère le collectif. Cela suppose que se diffusent au sein d'un même collectif ces préférences pour certaines formes de cohésion plutôt que d'autres – ce qui donne lieu, entre collectifs, à une diversité culturelle. Un comportement qui satisfait cette référence aux préférences collectives ne donne pas forcément le plaisir de punir, il donne surtout une reconnaissance sociale.

Il faut maintenant nous demander quel est le rapport des comportements de punition avec la constitution d'un réseau d'interactions sociales qui apparaît organisé par des règles, et comment peut y intervenir la reconnaissance sociale. Ensuite nous pourrons revenir sur l'interprétation donnée par De Quervain *et al.* de son expérience, et montrer que l'hypothèse de la recherche de reconnaissance sociale est plus satisfaisante que celle d'un plaisir de punir.

M

Des modèles économiques de théorie des jeux nous fournissent des processus d'émergence de ce genre de réseaux. Pour qu'il y ait vraiment émergence de règles et non pas normativité présupposée, il faut cependant construire des dispositifs complexes, qui mettent en jeu les relations entre coopération, exploitation de cette coopération, punition des exploiteurs, et aussi, ce qui peut sembler étrange, mais qui constitue pourtant un facteur essentiel, punition de ceux qui ne punissent pas les coopérateurs.

Commençons par des modèles plus simples, mais qui ne sont pas très plausibles socialement. Ainsi un modèle comme celui de Bowles et Gintis (2004), qui fait intervenir une punition des non-coopérateurs, exige de distinguer trois populations : 1) les coopérateurs purs, qui coopèrent sans jamais punir les non-coopérateurs, 2) ceux qui suivent leur propre intérêt (les *self-interested*), et 3) ceux qui punissent les coopérateurs inconditionnellement – sans faire dépendre leur punition d'un bénéfice personnel présent ou futur – et qu'ils appellent les « réciprocateurs » (notons qu'ils ne font pas partie des coopérateurs « conditionnels », puisqu'ils coopèrent sans condition, sinon celle de pouvoir punir les non-coopérateurs ; ces réciprocateurs pratiquent la « punition altruiste », d'après la typologie neuroéconomique). Les simulations du modèle montrent alors que dans des conditions plus défavorables pour la survie du groupe, la population des coopérateurs va être envahie par les *self-interested*, mais qu'il suffit d'une arrivée de « réciprocateurs » pour que les coopérateurs puissent se développer, le tout donnant finalement lieu à une population mixte entre les trois groupes. Le problème pour la plausibilité sociale de ce genre de modèle, c'est que, si l'on revient aux pratiques sociales, on ne voit pas quel intérêt social auraient les « réciprocateurs ». En effet ils ne seraient pas estimés par les coopérateurs – qui trouveraient qu'ils sont trop haineux –, ni par ceux qui s'occupent de leurs propres intérêts, qui les trouveraient irrationnels. Or on imagine difficilement qu'un groupe d'humains puisse se développer sans satisfaire ses motivations de reconnaissance sociale.

En revanche, situons-nous dans un collectif où existent des motivations conformistes. C'est en grande partie ce conformisme qui induit la coopération et la punition des non-coopérateurs. Ajoutons une possibilité de punir ceux qui ne punissent pas les coopérateurs. Cette punition peut consister en des coûts monétaires, mais tout aussi bien et plus sûrement en des pertes de réputation. Les simulations montrent alors qu'il peut suffire de punir ces gens trop tolérants (qui peuvent être des altruistes inconditionnels), qui atténuent la motivation à sanctionner (Hwang, 2012) pour que cela stabilise la coopération, car cela permet de ne pas voir s'éteindre le comportement de punition des non-coopérateurs quand arrivent des mutants, coopérateurs ou non, qui ne le pratiqueraient pas. Ce mécanisme par lui-même n'implique pas initialement de référence à une règle, puisque pour les conformistes il s'agit simplement de suivre les actions de leur groupe sans que se pose immédiatement la question de choisir entre des pratiques concurrentes. Il implique simplement une référence aux tiers. Mais il peut arriver à produire de véritables règles – qui sont aussi des normes puisqu'il y a sanction – parce qu'il assure bien finalement une régulation des conflits entre les manières d'agir des coopérateurs, des exploiteurs, et des coopérateurs punisseurs et non punisseurs (voire la revue de Chudek et les modèles de Henrich et Boyd, 2001 ; et Guzman *et al*., 2007).

Le conformisme peut se parer des atours du légalisme, voire du moralisme. Là aussi des simulations et des expériences montrent que n'autoriser à punir que les personnes supposées socialement vertueuses – parce qu'elles ont montré leurs capacités de coopération par le passé – renforce les usages coopératifs, en particulier si ceux qui veulent se faire bien voir disposent d'informations sur le niveau de coopération manifesté par les plus exemplaires (Faillo *et al*., 2013).

On notera que, dans ce genre de modèle, toutes les catégories d'individus peuvent obtenir une forme de reconnaissance sociale : les punisseurs réciprocateurs peuvent voir leur motivation de reconnaissance sociale satisfaite, et être appréciés des coopérateurs. Les punisseurs de coopérateurs peuvent se considérer comme supérieurs aux simples coopérateurs, parce qu'ils montrent en les punissant qu'ils jugent leur comportement socialement irresponsable, et être appréciés des punisseurs de non-coopérateurs. Les coopérateurs peuvent toujours se juger porteurs de valeurs plus hautes que les punisseurs de coopérateurs et rester appréciés par les punisseurs de non-coopérateurs. Ceux qui visent seulement leur propre intérêt ont une reconnaissance moindre, mais ils peuvent considérer tous les autres comme moins réalistes, et ils justifient l'activité des punisseurs de tout poil. Les tendances à la recherche de reconnaissance sociale permettent même de nourrir les aspects négatifs du modèle : la punition du *self-interested* peut être une perte de réputation, voire un ostracisme (comme dans le modèle de Bowles et Gintis), et celle du coopérateur qui ne punit pas les non-punisseurs peut être simplement qu'il soit mal considéré par les autres conformistes et que sa reconnaissance sociale soit diminuée d'autant.

Dans cette interprétation en termes de reconnaissance sociale, les punisseurs de non-punisseurs ont un comportement « altruiste » au sens évolutionniste, ou même peuvent trouver un bénéfice direct. Car si l'on fait le bilan de ce que ce comportement coûte et de ce qu'on y gagne, ceux qui l'adoptent sont en principe légèrement gagnants. En effet, en punissant les non-punisseurs ils créent une

différence en leur faveur entre eux et les autres membres du groupe en apparaissant comme plus exigeants, et c'est là pour eux une source de reconnaissance différentielle de statut social à l'intérieur du groupe.

Cette interprétation échappe aux critiques que Guala (2011) a adressées aux modèles de réciprocité forte du genre de celui de Bowles et Gintis. Il leur reproche, avec quelque fondement, de partir d'un comportement de punition qui est local, coûteux et non coordonné, alors que dans les sociétés réelles on observe des comportements de punition certes locaux, mais peu coûteux – parce que la punition est surtout symbolique, le maximum étant l'ostracisme – et coordonnés – puisque cette punition symbolique repose sur ce que se disent les uns aux autres les membres de la communauté de celui qu'ils pensent défaillant – les commérages et les rumeurs alimentant les réputations négatives. Or, dans notre interprétation de ces modèles de Henrich et Boyd, de Guzman *et al.*, ou des expériences de Faillo *et al.* en des termes de reconnaissance sociale, il s'agit bien de punitions « symboliques » en ce sens, et qui reposent sur des communications entre membres de la communauté.

Si on pousse un peu plus loin l'interprétation d'un tel modèle, on pourra penser que ce dispositif revient à transformer le statut de la punition – qui pourrait n'être qu'un comportement agressif réactif – en lui donnant une valeur sociale. La notion de valeur est ici introduite sans devoir présupposer des valeurs idéales, mais en reliant chaque valeur à des comportements, qui ont la particularité de constituer un signal adressé par un individu à ses partenaires, signal coûteux pour cet individu, mais bien discernable, précisément parce que coûteux, et apprécié par les tiers qui sont membres du groupe référent mais qui ne sont pas directement impliqués dans l'interaction en cause (à vrai, dire, l'estimation d'un coût est principalement comparative : par exemple, l'autre fait plus d'efforts que nous)[9].

C'est la synergie de toutes ces interactions en retour – évaluation par des tiers, similarité avec des membres d'un sous-groupe, différenciation avec ceux d'un autre sous-groupe, opposition commune aux comportements d'autres individus qui sont par là même exclus du groupe – qui constitue le groupe comme entité.

Il faut noter qu'il existe aussi des comportements collectifs de punition « antisociale », qui punissent les coopérateurs indépendamment de leur participation ou non à la punition des non-coopérateurs – mais cela dans des situations sociales particulières. Ils peuvent avoir un effet négatif sur les coopérations (Powers, 2012), puisqu'ils incitent les coopérateurs à la défection, mais ils ont moins d'influence, quand existent dans le réseau d'interactions des « indépendants », qui n'exploitent pas les autres ni ne les punissent, et sur lesquels les tentatives de sanctions n'ont pas de prise. Ces indépendants, dans l'évolution des interactions, vont assurer un pont entre les stades où les défections mutuelles sont dominantes et ceux où la coopération peut réapparaître (Garcia, 2012).

Des études interculturelles montrent que moins les gens ont confiance dans le respect des lois par les autres membres de la population, plus ils vont avoir tendance à punir les coopérateurs. Est-ce parce qu'ils craignent des comportements de vengeance des non-coopérateurs punis ? Cela expliquerait une diminution des punitions pour les non-coopérateurs, mais pas la punition des coopérateurs. En fait on retrouve ici au départ un comportement conformiste : on s'attend, dans ces sociétés où les normes de

punition collective des *free-riders* sont faiblement partagées, à ce que les gens ne coopèrent pas beaucoup. Des coopérateurs trop généreux pourraient obtenir une différence de statut au détriment des autres, qui ont donc intérêt à les punir pour en avoir trop fait, toujours dans une recherche de reconnaissance de leur propre statut. La recherche d'un statut social reconnu semble donc là encore le motif principal de cette punition des coopérateurs. De fait, dans ces sociétés, les personnalités dominantes et qui sont engagées dans une compétition pour une reconnaissance de premier plan sont les plus actives aussi bien pour punir les coopérateurs que pour punir les free-riders (Hermann *et al.*, 2008).

Par rapport à cette relative sophistication des modèles de théorie des jeux, les données neuroéconomiques semblent un peu sommaires. Ainsi on n'est pas surpris d'apprendre que si une récompense monétaire suscite une activation du striatum ventral, une récompense qui consiste en une reconnaissance sociale (par des tiers observateurs) suscite une activité comparable. On sera plus prudent que les auteurs de cette observation, qui concluent, du fait que dans les deux cas le striatum soit actif, qu'il réalise par là même une équivalence entre récompense monétaire et reconnaissance sociale (Izuma, 2008, 2009). Il faudrait pour cela au moins avoir réalisé des expériences dans des situations où les deux gratifications pourraient être en conflit, où un accroissement de récompense monétaire pourrait diminuer la reconnaissance sociale, ou bien la reconnaissance sociale impliquer une diminution monétaire, pour voir si l'activité dans le striatum diminue ou pas. Or, quand se présentent de tels conflits, on observe généralement une activation dans le préfrontal, ce qui implique non seulement que la comparaison est faite, mais qu'elle posait un problème, et qu'il fallait établir des éléments de comparabilité entre les deux options, donc que l'équivalence n'était pas déjà assurée et qu'il a fallu décider – ce qui peut se faire dans des sens différents selon les contextes. En revanche ces expériences ont clairement montré qu'à récompense monétaire égale, l'approbation supposée des autres (comme tiers observateurs) accroît la satisfaction.

Notons par ailleurs que, malgré la sophistication de ces processus de recherche d'une reconnaissance sociale, des données qui s'appuient sur des processus très basiques, comme l'influence forcément assez générique de facteurs hormonaux, peuvent cependant être intéressantes dans ce domaine parce qu'elles révèlent en quelque sorte « en creux » l'importance pour les humains de la reconnaissance de leur statut social. Nous avons déjà étudié le cas de l'ocytocine. Nous pouvons évoquer maintenant celui de la testostérone. Un modèle mécanique de l'influence de la testostérone conduirait à prédire que l'accroissement de testostérone augmente l'agressivité, et, pourrait-on en conclure, amène dans des jeux du genre ultimatum à n'offrir qu'un gain minimal au partenaire. Or c'est le contraire qui peut se passer, comme on l'a observé dans une expérience sur des femmes. On peut interpréter ce résultat par un souci d'éviter dans ce jeu de voir son offre refusée, ce qui pourrait faire « perdre la face », selon les termes de Goffman. C'est là une démonstration indirecte du fait que même des mécanismes biochimiques de fond sont modulés par notre recherche de reconnaissance sociale (voir Eisenegger, Haushofer et Fehr, 2011). Et ce alors même que plus généralement on a tendance à interpréter ces effets d'hormones comme bloquant des processus cognitifs plus

complexes : par exemple, un accroissement de testostérone peut nous rendre incapable d'une confiance dans des coopérations complexes quand nous sommes dans des situations de compétition ou, dans un autre sens, peut diminuer le stress dans ces situations de compétition (Eisenegger 2010). Il semble donc que même cette forme d'insensibilité qui peut bloquer des ajustements sociaux soit aussi liée à des tendances à maintenir des statuts sociaux, ou à les accroître.

Toujours en raisonnant de manière différentialiste, on peut encore utiliser des différences d'activation de l'insula antérieure, liées à des traitements de la douleur ou de la peine, comme révélateurs de l'effet des évaluations négatives liées à des différentiels dans une interaction sociale : quand les participants d'un dilemme du prisonnier pensent jouer avec des partenaires humains, et que ce qu'ils reçoivent dépend de la congruence de leur choix avec celui du partenaire, l'activation de l'insula antérieure est plus élevée, quand l'autre a fait défection et qu'ils se trouvent tous deux avoir coopéré, que lorsque dans ces mêmes conditions de coopération ou de défection on leur présente la situation comme une loterie sans partenaire humain.

Enfin la différence entre les attentes normatives et les règles représentées, comme aussi le rôle des règles émergentes dans la résolution de conflits entre des pratiques, peuvent trouver une confirmation « en creux » en neuroéconomie. L'activation du DPFC (cortex préfrontal dorsolatéral) est associée à des confrontations entre tendances et son activité amène des résolutions de leurs conflits – par exemple en inhibant une des tendances. Or si l'on diminue sa capacité de régulation, on peut bloquer les comportements d'obédience aux règles, alors même que malgré ce blocage le sujet sait toujours reconnaître ce qui est conforme à ces règles. Cela montre qu'on peut se représenter des règles sans pour autant suivre le type de résolution des conflits qu'elles impliquent, et, indirectement, que l'activation du DPFC est importante pour passer de la simple représentation normative à une conduite conforme à cette représentation. Cela montre aussi, s'il en était besoin, que les normes impliquent bien des résolutions de conflits – ici entre la conduite qui n'applique pas la norme et celle qui la met en pratique.

P

Revenons enfin sur l'expérience de De Quervain *et al.* (2004). Elle est souvent évoquée par Ernst Fehr quand il veut montrer non seulement que nous marquons une préférence pour punir les non-coopérateurs, mais surtout que cette préférence repose sur un calcul qui pèse le contre – le coût de l'activité de punition pour celui qui punit – et le pour – censé être le plaisir que l'on tire d'une punition dans ces circonstances, le plaisir de voir sanctionné celui qui ne nous a pas renvoyé l'ascenseur. Les données de l'imagerie cérébrale se mêlent ici à une interprétation économiste qui nous semble un peu rapide et simplificatrice. Nous allons tenter de montrer que la neuroéconomie est ici encore compatible avec des interprétations plus sociales, qui font leur part aux évaluations normatives qui passent par l'approbation ou la désapprobation supposées des tiers. Nous analyserons

ce cas exemplaire en détail, parce qu'il est significatif d'une tendance répandue des expériences de neuroéconomie à conjoindre sophistication du protocole expérimental et simplification des interprétations, par rapport à la complexité des processus sociaux.

Christian Schmidt a rappelé les 4 conditions de l'expérience. 1) IC : punition en condition intentionnelle, c'est-à-dire liée à l'intention supposée du receveur ou mandataire ; on dit à l'investisseur I que le receveur R a intentionnellement décidé de retenir l'argent, et la punition est coûteuse pour l'investisseur (pour 2 points infligés à R, I doit payer un point). 2) INC : intentionnelle et non coûteuse ; on dit encore que R a agi intentionnellement, mais la punition ne coûte rien à I. 3) IS : intentionnelle et « symbolique » ; les points de punitions ne coûtent rien ni à R ni à I. 4) NIC : non intentionnelle et coûteux ; c'est un dispositif au hasard qui décide pour R ; la punition est alors coûteuse pour les deux.

Les chercheurs font alors les prédictions suivantes, qu'il est nécessaire de rappeler de manière précise pour réexaminer leur interprétation des résultats : 1) dans la condition de punition symbolique IS, comparée à la condition INC, on ne doit pas trouver d'activation de récompense de I quand il punit, « puisque son action n'impose pas de coût à R », alors que les sites de récompenses seront activés dans INC. 2) Si punir donne une satisfaction, alors, dans le cas où cette satisfaction l'emporte sur le coût pour I, celui-ci devrait pouvoir punir même quand c'est coûteux. Plus les joueurs présentent une activation forte des aires liées à la récompense dans la condition de punition non coûteuse, plus élevés devraient être aussi les coûts de punition que ces joueurs peuvent accepter. 3) Si les joueurs recherchent une satisfaction dans l'activité de punir, on devrait aussi retrouver, en comparant IC et IS, une activation plus élevée dans les sites de récompense dans IC que dans IS. 4) Punir un dispositif aléatoire ne donne pas de satisfaction, donc dans la condition NIC on doit trouver moins d'activation de récompense que dans la condition IC. 5) Si on soustrait de la somme (IC + INC) la somme (IS + NIC), le solde d'activation dans le striatum, lié à la récompense, doit être positif, puisque les deux premières conditions sont toutes deux censées activer le circuit de la récompense, alors que la quatrième, NIC, doit donner peu d'activation puisqu'on n'a pas le désir de punir un objet dénué d'intentions, et que la troisième IS doit encore moins en donner.

Les résultats montrent que tous les sujets recourent à la punition dans la condition INC, que 12 sur 14 y recourent dans IC, et qu'il y a peu de punitions, mais tout de même quelques-unes (3 sur 14) dans la condition NIC (après tout certains d'entre nous tapent bien sur les écrous qui leur résistent). Le noyau caudé du striatum, une zone de la récompense, est actif dans tous les cas prévus. On y observe une activation plus forte quand l'investisseur inflige une plus forte punition.

Les auteurs se posent alors la question : est-ce l'espérance d'une plus grande satisfaction qui est la cause d'un investissement dans une punition ? Ou est-ce inversement cet engagement pour une punition plus forte qui est la cause de plus de satisfaction ? Ils examinent l'activation dans le noyau caudé des 11 joueurs qui punissent le plus, et à peu près au même niveau, si bien que leurs différences d'activation ne peuvent être dues à l'importance de la punition. Des deux explications ne reste donc en lice que la première. Ils considèrent alors les sujets pour lesquels, dans la condition INC, on observe une activation de récompense plus grande pour ce même niveau de punition. S'ils présentent aussi

plus d'activation dans la condition coûteuse, on en déduira que l'explication qui reste est la bonne, et que la cause de l'investissement dans la punition est une espérance de satisfaction plus élevée. Cette déduction ne pourrait être valide que si les deux explications étaient les seules que l'on puisse proposer, et si les sujets pouvaient projeter leurs évaluations propres à la condition INC sur ce qu'ils attendent dans la condition IC.

Dans la condition coûteuse, où il faudrait alors établir un bilan entre plaisir de la punition et coût de l'activité de punir, on observe une activation du cortex préfrontal ventromédian, qui est souvent activé quand il faut intégrer des opérations cognitives séparées, portant par exemple sur différents moyens utilisables dans la poursuite d'un but de rang plus élevé, et du cortex médian orbital. Il y a plus d'activation dans cette zone dans la condition IC que dans la condition INC.

La conclusion tirée est que « punir donne de la satisfaction, puisque si ce n'était pas le cas, on n'aurait eu aucun bénéfice à mettre en balance contre les coûts de punition et il n'y aurait pas eu d'activité d'intégration[10] ».

On notera à la fois l'intérêt de la méthode utilisée, et ses limites. L'une d'elles tient à ce que l'imagerie cérébrale ne permet pas de donner grand sens à une expression comme (IC + IF) – (IS + NIC). En effet, cette expression suppose que l'on puisse additionner des niveaux d'activation dans une condition avec ceux observés dans une autre condition, et qu'on puisse soustraire la somme (IS + NIC) de la somme (IC + INC) qui résume les inférences précédentes. Cela suppose que les valeurs mesurées des activations dans des conditions différentes aient des sens comparables, afin qu'on puisse donner un sens à ces additions. Or il faudrait pour ce faire admettre la possibilité que cela ait un sens que la deuxième somme soit égale à la première – et même qu'il y ait égalité entre deux différences (par exemple celle entre IC et IS et celle entre INC et NIC). Mais admettre la possibilité que de telles égalités font sens, c'est supposer qu'on peut s'appuyer sur des relations parfaitement quantitatives et que toutes les propriétés et opérations liées à ces relations quantitatives ont un sens pour ce qui concerne les rapports entre activations cérébrales et fonctions psychologiques (évaluation, comparaisons, émotions, décisions).

Or ni l'imagerie cérébrale ni les concepts psychologiques ne peuvent garantir qu'une telle exigence soit satisfaite. Ce qu'on désigne du nom d'activation dans une localisation cérébrale doit se comprendre comme une différence moyenne observée entre différentes activations globales de réseaux de processus neuronaux. Ces différentes activations sont elles-mêmes en fait le résultat de différences moyennes entre une activité au repos et une activité sous telle condition, ou plus généralement entre des activités sous deux conditions différentes. Comparer des activations sur une seule zone du cerveau ne permet donc pas de comparer deux quantités qui pourraient être égales, puisqu'il est probable que ces activations vont appartenir à des réseaux d'activations différents. Il en est de même des activités psychologiques : on ne peut pas observer un plaisir ou une peine de manière isolée du réseau particulier des activités de perception, d'évaluation, des activités motrices, de planification d'action, etc. au sein desquelles cette activation affective est insérée, et d'un réseau à l'autre, le sens à donner à ces comparaisons n'est pas clair.

Cependant, l'imagerie cérébrale nous offre des informations dont nous devons pouvoir tenir compte et qui sont précieuses parce qu'elles peuvent se croiser. Rappelons que pour donner un sens psychologique aux différences de localisation cérébrale de ces activités, il faut que les chercheurs s'appuient sur les différences entre de multiples tâches, non seulement celles de l'expérience proposée, mais celles mises en jeu dans toute la littérature neuropsychologique concernant les zones activées. Ces différences sont qualitatives. Par exemple citons les différences entre des tâches de décision où l'on peut fonder la décision sur un repère perceptif prégnant et celles où l'on doit procéder à des confrontations croisées entre des critères variés. Si on note régulièrement que telle zone du cortex préfrontal manifeste une activité différenciée quand il y a confrontation, et ne la manifeste pas quand on peut se laisser guider par le repère prégnant, on aura tendance à parler de localisation cérébrale de l'activité de comparaison et de confrontation dans cette zone. Mais ce sera une manière de parler, puisque dans la réalité on a toujours affaire à l'activation d'un réseau plus global, et qu'inversement les résolutions spatiales et temporelles de l'imagerie cérébrale ne permettent pas de tenir compte d'autres différences fines qui pourraient exister, et qui pourraient être signifiantes, entre deux réseaux locaux d'activations.

Pour l'instant, l'imagerie cérébrale ne nous permet de repérer comme significatives que des différences plus grossières que celles qui existent entre les tâches proposées dans l'expérience de De Quervain *et al.*, et certainement plus grossières que la différence entre sommes liées à la prédiction 5. Certes, cela n'empêche pas que ces différences entre activations cérébrales puissent cependant être parfois plus fines que les différences grossières entre fonctions psychologiques – rien n'assurant que les deux types de différences correspondent aux mêmes frontières ou transitions entre états. Inversement, et précisément parce que, *a priori*, nous n'avons pas de garantie que les différences entre activations cérébrales soient en correspondance avec les différences entre fonctions psychologiques, qui sont des reconstructions linguistiques de nos expériences, trouver malgré tout de telles correspondances est déjà d'un grand intérêt, toujours parce que cela nous permet de croiser des informations de sources variées. Et notre intérêt est encore plus élevé quand nous découvrons des différences cérébrales qui sont corrélées à des différences dans les tâches, mais qui ne sont pas en correspondance avec les schémas de localisation des opérations psychologiques que nous avions présupposés ou que nous avions élaborés à partir de la série des expériences de neuropsychologie déjà à notre disposition. Cela peut nous amener à ajouter de nouvelles distinctions entre opérations psychologiques, ou du moins à réajuster les définitions de ces opérations, ou encore à trouver des liens entre des processus psychologiques que nous n'avions pas pris en compte jusqu'alors.

Or l'interprétation des résultats de leur protocole d'expérience par De Quervain *et al.* opère, par rapport à tout ce dispositif méthodologique, plusieurs raccourcis qui sont discutables. Ainsi elle considère, dans la tradition utilitariste, comme admis que des coûts, supposés être des peines, ne peuvent être mis en balance qu'avec des plaisirs ou satisfactions, et que ces plaisirs sont directement liés à la punition comme telle. Mais c'est là supposer, premièrement, que toute activité d'intégration et de mise en balance est une activité qui oppose des peines et des plaisirs – or on peut vouloir

comparer des objets ou actions selon différents critères sans que ces critères soient réductibles à des peines et ou à des plaisirs ; deuxièmement, que tous les coûts sont des peines – or des efforts, par exemple, sont des coûts sans être forcément vécus comme des peines. Cela en particulier dans les interactions qui font référence à des tiers. On peut encore interpréter l'imposition de coûts sur l'activité de punition comme un signal de modération de notre désir de vengeance qui nous est donné par un tiers – en l'occurrence, celui qui a conçu l'expérience. Or cette modération ne se laisse sûrement pas réduire à une peine : nous pouvons être soulagés que quelqu'un d'extérieur nous aide à modérer notre colère.

C'est aussi supposer, troisièmement, que l'activité de punir nous donne une satisfaction, soit par son résultat (la perception de la peine de l'autre), soit par elle-même. Mais elle peut en donner d'autres manières, plus indirectes. Par exemple avoir la possibilité de punir peut nous indiquer que des tiers reconnaissent notre droit à nous défendre contre celui qui a exploité notre confiance. Nous disposons d'une contre-épreuve de cette hypothèse : attaquer autrui *via* son porte-monnaie sans que cela donne lieu à reconnaissance sociale par des tiers risque de ne pas nous donner beaucoup de satisfaction quant à notre statut social. Ly et *al*. (2011) ont montré que le striatum s'active aussi lors de la perception d'une amélioration du statut de la personne, donc de la reconnaissance qu'elle s'attire.

Plus généralement, les auteurs sont passés d'une approche différentielle, relationnelle et comparative assez justifiée (on punit plus dans certaines conditions que dans d'autres) à une approche directe (punir = satisfaction). Mais c'est passer sans précaution du conditionnel : « Si ce qui est coûteux est une peine et si punir donne une satisfaction, alors on va punir quand le coût ne dépasse pas la satisfaction » à une équivalence entre punir dans ces conditions et vérifier que punir donne une satisfaction, mouvement qui est cette fois logiquement douteux (passons sur le fait que l'antécédent de ce conditionnel est lui-même discutable et pourrait bien être parfois contrefactuel, car cela n'entamerait pas forcément la validité du conditionnel).

Tant que d'autres interprétations ne sont pas disponibles, les explications de De Quervain *et al*. peuvent passer pour une inférence à la meilleure explication. Mais d'autres interprétations de ces données comparatives sont possibles, plus en accord avec la prise en compte des tiers et l'influence des attributions d'intentions, et avec la nécessité, comme nous l'avons vu au chapitre précédent, de replacer les réactions des sujets dans le cadre d'une histoire où ils ont appris, en fonction des retours variés de différentes interactions, dans quelles conditions il est souhaitable de coopérer.

Supposons que l'attitude initiale des sujets soit celle d'une attente générale de réciprocité, ce que nous avons appelé une confiance-cadre. Puisqu'on les met en relation avec un partenaire anonyme, ils vont assez vite comprendre que dans cette expérience ils sont plutôt dans une situation de confiance-pari. Cela induit une différence de réactions. Quand notre confiance-cadre est déçue en une occasion, nous pouvons considérer cela comme une exception et donner une seconde chance à autrui. En revanche une fois découvert que la situation est celle d'une confiance-pari, et que notre confiance est déçue, nous sommes mécontents d'avoir conservé dans une situation de confiance-pari les attentes de notre confiance-cadre, qui était valable avec des proches, dans un domaine où nous aurions dû être

plus prudents – puisque nous découvrons que le supposé partenaire n'agit pas comme s'il était des nôtres. Nous déclenchons alors un autre scénario : il nous faut marquer qu'il n'est pas des nôtres, marquer que nous prenons en compte cette différence – sinon, ce membre d'un autre groupe non identifié va pouvoir prétendre « nous » avoir grugés. Ce serait non plus seulement une atteinte à nous, individus, mais aussi, indirectement, au « nous » qu'est le cercle de nos proches, qu'on aurait ainsi tenté d'envahir. « Nous » érigeons donc une défense. Cela est d'ordinaire coûteux, comme toute défense, mais ce n'est pas forcément vécu comme une peine. Nous vivrions la situation comme une peine seulement si nous ne pouvions pas nous défendre.

Par ailleurs, si dans l'expérience « on » nous permet de punir sans que cela nous coûte – « on », c'est-à-dire les tiers qui ont conçu l'expérience et qui en observent les résultats – c'est qu'on nous approuve. Cette approbation par des tiers induit une forme de soulagement, grâce à ce partage implicite des points de vue : les tiers, en nous donnant cette possibilité de punir, donnent implicitement le signal qu'ils sont d'accord avec nous pour penser que cette rupture de confiance ne correspond pas au cadre normal. Notre chute dans une situation où notre confiance-pari est déçue est compensée par le fait de retrouver une forme de confiance-cadre avec des tiers.

Maintenant, quand la punition est symbolique (IS), son autorisation vaut-elle accord du tiers ? Un accord formel, sans doute, mais qui laisse aussi penser que le dommage que nous avons subi n'est pas reconnu décisif. En effet, puisqu'il n'implique aucun coût de rétorsion, nous pouvons penser que les tiers jugent ce dommage peu coûteux. Dans ce cas, nous ne pouvons croire à une communauté d'appréciation avec ces tiers qui ne manifestent pas d'empathie avec nous. Que l'activation dans le striatum soit plus élevée dans IC que dans IS pourrait donc indiquer la satisfaction de voir notre propre déception partagée par un tiers dans la condition IC, alors que IS indique simplement que le tiers donne la permission formelle de punir, sans aller jusqu'à partager nos sentiments. On peut penser qu'il en serait tout autrement si la condition de punition « symbolique » impliquait une perte de réputation infligée au partenaire indélicat.

Comment expliquer que ceux dont le striatum est le plus activé quand la punition n'est pas coûteuse (INC) sont aussi ceux qui vont se lancer dans les punitions les plus coûteuses dans la condition IC ? Ce sont ceux qui dans les deux cas estiment avoir subi le plus de dommages, et jugent leur droit le plus bafoué, donc ceux chez qui la punition a le plus de satisfaction compensatoire à produire. Dans les deux cas, les tiers ont reconnu leur droit à punir. Une fois ce droit reconnu, son exercice peut aussi avoir un coût – celui d'une défense, et d'une réparation de cette atteinte à leur cercle de proches, à leur « nous ». Défendre cette frontière entre « nous » et « eux » est une activité qui procure de la reconnaissance sociale, en proportion de l'importance d'une défense justifiée. Cette reconnaissance, à ce stade, est simplement supposée par le sujet, qui estime que son cercle partagerait son évaluation. On est assez près d'une autoreconnaissance, mais elle est toujours aussi supposée intersubjectivement valide. Le coût de la défense peut alors être un signe de l'importance de cette reconnaissance (tout comme le prix d'un produit est parfois pris comme indice de sa qualité), en l'absence d'autres informations (voir le *lemon market* d'Akerlof). Les sujets en question sont donc

ceux qui estiment que ne pas coopérer avec eux est une faute dont la reconnaissance sociale doit être maximale.

Quelle interprétation donner alors de l'activité préfrontale, qui implique, surtout dans la situation de punition coûteuse, une confrontation entre différentes réactions possibles et une tentative d'intégration de divers critères qui peuvent aller en sens opposé ? Si le coût de la punition peut valoir comme indice d'une « autoreconnaissance » intersubjectivement partagée, où pourrait encore résider un possible conflit entre modes d'évaluation ? Mais il reste que cette défense a un coût, et que le sujet n'est pas certain d'obtenir une reconnaissance équivalant à ce coût. D'autres éléments restent aussi susceptibles d'interprétations opposées : le coût peut être le signal de l'importance du dommage subi et donc reconnu par le tiers, ou bien un signal de sa part d'avoir à modérer notre vengeance. On peut aussi se poser des questions sur le rapport entre l'importance de notre défense (est-elle suffisamment marquée ?) et l'importance de notre déception. On est donc avec IC dans un domaine d'incertitude interprétative, alors que la condition INC est plus simple : le tiers nous autorise à punir et reconnaît notre droit sans ambiguïté.

Dans cette perspective la condition NIC ne devrait-elle pas elle aussi susciter une certaine incertitude et perplexité, amenant une activité préfrontale ? Mais quand nous savons avoir affaire à une machine et qui plus est à un dispositif aléatoire, nous ne nous mettons guère en frais de discussions entre interprétations divergentes, et nous ne supposons pas non plus que des tiers puissent rentrer dans de telles subtilités. Il semble aussi inutile de nous lancer dans la multiplicité des interprétations à propos de la différence entre les deux sommes de deux conditions. Nous avons montré que les auteurs présupposaient qu'une telle mesure était signifiante, sans pouvoir l'assurer.

On voit que les expériences de neuroéconomie n'impliquent pas forcément d'en rester à des interprétations qui déconnectent les sujets de leur histoire d'interactions sociales, de leur prise en compte des opinions supposées des tiers et de leur recherche de reconnaissance sociale. Et c'est seulement si l'on prend en compte ces dimensions que l'on peut penser rendre compte de l'émergence dans une communauté de règles sociales.

La neuroéconomie peut d'ailleurs aussi mettre en évidence le rôle des attentes sociales dans l'activation de normativités. Marchetti *et al.* (2011) ont ainsi construit une expérience dans laquelle les participants sont d'abord l'objet d'évaluations par des tiers[11], qui les déclarent plus ou moins appréciés. Ces évaluations ont un caractère normatif, car elles servent de références pour ce qui est attendu. On aura d'ailleurs tendance, la plupart du temps, à tenir ce qui est communément attendu comme normatif, parce que l'on considère implicitement que c'est aussi attendu par des tiers. Sanfey soutient (Chang *et al.*, 2011), de manière complémentaire à Falk et Fischbacher, qui introduisaient dans leur modèle la prise en compte des intentions d'autrui à notre égard, que ces effets de dépendance par rapport à des attentes rendent mieux compte des comportements que l'attachement supposé à une norme d'équité (allusion à l'aversion pour l'iniquité, première théorie de Fehr). Effectivement, on peut penser que nous sommes sensibles à la reconnaissance des tiers avant même de parer nos évaluations d'une référence à l'équité. Formuler explicitement des valeurs et des normes comme celle de l'équité demande un pas supplémentaire dont nous étudierons les conditions dans le

chapitre suivant.

---

1. Borel était lui-même un grand joueur de bridge. Il rédigea avec Chéron, journaliste spécialiste de ce jeu, une *Théorie mathématique du bridge à la portée de tous* (1940). Plus largement, il consacra son cours de la faculté des sciences de Paris durant l'année universitaire 1936-1937 aux applications aux jeux de hasard (Borel et Ville, 1938).

2. On consultera sur cette question l'ouvrage de Osborne et Rubinstein, *Bargaining and Markets*, (1990).

3. La théorie des préférences révélées soulève beaucoup de questions qui tiennent principalement au caractère singulier des situations de choix. Doit-on, par exemple, induire du refus de choisir une indifférence traitée comme une équivalence ? Comment peut-on interpréter en termes de préférences individuelles un choix dit « stratégique », destiné à tromper l'autre dans une situation de jeu ? Sur l'ensemble des problèmes méthodologiques soulevés par la théorie des préférences révélées de Samuelson on renverra au livre de Wong, *The Foundations of Paul Samuelson Revealed Preference Theory* (1978).

4. Dès le début des années 2000, les théoriciens des jeux ont élaboré des modèles théoriques de jeux séquentiels de réciprocité pour lesquels ils ont conçu des concepts de solutions originaux définis en termes d'équilibre (Dufwenberg et Kirchteiger, 2003).

5. Par induction à rebours (*backward induction*), on entend un mode de raisonnement qui part de l'issue projetée d'une décision ou d'un système d'interactions pour remonter, séquence par séquence, jusqu'à son origine.

6. Le concept de standard de comportement que l'on trouve au cœur du livre fondateur de von Neumann et Morgenstern prouve que la trame initiale de la théorie des jeux n'était pas individualiste mais sociale (Schmidt, 2001).

7. Descombes (2004) insiste sur ce point dans *Le Complément de sujet*, Paris, Gallimard, « NRF Essais ».

8. *Cf.* Clavien, Chapuisat, 2013.

9. Cela permettrait de rapprocher les solutions « punitives » qui semblent procéder par « contrôle du partenaire » des solutions d'intérêt mutuel, par « choix du partenaire » dans un « marché biologique » (Baumard et Sperber, 2013 ; Nesse, 2007).

10. De Quervain *et al.* (2004), « The neural basis of altruistic punishment », *Science*, 305, p. 1257.

11. D'autres auteurs ont insisté sur cette importance du rapport aux tiers, par exemple Buckholtz, 2012 ; Chavez et Bicchieri, 2013 ; Nettle, 2013 ; Civai, 2013.

CHAPITRE 7

Règles explicites et normes morales

Les chapitres précédents nous ont montré que des comportements de coopération peuvent émerger dans l'évolution, par exemple quand ils sont fondés sur un noyau de relations de parenté et que des apprentissages peuvent étendre cette coopération à d'autres membres du groupe. La recherche de reconnaissance sociale qui devient alors possible renforce et stabilise cette coopération, si bien qu'on observe des comportements dont l'orientation vers la coopération est contrôlée par le souci d'être reconnu socialement, ce qui permet d'assister à l'émergence de comportements normatifs. Ces comportements peuvent émerger sans qu'il soit besoin de formuler des règles et de les édicter de manière que tous puissent les connaître, comme on l'a examiné au chapitre 6. Les règles sont alors implicites aux interactions qui les mettent en pratique. Mais il peut se révéler utile voire nécessaire d'expliciter les règles, parce que plusieurs pratiques différentes qui semblaient se rapporter à des règles similaires se révèlent pouvoir entrer en conflit.

L

Les théoriciens des jeux ont étudié les conditions de recours à ces règles. Nous esquisserons les différences entre les trois principaux travaux sur ce problème en les comparant à notre perspective, Christian Schmidt les analysera ensuite de manière plus fouillée. Le premier est celui de Binmore (1994, 1998). Il suppose que le jeu de l'évolution a déjà sélectionné certaines propensions pour certaines relations entre individus. Mais il reste encore plusieurs possibilités, et un problème de sélection entre ces possibilités se pose. Les individus vont alors négocier en ayant en tête des préférences « étendues » : ils préfèrent être une personne appartenant à tel type de collectivité. En fonction de ces préférences, ils vont négocier sous voile d'ignorance pour définir un contrat social. Les normes sociales sont le résultat de cette négociation.

Nous avons aussi admis que l'évolution peut favoriser les comportements altruistes, d'abord entre parents, puis entre semblables, et que cela incite à adopter un point de vue de groupe. Une fois ce point de vue adopté (ce qui revient aux préférences étendues ou « empathiques » de Binmore), on est amené à comparer les solutions offertes à des joueurs qui jouent selon leur intérêt strictement individuel et celles offertes à ceux qui prennent en compte ce point de vue de groupe. Prendre en compte la reconnaissance sociale, c'est de plus faire cette évaluation comparative du point de vue du groupe, à condition qu'il permette d'améliorer le sort des individus – ceux qui prennent ce point de vue de groupe. Cela peut amener à justifier des conduites de punition, mais alors le ressort est la reconnaissance sociale et non la punition. Un tel schéma évite d'avoir à imaginer une négociation, et encore moins une négociation sous voile d'ignorance, ce qui nous emmenait très loin de la vie sociale réelle, puisque dans cette vie, négociation et voile d'ignorance sont antinomiques.

Gintis (2010) a proposé de comprendre les normes dans le cadre d'un jeu où les joueurs n'arriveraient pas à se coordonner de manière satisfaisante s'ils ne corrélaient pas leur choix d'action à un signal commun. Pour que les normes sociales soient efficaces même en situation d'incertitude, Gintis doit de plus doter ses agents d'une « prédisposition normative » qui renforce dans ces situations leurs attentes de coordination. Mais Gintis ne nous renseigne pas sur le processus qui mène à cette explicitation. Alors que chez Binmore, le processus d'explicitation est irréaliste, chez Gintis, il est absent.

Les propositions de Bicchieri (2006, 2010) sont plus complexes. Ses agents sont dotés d'attentes empiriques – ils s'attendent à ce qu'un certain nombre d'autres agents suivent une norme – et d'attentes normatives – ils pensent que les autres ont tendance à obéir à la norme. Dans des situations où il peut y avoir conflit entre l'intérêt individuel et l'intérêt collectif – par exemple un dilemme des biens publics – et qui ne sont pas des jeux de coordination, la présence d'une norme sociale va transformer le jeu en un jeu de coordination, en changeant les utilités et donc les préférences. Mais la voie suivie par cette transformation va rester sensible au contexte – on ne peut définir une stratégie générale pour toute cette gamme de problèmes. Cependant, les agents ne persistent à suivre la norme que si leurs attentes sont satisfaites : leur adhésion à la norme reste conditionnelle. Ces propositions sont intéressantes et plus réalistes, mais on ne sait toujours pas comment la norme a pu devenir explicite, et quel effet est propre à cette explicitation – répondre à cette deuxième question peut nous guider pour répondre à la première, puisque si nous connaissons l'effet spécifique ainsi fourni, nous aurons une idée de son processus de sélection et d'émergence.

Les fonctions des règles explicites

Qu'apporte la formulation explicite de règles dans l'émergence de comportements normatifs ? On peut évoquer ici une thèse de Joëlle Proust, dans son étude sur la métacognition (2013), cette capacité qui nous permet, par exemple, de savoir que nous avons ou n'avons pas une connaissance,

sans pour autant avoir à l'activer immédiatement ou croire par erreur que nous la possédons. Elle soutient, en appuyant cette thèse sur des travaux de psychologie expérimentale et accessoirement d'imagerie cérébrale, que, contrairement au schéma qui nous vient en premier lieu à l'esprit, la métacognition n'implique pas d'entretenir des représentations explicites de second ordre concernant nos états ou processus cognitifs de premier ordre, mais qu'il suffit, pour être capable de métacognition, que nous disposions de *processus de contrôle*[1] de ces états ou processus cognitifs. Le contrôle consiste à pouvoir ajuster, modifier, réviser un processus cognitif, que ce soit pour lui assurer les préconditions de son bon fonctionnement, pour vérifier qu'il a bien atteint son but ou rempli sa fonction, et pour s'assurer de son déroulement correct – le contrôle intervient donc non seulement pour évaluer le résultat du processus cognitif, mais au cours de nos actions mentales.

Nous pouvons utiliser cette idée, une fois transposée, pour étudier le rapport entre comportements normatifs, qui suivent des règles implicites, et explicitation des normes. Certes bien des conduites sociales peuvent s'expliquer simplement en renvoyant à des normes « descriptives », c'est-à-dire en calquant nos comportements sur ceux le plus souvent exhibés par les autres, qui servent de normes implicites[2]. Mais plus on accepte cette idée, plus il devient nécessaire de trouver une explication à tout le travail déployé pour la production de normes explicites. Cela nous amène à l'hypothèse suivante : l'institution de règles explicites n'a pas pour principale fonction de nous permettre de nous donner une *représentation* d'une norme de conduite, mais bien celle de nous fournir des *moyens de contrôle* de la conformité de tel ou tel comportement à cette norme de conduite. Plus généralement, il s'agit d'étendre nos moyens de contrôle sur la normativité des comportements.

Certes, aucun contrôle ne peut garantir qu'un sujet a tel comportement dans l'intention de suivre telle règle, et pas pour d'autres raisons. En revanche, aucun être social ne peut prétendre avoir réussi à suivre une règle si son comportement sort des repères qui sont donnés par le processus de contrôle.

La tendance des philosophes du social est d'ailleurs trop souvent d'en rester aux représentations. Il en est ainsi de Searle, quand il soutient que pour créer une institution, il suffit d'une intentionnalité collective, d'une fonction donnant un statut (X compte comme F dans le contexte C ; par exemple, ce morceau de papier compte pour 10 euros dans le contexte français), et d'un acte de langage, une déclaration qui énonce cette fonction de statut. Une déclaration est un acte de langage qui invoque une représentation, qu'elle prétend transformer en un fait du monde par ce seul acte de déclarer (Searle, 1998). Cette théorie a le mérite de mettre l'accent sur le lien entre institution et règle explicite. Mais on peut faire toutes les déclarations du monde, cela ne suffit pas pour qu'une institution existe. Il faut encore que des activités effectives soient coordonnées par référence à cette institution.

L'explicitation des règles joue un rôle important dans cette mise en œuvre effective, qui est le test décisif de l'institution. Elle permet un contrôle bien plus étendu dans le temps et l'espace que les contrôles directs que chacun de nous est capable d'exercer sur une personne qui est à portée de main, ou avec qui il est en contact visuel ou auditif. La formulation de la règle se transmettant d'acteur en acteur, le contrôle qu'elle permet d'exercer passe en quelque sorte à travers les esprits de bien des acteurs pour finir par activer un ou plusieurs contrôles directs, bien loin de l'acteur qui a formé avec quelques autres le projet de ce contrôle. Ce type de contrôle, qu'on peut qualifier d'indirect, permet de

faire comme si tous les intermédiaires entre ce projet et les contrôles directs plus ou moins lointains qui en assurent l'efficience étaient contrôlés par référence à la norme explicite ainsi propagée.

On voit le lien entre admettre ces modalités de contrôle indirect et pouvoir prendre le point de vue du collectif, ou encore avoir confiance dans le collectif. Notons aussi un lien avec notre sensibilité à la reconnaissance sociale ; pour que je sois reconnu socialement, il faut au moins que les autres pensent ne pas avoir besoin de me contrôler directement pour pouvoir compter sur moi.

Les règles explicites ne surviennent pas seulement au départ d'un projet institutionnel, comme dans l'exemple de Searle ; elles peuvent aussi survenir en cours de coordination, voire à l'occasion de conflits de coordination. Mais quoi qu'il en soit, elles fournissent des potentialités de contrôle bien au-delà de leurs lieu et temps d'apparition et permettent de relier entre eux plusieurs points d'application de ces contrôles, plusieurs points de contact avec les activités effectives, ainsi reliés malgré leur distance. En chacun de ces points, cependant, tout contrôler serait impossible, si bien que le contrôle s'attache seulement à quelques traits de ces activités. À ces conditions, les institutions nous permettent d'établir des rapports directs entre des ancrages concrets pourtant séparés par bien d'autres activités. Cependant toutes les règles qu'on formule et propose ne parviennent pas à se propager ainsi, ou, quand elles se propagent, ne donnent pas forcément lieu à des coordinations qui puissent encore se référer à la règle initialement explicitée. Ces échecs montrent que les règles explicites ne peuvent pas se réduire au seul ancrage de leur déclaration et qu'elles doivent être mises à l'épreuve de multiples autres tests dans différents points de contrôles.

Les contrôles que permettent les règles explicites se déclinent selon plusieurs fonctions de contrôle. Tout d'abord, 1) les règles explicites permettent un *contrôle de la communication*, plus précisément de sa fiabilité et fidélité. Elles peuvent être communiquées à un grand nombre d'acteurs sans que leur formulation soit trop distordue d'un individu à un autre. Cette propriété, d'ailleurs, exige de formuler les règles de telle manière que leur représentation soit stable d'un individu à l'autre et que chacun puisse contrôler la fidélité de sa propre représentation au message qu'il a reçu. À ce stade, les fonctions de représentation et celles de contrôle sont intimement liées. On peut renvoyer aux travaux de Sperber sur l'épidémiologie des représentations pour des pistes de recherche[3] sur ce qui peut assurer la fiabilité de ce genre de transmission. Cependant, une règle, si explicite qu'elle soit, n'assure pas à elle seule que son application dans un contexte soit pertinente ou qu'elle ne dévie pas trop de ce que donne son application dans un autre contexte. Et il n'existe pas de métarègle qui puisse nous dire pour chaque contexte différent comment modifier l'application d'une règle explicite. La stabilité des représentations dans la communication ne suffit donc pas à assurer la bonne application des règles.

Une condition supplémentaire est donc requise. L'application des règles doit avoir des chances de résoudre des problèmes de coordination et de coopération. Ceux qui nous intéressent ici tiennent tout d'abord à ce que certaines tâches ou activités collectives requièrent des coordinations, et nous considérons celles de ces coordinations qui exigent elles-mêmes des efforts de coopération de la part des individus – dont certains pourraient espérer profiter des efforts coordonnés des autres sans y participer eux-mêmes. Mais de plus, même quand ces individus font réellement des efforts de

coopération, il se peut encore que leurs efforts ne soient pas bien coordonnés. Nous retrouvons ici des conditions posées par Bicchieri et par Gintis, mais nous pouvons les rendre cohérentes entre elles, alors que leurs formulations initiales s'opposent : il y a bien au départ un problème de coopération (pour Bicchieri), et la coordination est au départ incertaine (pour Gintis). Pour répondre à ces deux problèmes, les règles explicites ont une fonction de *contrôle des coopérations de second ordre*. En effet, ces coordinations se révèlent exiger des coopérations, et inversement ces coopérations doivent de plus être elles-mêmes correctement coordonnées. Nous avions signalé au chapitre précédent le problème des exploiteurs de second ordre, ces coopérateurs qui ne punissaient pas les exploiteurs. Notre problème est maintenant celui de coopérateurs qui pourraient ne pas coopérer pour coordonner correctement leurs efforts de coopération. Il s'agit donc d'un problème de coopération de second ordre. Il faut que des coordinations permettent à chacun de contrôler que les conduites d'agents différents vont bien pouvoir se coordonner de manière conforme à la norme, alors même que cette coordination exige des efforts de coopération et que chacun pourrait avoir une interprétation légèrement différente de ce qu'il faut faire pour coopérer. La simple conformité telle que l'évalue tel ou tel sujet n'est donc pas suffisante, il faut que les différentes conduites puissent s'interpréter comme corrigeant leurs déviances pour aller dans le même sens. Les règles explicites peuvent alors nous donner des repères pour déclarer un comportement non conforme et faire partager ce jugement aux autres coopérateurs. Elles permettent un contrôle des coopérateurs sur l'orientation des efforts de coopération de leurs partenaires.

Les règles explicites peuvent encore avoir une troisième fonction, celle 3) du *contrôle du respect collectif des finalités et des valeurs*. Elles ne se bornent plus alors à proposer seulement une recette ou des critères de contrôle, mais elles visent aussi une finalité, ou encore elles mettent la coordination sous l'invocation d'une valeur (qui peut aller de la productivité au respect de la démocratie en passant par la bonne entente). Cela a l'avantage de nous permettre de ne pas avoir à attendre, pour exercer notre contrôle, d'avoir observé le résultat final des comportements agrégés. En fonction de cette fin ou de cette valeur, nous pouvons contrôler en cours d'action, voire avant le démarrage de l'action, la tendance à la convergence des efforts de coopération. Pour ce faire, nous prenons en point de mire la finalité ou la valeur qui justifie la règle, et nous exigeons des autres et de nous-même que chacun révise sa conduite en fonction de cette finalité ou valeur. Ces révisions doivent se coordonner – leur coordination va servir de repères de contrôle – si bien que la définition de la révision que chacun a à faire va dépendre de l'orientation de la révision des autres. Chacun va ajuster sa conduite en cours d'action, voire ses manières de se préparer à l'action, de façon à parvenir à une correction mutuelle des divergences entre les partenaires par rapport à cette finalité ou à cette valeur.

Cela peut nous amener à modifier notre propre conduite quand, prise isolément, elle serait correcte par rapport aux repères donnés par la règle, mais que, prise dans l'interaction avec les interprétations de la règle par les autres, elle n'assure plus la convergence des actions. Nous devons alors nous ajuster pour compenser les divergences d'autres coopérateurs. Nous ne contrôlons plus seulement le fait que les comportements sont conformes à certains repères ou critères, nous contrôlons la tendance des coopérateurs à ajuster mutuellement leurs comportements de manière à maintenir

l'orientation de la coopération vers telle finalité. Nous ne contrôlons pas seulement la manière dont ces comportements se *conforment à* une règle, nous contrôlons, par ces repères tenant aux coordinations, leur tendance *à suivre collectivement* la règle.

Cette distinction revient, on le voit, à faire une différence entre un contrôle de premier ordre, qui porte sur l'atteinte de repères ou de buts, et un contrôle de second ordre. Celui-là porte sur la façon dont, pour assurer une orientation collective vers une fin ou dans le respect d'une valeur commune de coordination, chacun a contrôlé sa manière d'atteindre les repères et de satisfaire les critères ; et cela en fonction des manières des autres et en fonction de la satisfaction commune de la finalité souhaitée, ou du respect de la valeur affichée. Ce faisant, ces contrôles de second ordre ont assuré un enracinement mutuel et collectif de cette finalité ou de cette valeur. Ainsi, dans un débat juridique contradictoire, le fait qu'on puisse disputer des arguments montrant qu'on a satisfait ou violé telle valeur, montre bien qu'on débat alors du suivi des règles, même si on se donne comme garde-fous des indices de conformité aux repères légaux.

Ces trois premiers effets de l'explication des normes suffisent déjà à expliquer leur émergence. Pouvoir exercer sur les autres un contrôle indirect mais plus étendu, doublé d'un contrôle en continu de la conformité du comportement à des valeurs ainsi que d'un contrôle de leur intention de suivre une règle, donne à chacun un pouvoir et fournit une source de reconnaissance sociale, à laquelle les interactions avec autrui et les tiers qui ont constitué notre subjectivité nous ont rendu sensible. Un repère qui permet ces contrôles va donc être un attracteur pour les acteurs sociaux, ce qui va augmenter la puissance de cet attracteur en fonction du nombre mais aussi de l'investissement des acteurs qui s'y réfèrent.

Or nous pouvons aller encore plus loin, puisque nous pouvons vouloir faire porter notre contrôle sur la formulation même de la règle, puisqu'il s'agit d'une règle explicite. Cette fonction 4) *de contrôle des règles elles-mêmes* peut d'ailleurs se dédoubler. D'une part nous pouvons tenter de savoir si, en nous proposant tel et tel repère, la règle, quand on examine les conséquences de son application, assure bien la finalité qui la motive. D'autre part nous pouvons contrôler la cohérence de la finalité de la règle avec d'autres motivations et valeurs dont nous pourrions souhaiter qu'elles priment sur celles que la règle invoque directement. Dans le premier cas, nous adoptons un contrôle du même genre que celui qui porte sur le choix de conventions, voire de règles d'étiquette (par exemple : est-il plus courtois de poser les fourchettes les pointes en l'air, ce qui peut être un peu agressif, ou vers le bas ? Ou encore : les ronds-points régulent-ils de manière plus socialisée et plus fluide le problème des croisements que les feux rouges ?). Dans le second, nous discutons de la subordination d'une norme à d'autres normes au nom de la priorité à donner à telle exigence sur d'autres, et cela peut nous conduire dans le domaine des normes morales.

C

Les études de psychologie cognitive montrent que nos évaluations diffèrent – très tôt dans le développement – selon que nous avons affaire à des règles qui sont de convention sociale, ou à des règles morales. Nous reconnaissons assez vite que des conventions sociales, comme les manières de mettre le couvert ou les règles de politesse, peuvent être autres qu'elles ne sont. Nous avons tendance inversement à penser que les règles morales, comme ne pas tuer ou dire la vérité, peuvent être tenues pour universelles voire inconditionnelles. De fait, des enfants même peu âgés sont capables de faire la différence entre les règles d'étiquette et les normes que nous pensons être des normes morales.

Cette stabilité va cependant de pair avec de nettes évolutions au cours du développement. Une étude (Fehr, Bernhard et Rockenbach, 2008) montre que quand on confronte des enfants à des situations qui proposent des choix entre coopérer ou ne pas le faire, leur conduite évolue, en grandissant, d'une manière complexe : ils peuvent commencer par des comportements de don qui nous semblent exagérés, pour ensuite ne tenir compte que de leurs propres désirs d'appropriation, et finir par combiner, selon les situations, coopération ou poursuite de leur intérêt propre, d'une manière similaire aux adultes. Il semble donc qu'ici l'éducation sociale joue un grand rôle et que les conduites que nous considérons comme morales s'édifient sur une base d'interaction sociale[4].

Ces apprentissages sociaux vont pouvoir orienter les évaluations morales en fonction des contextes culturels, et on peut relier ces différences d'évaluation à des différences d'activations cérébrales. Ainsi Han *et al*. (2013) montrent que les Américains non seulement prennent plus de temps à répondre à des dilemmes moraux (mettant en jeu des personnes, comparés à des dilemmes qui ne mettent pas en jeu des personnes) alors que les Sud-Coréens répondent plus vite, mais aussi que les premiers activent davantage la zone du cortex cingulaire antérieur (ACC), liée à des résolutions de conflits, alors que les seconds activent davantage celle du cortex préfrontal dorsolatéral et le putamen (lié à des réactions rapides concernant la satisfaction). Il se pourrait que les Coréens visent non pas à résoudre un conflit, mais à trouver le contrôle qui les satisfasse socialement.

Si nous pouvons distinguer entre des règles de politesse et des règles morales (comme ne pas tuer ou ne pas mentir), nous pouvons aussi distinguer entre reconnaître la valeur morale d'un acte et choisir d'accomplir cet acte. En effet, nous pouvons reconnaître qu'un acte est bon ou mauvais, et cependant ne pas déclencher la conduite appropriée (Moll, 2005) ! FeldmanHall *et al.* (2012) ont montré qu'il y a une grande différence entre les anticipations des sujets sur leur conduite et leur comportement effectif dans une expérience (version atténuée de l'expérience de Milgram) où on leur donne le choix de soumettre d'autres sujets à un choc électrique plus ou moins élevé (tous les chocs ici sont supportables), ou bien de payer pour leur éviter le choc proposé. Les sujets pensent que les participants vont payer pour l'éviter aux autres, mais, quand ils agissent, ils choisissent plutôt de conserver l'argent pour eux[5]. Nous pouvons aussi faire la différence entre des normes morales qui régissent les relations entre proches, et celles qui régissent notre groupe social. Nous en arrivons même, lorsque nous sortons de notre groupe ou que nous sommes amenés à vivre des situations nouvelles ou plus complexes que prévu, à pouvoir critiquer les normes valant pour nos proches et celles valant pour notre groupe. On pourrait penser que c'est en étendant notre champ d'expérience à

d'autres groupes que nous sommes amenés à prétendre à l'universalité de certaines normes, mais il semble que ce soit plutôt en tentant d'étendre les normes de notre groupe à d'autres groupes ! Inversement, quand nous devenons plus sensibles à la complexité des situations, cela nous amène à porter un regard critique sur des normes qui seraient insensibles à la singularité de ces situations. Enfin, quand nous avons à choisir entre ces différents types de normativités dans une situation donnée, nous pouvons dans notre réflexion nous référer non plus à des normes directement applicables, mais à des orientations que nous pouvons choisir pour nous guider (mais non pas nous déterminer) et à des principes qui sont communs à ces orientations.

Ce que des auteurs comme Parfit ont appelé la morale de sens commun[6] combine la « morale des proches » (qui nous conseille de nous soucier surtout de nos proches) et une prise en compte partielle des limites de cette première morale quand nous étendons nos relations à des personnes extérieures à ce cercle. Cette morale de sens commun trouve à son tour ses limites, quand, par exemple, elle conduit à ne pas prêter attention à des petites défaillances dont l'accumulation sur des collectifs importants peut avoir des conséquences néfastes, attention à laquelle nous incite en revanche une morale utilitariste. L'éthique du *care* se souciant de ceux que nous rencontrons semble proposer une extension de la morale des proches à toute personne, mais elle exige de tenir compte chaque fois de la singularité des situations, si bien qu'elle semble devoir renoncer à nous fournir des règles explicites. Elle a surtout une fonction de critique d'autres morales comme la morale kantienne, qui prétend suivre des règles rigides parce que universelles, ou comme l'utilitarisme, qui prétend tout faire dépendre d'un calcul des conséquences collectives. Nous pourrions ainsi distinguer entre une morale des proches, une morale sociale ou morale du groupe, et diverses morales généralistes. Nous pourrions alors réserver le terme d'éthique pour une réflexion sur ces différentes morales, pour des critiques qui peuvent en être faites au nom d'une exigence d'universalité, au nom d'une prise en compte des conséquences plus lointaines, ou encore au nom de la particularité des situations, ou encore pour une discussion argumentée sur ce qui peut justifier l'appel à tel ou tel principe.

D

Les neurosciences et la neuroéconomie peuvent prétendre avoir aussi des choses à dire sur les activités cérébrales qui sous-tendent les jugements portant sur la manière dont une conduite peut ou non satisfaire soit des conventions, soit des règles morales. Moll *et al*. (2005) notent que les évaluations morales exigent, pour juger des actions, d'avoir des perspectives plus larges qu'une visée instrumentale et de prendre en compte des conséquences à long terme, si bien que ces évaluations vont activer le cortex préfrontal qui permet cette prise en compte. Ils proposent de comprendre notre sensibilité morale comme résultant de la combinaison de divers traitements (évaluations à long terme et évaluations émotionnelles, qui activent le système limbique et le cortex orbitofrontal pour des associations entre situations sociales et émotions). Certes, mais c'est là simplement transposer dans le

vocabulaire des neurosciences des associations de catégories que nous faisons spontanément à propos de la morale. Ces activations, on les retrouve d'ailleurs aussi dans des évaluations qui ne sont pas morales, mais esthétiques[7]. On pourrait ainsi objecter que les dispositifs expérimentaux actuels en neurosciences ne sont pas d'un grain assez fin pour tenir compte des distinctions que nous avons faites entre des contrôles de différents ordres. C'est vrai en général, mais les différences observées expérimentalement peuvent cependant nourrir la réflexion sur ces distinctions plus fines. Nous partirons de deux types d'études qui peuvent alimenter ce genre de discussion, celle de Grygolec, Coricelli et Rustichini (2007) sur le rapport entre responsabilité personnelle et observabilité sociale, et celles de Greene (2001, 2003, 2004) sur les rapports entre raisonnement sur les conséquences et évaluations liées aux émotions.

La première étude ne porte pas spécifiquement sur la responsabilité morale et met plutôt en jeu des questions de statut social. Mais elle peut nous servir de « contrôle » pour détecter par différence les évaluations et choix proprement moraux. Le dispositif expérimental montre que la différence entre ce qu'un sujet a obtenu effectivement et ce qu'il aurait pu obtenir a plus d'importance pour le sujet quand il est arrivé à ce résultat par une action qu'il a lui-même décidée que lorsque le résultat ne dépend pas de lui mais d'une loterie ou d'éléments extérieurs. Le fait que d'autres personnes puissent observer et donc évaluer ce résultat accroît aussi son importance. Ces différences d'importance sont corrélées aux différences d'activation du striatum ventral – qui est comme on l'a vu un élément central du circuit lié à la récompense – et du cortex orbitofrontal, lié aux comparaisons entre une donnée réelle et la représentation d'un état auquel on aurait pu parvenir mais qui ne s'est pas réalisé et est resté contrefactuel.

La conjugaison des deux activations, peut-on penser, montre que notre satisfaction est liée à cette comparaison entre ce qui est et ce qui aurait pu être. Cette sensibilité qui ne se borne pas à ce qui existe est évidemment un composant nécessaire de toute évaluation morale aussi bien que sociale. Il est clair d'ailleurs que les différentes catégories de normes morales que nous avons distinguées exigent des modalités de comparaison contrefactuelle de plus en plus complexes. Ainsi pour pouvoir proposer des principes éthiques, qui ne définissent pas des règles d'action, mais seulement des orientations qui peuvent nous guider dans nos évaluations morales, il faut avoir comparé ce qu'a donné l'application d'une norme morale dans telle situation à ce qu'y aurait donné (contrefactuellement) l'application d'autres normes, et avoir extrait de ces comparaisons des orientations qui seraient pertinentes pour la situation contrefactuelle comme pour la factuelle, et qui pourraient être évoquées avec profit pour des situations similaires.

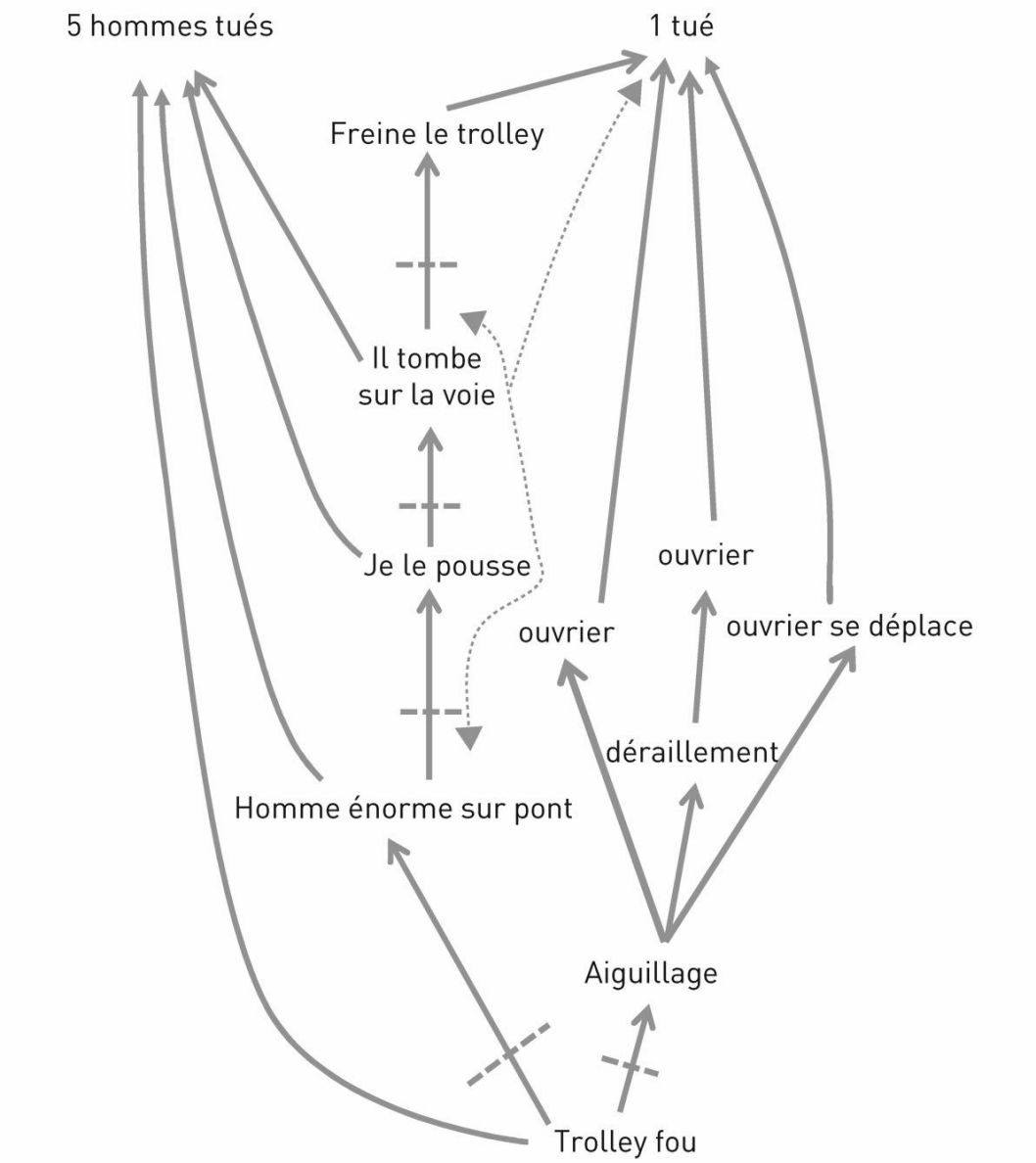
Cependant, si nous revenons à cette expérience, l'activation du striatum ventral et du cortex orbitofrontal ne peut y correspondre qu'à des évaluations faites au sein d'un circuit activé dans une situation donnée. Il serait alors nécessaire, pour aller jusqu'à l'éthique, de montrer des activités de comparaison qui impliquent de relier différents circuits activés pour des situations différentes.

Les expériences de Greene ne vont pas jusque-là, mais elles repèrent des différences entre circuits. Il a analysé en imagerie cérébrale les différences entre les activations provoquées par les présentations des deux scénarios les plus classiques de ce qu'on appelle le problème du trolley fou. Un

trolley devenu incontrôlable va écraser cinq ouvriers qui travaillent sur une ligne, à moins que vous aiguilliez le trolley sur une autre ligne où il ne va écraser qu'une seule personne, soit vous poussiez un énorme personnage du haut d'un pont de telle manière qu'en tombant sur la ligne il fasse dérailler le trolley. L'imagerie cérébrale montre dans ce second cas une activation supérieure de zones liées aux émotions – si bien que les circuits cérébraux peuvent différer dans les deux cas. Les sujets disent préférer manœuvrer l'aiguillage plutôt que pousser le gros homme, alors que du point de vue utilitariste, sensible aux seules conséquences[8], les deux choix devraient être indifférents puisqu'il y a dans les deux cas mort d'un homme. Selon Greene, pour que l'on arrive à la position utilitariste il faudrait que le cortex préfrontal (particulièrement dorsolatéral, DPFC) ainsi que le cortex cingulaire antérieur (ACC), associés pour le premier à une forme de calcul comparatif et pour le second à la difficulté de l'évaluation, l'emportent sur les zones émotionnelles qui déclenchent une réponse plus immédiate[9].

Cependant d'autres chercheurs en neurosciences (*cf.* Moll *et al.*, 2005) sont quelque peu sceptiques à l'égard de l'identification entre calcul utilitariste et activation du DPFC, comme à l'égard d'une opposition trop dualiste entre cognition calculatrice et émotions. Pour prendre un exemple, l'amygdale, censée traiter l'émotion de peur, est aussi active dans la détection d'éléments surprenants et informatifs, et a donc un rôle de motivation cognitive important.

Nous pouvons ajouter à cette perplexité un autre scepticisme, qui concerne cette fois le caractère principalement moral des motivations qui nous amènent à préférer manœuvrer un aiguillage plutôt que de pousser un gros bonhomme. Il est en effet possible qu'il s'agisse de considérations qui seraient à la fois pragmatiques, sociales et morales (mais pas seulement morales), et qui auraient trait au problème de l'assignation de la responsabilité d'une action. Une revue de quelque 17 variantes différentes du scénario du trolley fou disponibles dans la littérature, que nous avons esquissée à l'occasion d'une communication (Livet, 2011), semble bien indiquer que ce qui compte avant tout pour les sujets soumis à ces expériences, c'est de pouvoir insérer le plus d'étapes et donc de distance possible entre leur intervention directe, considérée comme le premier segment d'une suite d'étapes qui s'enchaînent mais qui peuvent aussi présenter des bifurcations, et les conséquences fatales qui en résultent pour les ouvriers et autres gros bonshommes (en y ajoutant dans quelques scénarios plusieurs trains sur différentes voies).



Dans ce graphe, une flèche barrée en tirets indique une succession qui n'intervient pas nécessairement, mais demande une intervention de l'agent et implique sa responsabilité. L'agent préfère entre deux scénarios conduisant au même nombre de morts, celui qui implique le moins d'interventions possible. Les cours-circuits de responsabilité sont en pointillé serré.

Pour ne prendre que les deux variantes que nous avons citées, le deuxième scénario nous assure qu'en poussant le gros homme, il sera nécessairement tué. Dès lors c'est notre poussée, combinée avec l'arrivée du trolley, qui cause sa mort. Circonstance aggravante, il a été nécessaire que nous soyons attentifs à assurer cette coïncidence entre sa chute et l'arrivée du trolley – il nous fallait déclencher notre poussée au bon moment et pousser dans la bonne direction. Si bien que la « réussite » de cet acte, et donc sa conséquence, la mort d'un homme, remonte aussi directement que possible à nos efforts et à notre intention. En revanche quand nous manœuvrons l'aiguillage, cette action déclenche un mécanisme, qui peut fonctionner plus ou moins bien, sans que nous puissions rien y faire – par exemple rien n'exclut que le trolley déraille au lieu d'aller sur l'une des voies qui conduisent à des

résultats mortels. Cela introduit une bifurcation possible des événements, et donc une étape supplémentaire, celle qui est nécessaire pour assurer le bon fonctionnement de la déviation sur l'autre ligne au lieu du déraillement. Enfin, il reste dans nos esprits une certaine incertitude sur le comportement des ouvriers ou de l'ouvrier, ce qui introduit éventuellement encore une autre bifurcation possible et une étape supplémentaire. Toutes ces possibilités, selon le scénario proposé, ne sont pas accordées au gros homme, que nous sommes censés contraindre à tomber selon la direction que nous lui avons donnée.

On peut alors raisonner ainsi : quand nous poussons le gros homme, notre responsabilité est engagée aussi dans le blocage des bifurcations possibles de la situation – les scénarios qui proposeraient d'autres déroulements que sa chute au point voulu. En revanche, une fois l'aiguillage déclenché, nous n'avons plus de responsabilité concernant les bifurcations possibles des événements. On peut représenter ces séquences d'étapes selon un graphe, dont le point de départ est notre poussée et le terme est la collision entre le gros bonhomme et le trolley, les bifurcations possibles étant alors représentées par des nœuds intermédiaires, d'où partent les embranchements des autres scénarios. Bloquer ces bifurcations pour assurer une chute bien ciblée peut alors être transcrit par un arc qui court-circuite les différents nœuds de bifurcation pour parvenir au but poursuivi. En effet si, une fois exercée une poussée sur le gros homme, nous pouvons réintervenir directement quand nous voyons un instant plus tard qu'il ne tombe pas ou que sa trajectoire risque de ne pas être la bonne, cela revient à dire que cette deuxième étape est toujours sous le contrôle direct de notre première action, donc du premier nœud du graphe. Différentes étapes successives peuvent être ainsi court-circuitées, ce qui redirige la responsabilité directement sur notre intervention. On comprend que lorsque la conséquence finale est désastreuse, nous ayons une certaine aversion pour une remontée de responsabilité aussi directe. Que le scénario qui conduit à la remontée la plus rapide de la responsabilité d'une conséquence négative vers notre décision d'action suscite en nous des émotions négatives n'a par ailleurs rien d'étonnant.

Or cette aversion n'a rien de particulièrement moral, sinon éviter ses responsabilités serait devenu une vertu morale. Elle a tout à voir, en revanche, avec les problèmes de contrôle dont nous avons fait l'hypothèse, en reprenant l'idée de J. Proust, qu'ils sont au cœur de la référence à des normes explicites. Certes, dans cet exemple, les normes suivies par les acteurs n'ont pas été explicitées, mais les évaluations que l'on demande aux sujets des expériences, elles, sont explicites. Elles visent bien à être communiquées, partagées, elles peuvent être critiquées par les autres, s'appliquer à d'autres conduites dans des situations similaires, etc. Les modalités de contrôle de notre action – et de celles des autres – sont donc au cœur de nos évaluations explicites, qui elles-mêmes nourrissent nos manières d'expliciter des normes.

Dans notre vie sociale, cette tendance à éviter de contrôler directement une action dont les conséquences négatives sont assurées nous évite beaucoup de problèmes. Ceux qui se hérissent à l'idée de pouvoir dissocier vivre en personne morale et assumer ses responsabilités la trouveront immorale. On pourrait aussi simplement la trouver amorale. Elle n'est cependant pas dépourvue de rapport avec le domaine de la moralité : c'est bien l'évaluation morale de notre action et de notre

personne, évaluation à laquelle conduirait une telle remontée de responsabilité, que nous souhaitons éviter. Mais ici, comme souvent, se mêlent désir de reconnaissance sociale, attentes normatives et appréciations morales.

Il ne serait pas impossible de mettre ce genre d'hypothèse à l'épreuve de l'imagerie cérébrale. Après tout, cela pourrait se rapprocher de résultats (Blair, 2000, 2004, suivant Rolls) qui montrent que, quand le cortex orbitofrontal est endommagé, on arrive moins bien à inverser une tendance initiale dont on aperçoit les conséquences négatives (*response-reversal*). On est alors aussi moins sensible aux réactions des autres, manifestées par des expressions de colère. On pourrait donc étudier comment une tendance, non plus cette fois à inverser la réponse, mais simplement à la rendre moins directe, peut être reliée ou non à une sensibilité à la désapprobation d'autrui. Il faudrait d'ailleurs distinguer ici la sensibilité émotionnelle et la capacité à comprendre les intentions d'autrui (que l'on dit lié à la théorie de l'esprit, TOM) puisque les psychopathes conservent la seconde alors que la première est chez eux quasiment absente (Richell *et al.*, 2003). Or pouvoir capter des indices de reconnaissance sociale exige à la fois la première et la seconde.

Ces expériences nous montrent bien que nos conduites morales mobilisent différents circuits d'évaluation. Nous faisons appel à un circuit émotionnel, à un circuit d'interprétation des intentions d'autrui. Nous les associons à des circuits qui activent des comparaisons entre ce qui se produit et ce qu'on pouvait attendre comme autre scénario, ou entre ce qu'on pouvait attendre comme émotion d'autrui et celles qu'il manifeste, et à des circuits qui intègrent ces différents modes de comparaison. Cette intégration est délicate, parce que ces circuits ne fonctionnent pas tous à la même vitesse. Les circuits émotionnels peuvent être plus rapides, les comparaisons qui impliquent l'activation du cortex orbitofrontal ou du cortex préfrontal dorsolatéral sont plus lentes et peuvent n'avoir d'effet inhibiteur qu'après quelque temps. Enfin nos évaluations continuent même après notre décision et notre acte, nourrissant alors des comparaisons avec d'autres actes que nous aurions pu accomplir, comparaisons qui semblent être essentielles pour un apprentissage moral et encore davantage pour des réflexions éthiques. Il est clair que notre réflexion éthique philosophique doit tenir compte des conditions de fonctionnement concrètes de ces processus mentaux. Cela conduit à relativiser les prétentions d'une morale qui ne proposerait que des règles universelles ou que des calculs conséquentialistes, et à devenir plus sensible à des considérations éthiques qui tiennent compte de la mixité des sources de nos décisions morales.

C

Les évaluations morales combinent donc des réactions immédiates – des réactions émotionnelles, mais aussi des orientations pragmatiques ou encore des évaluations non réfléchies – et des estimations plus réfléchies et plus élaborées – qui peuvent impliquer une prise de position critique par rapport aux normes sociales qui sont devenues routinières. Elles impliquent des capacités de comparaison entre ce

qui existe et des possibilités qui pourraient exister, mais qui ne sont pas réalisées (des contrefactuels). Or de récentes études qui portent sur différents types de croyances, dont les croyances religieuses, montrent que l'on peut non seulement comparer une situation réelle à une situation contrefactuelle, mais que l'on peut aussi, en particulier dans le domaine religieux, entretenir en parallèle ou de manière superposée des croyances qui supposent une sorte de coexistence du factuel et du contrefactuel. On peut alors se demander en quoi ces croyances religieuses diffèrent des croyances morales et en quoi elles leur sont similaires, et aussi dans quelle mesure d'autres croyances que les croyances religieuses peuvent présenter des dispositifs similaires. Cela n'implique pas que les religions soient réductibles à des croyances, puisque leurs liens avec les émotions, et les émotions collectives en particulier, sont bien connus. Mais si, dans le domaine des émotions, les émotions religieuses nous paraissent aller de soi, la cohérence des croyances religieuses et des croyances ordinaires n'est pas évidente.

Sur ce terrain, la combinaison entre anthropologie, psychologie cognitive et neurosciences permet, tout comme sur celui des évaluations morales, de montrer la complexité des phénomènes. Paul Harris et Rita Astuti ont montré, Harris chez les catholiques, puis Astuti et Harris chez les Vezo de Madagascar (2008) que coexistent deux croyances apparemment contradictoires sur la mort, la première reconnaissant que toutes les fonctions, corporelles et mentales, cessent avec la mort, la seconde, qui chez les Vezo réunit les morts aux ancêtres, envisageant leur présence possible dans la communauté après leur mort – alors que chez les catholiques, il s'agit d'une survie de l'âme, que le corps ne doit rejoindre qu'au jugement dernier. Cette croyance à un mode de relation des ancêtres avec la communauté après leur mort n'empêche pas qu'en préparant les morts pour l'inhumation, on n'hésite pas à utiliser de l'eau très chaude qui pourrait les ébouillanter, ou à leur tirer les cheveux si nécessaire, en donnant pour raison de ces traitements cavaliers qu'« ils ne sentent plus rien ». Ces croyances religieuses semblent donc jouer sur un registre qui se superpose à celui de la croyance « naturaliste » qui voit dans la mort une cessation d'existence physique et mentale. Ce registre est déclenché dans le contexte d'histoires communément acceptées et qui concernent les ancêtres, histoires qui ont été socialement apprises. On observe d'ailleurs que les jeunes enfants produisent plus spontanément la croyance qu'après la mort l'esprit et le corps cessent de fonctionner que la croyance liée à la présence des ancêtres, qui est en revanche entretenue par les adultes et les enfants plus âgés.

Certaines de ces croyances acquises par confiance dans les assertions des autorités sociales, des philosophes (à commencer par Jonathan Cohen, 1992) les ont appelées des « acceptations » (acceptances) – on peut renvoyer à *Va savoir* de P. Engel (2003) pour plus de précisions. Il s'agit là seulement de ces croyances que l'on peut acquérir volontairement, à la différence des croyances au sens épistémique du terme (au sens où l'on croit une proposition p si l'on pense que p est vrai, que p est le cas, sans pouvoir définitivement exclure que p soit faux). En ce dernier sens, si nous étions conscients que notre croyance en une proposition tient simplement à notre souhait volontaire qu'elle soit vraie, nous n'arriverions justement pas à la croire – au sens épistémique du terme. Or on constate que croyances religieuses et acceptations se mêlent assez intimement.

Cependant, ce n'est pas par un effet de volonté que l'on croit à la vie éternelle ou au retour des

ancêtres, mais d'abord parce qu'on croit à ce que d'autres croient. Pour autant il ne s'agit pas d'une croyance qu'on ait pu obtenir simplement à partir d'information sur des faits, sans intervention de souhaits et de volonté (James, on le sait, parle pour la croyance en Dieu de « *will to believe* », de vouloir croire), et c'est ce qui la rapproche d'une acceptation. Les croyances religieuses impliquent le plus souvent l'acceptation de l'autorité des membres influents d'une certaine communauté. Mais une fois inculquées, les croyances religieuses se présentent comme des croyances qui s'imposent à nous sans intervention de notre volonté. La trace de cette origine culturelle et sociale se manifeste seulement dans cette superposition ou coexistence entre les croyances religieuses et les conceptions basiques des régularités du monde qui nous entoure – et une de ces régularités naturelles nous montre que le corps des morts ne fonctionne pas.

Il nous semble que la notion de contrôle est plus appropriée ici que celle de volonté ou de désir. En effet elle permet de relier l'aspect social des religions et leur aspect cognitif. L'anthropologie cognitive (Sperber, Boyer, Atran[10]) propose de voir dans les religions un effet collatéral d'une combinaison entre deux capacités cognitives de base (que nous utilisons dans bien d'autres domaines que celui des croyances religieuses) : la capacité de pouvoir identifier un agent comme source d'un changement et celle de pouvoir se représenter des entités directement inaccessibles – comme les pensées des autres, des actions contrefactuelles, etc. Les religions nous proposent donc de trouver la source des changements qui nous affectent dans des entités personnifiées mais auxquelles nous n'avons pas d'accès direct, et dont nous pouvons imaginer les intentions[11]. Or, ajouterons-nous, ce qui permet à un agent d'être source d'un changement, c'est qu'il peut avoir par son action un certain contrôle sur les événements. De même, deviner les intentions des autres nous permet d'avoir un certain contrôle sur la manière dont ils orientent leurs actions.

Or pouvoir formuler explicitement certaines règles qui régissent les interactions des divinités entre elles et avec nous ouvre la possibilité d'un double contrôle. Par similitude avec notre propre usage des règles explicites, cela donne aux divinités ou puissances un moyen de contrôler la conformité de nos actions à leurs exigences (contrôle de premier ordre). Inversement, en contrôlant nous-même cette conformité de nos comportements, nous pouvons penser satisfaire ces exigences, et donc croire assurer une stabilité dans nos relations avec ces divinités. Ces puissances représentent aussi des finalités ou des valeurs, et nous pouvons donc être motivés par le souci de suivre les règles qu'elles nous donnent (contrôle de deuxième ordre). Les rites conjoignent les deux ordres de contrôle, puisque non seulement les comportements des pratiquants d'une religion doivent être conformes aux rites, mais qu'ils doivent être accomplis avec un engagement authentique.

Bien des religions mettent encore en jeu un autre dispositif de contrôle, celui exercé par d'éventuels intermédiaires entre les croyants et les divinités. Ces intermédiaires sont censés mieux identifier les intentions des divinités que la plupart des croyants, et ils vont aussi s'arroger le rôle de contrôler la conformité des conduites des autres aux rites, voire le suivi des règles religieuses. Ces religieux sont d'ailleurs supposés exercer sur leurs propres conduites un contrôle supérieur à celui des autres membres de la société – ce contrôle s'appliquant dans certaines sociétés jusqu'à des processus

corporels que l'on ne penserait guère soumis à la volonté – contrôle de processus gérés par le système vagal comme la respiration.

Par ailleurs les religions dont les dogmes centraux sont liés à des préceptes comme la règle d'or (« fais à autrui ce que tu voudrais qu'on te fît ») et qu'on peut qualifier de « religions morales » apparaissent quand les techniques de production permettent d'obtenir des surplus si on y consacre plus d'efforts, ce qui pose des problèmes de répartition, si bien que l'apparition de ces religions pourrait être liée à la recherche d'une règle proportionnelle de répartition, qui permet de maintenir des coopérations entre des individus qui se livrent à des activités différentes (Baumard et Boyer, 2013). Il s'agit bien là d'un mode de contrôle des coopérations. C'est un noyau qui demeure chez des athées : ainsi les principes laïques sont liés au souci de contrôler l'égalité des conditions entre les croyances.

Enfin les religions nous fournissent des moyens de contrôler notre tendance à surestimer les satisfactions dans un futur proche (le discount temporel), contrôle qui atteint son plus haut point chez les calvinistes – dont la doctrine rapporte toute action, à tout moment, à notre destinée ultime – alors que chez les catholiques, on peut toujours espérer un pardon *in extremis* (Paglieri *et al.*, 2013 ; Carter *et al.*, 2013).

Il se pourrait donc bien que les religions se soient construites comme des mises en œuvre sociales de tous ces niveaux de contrôle, dont les modes les plus complexes ne sont devenus disponibles que par la formulation de règles explicites[12].

Plusieurs expériences en neurosciences peuvent d'ailleurs corroborer cette approche socialisante des croyances religieuses. Ainsi Harris *et al.* (2009) ont comparé les activations cérébrales liées aux croyances religieuses et non religieuses, ainsi qu'aux jugements portés par des croyants et des non-croyants sur des propositions comportant des contenus religieux. Ils n'ont pas trouvé de différences – de zones activées préférentiellement – entre l'activité de croire et celle de ne pas accorder sa croyance (*disbelief*), et cela que l'on soit croyant ou non. En revanche, dans nos sociétés, où les croyants savent qu'il existe des non-croyants et réciproquement, les propositions religieuses sont plus longues à juger vraies ou fausses que les non-religieuses. Les propositions contraires à la religion (ici la religion chrétienne) activent, aussi bien chez les croyants qui les refusent que chez les non-croyants qui les jugent vraies, les zones de l'insula antérieure (liée à des réactions émotionnelles négatives) mais aussi la zone du striatum ventral, liée ordinairement à la récompense.

Ce dernier point, qui semble à première vue étrange dans le cas des croyants, peut s'expliquer de la même manière que dans notre réinterprétation de l'expérience de De Quervain *et al.*, où l'activité dans le striatum semblait liée à des effets de reconnaissance sociale. Dans une société où chacun sait qu'il y a des croyants et des non-croyants, un jugement sur des propositions religieuses vaut adhésion à l'un ou l'autre groupe, et donc peut éveiller un sentiment lié à la reconnaissance sociale interne à ce groupe dans les deux cas, aussi bien qu'un sentiment négatif à l'idée que l'autre groupe ne partage pas nos opinions.

Un travail de Kapogiannis *et al.* (2009) montre de même que les différentes régions activées par des propositions religieuses soit théoriques, soit pratiques, ou encore indiquant une implication de Dieu (de manière positive ou négative), ou enfin indiquant les émotions divines (soit positives, soit

négatives) activent les réseaux auxquels on pourrait s'attendre pour des propositions de chacun de ces types, mais qui ne porteraient ni sur la religion ni sur Dieu, mais sur des acteurs humains usuels.

Enfin deux expériences de Newberg (Newberg et d'Aquili, 2001 ; Newberg et Waldman, 2006), l'une portant sur les états de méditation de moines tibétains, l'autre sur les activités de prière de religieuses franciscaines, montrent que ces états impliquent tous deux des activations cérébrales spécifiques mais similaires, en particulier une baisse d'activité dans le lobe pariétal latéral supérieur postérieur. Or, alors que pour les moines tibétains cette baisse est liée au sentiment de n'être plus situé nulle part, pour les religieuses franciscaines elle pourrait être liée à celui d'être en interaction avec Dieu – qui il est vrai est ubiquiste, mais ce n'est pas cet aspect que les religieuses mettent en avant. Cela montre d'une part que les expériences de ce genre, qui sont liées à des institutions qu'on peut qualifier de religieuses, ne dépendent pas dans leur spécificité cérébrale de la religion considérée, et pourraient ne pas être spécifiquement « religieuses » (ce sont les Occidentaux qui voient dans le bouddhisme une religion), mais d'autre part que ce sont bien des éducations sociales différentes qui permettent de faire diverger les interprétations qui sont liées à ces activations particulières. Notre cerveau interprète ses propres activations selon nos acquis culturels !

Autrement dit il est possible de voir dans les interactions dites religieuses (avec Dieu, ou plongé dans une forme de vide qui est au-delà ou en deçà des particularités de ce monde) des manières d'exploiter autrement nos capacités sociales d'interaction, en les vivant sur un mode virtuel, qui a l'avantage de pouvoir être plus aisément contrôlé que ne le pourraient être des interactions sociales effectives. Les religions nous entraîneraient ainsi au contrôle de nos interactions sociales, dans la mesure où elles nous permettraient de nous y exercer. Mais elles pourraient aussi induire des conduites plus radicales, voire plus violentes, quand des croyants tenteraient de rendre les comportements sociaux effectifs conformes à des conduites dont le contrôle n'a été efficient au départ que dans un monde seulement virtuel.

N

Sur la base de cette approche, nous pouvons esquisser une distinction entre différents types de normes, à commencer par les normes religieuses et les normes scientifiques. Les premières gèrent ce problème du contrôle des conduites en utilisant comme repères des rites, reliés à l'apprentissage d'une hiérarchie stabilisée de valeurs. Les secondes proposent aussi des repères, mais donnés par des méthodes qui puissent, tout en maintenant une forme de cumulativité des savoirs, assurer aussi des changements de priorités conceptuelles – Kuhn a nommé cela un changement de paradigme, ce qui est sans doute excessif.

Les normes politiques, notons-le, ont un statut quelque peu intermédiaire. Elles régulent les prises de décisions, décisions qui apportent des changements dans les interactions sociales. Mais elles n'impliquent pas nécessairement de changement conceptuel, ou si elles en produisent – par exemple

lors d'une révolution – elles ne se soucient guère de cumulativité.

Les normes juridiques proposent un régime mixte. D'une part les normes légales sont supposées intangibles et donc cumulatives en un sens additif, avec pour idéal qu'elles forment un système cohérent (dans la perspective de Kelsen) ou du moins, dans la perspective de la *Common law,* qu'elles puissent être inférées de principes communs (le juge « herculéen » de Dworkin est censé extraire ces principes des jugements précédents). Lors des procès, en revanche, elles fonctionnent en conjonction avec une intense activité de réinterprétation de la part des parties, qui chacune rapportent les faits à des normes différentes, et de la part des juges, qui interprètent la jurisprudence. On notera d'ailleurs que quand les juristes (plus audacieux que les juges, qui craindraient une ingérence extérieure) s'intéressent aux neurosciences, c'est pour tenter d'y trouver des indicateurs de psychopathologies supposés plus fiables que les évaluations des experts psychiatres, et pour pouvoir plus facilement décider de considérer un criminel comme irresponsable (Vincent, 2011 ; Nadelhoffer *et al.*, 2012). Autrement dit, il s'agit pour eux de pouvoir opposer aux émotions des plaignants une garantie moins contestable.

D'autre part, et inversement, l'activité du législateur consiste à partir de normes sociales émergentes pour les réguler par des lois explicites, quitte à réviser les normes juridiques précédentes. Nous retrouvons donc dans le fonctionnement des normes juridiques tous les types de contrôles que permet l'explicitation des normes : contrôle de la stabilité des formulations, contrôle de la coordination correcte de la coopération entre plaignant, défenseur, accusateur et juge, contrôle du respect des finalités et des valeurs, et contrôle des règles elles-mêmes (par le législateur).

La comparaison de la croyance religieuse avec la croyance que le profane peut avoir dans les théories scientifiques est ici intéressante. Le profane croit les savants ; sous cet aspect, il croit, comme le croyant religieux, à ce que d'autres croient. L'étude de cette relation des connaissances scientifiques avec certains modes d'autorité fait d'ailleurs partie du programme de ce qu'on appelle l'épistémologie sociale (*cf.* Goldman, 1987, 2006). Une différence entre scientifique et religieux peut se faire jour. Les croyances du savant restent toujours opaques pour le profane qui ne peut intégrer toute leur complexité. Au contraire, le croyant religieux, dès qu'il a terminé son initiation, partage en gros les mêmes croyances que celles de celui qui les lui a inculquées (cela du moins quand la religion n'est pas doublée de « sciences du religieux », recourant à l'expertise de théologiens, exégètes et spécialistes des religions). La différence essentielle, cependant, est que celui qui doute de ses croyances religieuses risque d'être exclu de la communauté, alors qu'il est admis par la communauté des savants qu'ils puissent en venir à mettre en question certaines conclusions. Ce n'est toutefois jugé admissible – et avec résistance – que si de tels hétérodoxes ne doutent pas de tout et si leurs travaux permettent d'avoir accès à de nouvelles branches du savoir. Quand ces travaux amènent des révisions des connaissances, certaines croyances scientifiques peuvent se révéler fausses – mais pas toutes, et l'ensemble constitué par les croyances restantes et par les nouvelles croyances doit permettre d'explorer des domaines nouveaux. C'est là une forme de cumulativité qui n'est pas simplement additive, qui ne se borne pas à ajouter de nouvelles propositions sans rien retrancher, et qui peut même amener à modifier le statut des croyances persistantes, lesquelles peuvent devenir moins centrales,

voire valides seulement dans un domaine restreint. Non seulement on révise les croyances, mais on peut aussi réviser leur statut de croyances prioritaires et fondamentales[13]. Certes, là encore, les dispositifs de contrôle de ces révisions par différentes autorités de l'institution scientifique sont nombreux. Mais ces autorités doivent elles-mêmes satisfaire une certaine contrainte : leur contrôle doit pouvoir donner lieu lui-même à contrôle, ce qui exige qu'elles donnent des raisons explicites de leurs décisions (même si elles ne donnent pas forcément toutes leurs raisons, dont certaines peuvent être peu scientifiques).

L'épistémologie sociale semble nous mener loin de l'économie[14]. Pourtant dans ses développements ce courant de recherches s'est fortement inspiré non seulement de la psychologie cognitive, mais aussi de l'économie. List et Pettit ont ainsi repris dans ce domaine la stratégie construite pour produire les théorèmes d'économie qui démontrent, comme le font le théorème d'Arrow ou celui de Sen, l'impossibilité de définir une fonction de choix collectif qui satisfasse certaines exigences pourtant souhaitables. On pourrait peut-être étendre aussi ces théorèmes au domaine religieux : il semble impossible de trouver pour une collectivité une fonction de choix d'une religion unique, religion qui permette cependant à chacun de se forger sa propre croyance, qui s'appuie sur des comparaisons entre religions faites deux à deux, qui respecte la préférence unanime de tous pour une doctrine religieuse plutôt qu'une autre, mais qui évite que tous se soumettent à la religion d'un seul.

On pourrait avec quelque malice rappeler que l'économie combine des traits des connaissances scientifiques et des traits des religions. D'une part les découvertes en économie théorique sont bien soumises à la critique scientifique et restent par ailleurs opaques pour les profanes ; et les études d'économétrie et celles d'économie expérimentale ou de neuroéconomie présentent aussi quelques-uns de ces traits. Mais d'autre part, même si les économistes théoriciens préféreraient moins de distorsions entre leurs résultats théoriques et les phénomènes économiques, ils sont capables d'y rester peu sensibles. Dès lors la dualité entre les situations idéalisées étudiées dans la théorie et les situations économiques de fait peut s'installer assez durablement, sur un mode qui présente quelques similarités avec la coexistence ou superposition entre les croyances de base et les croyances religieuses, que nous avons évoquée à propos du sort des morts chez les Vezos.

Quelques études d'imagerie cérébrale ont suggéré que lors d'activités mentales liées à des évocations religieuses est activée une combinaison de zones (le cortex préfrontal dorsal, le cortex frontal dorsomédian et le cortex pariétal médian) que d'autres expériences ont trouvée associée à une activité d'évaluation réflexive, dans un retour du sujet sur lui-même (Azari *et al.*, 2001). On note aussi une réduction de l'activité du cortex cingulaire antérieur (ACC), qui peut être liée à une diminution de l'anxiété (Inzlicht *et al.*, 2009). Cependant, ajouterons-nous, le caractère réflexif de ces activités n'est pas forcément conscient chez les sujets, leur expérience étant plutôt celle d'une relation immédiate à ce qu'ils croient – même si la recherche de la paix de l'âme semble un leitmotiv de bien des religions.

Les croyances religieuses apparaissent donc impliquer une dualité ou une superposition entre inculcation sociale et croyances « naturalistes » de base, comme entre réflexivité et adhésion

immédiate. Mais cette dualité est la plupart du temps mise entre parenthèses, sinon effacée : la source externe de l'inculcation est pleinement intériorisée, la réflexivité et l'introspection sont vécues comme sentiment ou intuition immédiate. À l'inverse, les croyances scientifiques remettent sans cesse en tension la dualité entre d'une part les théories admises ou supposées établies, et d'autre part la réflexion critique personnelle ou les discussions entre pairs. La réflexivité est présente dans les deux régimes de croyance, mais dans les croyances religieuses elle est reprise dans un mode implicite, alors que dans les croyances scientifiques elle doit donner lieu à explicitation.

Or la réflexivité est liée au contrôle. Dans la religion et dans la morale, elle donne lieu à un autocontrôle, qui amène la personne responsable à s'infliger à elle-même un blâme quand il y a eu défaillance, contrôle qui lui rappelle les valeurs qu'il est désirable de respecter. Dans le travail scientifique, on tente d'éliminer les sources de blâme, par exemple les fautes techniques. En revanche on aurait plutôt tendance à tenter, dans les protocoles expérimentaux, de renforcer les sources de résistance de la nature aux hypothèses du chercheur – ces hypothèses qui sont ce à quoi il accorde une valeur toute particulière. La religion met le croyant à l'épreuve des valeurs, l'expérimentation scientifique met les hypothèses valorisées par le chercheur à l'épreuve.

Le contrôle se décale. Dans la religion, il porte d'une part sur le respect des conduites rituelles, d'autre part sur les motivations du croyant à suivre les pratiques religieuses et sur la conformité de ses formulations de ces motivations avec les formulations de sa religion. Dans les deux cas de figure, nous retrouvons nos trois premières fonctions de contrôle (contrôle de la communication, contrôle des coopérations de second ordre, contrôle du respect de la finalité des règles).

Dans la recherche scientifique, le contrôle porte sur les hypothèses, qui sont des formulations des attentes du chercheur contrôlables par sa communauté scientifique. Or on peut aussi voir les hypothèses comme des règles explicites : elles formulent en effet le lien conditionnel entre les conditions de départ de l'expérience et les résultats attendus. Comme toutes les règles explicites, elles établissent ainsi des rapports directs entre des points de contrôle qui peuvent être lointains et demander pour être atteints bien des médiations. Mais si les hypothèses sont des règles explicites, alors le contrôle de ces hypothèses relève bien aussi de notre quatrième fonction de contrôle : on contrôle les règles elles-mêmes – du moins certaines d'entre elles. Alors que le troisième type de contrôle pouvait amener le croyant religieux, dans un retour sur lui-même, à s'interroger sur la conformité de ses motivations personnelles avec les finalités ou valeurs de sa communauté, le quatrième type de contrôle renvoie le contrôle des hypothèses scientifiques à l'extérieur, à la résistance de la nature, et à cette série de points d'ancrage que sont les expériences ultérieures et les remises en cause par d'autres théories et modèles. Cette position peut sembler donner l'avantage aux « réfutationnistes » à la Popper sur des partisans de justifications positives. Mais pour une règle passer avec succès une mise à l'épreuve est le meilleur ersatz que nous puissions avoir d'une justification.

On comprend alors que la recherche expérimentale fasse partie de ces pratiques institutionnelles qui font porter le contrôle jusque sur les règles explicites elles-mêmes. Font aussi partie de ces pratiques les discussions politiques et éthiques, qui se donnent, elles aussi, les moyens de remettre en cause certaines règles explicites tout en continuant à s'appuyer sur d'autres normes, jugées plus

fondamentales.

Cette remarque nous permet finalement de trouver un trait commun aux différents types de normes explicites : elles permettent de faire fonctionner en parallèle deux régimes dont chacun répond aux défaillances possibles de l'autre. L'explicitation des normes religieuses fixe des repères communautaires et rituels externes, alors même que les croyances religieuses font fonctionner des formes de réflexion et d'introspection sans les afficher comme telles, en effaçant leur aspect de découplage. L'explicitation des normes juridiques donne un cadre aux échanges sociaux et aux procès, alors même que ces échanges font émerger de nouvelles normes et que les procès déploient la pluralité des interprétations. L'explicitation des normes politiques donne un cadre aux interactions sociales tout en permettant une critique des règles proposées. L'explicitation des normes scientifiques donne un cadre doctrinal aux chercheurs, tout en permettant aussi la critique des résultats proclamés et des programmes de recherche. L'explicitation des normes morales permet à l'éthique de mieux définir le cadre des exigences morales tout en développant la discussion sur le contenu à donner aux valeurs.

Commentaire Émergence des normes et interprétations des règles

La notion de norme n'est pas aisée à saisir et, moins encore à définir. Ce terme renvoie d'abord à l'idée de normalité. Le normal, en médecine, s'oppose au pathologique. Lorsqu'il s'agit de produits et de technologies de fabrication, la normalisation vise une standardisation. En droit, la norme, selon la jolie formule de Catherine Thibierge sert à la fois « de tracé et de mesure » (Thibierge, 2008). En matière de morale, cette référence à la normalité n'est plus suffisante, puisque les normes s'appuient, cette fois, sur des principes qui les fondent. Par-delà cette polysémie, un trait leur est commun : l'existence de normes se trouve étroitement liée à celle de règles. Mais la relation entre normes et règles est, elle-même, ambivalente puisqu'elle peut s'interpréter dans les deux sens. D'un côté, comme c'est le cas pour la morale, on trouve des normes, en amont des règles, dont elles constituent les fondements abstraits ; d'un autre côté, les règles permettent aux acteurs sociaux de transformer les principes qui les inspirent en normes comportementales concrètes. Cette relation complexe entre les règles et les normes sert de trame au chapitre 7[15].

Pierre Livet prend toutefois le soin de préciser, dès l'introduction, qu'il ne traite, dans ce chapitre, que des normes qui sont reliées à des règles « explicites ». Par règles explicites, il faut entendre ici des règles dont l'information est publique, et par conséquent, en théorie au moins, accessibles à tous. L'énoncé des priorités, en matière de code de la route, ou l'interdiction de fumer dans certains emplacements, fournit deux exemples de règles explicites. Les normes associées aux règles explicites sont supposées, dans les deux cas, pouvoir être déduites de ces règles, dont les individus ont une connaissance directe ; d'où l'importance de leur formulation et de leurs interprétations. Cette connaissance des règles n'est cependant pas suffisante pour garantir leur transformation en normes comportementales par tous les individus. Il faut encore, en effet, que chacun en accepte les principes, et peut-être même, sache (ou seulement croie) que les autres les acceptent aussi. Ainsi, dans l'exemple de l'interdiction de fumer, il n'est pas certain que le fumeur invétéré accepte le principe d'hygiène qui motive cette règle et qu'il ne lui oppose pas un principe libertarien

de choix individuel, en contradiction ici avec le principe dans lequel cette règle puise sa légitimité. Une autre conséquence de cette limitation aux règles explicites est l'exclusion des règles implicites de cette investigation. Elle entraîne, *de facto*, l'élimination des simples conventions qui ont été traitées dans le chapitre 6. Les développements du chapitre 7 confirment notre hypothèse que la formulation de ces règles explicites, et leurs incidences sur les comportements, soulèvent des questions distinctes de celles qui ont été étudiées à propos des conventions tacites.

L

On comprend mieux, dans ces conditions, que rechercher l'émergence de normes morales, à partir de la sélection d'un équilibre dans un jeu qui en possède plusieurs, comme l'ont proposé de le faire, selon des modalités différentes, Binmore, Bicchieri, Gintis et bien d'autres, n'ait guère fait avancer les questions qui sont étudiées dans ce chapitre. Cette formulation de la question des normes en termes de jeu de coordination entraîne que la norme recherchée revêt presque toujours, pour ces auteurs, la forme technique d'un équilibre de Nash, au sens où un tel équilibre correspond aux meilleures réponses stratégiques de chaque joueur aux stratégies de meilleures réponses de chacun des autres joueurs. Or nous savons que certains équilibres de Nash peuvent contredire des normes sociales souvent considérées pourtant comme évidentes. Pour rester dans une tradition utilitariste chère à beaucoup d'économistes, nombre d'équilibres de Nash ne sont pas optimaux, parce qu'ils ne maximisent pas l'utilité collective des joueurs. Pis encore, dans le célèbre dilemme du prisonnier, la solution coopérative, qui pourrait coïncider avec une norme sociale, n'est précisément pas un équilibre au sens de Nash.

Certes, Bicchieri (2006), à la différence des deux autres auteurs mentionnés, souligne que la question des normes ne se présente pas en théorie des jeux, du fait de l'existence d'un problème de coordination, mais du fait de situations caractérisées par des conflits d'intérêts qu'elle nomme « jeux d'intérêts mixtes ». Mais elle reconnaît que de telles situations entraînent nécessairement un problème de coordination. Même si, par conséquent pour elle, la coordination n'est pas à l'origine de l'émergence de normes sociales (ou morales), c'est bien cette contrainte de coordination qui permet *in fine* de résoudre ce (ou ces) problème(s) posé(s) par les normes[16].

On peut noter également une petite différence entre Gintis (2010) et les deux autres auteurs qui ont été mentionnés. Pour Gintis, ce n'est pas la sélection d'un équilibre de Nash qui constitue le cœur du raisonnement, mais l'établissement d'un équilibre défini de manière légèrement différente, sous l'appellation d'équilibre corrélé. Dans un équilibre corrélé, les joueurs ne disposent pas, au moment de leur décision, d'une information complète sur la stratégie qui sera choisie par les autres joueurs, mais seulement d'une distribution de probabilités sur leurs stratégies (Aumann, 1987). L'idée d'appliquer cet équilibre corrélé au problème posé par la sélection d'un équilibre avait déjà été évoquée par Carlson et Van Damme (1993). La novation de Gintis est ici d'interpréter la norme

sociale comme un signal qui favorise, dans cette perspective, la coordination de leurs stratégies dans une direction coopérative. Pour préciser sa pensée, Gintis, compare ce signal à celui du chorégraphe lorsqu'il s'agit de coordonner les mouvements d'une troupe de danseurs (Gintis, 2010).

Mais ni un équilibre de Nash ni l'équilibre corrélé, qui en représente une forme affaiblie, ne sont nécessairement optimaux. Quant au système de sélection par lequel la norme sociale se trouve ici introduite, soit il fait intervenir des fictions (le voile de l'ignorance emprunté à Rawls par Binmore, la métaphore du chorégraphe imaginée par Gintis), soit il n'est pas suffisant, en lui-même, pour garantir toujours la coordination, car la transformation des normes en solution d'un problème de coordination imaginée par Bicchieri n'est valable que pour la catégorie des jeux d'intérêts mixtes. Or tous les jeux où la référence à une norme sociale pourrait résoudre le problème posé par la coordination des joueurs ne sont pas des jeux d'intérêt mixtes au sens de Bicchieri. On assiste enfin, chez tous ces auteurs au glissement, jamais explicité, d'une norme sociale (maintien et survie du groupe), à une norme morale qui reste non définie.

Cette approche par la coordination apparaît, pour toutes ces raisons, mal adaptée à l'énoncé des conditions d'émergence de normes morales au cours des interactions sociales. Le jeu de la chasse au cerf, qui est l'un des archétypes des jeux de coordination, fournit, sous cet angle, une manière de contre-exemple. Son interprétation évolutionniste, qui a pourtant la faveur de Binmore, de Gintis et, à moindre degré, de Bicchieri, a justement montré, modélisation à l'appui, qu'en la circonstance, ce serait plutôt l'équilibre sous-optimal (chacun chasse un lièvre de son côté) qui serait plus probablement sélectionné (Kandori, Mailath, Rob, 1993 ; Samuelson, 1997). L'explication intuitive en est facile. Supposons une population suffisamment nombreuse, partagée entre des coopérateurs, prêts à chasser le cerf, et des chasseurs averses au risque, qui, dans une région supposée giboyeuse, préfèrent capturer tout seul un lièvre. Au bout d'un temps suffisant, un nombre croissant de coopérateurs, déçus par la défection de leurs partenaires, grossiront progressivement les rangs des chasseurs de lièvre, au point d'éliminer les partisans de la chasse au cerf. Le mécanisme évolutionniste aura ainsi sélectionné l'équilibre résultant d'un comportement du « chacun pour soi ». Faut-il considérer, dès lors, le « chacun pour soi » comme une norme sociale ? On aurait en tout cas quelques raisons de douter de son efficacité sociale, puisque les chasseurs obtiendraient tous davantage en chassant, ensemble, le cerf. Quant à sa portée morale, on imagine, plus difficilement encore, quel principe moral pourrait le justifier.

Au total, nous savions déjà que certaines normes de comportement social pouvaient faciliter, voire assurer la coordination des joueurs. La théorie des jeux telle qu'elle est utilisée par ces trois théoriciens, nous apprend donc, en définitive, peu de chose sur la relation entre des normes sociales ainsi entendues et des normes morales, et moins encore sur leur transformation en règles explicites.

Il serait pour autant excessif et même inexact d'induire de cet exemple de la chasse au cerf, et d'autres du même genre, que la théorie des jeux est un cadre inadéquat pour étudier les relations entre normes morales et règles explicites. Ce qu'il met en évidence, avec d'autres exemples du même type, c'est seulement le fait que la question des normes, dans leurs liens à des règles explicites, n'est pas réductible aux problèmes de coordination rencontrés par les agents individuels dans leurs interactions

sociales. Quant au concept d'équilibre de Nash, qui sert, directement ou indirectement, de pivot au raisonnement de ces auteurs, il ne représente ni une règle explicite des jeux sociaux, ni une norme morale qui s'imposerait aux joueurs. C'est donc dans une direction différente de la théorie des jeux qu'il faut porter nos regards, en repartant du concept de standards de comportements « acceptés », introduit, dès l'origine, par von Neumann et Morgenstern, et en convoquant d'autres concepts de solution inspirés, le plus souvent, de jeux dits coopératifs, comme l'équilibre de Berge évoqué au chapitre 6. Nous reviendrons sur cette question en abordant la question des règles de partage qui se trouve au centre de beaucoup des jeux sur lesquels ont porté les expérimentations et les travaux de neurosciences. Mais il nous faut auparavant discuter brièvement la manière dont la science économique présente la question des relations entre les règles explicites et les normes sociales et morales, avant d'examiner comment elle a cherché à la résoudre.

L

La science économique s'est développée en distinguant assez nettement deux branches séparées de la connaissance économique, celle de l'économie positive, qui se propose d'expliquer les phénomènes économiques observés, et celle de l'économie normative qui construit des modèles cohérents de comportements et d'institutions économiques, sur la base de principes moraux abstraits, comme la justice, l'équité, le bien-être, voire, plus récemment, le bonheur… Dans une visée plus ambitieuse, l'économie normative est allée, parfois, jusqu'à inverser le sens de cette démarche, en proposant, par exemple, de dériver directement des théories économiques de la justice, de l'équité, ou du bien-être, à partir de concepts appartenant à l'analyse économique (les choix rationnels, la maximisation de l'utilité, l'optimalité…). On retrouve en tout cas, dans cette dichotomie entre l'économie positive et l'économie normative, un peu des deux versants de la relation règles/normes précédemment présentée. L'économie positive, tout au moins dans sa dimension théorique, s'efforce de dégager des lois de fonctionnement des phénomènes observés (ou observables). Elle cherche ensuite, à tirer de ces lois certaines règles de comportements des agents, ayant vocation à se transformer, en normes comportementales. L'économie normative part, au contraire, de normes sociales ou morales dont elle cherche à donner un modèle à l'aide de concepts économiques. Elle s'efforce ensuite, d'en tirer des règles explicites, capables d'induire, chez les agents, des comportements conformes à ces normes. L'une et l'autre de ces deux branches du savoir économique entretiennent une relation, certes différente, avec la réalité économique. Mais la confrontation des résultats ainsi obtenus dans chacune de ces deux branches de la connaissance économique n'est pas évidente. Rien ne garantit, en effet, la coïncidence, et même la comparabilité, des normes comportementales mises en évidence au terme de chacune de ces deux approches.

Quelques économistes se sont risqués à promouvoir une construction intellectuelle, voire à développer une véritable doctrine, qui rendrait intelligible le passage d'une approche à l'autre, et

éclairerait, de cette manière, les relations entre règles et normes en économie. Nous mentionnerons ici deux de ces tentatives les plus connues et, à bien des égards, opposées : l'utilitarisme de règles (*rule utilitarism*), élaboré par Harsanyi (Harsanyi, 1977, 1982), et l'ordre spontané (*spontaneous order*) proposé par Hayek (Hayek, 1975).

La doctrine d'Harsanyi repose sur un utilitarisme économique, au sens où l'objectif poursuivi par l'activité économique dans son ensemble serait une maximisation de l'utilité collective, et où sa réalisation résulterait des choix des individus. Cette version de l'utilitarisme de règles, prônée par Harsanyi, se distingue cependant de l'utilitarisme de choix. Le choix utilitariste des agents ne porte plus, dans le premier cas, sur des actions individuelles, mais sur des règles, même si l'objectif utilitariste est toujours celui de la maximisation de l'utilité collective. Il en résulte, notamment, que ce seront les règles ainsi choisies qui permettront de transformer en normes de comportements les normes morales initialement contenues dans l'utilitarisme. Harsanyi y voit ainsi une facilitation dans la coordination des actions. Mais cela n'est pas l'objet premier de sa construction. Il emprunte à l'appareil analytique de l'économie positive la théorie du choix rationnel dérivé de la maximisation de l'utilité, et fonde les principes moraux qui inspirent les normes utilitaristes sur la rationalité de ces choix. L'exercice est habile, sans être tout à fait convaincant, car il se heurte à une objection majeure. On connaît les difficultés rencontrées par les économistes pour définir le contenu précis de cette utilité collective. L'adhésion des agents, même rationnels, aux principes moraux de l'utilitarisme n'est donc nullement assurée. Les chances de transformer ces principes en normes comportementales, par l'intermédiaire de règles qui seraient acceptées par tous (ou presque tous), s'en trouvent d'autant amoindries. Or est-il encore rationnel, dans une optique utilitariste, de choisir des règles utilitaristes, si l'on doute, avec quelques arguments, que de telles règles soient également choisies par les autres ?

La doctrine de l'ordre spontané d'Hayek part d'une hypothèse inverse de celle d'Harsanyi. L'économie positive traite les phénomènes économiques et sociaux comme des processus organisés. Mais leur ordre, qui émane des actions des individus, n'est pas, pour Hayek, celui des buts qu'ils poursuivent chacun, bien au contraire. Hayek élimine ainsi l'idée d'Harsanyi d'ancrer la relation entre les normes et les règles sur une quelconque théorie des choix rationnels. C'est en ce sens qu'il faut entendre l'idée de spontanéité présente dans la notion d'« ordre spontané ». Pour autant, cet ordre s'appuie sur des règles qui permettent d'éviter, ou d'évacuer, les désordres susceptibles d'intervenir au cours de ces processus d'organisation. Le rôle de ces règles est, en définitive, de fournir ici aux agents des normes comportementales pour orienter leurs actions. Ces règles favorisent leur coordination, mais il s'agit, là encore, d'une simple implication. Cette construction peut paraître séduisante. Mais le statut exact de ces règles qui rendent possible le développement d'un ordre spontané reste problématique. Elles sont, à la fois indépendantes de toute intentionnalité, et identiques pour tous ; mais elles servent, en même temps, de référentiels pour les actions décidées par chacun. Les premières de ces propriétés laisseraient à penser qu'il s'agit de règles abstraites, et donc explicites. Hayek précise cependant qu'elles ne sont pas nécessairement « verbalisées », ou même seulement connues, au sens d'une connaissance consciente, par les acteurs qui s'y réfèrent dans leurs actions. Il ne pourrait s'agir, dès lors, que de règles implicites, dont Hayek reconnaît lui-même les

difficultés rencontrées lorsqu'on cherche à expliquer comment elles sont transformées en règles explicites (Hayek, 1980).

L'utilitarisme de règles et l'ordre spontané ne représentent, en définitive, que deux fictions philosophico-économiques destinées à fournir un pont entre l'économie positive et l'économie normative, en intégrant normes et règles dans une formalisation positive des activités économiques. Mais elles ont le mérite de mettre chacune l'accent sur une dimension différente, souvent omise ou négligée, de la relation entre règles et normes. L'utilitarisme de règles montre qu'il ne suffit pas, pour donner vie à des normes, de conformer ses choix à des règles de conduite correspondant à ces normes. Ainsi, ce n'est pas parce que chacun fera ses choix par rapport au critère utilitariste que la norme utilitariste sera instituée. L'ordre spontané révèle que la simple connaissance de règles, au sens d'une conscience réfléchie, n'est ni suffisante ni même peut-être nécessaire pour faire émerger des normes de comportement qui contribuent à un ordre social. La simple formalisation de cet ordre au moyen de règles formelles ne garantit pas, en effet, le respect de sa traduction en termes de normes de comportement. Ces deux constats peuvent paraître, à première vue, contradictoires. Nous montrerons, à la lumière des résultats de plusieurs travaux en neurosciences, qu'ils se révèlent, par certains aspects, complémentaires. Ils conduisent, en tout cas, à orienter nos recherches en séparant deux questions distinctes, même si elles ne sont pas sans liens. La première porte sur les conditions d'émergence de règles sociales incorporant des normes morales ; la seconde concerne la manière par laquelle l'existence de ces règles agit sur les comportements des individus, à travers leur appréhension et les modalités de leur connaissance.

N

La première de ces questions a déjà été discutée au chapitre 6, dans le cadre de jeux non coopératifs dont les études ont bénéficié des acquis récents de l'économie expérimentale et de la neuroéconomie. Les différents protocoles expérimentaux qui ont été examinés concernent le partage d'un bien divisible, imposé (jeu du dictateur), ou seulement proposé (jeu de l'ultimatum), par un joueur à un autre joueur. Différentes variantes en ont été développées avec plusieurs joueurs, dans le rôle du joueur en second, ou même dans les deux rôles. Des versions répétées avec ou sans bornes temporelles ont également été introduites.

Bicchieri fait observer avec raison que ces jeux sont différents pour les joueurs. Dans le jeu du dictateur, le partage est imposé par le premier joueur au second, pour qui l'alternative est seulement entre « prendre » ou « laisser ». Ce n'est plus le cas du jeu de l'ultimatum, lorsqu'il est répété. Le partage devient alors le résultat d'un jeu stratégique entre les deux joueurs dont l'issue dépend de leurs anticipations respectives et des niveaux de ces anticipations. Considérés toutefois du point de vue du théoricien, ces jeux posent tous un même problème de partage : il s'agit, en effet pour les joueurs de partager un bien (généralement une somme d'argent) entre eux, au terme des différents

protocoles envisagés. Or le principal résultat mis en évidence par ces expériences est, pour le théoricien des jeux, que le partage retenu par les sujets de ces expériences ne coïncide presque jamais avec la solution théorique tirée d'un équilibre de Nash, même si, dans le cas d'un jeu de l'ultimatum répété, les résultats obtenus peuvent être justifiés après coup dans une logique nashienne. Du point de vue de la théorie positive, par conséquent, l'équilibre de Nash n'est pas le concept de solution pertinent pour comprendre ces jeux de partage.

Face au défi posé par ces résultats expérimentaux, les réactions des théoriciens des jeux se sont partagées en deux courants distincts.

Certains, comme Binmore et Shaked, ont préféré garder le concept d'équilibre de Nash comme solution de référence. Ils ont alors invoqué l'existence de « normes sociales comportementales » pour expliquer les écarts observés entre les partages retenus et le strict équilibre de Nash défini par la théorie. Pour ces auteurs, la solution de ces jeux est le plus souvent compatible avec un ensemble de points d'équilibre. Il existerait un ensemble de normes sociales à la disposition des joueurs qui leur permettrait de les orienter dans le choix d'un de ces équilibres en fonction des environnements sociaux propres à chacun de ces jeux. On comprend mieux dans ces conditions comment ils ont pu ramener, par ce biais, la question du partage à un problème de coordination sociale (Binmore, Shaked et Sutton, 1985). C'est sur cette base que Binmore a développé ultérieurement sa théorie évolutionniste des normes sociales, et en particulier celle de la justice.

D'autres, comme Fehr et Schmidt ont, au contraire, élaboré, sur la base des résultats expérimentaux obtenus dans ces différents jeux de partage, un modèle alternatif pour en rendre compte. Partant d'une interprétation des rejets des propositions manifestés par le joueur en second et craints par le joueur en premier, ils ont formulé l'hypothèse générale d'une aversion à l'iniquité qui caractériserait les préférences des joueurs. Forts des résultats empiriques de leur modèle appliqué à d'autres jeux, comme certains jeux de marché, ils lui ont conféré un statut de théorie, sans pour autant remettre en question son concept de solution (Fehr et Schmidt, 1999).

Des débats méthodologiques violents s'en sont ensuite suivis entre les deux camps, qui n'ont guère permis, jusqu'à présent encore, de clarifier, dans le cadre de la théorie des jeux, la question de l'émergence de normes de partage[17].

Il existe pourtant, en théorie des jeux, d'autres concepts de solution que l'équilibre de Nash, parfois définis en termes axiomatiques, qui s'attachent à déterminer des règles de partage équitable. Mais de tels concepts ont souvent été élaborés dans le cadre de jeux coopératifs, sur la base précisément de normes sociales, si ce n'est morales. Par jeux coopératifs, il faut entendre ici la recherche par les joueurs de gains communs obtenus à travers ce que la théorie des jeux appelle des coalitions, qui désignent les ententes susceptibles d'être passées entre les joueurs. Une fois déterminé le gain collectif qui résulterait de l'entente de tous les joueurs, formant ce que l'on nomme la « grande coalition », le problème posé est de partager ce gain entre les joueurs. C'est là qu'intervient l'introduction d'une norme dans la définition des concepts de solution.

Au premier rang des normes de partage équitable conçues par les théoriciens des jeux figure la valeur de Shapley. Cette solution propose une formule mathématique de division/partage, fondée sur

une évaluation économiquement équitable des contributions respectives de chaque joueur au gain collectif de cette grande coalition (Shapley, 1953). La valeur de Shapley a été appliquée à la résolution d'un très grand nombre de problèmes concrets, concernant en particulier le partage d'un bien public ; depuis le partage des coûts d'un aéroport jusqu'à l'allocation de l'usage des eaux entre pays riverains. La valeur de Shapley a le mérite de relier étroitement la solution du jeu à une norme explicite de partage. Elle ne permet pas, en revanche, d'expliquer selon quel mécanisme (*mechanism design*), elle peut être mise en œuvre par les joueurs, ce que l'on nomme la question de son *implementation*. Cette difficulté explique en partie pourquoi une majorité des théoriciens des jeux s'en est aujourd'hui un peu détournée pour résoudre notre problème de norme de partage.

Différents modèles de négociation ont néanmoins été proposés pour rendre compte de cette *implementation* de la valeur de Shapley. L'un des plus intéressants, pour notre sujet, a été mis en évidence dans le cadre d'un jeu de l'ultimatum répété entre une population de joueurs. L'offreur qui joue en premier y est tiré au sort, il fait une proposition de partage aux autres joueurs et il suffit qu'un seul des autres joueurs refuse sa proposition pour que cette proposition soit abandonnée ; on tire au hasard si son auteur doit être éliminé. C'est alors à un second joueur, également tiré au sort, de formuler une autre proposition, soumise, à son tour, dans les mêmes conditions, au verdict des autres joueurs, et ainsi de suite… Au terme de cette procédure itérative, il a été démontré que le jeu tend à converger sur une allocation équitable et stable, qui correspond exactement à la valeur de Shapley (Hart et Mas-Collel, 1992, 2001 ; Moulin, 1995). À la norme d'équité contenue dans la valeur de Shapley peuvent donc être associées des règles cohérentes d'un partage négocié.

Ce modèle théorique n'a fait l'objet d'aucun test expérimental. Les seules données empiriques dont nous disposons pour confronter son résultat à celui des jeux de partage sur lesquels ont porté les expériences sont de nature statistique. Les relevés statistiques effectués par Camerer à partir d'un très large échantillon d'expériences de jeux de partage englobent la plupart des aires culturelles du monde. Selon ses calculs, la médiane et le mode des offres acceptées varient entre 40 et 50 % du montant total, la moyenne étant un peu inférieure, entre 30 et 40 % (Camerer, 2003). Une évaluation des bornes d'acceptabilité permet de préciser encore les limites de cette solution. La norme de partage communément acceptée par les joueurs serait ainsi assez voisine de celle déduite de la valeur de Shapley.

Il faut comprendre maintenant comment cette norme peut s'imposer aux joueurs au cours du déroulement de ces jeux de partage et comment les règles comportementales auxquelles elle peut être associée sont susceptibles de contribuer et de soutenir sa mise en œuvre. Pour y parvenir nous ne sommes plus contraints de formuler une hypothèse *a priori* sur les préférences des joueurs (aversion à l'iniquité). Le partage équitable version Shapley est ici un résultat final accepté par les joueurs. Il peut donc se révéler éloigné de l'idée différente que chaque joueur peut initialement se faire de cette équité. Il n'est pas davantage nécessaire d'invoquer, chaque fois, des normes *ad hoc* pour le justifier.

Plusieurs résultats obtenus dans des travaux récents de neurosciences fournissent déjà quelques informations sur les mécanismes mentaux qui régulent ces règles de partage. Ils nous mettent sur la

voie de la réponse à notre question. Chaque joueur estimerait l'action, ou la réaction de l'autre, par rapport à une anticipation qu'il effectue, en référence à sa propre norme d'équité (Chang et Sanfey, 2011). Ces normes personnelles d'équité, dont la satisfaction et la violation sont génératrices d'émotions, sont, elles-mêmes, modifiées par les informations recueillies dans le cours du jeu. Certaines d'entre elles résultent des propositions ou de la conduite de l'autre joueur avec lequel le joueur concerné se trouve en interaction. La proposition émanant du joueur en premier est plus généreuse (ou moins généreuse), par rapport à la norme attendue du joueur en second. Le joueur en premier, sur la base de l'anticipation de sa norme, ne s'attendait pas à un rejet de sa proposition par le joueur en second, ou, au contraire, est surpris par son acceptation. D'autres informations sont extérieures au jeu et proviennent de l'environnement, comme, par exemple, des photos du joueur avec lequel on interagit.

Sur cette trame très générale, un système de régulation des émotions, guidant les joueurs vers une norme socialement acceptée par l'ensemble, a commencé à être mis en évidence. Une équipe de chercheurs s'est d'abord efforcé d'isoler le rôle de ces émotions en demandant aux joueurs du jeu de l'ultimatum de réexaminer leurs réactions, a) en réactivant leur émotion potentielle, et b) en supprimant autant qu'ils le pouvaient l'émotion qui leur était associée. Ces recherches préparatoires ont ensuite permis de dégager les lignes directrices d'un mécanisme progressif de régulation des émotions par ré-estimation successives des décisions. En remontant et en abaissant ainsi l'estimation initiale des joueurs, qui dépend, comme on l'a vu, de leurs normes personnelles, ce système modulerait leurs comportements autour d'une norme d'équité socialement acceptée par le groupe. Les substrats neuronaux de ce système ont été identifiés. Ils activent et désactivent, en particulier, plusieurs régions du gyrus frontal et du cortex préfrontal médian, ainsi que certaines parties de l'insula. Son fonctionnement révèle ainsi que cette modulation fait intervenir une coordination entre des fonctions proprement émotionnelles (insula) et des fonctions cognitives du cerveau (gyrus frontal median) (Greccuci *et al.*, 2013). Il reste encore à modéliser la dynamique de ce système.

Cette approche ressuscite, comme nous l'avons montré, une intuition féconde des jeux coopératifs, selon laquelle on partage dans l'interaction ce que cette interaction permet de produire et qui, pour cette raison est, d'une certaine manière, nécessairement commun aux joueurs. Mais elle traite cette évidence comme une réalité dynamique dont elle entreprend de dégager les ressorts cérébraux.

R

L'autre question posée par les règles est de nature différente. On suppose, cette fois, l'existence de règles explicites, c'est-à-dire formelles et publiques, comme le sont, par exemple, les règles de droit. Mais il peut également s'agir de règles morales, d'origine laïque ou religieuse, voire de simples instructions codifiées. Confrontés à ces règles, les sujets doivent répondre par des comportements

adaptés aux situations concrètes particulières dans lesquelles ils se trouvent placés. Cela implique, de leur part, une intelligence de ces règles et une aptitude à les traduire par des comportements correspondant à la visée de ces règles, ce qui peut s'avérer délicat lorsque ces règles se révèlent compatibles avec l'adoption de plusieurs comportements différents, voire opposés.

Le problème du trolley fou, qui a été discuté dans ce chapitre, en fournit une illustration extrême. Certes, l'impératif moral de sauver la vie de cinq ouvriers au prix d'une autre vie humaine ne prend pas tout à fait la forme d'une règle explicite. L'évidence de la norme sociale et morale qui l'inspire permet néanmoins de l'assimiler à une règle commune. Or, comme l'a rappelé Pierre Livet, cette règle peut être satisfaite au moyen d'un nombre très élevé de scénarios qui correspondent à autant de comportements possibles de la part des intervenants. D'où l'apparition de nouveaux dilemmes à résoudre par le conducteur, sous contrainte de temps. Ce genre de situations a donné lieu à une abondante littérature consacrée précisément aux dilemmes moraux.

Signalons, en outre, que les règles abstraites formulées de manière explicite doivent encore, pour être traduites en comportements conformes aux normes qui les inspirent, être admises par les individus auxquelles elles s'adressent. Il est clair que certains sujets peuvent, pour une série de raisons, les rejeter, plus ou moins explicitement, et même les violer, en adoptant des comportements antisociaux. L'examen de telles situations non seulement relève du champ étudié, mais les informations recueillies à leur sujet enrichissent la compréhension, notamment en terme neuronal, de notre appréhension des règles explicites par les sujets.

On peut considérer, en première approximation, cette transcription pragmatique des propriétés associées aux règles explicites comme une opération de métacognition. Pour éviter le problème logique que soulève son interprétation en forme de métarègles, Pierre Livet préfère introduire ici une notion de « contrôle », qui n'exige pas des sujets de disposer d'une représentation abstraite de second degré par rapport à la règle. Je préfère, pour ma part, parler ici de « mode d'emploi » de la règle explicite. Par mode d'emploi il faut entendre ici une forme de connaissance pratique, certes associée à la règle, mais qu'il revient au sujet en situation de dégager de la règle en fonction des données dont il dispose sur l'environnement où se présente son application. C'est cette connaissance « mode d'emploi » qui guide le « contrôle ». Elle doit, pour y parvenir, mobiliser différentes méthodes. Certaines d'entre elles consistent simplement dans la recherche d'une correspondance termes à termes entre les composantes de la règle et les données fournies par la perception de l'environnement. D'autres introduisent des modalités plus élaborées de classement des données de l'environnement, en vue de faciliter leur confrontation à la règle. Leur nombre n'est pas déterminé et l'on peut en construire de nouvelles, soit par combinaison, soit par hiérarchisation des procédés utilisés. Le recours à ces méthodes peut, du reste, révéler, à cette occasion, des faiblesses, voire des lacunes, de la règle elle-même ; d'où la possibilité, dans certains cas, de retour sur la règle.

L'appréhension, le traitement et la mémorisation de ces « modes d'emploi », des règles et les méthodes mises en œuvre par le cerveau font intervenir différentes régions cérébrales reliées entre elles selon des modalités complexes. Les neurosciences sociales commencent seulement à les étudier. Les quelques travaux qui leur sont consacrés fournissent de premières informations à leur sujet. Ces

études ne révèlent rien de très nouveau quant aux zones activées (parties latérales du cortex frontal et du cortex préfrontal, nucleus caudatus, striatum…). Elles nous informent, en revanche, sur plusieurs particularités de la dynamique de leur coordination. Il semblerait ainsi que l'on puisse distinguer, d'une part, une activation simultanée des parties latérales des cortex frontaux et préfrontaux et, d'autre part, l'activation du noyau caudé ; chacun des deux mécanismes pouvant être rapproché de méthodes différentes, dans le mode d'emploi utilisé pour mettre en œuvre la règle (Ruge *et al*., 2010). En outre, une dynamique d'activation réciproque du cortex frontal et du striatum pourrait être à l'origine des différents niveaux rapidement explorés par le cerveau en quête d'applications de la règle. Cette aptitude cérébrale serait peut-être à rapprocher de l'architecture rostro-caudale du cortex frontal des humains (Badre *et al*., 2010). Il reste encore à explorer les voies par lesquelles ce système du cortex frontal élaboré hiérarchiquement communique avec le système émotionnel de ganglions basal au cours des séquences de renforcement qui concourent à l'appropriation de la règle.

D'autres informations, plus indirectes, fournies par des simulations réalisées sur des cerveaux artificiels (« réseaux neuronaux ») mettent, de leur côté, en évidence le fonctionnement de procédés dynamiques permettant de passer d'une règle à une autre (Maniadakis *et al*., 2009)[18].

Une question différente relative à la transformation de règles abstraites en instructions comportementales concrètes concerne le rôle exact joué par les mécanismes d'essais et d'erreurs dans cette transmission. Une majorité d'études, d'inspiration évolutionniste, a eu tendance à éclipser le travail accompli en amont par le cerveau, avant que ne s'enclenchent ces processus d'essais et d'erreurs. Plusieurs expériences ont été effectuées portant sur la traduction comportementale de règles abstraites présentées aux sujets sous la forme de nouvelles instructions, sans phase préalable d'essais et d'erreurs. Elles ont mis en évidence l'activation d'un système dynamique complexe, allant des zones antérieures aux zones postérieures du cortex prémoteur, dans un laps de temps très bref (Ruge et Wolfensteller, 2010). La présentation d'une règle nouvelle enclencherait donc un travail cognitif antérieur à toute mise à l'épreuve dans l'environnement considéré. À côté du mécanisme classique par essais et erreurs, l'apprentissage de nouvelles règles pourrait également mobiliser une mémoire procédurale stockée dans le cortex latéral préfrontal qui fonctionnerait indépendamment des informations fournies par l'environnement (Ruge *et al*., 2013).

On objectera peut-être que les expériences et les simulations qui ont été rapportées ici concernent l'appréhension de règles, certes explicites, mais sans portées sociales ou morales particulières. Comme toujours avec les neurosciences, les travaux portent d'abord sur les opérations les plus élémentaires de la catégorie étudiée. Ces résultats, s'ils sont confirmés, peuvent ensuite être prolongés et étendus à des opérations plus compliquées qui intéressent davantage les économistes et les chercheurs en sciences sociales. Mais ce qui a été mis en évidence à l'occasion de l'appréhension de ces « règles/instructions » concernant leur traduction en comportements suggère une hypothèse qui intéresse également les règles dérivées d'une norme morale. Les modalités du travail cérébral mises en évidence dans ces expériences devraient se trouver facilitées par la référence de ces règles à une norme morale, ou tout au moins sociale. En renvoyant explicitement à une norme, de nature

nécessairement abstraite, comme la justice ou l'équité, la portée de la règle correspondante gagne en niveau d'abstraction, ce qui accroît, selon notre interprétation, l'étendue de sa métacognition. De plus, le caractère explicite et public de cette norme morale intégrée dans la règle tend, chez les sujets, à leur faire inclure les autres et leur connaissance de cette norme dans cette hiérarchie par niveaux, selon laquelle s'organise prioritairement ici leur propre connaissance de la règle. Si l'on adopte l'interprétation de la métacognition développée par Joëlle Proust, il n'est pas nécessaire que les sujets disposent d'une représentation logique de ces différents niveaux pour que la norme morale, présente dans la règle, élargisse son épaisseur cognitive. Tout cela reste évidemment à étayer sur la base de nouvelles expériences et de nouvelles investigations du fonctionnement du cerveau.

Ce chapitre réserve un traitement particulier à la notion de contrôle, en relation avec la dimension sociale que revêtent les règles explicites. Si, comme cela a été développé, les règles explicites favorisent différents modes de contrôle, l'acceptation de ces règles et leur mise en œuvre par les sujets requierent également, de leur part, une forme d'autocontrôle. Nous avons montré la complexité des tâches mentales qui permettent les transformations de ces règles abstraites en comportements concrets. Leur organisation et l'ordre de leur déroulement sont, en effet, soumis à des protocoles très précis. Plusieurs expériences ont montré que le rejet de ces règles, pouvant aller jusqu'à la violation des normes qui les inspirent, est à rapprocher de différents déficits dans l'autocontrôle de ces opérations observés chez leurs auteurs (Gaillot *et al.*, 2012). Ces déficits peuvent, dans certains cas, prendre des formes pathologiques. Mais le principal apport de ces travaux, dans la perspective de notre enquête sur les interactions, est de montrer que le succès des formes de contrôle social exercé sur les individus par les règles explicites dépend aussi du contrôle que les individus sont capables d'exercer sur eux-mêmes. En contrôlant l'usage des règles, le sujet se les approprie, mais cette appropriation exerce, à son tour, un contrôle sur lui-même.

---

1. Qui est pour Jeannerod un critère central d'une action (*cf.* Frith, 2013).

2. *Cf.* Chang (2013) et Irwin (2013).

3. Sperber part cependant de présupposés différents de ceux de Proust, puisqu'il donne priorité aux représentations, et conçoit la métacognition en termes de métareprésentations (Sperber, 1996).

4. Par ailleurs, les localisations des activités cérébrales déclenchées par des situations de souffrance et des jugements condamnant une mauvaise action se modifient avec l'âge (Decety, 2013).

5. C'est surtout quand on leur fait subir un petit choc d'essai avant de leur demander leurs anticipations – et qu'ils voient que ce choc est faible – qu'ils anticipent leur comportement effectif. Ce qui confirme notre conviction de la prédominance des effets différentiels.

6. Parfit (1986), *Reasons and Persons*, Oxford, Oxford University Press.

7. Avram (2013). Cependant les zones liées à la représentation des intentions d'autrui (*Theory of Mind*) sont davantage activées dans les jugements moraux.

8. S'il ne s'agit pas d'un utilitarisme des actes.

9. Christensen et Gomila (2012) font une revue de ces dilemmes moraux et des travaux d'imagerie cérébrale en déplorant que leurs différents protocoles les rendent difficiles à comparer.

10. Par exemple dans S. Atran, Ara Norenzayan (2004), « Religions' evolutionary landscape : Counterintuition, commitment, compassion, communion », *Behavioral and Brain Sciences*, 27 ; ou encore Willard, Norenzayan (2013).

11. Norenzayan (2012) voit une source de passage à l'incroyance dans la difficulté à imaginer les capacités des dieux, combinée à une culture qui passe au crible ces capacités, comme dans un mode d'existence moins soumis à l'insécurité (donc mieux contrôlé).

12. Ce qui ne veut pas dire transmises à tous, puisque le cœur des mystères religieux est souvent d'un accès restreint aux seuls initiés.

13. Ainsi, la géométrie euclidienne devient un cas particulier parmi d'autres géométries.

14. La neuroéconomie a l'avantage de s'interroger sur le lien entre nos évaluations sociales ordinaires et nos évaluations en termes économiques. Ainsi Lebreton *et al.* (2009) se sont demandé quels étaient les liens entre évaluations subjectives (les préférences des sujets, en termes économiques) et leurs évaluations marchandes. Leurs résultats plaident pour un système cérébral d'évaluation commun, dans lequel les évaluations subjectives ou préférences interviendraient en premier.

15. Pour une vue d'ensemble des questions posées par les normes, voir *Sciences et sociétés. Les normes en question*, Paris, IHEST/Actes Sud, « Questions vivres », 2014.

16. Bicchieri a raison de préciser que ce n'est pas la coordination qui est à l'origine de l'émergence de ces normes sociales. Cependant de nombreuses situations de jeux, relevant à première vue de sa catégorie des intérêts mixtes, ne sont pas solubles par la seule transformation de normes sociales en règles de coordination. On songe ici aux jeux politiques, pour lesquels des théorèmes économiques comme ceux dits d'impossibilité constituent une impasse pour l'application de la solution de Bicchieri. Si, en effet, ces normes sociales préexistent à leur usage à des fins de coordination, comme le soutient Bicchieri, encore faut-il que de telles normes puissent exister au sein de populations d'individus culturellement disparates. On retrouve ici un problème évoqué dans ce chapitre par Pierre Livet, à propos des normes issues de croyances religieuses. La signification exacte donnée par Bicchieri à ces jeux d'intérêts mixtes mériterait du reste d'être précisée. Reflètent-ils des dilemmes liés au collectif (chaque joueur a des intérêts différents, et même des valeurs différentes des autres), des dilemmes intrapersonnels (il existe en chaque joueur des intérêts différents et des valeurs différentes), ou des deux à la fois ?

17. De 2005 à 2010, les défenseurs de ces deux approches des normes de partage ont multiplié leurs critiques réciproques dans un débat méthodologique peu intelligible. L'essentiel de leurs arguments est, en effet, tiré de leurs deux interprétations divergentes, mais également possibles, des résultats expérimentaux du jeu de l'ultimatum (Fehr et Schmidt, 2005 ; Binmore et Shaked, 2010 ; Fehr et Schmidt, 2010).

18. Pour y parvenir ces chercheurs de la dynamique de la cognition ont utilisé un robot configuré à partir d'un modèle de réseaux neuronaux de type CTRNN. Les résultats qu'ils ont ainsi obtenus valident l'intérêt d'appréhender par une dynamique des systèmes les mécanismes de métacognition qui interviennent dans les opérations de changements de règles. Leur application au fonctionnement des cerveaux humains dépend toutefois du degré de réalisme de l'architecture des réseaux neuronaux artificiels utilisés. On renverra sur ce point à la discussion méthodologique développée à la fin du chapitre 4.

Bibliographie

AIMAR T. (2009), *The Economics of Ignorance and Coordination*, Cheltenham (UK), Edward Elgar Publishing.

AINSLIE G. (1975), « Specious reward : Behavioral theory of impulsiveness and impulse control », *Psychological Bulletin*, 82 (4), p. 463-496.

AINSLIE G. (1991), « Derivation of rational economic behavior from hyperbolic discount curves », *American Economic Review*, 81 (2), p. 334-340.

AINSLIE G., MONTEROSSO J. (2003), « Building blocks of self-control : Increased tolerance for delay with bundled rewards », *Journal of the Experimental Analysis of Behavior*, 79 (1), p. 37-48.

AINSLIE G. (2012), « Pure hyperbolic discount curves predict "eyes open" self-control », *Theory Dec*., 73, p. 3-34.

ALBRECHT K. *et al*. (2011), « What is for me is not for you : Brain correlates of intertemporal choice for self and other », *Soc. Cogn. Affect. Neurosci.*, 6 (2), p. 218-225.

ALLAIS M. (1953) « Le comportement de l'homme rationnel devant le risque. Critique des postulats et axiomes de l'École américaine », *Econometrica*, 21 (4), p. 503-546.

ANDRAS P. (2011), « Networks of artificial social interactions », *in* G. Kampis G., I. Karsai, Szathmáry E. (éds.), *ECAL 2009, Part II*, LNCS 5778, Berlin, Springer Verlag p. 383-390.

ANGELETOS G. M., LAIBSON D., REPETTO A. (2001), « The hyperbolic consumption model : Calibration, simulation, and empirical evaluation », *The Journal of Economics Perspectives*, 15 (3), p. 47-68.

ARROW K. J., *Social Choice and Andivudual Values*, New York, J. Wiley.

ASTUTI R., HARRIS P. L. (2008), « Understanding mortality and the life of the ancestors in rural Madagascar », *Cognitive Science*, 32, p. 713-740.

ATRAN S., NORENZAYAN A. (2004), « Religion's evolutionary landscape : Counterintuition, commitment, compassion, communion », *Behavioral and Brain Sciences*, 27.

AUMANN R. J. (1987), « Correlated equilibrium as an expression of Bayesian rationality », *Econometrica*, 55 (1), p. 1-18.

AUMANN R. J. (1990), « Nash equilibria are not self-enforcing », *in Economic Decision Making : Games, Econometrics and Optimization*, Amsterdam, Elsevier.

AUMANN R. J. , BRANDENBERGER A. (1995), « Epistemic conditions for Nash equilibrium », *Econometrica*, 63 (5), p. 1161-1180.

AUMANN R. J. (1999), « Interactive epistemology », *International journal of Game Theory*, 28, p. 301-314.

AVRAM M. *et al*. (2013), « Neurofunctional correlates of esthetic and moral judgments », *Neuroscience Letters*, 534, p. 128-132.

AXELROD R. (1997), *The Complexity of Competition*, Princeton (NJ), Princeton University Press.

AZARI N. P. *et al*. (2002), « Neural correlates of religious experience », *European Journal of Neuroscience*, 13, p. 1649-1652.

BACHARACH M. (1993), « Variable universe Games », *in Frontiers of Game Theory*, Cambridge (MA), The MIT Press.

BACHARACH M., BERNASCONI M. (1997), « The variable frame theory of focal points : An experimental study », *Games and Economic Behavior*, 19 (1), p. 1-45.

BACHARACH M. (1999), « Interactive team reasoning : A contribution to the theory of cooperation », *Research in Economics*, 53 (2), p. 117-147.

BACHARACH M. (2006), *Beyond Individual Choice. Teams and Frames in Game Theory*, sous la direction de N. Gold, R. Sugden (éds), Princeton (NJ), Princeton University Press.

BADRE D., KAYSER A. S., D'ESPOSITO M. (2010), « Frontal cortex and the discovery of abstract action rules », *Neuron*, 66 (2), p. 315-326.

BAILLARGEON R. , SCOTT R. M. , HE Z. (2010), « False belief understanding in infants », *Trends in Cognitive Sciences*, 14, p. 110-118.

BALARD K., KNUTSON B. (2009), « Dissociable representation of future reward magnitude and delay during temporal discounting », *NeuroImage*, 45 (1), p. 143-150.

BARDSLEY N., METHA J., STARMER C., SUGDEN R. (2008), « Explaining focal points : Cognitive hierarchy theory versus team reasoning », *CeDex Discussion paper n° 2008-17*, Londres, The University of Nottingham.

BARMETTLER F., FEHR E., ZEHNDER C. (2012), « Big experimenter is watching you ! Anonymity and prosocial behavior in the laboratory », *Games and Economic Behavior*, 75 (1), p. 17-34.

BAULT N., JOFFILY M., RUSTICHINI A., CORICELLI G. (2011), « Medial prefrontal cortex and striatum mediate the influence of social comparison on the decision process », *PNAS*, 20 septembre, 108 (38).

BAUMARD N., ANDRÉ J.-B., SPERBER D. (2013), « A mutualistic approach to morality : The evolution of fairness by partner choice », *Behavioral and Brain Sciences*, 36, p. 59-122.

BAUMARD N., BOYER P. (2013), « Explaining moral religions », *Trends in Cognitive Sciences*, 17 (6), p. 272-280.

BAUMGARTNER T., HEINRICHS M., VONLANTHEN A., FISCHBACHER U., FEHR E. (2008), « Oxytocin shapes the neural circuitry of trust and trust adaptation in humans », *Neuron*, 22 mai, 58, p. 639-650.

BAUMGARTNER T., FISCHBACHER U., FEIERABEND A., LUTZ K., FEHR E. (2009), « The neural circuitry of a broken promise », *Neuron*, 10 décembre, 64, p. 756-770.

BAUMGARTNER T . *et al*. (2011), « Dorsolateral and ventromedial prefrontal cortex orchestrate normative choice », *Nature Neuroscience*, 14, p. 1468-1474.

BAUMGARTNER T., GIANOTTI L. R. R., KNOCH D. (2013), « Who is honest and why : Baseline activation in anterior insula predicts inter-individual differences in deceptive behavior », *Biological Psychology*, 94, p. 192-197.

BEAUREGARD M. , PAQUETTE V. (2008), « EEG activity in Carmelite nuns during a mystical experience », *Neuroscience Letters*, 444, p. 1-4.

BERECZKEI T. , DEAK A. , PAPP P. , PERLAKI G. , ORSI G. (2013), « Neural correlates of Machiavellian strategies in a social dilemma task », *Brain and Cognition*, 82, p. 108-116.

BERG J. , DICKHAUT J. , MCCABE K. (1995), « Trust, reciprocity, and social history », *Games and Economic Behavior*, 10, p. 122-142.

BERNHEIM B. D. (1986), « Axiomatic characterization of rational choice in strategic environments », *Scandinavian Journal of Economics*, 88 (3), p. 473-488.

BERNHEIM B. D. (1994), « A theory of conformity », *Journal of Political Economy*, 102 (5), p. 841-877.

BERNS G. S. , LAIBSON D. , LOEWENSTEIN G. (2007), « Intertemporal choice – toward an integrative framework », *Trends in Cognitive Sciences*, 11 (11), p. 482-488.

BERTHOZ A. (1997), *Le Sens du mouvement*, Paris, Odile Jacob.

BERTHOZ A. (2003), *La Décision*, Paris, Odile Jacob.

BHATT M. A., LOHRENZ T., CAMERER C. F., MONTAGUE P. R. (2010), « Neural signatures of strategic types in a two-person bargaining game », *PNAS*, 16 novembre, 107 (46).

BICCHIERI C. (2006), *The Grammar of Society, The Nature and Dynamics of Social Norms*, Cambridge, Cambridge University Press.

BICCHIERI C. (2010), « Norms, preferences, and conditional behavior », *Politics, Philosophy and Economics*, 9 (3), p. 297-313.

BICKART K. , HOLLENBECK M. , BARRET, L. F. , DICKERSON B. (2012), « Intrinsic amygdalia-cortical functional connectivity predicts social network size in human », *Journal of Neurosciences*, 32 (42), p. 14729-14741.

BINMORE K., SHAKED A., SUTTON J. (1985), « Testing noncooperative bargaining theory : A preliminary study », *The American Economic Review*, 75 (5), p. 1178-1180.

BINMORE K. (1994-1998), *Game Theory and the Social Contract*, tome 1 : *Playing Fair*, tome 2 : *Just Playing*, Cambrige (MA), MIT Press.

BINMORE K. , SHAKED A. (2010), « Experimental economics. Where next ? », *Journal of Economic Organization*, 73, p. 120-121.

BLAIR R. J. , CIPOLOTTI L. (2000), « Impaired social response reversal. A case of “acquired sociopathy” », *Brain*, 123, p. 1122-1141.

BLAIR R. J. (2004), « The roles of orbital frontal cortex in the modulation of antisocial behavior », *Brain Cogn.*, 55, p. 198-208.

BOESCH C., TOMASELLO M. (1998), « Chimpanzee and human cultures », *Current anthropology*, 19 (5), p. 591-604.

BOHM-BAWERK E. VON (1889), *Positive Theory of Capital*, New York, MacMillan and Co.

BOREL É., VILLE J. (1938), *Applications des jeux de hasard*, Paris, Gauthier-Villars.

BOREL É., CHÉRON A. (1940), *Théorie mathématique du bridge à la portée de tous*, Paris, Gauthier-Villars.

BOUDON R. (1973), *L’Inégalité des chances*, Paris, Armand Colin.

BOURDIEU P. (1980), *Le Sens pratique*, Paris, Éditions de Minuit.

BOURGEOIS-GIRONDE S. (2010), « Is neuroeconomics doomed by the reverse inference fallacy ? », *Mind Soc.*, 9, p. 229-249.

BOWLES S. , GINTIS H. (2004), « The evolution of strong reciprocity : Cooperation in heterogeneous populations », *Theoretical Population Biology*, 65 (1), p. 17-28.

BOWLES S. , GINTIS H., (2009), « Beyond enlightened self-interest : Social norms, other-regarding preferences, and cooperative behavior », *in* S. A. Levin (éd.), *Games, Groups, and the Global Good*, Springer Series in Game Theory, Berlin, Springer Verlag.

BOWLES S., GINTIS H. (2011), *A Cooperative Species : Human Reciprocity and its Evolution*, Princeton, Princeton University Press.

BRAADBAAT L. (2014), « The shared neural basis of empathy and facial imitation accuracy », *NeuroImage*, 84, p. 367-375.

BRAMOULÉ Y. (2007), « Anti-coordination and social interactions », *Games and Economic Behavior*, 58 (1), p. 30-49.

BRAMS S. (1990, 2003), *Negotiation Game. Applying Game Theory to Bargaining and Arbitration*, Londres, Routledge.

BÜCHNER S. , CORICELLI G. , GREINER B. (2007), « Self-centered and other-regarding behavior in the solidarity game », *Journal of Economic Behavior and Organization*, 62, p. 293-303.

BUCKHOLTZ J. W., ASPLUND C., DUX P., ZALD D., GORE J., MAROIS R. (2008), « The neural correlates of the third- party punishment », *Neuron*, 60 (5) p. 930-940.

BUCKHOLTZ J. W., MAROIS R. (2012), « The roots of modern justice : cognitive and neural foundations of social norms and their enforcement », *Nature Neuroscience*, 15, p. 655-661.

BULBULIA J. (2012), « Spreading order : Religion, cooperative niche construction, and risky coordination problems », *Biol. Philos.*, 27, p. 1-27.

BULLMORE E., SPORNS O. (2009), « Complex brain networks, graph theoretical analysis of structural and functional systems », *Nature Reviews Neuroscience*, 10, p. 186-198.

BURGESS A. (2011), « What can neuroscience tell us about religious consciousness ? A complex adaptive systems framework for understanding the religious brain », *American Political Science Association Conference Current Research in Bio-politics*, septembre 01.

CAMERER C. (2003), *Behavioral Game Theory. Experiments in Strategic Interactions*, Princeton, Princeton University Press.

CAMERER C., HO T. H., CHONG J. K. (2004), « A cognitive hierarchy of games », *The Quarterly Journal of Economics*, 119 (3), p. 861-898.

CAMERER C., LOEWENSTEIN G., PRELEC D. (2005), « How neuroscience can inform economics », *Journal of Economic Literature*, 43 (1), p. 9-64.

CARLSSON H., VAN DAME E. (1993), « Global games and equilibrium selection », *Econometrica*, 61 (5), p. 989-1018.

CAPPELLETTI D. , GÜTH W. , PLONER M., (2011), « Being of two minds : Ultimatum offers under cognitive constraints », *Journal of Economic Psychology*, 32, p. 940-950.

CARTER E. C. *et al*. (2012), « Religious people discount the future less », *Evolution and Human Behavior*, 33 (3), p. 224-231.

CASSAR A. (2007), « Coordination and cooperation in local, random and small world networks : Experimental evidence », *Games and Economic Behavior*, 58 (2), p. 209-230.

CHANG L. J. *et al*. (2010), « Seeing is believing : Trustworthiness as a dynamic belief », *Cognitive Psychology*, 61, p. 87-105.

CHANG L. J. , SMITH A. , DUFWENBERG M. , SANFEY A. G. (2011), « Triangulating the neural, psychological, and economic bases of guilt aversion », *Neuron*, 12 mai, 70, p. 560-572.

CHANG L. J., SANFEY A. G. (2011), « Great expectations : Neural computations underlying the use of social norms in decision-making », *Social Cognitive and Affective Neuroscience*, 8 (3), p. 277.

CHANG L. J., KOBAN L. (2013), « Modeling emotion and learning of norms in social interactions », *The Journal of Neuroscience*, 33 (18), p. 7615-7617.

CHARNESS G., RUSTICHINI A. (2011), « Gender differences in cooperation with group membership », *Games and Economic Behavior*, 72, p. 77-85.

CHAVEZ A. K., BICCHIERI C. (2013), « Third-party sanctioning and compensation behavior : Findings from the ultimatum game », *Journal of Economic Psychology*, 39, p. 268-277.

CHEON B. K. *et al*. (2011), « Cultural influences on neural basis of intergroup empathy », *NeuroImage*, 57, p. 642-650.

CHERNIAWSKY A. , HOLROYD B. (2013), « High temporal discounters overvalue immediate rewards rather than undervalue future rewards : An event-related brain potential study », *Cognitive and Affective Behavioral Neurosciences*, 13 (1), p. 36-45.

CHOI J.-K., AHN T. K. (2013), « Strategic reward and altruistic punishment support cooperation in a public goods game experiment », *Journal of Economic Psychology*, 35, p. 17-30.

CHRISTENSEN J. F., GOMILA A. (2012), « Moral dilemmas in cognitive neuroscience of moral decision-

making : A principled review », *Neuroscience and Biobehavioral Reviews*, 36 (4), p. 1249-1264.

CHUDEK M., HENRICH J. (2011), « Culture-gene coevolution, norm-psychology and the emergence of human prosociality », *Trends in Cognitive Sciences*, 15 (5), p. 218-226.

CHURCHLAND P. S. (2011), *Braintrust : What Neuroscience Tells Us about Morality*, Princeton (NJ), Princeton University Press.

CHURCHLAND P. S., WINKIELMAN P. (2012), « Modulating social behavior with oxytocin : How does it work ? What does it mean ? », *Hormones and Behavior*, 61, p. 392-399.

CIVAI C., CRESCENTINI C., RUSTICHINI A., RUMIATI R. I. (2012), « Equality versus self-interest in the brain : Differential roles of anterior insula and medial prefrontal cortex », *Neuroimage*, 62 (1), p. 102-112.

CIVAI C., CORRADI-DELL ACQUA C., RUMIATI R., FINK G. (2013), « Disentangling self-and fairness-related neural mechanism involved in the ultimatum game », *Cognitive and Affective Neurosciences*, 8 (4), p. 424-431.

CIVAI C., RUMIATI R. I., RUSTICHINI A. (2013), « More equal than others : Equity norms as an integration of cognitive heuristics and contextual cues in bargaining games », *Acta Psychologica*, 144, p. 12-18.

CLAVIEN C., CHAPUISAT M. (2013), « Altruism across disciplines : One word, multiple meanings », *Biol. Philos.*, 28, p. 125-140.

COLEMAN J. S. (1990), *Foundations of Social Theory*, Cambridge (MA), The Belknap Press of Harvard University Press.

COLOMBO M., SERIES P. (2012), « Bayes in the brain : On Bayesian modelling neurosciences », *The British Journal of Philosophy*, 63 (3), doi:10.1093/bjps/axr043.

CONLISK J. (2011), « Professor Zak's empirical studies on trust and oxytocin », *Journal of Economic Behavior and Organization*, 78 (1-2), p. 160-166.

CORICELLI G., DOLAN R. J., SIRIGU A. (2007), « Brain, emotion and decision making : the paradigmatic example of regret », *Trends in Cognitive Sciences*, 11 (6), p. 258-265.

CORRADINI A., ANTONIETTI A. (2013), « Mirror neurons and their function in cognitively understood empathy », *Consciousness and Cognition*, 22, p. 1152-1161.

COSMIDES L., TOOBY J. (2005), « Neurocognitive adatations designer for social exchange », in Buss M. D. (éd), *The Handbook of Evolutionnary Psychology*, New York, J. Wiley, p. 584-667.

COUTLEE C. G., HUETTEL S. A. (2011), « The functional neuroanatomy of decision making : Prefrontal control of thought and action », *Brain Research*, 1428, p. 3-12.

COWAN R., JONARD N., ZIMMERMANN J.-B. (2006), « Evolving networks of inventors », *Journal of Evolutionary Economic*, 16 (1-2), p. 155-174.

CROCKETT M., CLARK L., LIEBERMAN M., TABIBNIA G., ROBBINS T. (2010), « Impulsive choice and altruistic punishment are correlated and increase in tandem with serotonin depletion », *Emotion*, 10 (6), p. 855-862.

CROCKETT M., APERGIS-SCHOUTE A., HERMANN B., LIEBERMAN M., MULLER, U., ROBBINS T., CLARK L. (2013), « Serotonin modulates striatal responses to fairness and retaliation in humans », *The Journal of Neurosciences*, 33 (8), p. 3505-3513.

DALY A. (2014), « Primary intersubjectivity : Empathy, affective reversibility, "Self-affection" and the primordial "We" », *Topoi*, 33 (1), p. 227-241.

DAMASIO A. (1995), *L'Erreur de Descartes*, Paris, Odile Jacob.

DAMASIO A. (1999), *Le Sentiment même de soi*, Paris, Odile Jacob.

DAVIDSON D. (1960), *Essays on Actions and Events*, Oxford, Oxford University Press.

DAWES C., LOEWEN P. J., SCHREIBER D., SIMMONS A., FLAGON T., MCELREATHG E., BOKEMPERH S., FOWLES J., PAULUS M. (2012), « Neural basis of egalitarian behavior », *PNAS*, 109 (17), p. 6479-6483.

DAYAN P., ABBOT L. F. (2001), *Theoretical Neuroscience*, Cambridge (MA), The MIT Press.

DAYAN P., YU A. L. (2006), « Phasic norepinephrine. A neural interrupt signal for unexpected events », *Network: Computation in neural systems*, 17 (4), p. 335-350.

DECETY J., CHAMINADE T. (2003), « Neural correlates of feeling sympathy », *Neuropsychologia*, 41, p. 127-138.

DECETY J., SOMMERVILLE J. A. (2003), « Shared representations between self and other : a social cognitive neuroscience view », *Trends in Cognitive Sciences*, 7 (12), p. 527-533.

DECETY J., CHAMINADE T. (2003), « When the self represents the other : A new cognitive neuroscience view on psychological identification », *Consciousness and Cognition*, 12, p. 577-596.

DECETY J., GRÈZES J. (2006), « The power of simulation : Imagining one's own and other's behavior », *Brain Research*, 1079, p. 4-14.

DECETY J., PORGES E. C. (2011), « Imagining being the agent of actions that carry different moral consequences : An fMRI study », *Neuropsychologia*, 49, p. 2994-3001.

DECETY J., HOWARD L. H. (2013), « The role of affect in the neurodevelopment of morality », *Child Development Perspectives*, 7 (1), p. 49-54.

DECLERCK C. H., BOONE C., EMONDS G. (2013), « When do people cooperate ? The neuroeconomics of prosocial decision making », *Brain and Cognition*, 81, p. 95-117.

DE DREU CARSTEN K. W. (2012), « Oxytocin modulates cooperation within and competition between groups : An integrative review and research agenda », *Hormones and Behavior*, 61, p. 419-428.

DELGADO M. R., FRANK R. H., PHELPS E. A. (2005), « Perceptions of moral character modulate the neural systems of reward during the trust game », *Nature Neuroscience*, 8 (11), p. 1611-1618.

DE N OUDEN H. E. M, FRITH U., FRITH C., BLAKEMORE S.-J. (2005), « Thinking about intentions », *NeuroImage*, 28, p. 787-796.

DE POSSEL R., (1936), *Sur la théorie mathématique des jeux de hasard et de réflexion*, Paris, Hermann.

DEPRAZ N. (2010), « De l'"inter-attention" à l'attention inter-relationnelle. Le croisement de l'attention et de l'intersubjectivité à la lumière de l'attention conjointe », *Symposium*, 14 (1), p. 104-118.

DE QUERVAIN D. , FISCHBACHER U. , TREYER V. , SCHELLHAMMER M. , SCHNYDER U. , BUCK A. , FEHR E. (2004), « The neural basis of altruistic punishment », *Science* 305 (5688), p. 1254-1258.

DESCOMBES V. (2004), *Le Complément du sujet*, Paris, Gallimard.

DE SOUZA R. (2009), « Epistemic feelings », *Mind and Matter*, 7 (2), p. 139-161.

DEWEY J. (1938), *The Logic of Inquiry*, New York, Henry Holt.

DOHA K., ISHII S., POUGET A., RAO R. P. V. (éds.) (2006), *Bayesian Brain : Probabilistic Approaches to Neural Coding Computational Neurosciences*, Cambridge (MA), The MIT Press.

DOHMEN T., FALK A., FLIESSBACH K., SUNDE U., WEBER B. (2011), « Relative versus absolute income, joy of winning, and gender : Brain imaging evidence », *Journal of Public Economics*, 95, p. 279-285.

DU W. , GREEN L. , MYERSON J. (2002), « Cross-cultural comparison of discounting delayed and probabilistic rewards », *The Psychological Record*, 52, p. 479-492.

DUFWENBERG M., KIRCHSTEIGER G. (2004), « A theory of sequential reciprocity », *Games and Economic Behavior*, 47, p. 268-298.

DURKHEIM É. (1924, 2002), *Sociologie et Philosophie*, Paris, PUF.

EDGEWORTH F. Y. (1881), *Mathematical Psychics*, Londres, Keagan Paul.

EISENEGGER C., HAUSHOFER J., FEHR E. (2011), « The role of testosterone in social interaction », *Trends in Cognitive Sciences*, 15 (6), p. 263-271.

ELLINGSEN D.-M. (2014), « In touch with your emotions : Oxytocin and touch change social impressions while others' facial expressions can alter touch », *Psychoendocrinology*, 39, p. 11-20.

ELSTER J. (2007), *Explaining Social Behavior*, Cambridge, Cambridge University Press.

EMONDS G., DECLERCK C. H., BOONE C., VANDERVLIET E., PARIZEL J. M. (2012), « The cognitive demands on cooperation in social dilemmas : An fMRI study », *Social Neuroscience*, 7 (5), p. 494-509.

EMONDS G., DECLERCK C. H., BOONE C., VANDERVLIET E., PARIZEL J. M. (2011), « Comparing the neural basis of decision making in social dilemmas of people with different social value orientations, a fMRI study », *Journal of Neuroscience, Psychology and Economics*, 4 (11), p. 11-24.

ENGEL P. (2007), *Va savoir ! De la connaissance en général*, Paris, Hermann.

ENGELMANN D., FISCHBACHER U. (2009), « Indirect reciprocity and strategic reputation building in an experimental helping game », *Games and Economic Behavior*, 67, p. 399-407.

ENGELMANN J. B. , HEIN G. (2013), « Contextual and social influences on valuation and choice », *Progress in Brain Research*, 202, p. 215-237.

ENGEN H. G., SINGER T. (2013), « Empathy circuits », *Current Opinion in Neurobiology*, 23, p. 275-282.

ERES R. , MOLENBERGHS P. (2013), « The influence of group membership on the neural correlates involved in empathy », *Frontiers in Human Neuroscience*, mai, 7 (176).

EVANS S. L. *et al.* (2013), « Intranasal oxytocin effects on social cognition : A critique », *Brain*

Research, e-pub.

FAILLO M., GRIECO D., ZARRI L. (2013), « Legitimate punishment, feedback, and the enforcement of cooperation », *Games and Economic Behavior*, 77, p. 271-283.

FALK A., FISCHBACHER U. (2006), « A theory of reciprocity », *Games and Economic Behavior*, 54, p. 293-315.

FALK A., FEHR E., FISCHBACHER U. (2008), « Testing theories of fairness-intentions matter », *Games and Economic Behavior*, 62, p. 287-303.

FAN Y., DUNCAN N. W., GRECKC M., NORTHOFFA G., (2011), « Is there a core neural network in empathy ? An fMRI based quantitative meta-analysis », *Neuroscience and Biobehavioral Reviews*, 35, p. 903-911.

FECHNER G. T. (1860), *Elemente der Psychophysics*, Leipzig, Bretkopft Härtel.

FEHR E., SCHMIDT K. (1999), « A theory of fairness, competition and cooperation », *The Quaterly Journal of Economics*, 114 (3), p. 817-868.

FERH E., SCHMIDT K. (2010), « On inequity aversion : Reply to Binmore and Shaked », *Journal of Economic Behavior and Organization*, 73 (1), p. 101-108.

FEHR E., FISCHBACHER U., GÄCHTER S. (2002), « Strong reciprocity, human cooperation and enforcement of social norms », *Human Nature*, 13, p. 1-25.

FEHR E., FISCHBACHER U. (2004), « Social norms and human cooperation », *Trends in Cognitive Sciences*, 8 (4), p. 185-190.

FEHR E., FISCHBACHER U. (2004), « Third-party punishment and social norms », *Evolution and Human Behavior*, 25, p. 63-87.

FEHR E., ROCKENBACH B. (2004), « Human altruism : Economic, neural, and evolutionary perspectives », *Current Opinion in Neurobiology*, 14, p. 784-790.

FEHR E., CAMERER C. F. (2007), « Social neuroeconomics : The neural circuitry of social preferences », *Trends in Cognitive Sciences*, 11 (10), p. 419-427.

FEHR E., RANGEL A. (2011), « Neuroeconomic foundations of economic choice – Recent advances », *Journal of Economic Perspectives*, 25 (4), p. 3-30.

FELDMANHALL O. *et al*. (2012), « What we say and what we do : The relationship between real and hypothetical moral choices », *Cognition*, 123, p. 434-441.

FINDLEY T. S., CALIENDO F. N. (2014), « Interacting mechanisms of time inconsistency », *Journal of Economic Psychology*, 41, p. 68-76.

FOGASSI L., (2011), « The mirror neuron system : How cognitive functions emerge from motor organization, », *Journal of Economic Behavior and Organization*, 77, p. 66-75.

FORBES G. E, POORE J. C, SALOMON J., LIPSKY R. H., HODKINSON C. A., GOLDMAN D., GRAFMAN J. (2012), « BDNF polymorphism-dependant OFC and DLPFC plasticity differently moderates implicit and explicit bias », *Cereb. Cortex*, 22 (11), p. 2602-2609.

FOSTER D. P. YOUNG H. P. (2003), « Learning hypothesis testing, and Nash equilibrium », *Games and*

Economic Behavior, 45 (1), p. 73-96.

FOSTER D. P., YOUNG H. P. (2006), « Regret testing learning to play Nash equilibrium without knowing you have an opponent », *Theoretical Economics*, 1, p. 341-367.

FREDERICK S., LOEWENSTEIN G., O'DONOGHUE T. (2002), « Time discounting and time preference : A critical review », *Journal of Economic Literature*, 40 (2), p. 351-401.

FRIEDMAN M. (1953), *Essays on Positive Economics*, Chicago (Ill), Chicago University Press.

FRISTON K. (2010), « The free-energy principle : A unified brain theory ? », *Nature Reviews Neuroscience*, 11, p. 127-138.

FRISTON K. *et al*. (2013), « The anatomy of choice : Active inference and agency », *Frontiers in Human Neuroscience*, septembre, 7 (598).

FRITH C. D., FRITH U. (2003), « Development and neurophysiology of mentalizing », *Philos. Trans. R. Soc., Lond. B. Biol. Sci.*, 358 (1431), p. 459-473.

FRITH C. D., FRITH U. (2006), « The neural basis of mentalizing », *Neuron*, 50 (4), p. 531-534.

FRITH C. D., FRITH U. (2006), « How we predict what other people are going to do », *Brain Research*, 1078, p. 36-46.

FRITH C. D., FRITH U., (2007), « Social cognition in humans », *Current Biology*, 21 août, 17, R724-R732.

FRITH C. D. , FRITH U. (2008), « Implicit and explicit processes in social cognition », *Neuron*, 6 novembre, 60, p. 503-510.

FRITH C. D. (2013), « Action, agency and responsibility », *Neuropsychologia*, 55, p. 137-142.

GALE J., BINMORE K. G., SAMUELSON L. (1995), « Learning to be imperfect : The ultimatum game », *Games and Economic Behavior*, 8, p. 56-90.

GAILLOT M. T., GITTER S. A., BAKER M. D. (2012), « Breaking rules : Low trait or state self-control increases social norms violations », *Psychology*, 3 (12), p. 1074-1083.

GALLAGHER H. L., JACK A. I., ROEPSTORFF A., FRITH C. D. (2002), « Imaging the intentional stance in a competitive game », *NeuroImage*, 16, p. 814-821.

GALLAGHER H. L., FRITH C. D. (2003), « Functional imaging of "theory of mind" », *Trends in Cognitive Sciences*, 7 (2), p. 77-83.

GARCIA J., TRAULSEN A. (2012), « Leaving the loners alone : Evolution of cooperation in the presence of antisocial punishment », *Journal of Theoretical Biology*, 3407, p. 168-173.

GEANAKOPLOS J., PEARCE D., STACCHETTI E. (1989), « Psychological games and sequential rationality », *Games and Economic Behavior*, 1 (1), p. 60-79.

GIGERENZER G. (2000), *Adaptative Thinking : Rationality in the Real World*, Oxfort, Oxford university Press.

GIGERENZER G. , GAISSMAIER W. (2011), « Heuristic decision making », *The Annual Review of Psychology*, 62, p. 451-482.

GILOVICH T., GRIFFIN DA., KAHNEMAN D., (2002), *Heuristics and Biases. The Psychology of Intuitive*

Judgment, New York, Cambridge University Press.

GINTIS H., BOWLES S., BOYD R., FEHR E. (2003), « Explaining altruistic behavior in humans », *Evolution and Human Behavior*, 24, p. 153-172.

GINTIS H. (2004), « The genetic side of gene-culture coevolution : Internalization of norms and prosocial emotions », *Journal of Economic Behavior and Organization*, 53, p. 57-67.

GINTIS H., HENRICH J., BOWLES S., BOYD R., FEHR E. (2008), « Strong reciprocity and the roots of human morality », *Soc. Just. Res.*, 21, p. 241-253.

GINTIS H. (2010), « Social norms as choreography », *Politics, Philosophy and Economics*, 9 (3), p. 251-264.

GIORGETTA C. *et al.* (2013), « Waves of regret : A meg study of emotion and decision-making », *Neuropsychologia*, 51, p. 38-51.

GLIMCHER P. W. (2003), *Decisions, Uncertainty, and the Brain. The Science of Neuroeconomics*, Cambridge, The MIT Press.

GLIMCHER P. W., CAMERER C. F., FEHR E., POLDRACK R. A. (2009), *Neuroeconomics : Decision Making and the Brain*, New York, Academic Press (nouvelle édition 2014).

GOFFMAN E. (1973), *La Mise en scène de la vie quotidienne*, Paris, Éditions de Minuit.

GOLD N., SUGDEN R. (2007), « Collective intentions and team agency », *Journal of Philosophy*, 104 (3), p. 109-137.

GOLDMAN A. (1987), « Foundations of social epistemics », *Synthese*, 73, p. 109-144.

GOLDMAN A. (2006), « Social epistemology », *Stanford Encyclopedia of Philosophy*.

GONZALEZ-LIENCRES C., SHAMAY-TSOORY S. G., BRÜ M. (2013), « Towards a neuroscience of empathy : Ontogeny, phylogeny, brain mechanisms, context and psychopathology », *Neuroscience and Biobehavioral Reviews*, 37, p. 1537-1548.

GOYAL S. , VEGA-REDONDO F. (2005), « Network formation and social coordination », *Games and Economic Behavior*, 50, p. 178-207.

GRABENHORST F., ROLLS E. T. (2009), « Different representations of relative and absolute subjective value in the human brain », *NeuroImage*, 48, p. 258-268.

GRACIA-LÁZARO C. , CUESTA J. A. , SÁNCHEZ Y. (2012), « Human behavior in prisoners' dilemma experiments suppresses network reciprocity », *Scientific Report*, 2, art. 325, doi:10.1038/srep00325.

GRECCUCI A. , GIORGETTA C. , BONONO N. (2012), « Living emotion, avoiding emotion behavior investigation of the regulation of socially driven emotion », *Frontiers in Psychology*, 3 (616), doi:10.3389.

GREENBERG J. (1990), *The Theory of Social Situation. An alternance Game-Theoretic Approach*, Cambridge, Cambridge University Press.

GREENE J. (2003), « From neural "is" to moral "ought" : What are the moral implications of neuroscientific moral psychology ? », *Nature Reviews. Neuroscience*, 4, p. 847-850.

GREENE J., NYSTROM LEIGH E., ENGELL ANDREW D., DARLEY JOHN M., COHEN JONATHAN D. (2004), « The neural bases of cognitive conflict and control in moral judgment », *Neuron*, 14 octobre, 44, p. 389-400.

GROSE J., PATERNOTTE C. (2013), « Social norms : Repeated interactions, punishment, and context dependence », *Public Reason*, 1, p. 3-13.

GRUJIC J., FOSCO C., ARAUJO L., CUESTA J. A., SÁNCHEZ A. (2010), « Social experiments in the mesoscale : Humans playing spatial prisoner's dilemma », *PLoS One*, 5 (11), p. 1.

GRYGOLEC J., CORICELLI G., RUSTICHINI A. (2012), « A neuroeconomic study of social observability and personal responsibility in decision making : An fMRI experiment », *Front Psychol.*, 3, p. 25.

GUALA F. (2012), « Reciprocity : Weak or strong ? What punishment experiments do (and do not), demonstrate », *Behavioral and Brain Sciences*, 35 (1), p. 1-15.

GUZMAN R. A. RODRIGUEZ-SICKERT C., ROWTHORN R. (2007), « When in Rome, do as the Romans do : the coevolution of altruistic punishment, conformist learning, and cooperation », *Evol. Hum. Behav.*, 28, p. 112-117.

HALKO M.-L., HLUSHCHUK Y., HARI R., SCHÜRMANN M. (2009), « Competing with peers : Mentalizing-related brain activity reflects what is at stake », *NeuroImage*, 46, p. 542-548.

HAMMOND P. (1976), « Changing tastes and coherent dynamic choice », *The Review of Economics Studies*, 1976, 43 (1), p. 159-173.

HAMMOND P. (1988), « Consequentialist foundations for expected utility », *Theory and Decision*, 25 (1), p. 25-78.

HAMMOND P. (1994), « Consequentialism, non-archimedian probabilities and lexicographic expected utility », *Operation Research*, 93, p. 219-250.

HA N R., TAKAHASHI T. (2012), « Psychophysics of time perception and valuation in temporal discouting of gain and loss », *Physica A*, 391 (24), p. 6568-6576.

HAN H., GLOVER G. H., JEONG C. (2013), « Cultural influences on the neural correlate of moral decision making processes », *Behavioural Brain Research*, 259, p. 215-228.

HARBAUGH W. T., MAYR U., BURGHART D. R. (2007), « Neural responses to taxation and voluntary giving reveal motives for charitable donations », *Science*, 316 (5831), 1622-1625.

HARENSKI C. L. *et al.* (2012), « Neural development of mentalizing in moral judgment from adolescence to adulthood », *Developmental Cognitive Neuroscience*, 2, p. 162-173.

HARSANYI J. (1977), « Rule utilitarianism and decision-theory », *Erkenntnis*, 11 (1), p. 25-53.

HARSANYI J. (1982), « Morality and the theory of rational behavior », *in* Sen A., Williams B. (éds.), *Utilitarism and Beyond*, Cambridge, Cambridge University Press.

HARSANYI J., SELTEN R. (1988), *A General Theory of Equilibrium in Games*, Cambridge, Cambridge University Press.

HARRIS S., KAPLAN J. T., CURIEL A., BOOKHEIMER S. Y., IACOBONI M., COHEN M. S. (2009), « The neural correlates of religious and nonreligious belief », *PLoS ONE*, 4 (10), e7272.

HARRISON G. W. (2008), « Neuroeconomics : A critical reconsideration », *Economics and Philosophy*, 24, p. 303-344.

HARUNO M., FRITH CHRIS D. (2010), « Activity in the amygdala elicited by unfair divisions predicts social value orientation », *Nat. Neurosci.*, 13 (2), p. 160-161.

HART S., MAS-COLELL A. (1996), « Bargaining and value », *Econometrica*, 64 (2), p. 357-380.

HART S., MAS-COLELL A. (2000), « A simple adaptative procedure leading to correlated equilibrium », *Econometrica*, 68 (5), p. 1127-1150.

HAUSHOFER J., FEHR E. (2008), « You shouldn't have : Your brain on others' crimes », *Neuron*, 11 décembre, 60, p. 738-740.

HAYEK F. A. (1980) [1973], *Droit, législation et liberté*, Paris, PUF, vol.1.

HAYEK F. A. (2001) [1952], *L'Ordre sensoriel*, Paris, Éditions du CNRS.

HE J. M., HUANG X. T., YUAN H., CHEN Y. G. (2012), « Neural activity in relation to temporal distance : Differences in past and future temporal discounting », *Consciousness and Cognition*, 21, p. 1662-1672.

HEIN G., SILANI G., PREUSCHOFF K., BATSON C. D., SINGER T. (2010), « Neural responses to ingroup and outgroup members' suffering predict individual differences in costly helping », *Neuron*, 7 octobre, 68, p. 149-160.

HENRICH J. *et al.* (2006), « Costly punishment across human societies », *Science*, 312 (5781), p. 1767-1770.

HERRMANN B., THÖNI C., GÄCHTER S. (2008), « Antisocial punishment across societies », *Science*, 319 (5868), p. 1362-1367.

HOFFMAN M. B. (2004), « The neuroeconomic path of the law », *Phil. Trans. R. Soc. Lond. B.*, 359, p. 1667-1676.

HOMBERT J.-M. (sous la dir.) (2009), *Aux origines des langues et du langage*, Paris, Fayard.

HOMBERT J.-M., LENCLUD G. (2014), *Comment le langage est venu à l'homme*, Paris, Fayard.

HOUSER D., SHUNK D., WINTER J. (2010), « Distinguishing trust from risk : An anatomy of the investment game », *Journal of Economic Behavior and Organization*, 74 (1-2), p. 72-81.

HUSSERL E. (2009) [1893-1912], *Phénoménologie de l'attention*, Paris, Vrin.

Hwang S.-H., Bowles S. (2012), « Is altruism bad for cooperation ? », *Journal of Economic Behavior and Organization*, 83, p. 330- 341.

IACCOBONI M., MOLNAR-SZACKAS J., GALLESE V., BUCCINO G., MAZZIOTA J. C., RIZZOLATTI G. (2005), « Grasping the intentions of others with one's own mirror neuron system », *PLoS Biol.*, 3 (3), e79.

IACCOBONI M. (2009), « Neurobiology of imitation », *Current Opinion in Neurobiology*, 19, p. 661-665.

INZLICHT M., MCGREGOR I., HIRSH J. B., NASH K. (2009), « Neural markers of religious conviction », *Psychological Science*, 20 (3), p. 382-392.

IRWIN K., HORNE C. (2013), « A normative explanation of antisocial punishment », *Social Science Research*, 42, p. 562-570.

ISRAEL S., WEISEL O., EBSTEIN R. P., BORNSTEIN G. (2012), « Oxytocin, but not vasopressin, increases both parochial and universal altruism », *Psychoneuroendocrinology*, 37, p. 1341-1344.

IZUMA K., SAITO D. N., SADATO N. (2008), « Processing of social and monetary rewards in the human striatum », *Neuron*, 24 avril, 58, p. 284-294.

IZUMA K., SAITO D. N., SADATO N. (2010), « Processing of the incentive for social approval in the ventral striatum during charitable donation », *Journal of Cognitive Neuroscience*, 22 (4), p. 621-631.

IZUMA K. (2013), « The social neuroscience of reputation », *Social Science Research*, 42, p. 562-570.

JACKSON P. L., MELTZOFF A. N., DECETY J. (2006), « Neural circuits involved in imitation and perspective-taking », *NeuroImage*, 31, p. 429-439.

JAMAL A. M. M., SUNDAR C. (2011), « Modeling exchange rates with neural networks », *Journal of Applied Business Research*, 14 (1), p. 1-6.

JEANNEROD M. (1983), *Le Cerveau-machine. Physiologie de la volonté*. Paris, Arthème Fayard.

JEANNEROD M., DECETY J. (1995), « Mental motor imagery : A window into the representational stages of action », *Current Opinion in Neurobiology*, 5, p. 727-732.

JEFFREY R. C. (1965), *The Logic of Decision*, Chicago, University of Chicago Press.

JONES B. A., RACHLIN H. (2009), « Delay, probability and social discounting in a public goods game », *Journal of the Experimental Analysis of Behavior*, 91 (1), p. 61-73.

JONES R. M. (1994), « Curricula focused on behavioral deviance », *in* Archer S. L. (éd.), *Interventions for Identity Development*, Newbury Park (CA), Sage Publications, Inc., p. 174-190.

KABLE J. W., GLIMCHER P. W. (2009), « The neurobiology of decision : Consensus and controversy », *Neuron*, 24 septembre, 63.

KAHNEMAN D. (2011), *Thinking, Fast and Slow*, Londres, FSG.

KAHNEMAN D., WAKKER P. P., SARIN R. (1997), « Back to Bentham ? Explorations of experienced utility », *Quarterly Journal of Economics*, 112 (2), p. 375-406.

KAHNEMAN D. (2012) [2011], *Système 1, Système 2. Les deux systèmes de la pensée*, Paris, Flammarion.

KAHNEMAN D., TVERSKY A. (1979), « Prospect theory : An analysis of decision under risk », *Econometrica*, 47 (2), p. 263-292.

KALISCH M., NASH J. F., NERING E. D. (1954), « Some experimental n-person Games », *in* Thrall R. M., Coombs C. H., Davis R.L (éds.), Decision Processes, New York, John Wiley and Sons, Inc.

KANDORI M., MAILATH G. J., ROB R. (1993), « Learning, mutation and long run equilibria in games », *Econometrica*, 61 (14), p. 29-56.

KAPOGIANNIS D., BARBEY A. K., SU M., ZAMBONI G., KRUEGER F., GRAFMAN J.(2009), « Cognitive and neural foundations of religious belief », *PNAS*, 106 (12), p. 4876-4881.

KÉRI S., KISS I. (2011), « Oxytocin response in a trust game and habituation of arousal », *Physiology and Behavior*, 102, p. 221-224.

KIESLING L. L. (2012), « Mirror neuron research and Adam Smith's concept of sympathy : Three points of correspondence », *Rev. Austrian. Econ.*, 25, p. 299-313.

KING-CASAS B., TOMLIN D., ANEN C., CAMERER C. F., QUARTZ S. R., MONTAGUE P. R. (2005), « Getting to know you : Reputation and trust in a two-persons economic exchange », *Science*, 308 (5718), p. 78-83.

KIRMAN A., MARKOSE S., GIANSANTE S., PIN P. (2007), « Marginal contribution, reciprocity and equity in segregated groups : Bounded rationality and self-organization in social networks », *Journal of Economic Dynamics and Control*, 31, p. 2085-2107.

KLACKLA J., PFUNDMAIR M., AGROSKIN D., JONAS E. (2013), « Who is to blame ? Oxytocin promotes nonpersonalistic attributions in response to a trust betrayal », *Biological Psychology*, 92, p. 387-394.

KNILL D. C., POUGET A. (2004), « The Bayesian brain : The role of uncertainty in neural coding and computation », *Trends in Neurosciences*, 27 (12), p. 712-719.

KNOCH D., PASCUAL-LEONE A., MEYER K., TREYER V., FEHR E. (2006), « Diminishing reciprocal fairness by disrupting the right prefrontal cortex », *Science*, 314 (5800), p. 829-832.

KNOCH D. , NITSCHE M. A. , FISCHBACHER U. , EISENEGGER C. , PASCUAL-LEONE A. , FEHR E. (2008), « Studying the neurobiology of social interaction with transcranial direct current stimulation – The example of punishing unfairness », *Cereb. Cortex*, 18 (9), p. 1987-1990.

KOFFKA K. (1955), *Principles of Gestalt Psychology*, New York, Harcourt and Brace.

KOSFELD M., HEINRICHS M., ZALK J. P., FISHBACHER U., FEHR E. (2005), « Oxytocin increases trust in humans », *Nature*, 435, p. 673-676.

KÖSZEGI B., RABIN M. (2007), « Reference-dependent risk attitudes », *The American Economic Review*, 97 (4), p. 1047-1073.

KOVÁCZ Á. M., TÉGLÁS E., ENDRESS A. D. (2010), « The social sense susceptibility to others' beliefs in humain infants and adults », Science, 330 (6012), p. 1830-1834

KRUEGER F. *et al.* (2007), « Neural correlates of trust », *PNAS*, 104 (50), p. 20084-20089.

KUO W.-J., SJÖSTRÖM T., CHEN Y.-P., WANG Y.-H., HUANG C.-Y. (2009), « Intuition and deliberation : Two systems for strategizing in the brain », *Science*, 324 (5926), p. 519-522.

LAIBSON D. (1997), « Golden eggs and hyperbolic discounting », *The Quarterly journal of Economics*, 112 (2), p. 443-477.

LAMBERT B. , DECLERCK C. H. , BOONE C. (2014), « Oxytocin does not make a face appear more trustworthy but improves the accuracy of trustworthiness judgments », *Psychoendocrinology*, 40, p. 60-68.

LAMM C., PORGES E. C., CACIOPPO J. T., DECETY J ; (2008), « Perspective taking is associated with specific facial responses during empathy for pain », *Brain Research*, 1227, p. 153-161.

LAMM C., SINGER T., (2010), « The role of anterior insular cortex in social emotions », *Brain Struct. Funct.*, 214, p. 579-591.

LAMM C. , DECETY J. , SINGER T. (2011), « Meta-analytic evidence for common and distinct neural networks associated with directly experienced pain and empathy for pain », *NeuroImage*, 54, p. 2492-2502.

LANGERGRABER K. E., BOESCH C. (2011), « Genetic and "cultural" similarity in wild chimpanzees », *Proceedings of the Royal Society*, 278 (1704), p. 408-416.

LARRIEU P. (2012), « Le droit à l'ère des neurosciences », *Médecine et Droit*, 115, p. 106-110.

LEBRETON M. , JORGE S. , MICHEL V. , THIRION B. , PESSIGLIONE M. (2009), « An automatic valuation system in the human brain : Evidence from functional neuroimaging », *Neuron*, 12 novembre, 64, p. 431-439.

LEGRAND D. (2006), *Phenomenology and the Cognitive Science*, Heidelberg, Springer.

LEVINE D. K., PESENDORFER W. (2007), « The evolution of cooperation through imitation », *Games and Economic Behavior*, 58 (2), p. 293-315.

LEWIS D. (1969), *Convention. A Philosophical Study*, Cambridge, Harvard University Press.

LICHENSTEIN S. , SLONIC P. (1971), « Reversals preferences between bits and choice in Gambling divisions », *Jounal of Experimental Psychology*, 89 (1), p. 45-55.

LIPPS T. (1905), *Das Wissen von fremden Ichen*, *in* T. Lipps (éd.), *Psychologische Untersuchungen*, Leipzig, Engelmann.

LIU L., FENG T., WANG J., LI H. (2012), « The neural dissociation of subjective valuation from choice processes in intertemporal choice », *Behavioural Brain Research*, 231, p. 40-47.

LIVET P. (2002), *Émotions et rationalité morale*, Paris, PUF.

LIVET P., NEF F. (2009), *Les Êtres sociaux*, Paris, Hermann.

LIVET P. (2010), « Rational choice, neuroeconomy and mixed emotions », *Phil. Trans. R. Soc. B*, 365, p. 259-269.

LOEWENSTEIN G. (mai 2000), « Emotions in economic theory and economic behavior », *The American Economic Review*, 90 (2), *Papers and Proceedings of the One Hundred Twelfth Annual Meeting of the American Economic Association*, p. 426-432.

LONG Y. , JIANG X. , ZHOU X. (2012), « To believe or not to believe : Trust choice modulates brain responses in outcome evaluation », *Neuroscience*, 200, p. 50-58.

LOOMES G. , SUGDEN R. (1982), « Regret theory : An alternative theory of rational choice under uncertainty », *Economical Journal*, 92, p. 805-824.

LUCE R. D. (1956), « Semiorders and a theory of utility discrimination », *Econometrica*, 24, p. 178-191.

LUCE R. D. (1990), « "On the possible psychophysical laws" revisited : Remarks on cross-modal matching », *Psychological Review*, 97 (1), p. 66-77.

LUCE R. D (2002), « A psychophysical theory of intensity proportions, joint presentations and

matches », *Psychological Review*, 109 (3), p. 520-532.

LY M., HAYNES M. R., BARTER J. W., WEINBERGER D. R., ZINK C. F. (2011), « Subjective socioeconomic status predicts human ventral striatal responses to social status information », *Current Biology*, 21 (9), p. 794-797.

MACDONALD K., MACDONALD T. M. (2010), « The peptide that binds : A systematic review of oxytocin and its prosocial effects in humans », *Harvard Review of Psychiatry*, 18 (1), p. 1-21.

MANIADAKIS M., TRAHANIAS P., TANI J. (2009), « Explorations on artificial time perception », *Neural Networks*, 22 (5-6), p. 509-517.

MARAZZITI D. *et al*. (2013), « The neurobiology of moral sense : facts or hypotheses ? », *Annals of General Psychiatry*, 12 (1), p. 6.

MARCHETTI A., CASTELLI I., HARLÉ K. M., SANFEY A. G. (2011), « Expectations and outcome : The role of proposer features in the Ultimatum game », *Journal of Economic Psychology*, 32, p. 446-449.

MARTÍNEZ M. M. *et al*. (2012), « Concerns about cultural neurosciences : A critical analysis », *Neuroscience and Biobehavioral Reviews*, 36, p. 152-161.

MASTEN C. L., MORELLI S. A., EISENBERGER N. I. (2011), « An fMRI investigation of empathy for "social pain" and subsequent prosocial behavior », *NeuroImage*, 55, p. 381-388.

MAUSS M. (1950, 2004), *Sociologie et anthropologie*, Paris, PUF.

MAYNARD S. J. (1981), *Evolution and the Theory of Games*, Cambridge, Cambridge University Press.

MCLURE S. M., LAIBSON D. I., LOWENSTEIN G., COHEN J. D. (2004), « Separate neural systems value immediate and delayed monetary rewards », *Science*, 306 (5695), p. 503-507.

MCLURE S. M., ERICSON K. M., LAIBSON D. I., LOWENSTEIN G., COHEN J. D. (2007), « Time discounting for primary rewards », *The Journal of Neuroscience*, 27 (21), p. 5796-5804.

MCCABE K., HOUSER D., RYAN L., SMITH V., TROUARD T. (2001), « A functional imaging study of cooperation in two-person reciprocal exchange », *PNAS*, 98 (20), p. 11832-11835.

MCCABE K., RIGDON M., SMITH V. (2003), « Positive reciprocity and intentions in trust games », *Journal of Economic Behavior and Organization*, 52 (2), p. 267-275.

MCMILLAN C. T., RASCOVSKY K. M., KHELLA C., CLARK R., GROSSMAN M. (2011), « The neural basis for establishing a focal point in pure coordination games », *Soc. Cogn. Affect. Neurosci.*, 7 (8), p. 881-887.

MERLEAU-PONTY M. (1996) [1946], *Le Primat de la perception et ses conséquences philosophiques*, Paris, Verdier.

MEYER MEGHAN L. *et al*. (2012), « Empathy for the social suffering of friends and strangers recruits distinct patterns of brain activation », *Social Cognitive and Affective Neuroscience (SCAN)*, 8 (4), p. 446-454.

MOLENBERGHS P. (2013), « The neuroscience of in-group bias », *Neuroscience and Biobehavioral Reviews*, 37, p. 1530-1536.

MOLL J., ZAHN R., OLIVEIRA-SOUZA R. KRUEGER F., GRAFMAN J. (2005), « The neural basis of human

moral cognition », *Nature Reviews Neuroscience*, 6, p. 799-809.

MONDERER D., SAMET D. (1989), « Approximating common knowledge with common beliefs », *Games and Economic Behavior*, 1 (2), p. 170-190.

MORAN J. M., LEE S. M., GABRIELI J. D. E. (2011), « Dissociable neural systems supporting knowledge about human character and appearance in ourselves and others », *Journal of Cognitive Neurosciences*, 23 (9), p. 2222-2230.

MORRISON S. , DECETY J. , MOLENBERGHS P. (2012), « The neuroscience of group membership », *Neurospychologia*, 50, p. 2114-2120.

MOSCOVICI S. (1979), *Psychologie des minorités actives*, Paris, PUF.

NADELHOFFER T. (2012), « Neuroprediction, violence, and the law : Setting the stage », *Neuroethics*, 5, p. 67-99.

NASH J. (1950), « The bargaining problem », *Econometrica*, 18 (2), p. 155-162.

NASH J. (1951), « Non-cooperative games », *The Annals of Mathematics*, Second Series, 54 (2), p. 286-295.

NASH J. (1953), « Two-person cooperative games », *Econometrica*, 21 (1), p. 128-140.

NESSE R. M. (2007), « Runaway social selection for displays of partner value and altruism », *Biological Theory*, 2 (2), p. 143-155.

NETTLE D. *et al*. (2013), « The watching eyes effect in the Dictator game : It's not how much you give, it's being seen to give something », *Evolution and Human Behavior*, 34, p. 35-40.

NEWBERG A., D'AQUILI E. (2001), *Why God Won't Go Away : Brain Science and the Biology of belief* , New York, Random House Ballantines Books.

NEWBERG A., WALDMAN M. R. (2006), *Why We Believe That We Believe*, New York, Free press.

NIELSEN L., MATHER M. (2011), « Emerging perspectives in social neuroscience and neureconomics of aging », *Soc. Cogn. Affect. Neurosci.*, 6 (2), p. 149-164.

NORENZAYAN A., GERVAIS W. M. (2013), « The origins of religious disbelief », *Trends in Cognitive Sciences*, 17 (1), p. 20-25.

O'CONNELL G. , CHRISTAKOU A. , HAFFEY A. T. , CHAKRBARTI B., (2013), « The role of empathy in choosing rewards from another's perspective », *Frontiers in Human Neuroscience*, 7 (174), p. 1-5.

ONISHI K. H., BAILLARGEON R. (2005), « Do 15-month-old infants understand false beliefs ? », *Science*, 308 (5719), p. 255-258.

OSBORNE M. L., RUBINSTEIN A. (1990), *Bargaining and Markets*, San Diego, Academic Press.

OTTI A. *et al*. (2010), « I know the pain you feel – how the human brain's default mode predicts our resonance to another's suffering », *Neuroscience*, 169, p. 143-148.

OULLIER O. , KELSO J. A. S. , KIRMAN A. P. (2008), « Social neuroeconomics : A dynamical systems perspective », *Revue d'économie politique*, 118 (1), p. 51-62.

OULLIER O. , BASSO F. (2010), « Embodied economics : How bodily information shapes the social

coordination dynamics of decision-making », *Philosophical Transactions of the Royal society B.*, 365 (1538), p. 291-301.

PAGLIERI F. *et al.* (2013), « Heaven can wait. How religion modulates temporal discounting », *Psychological Research*, 77, p. 738-747.

PARDO M. S., PATTERSON D. (2011), « Minds, brains, and norms », *Neuroethics*, 4, p. 179-190.

PARFIT D. (1986), *Reasons and Persons*, Oxford University Press.

PARSONS T. , SHILS E. A. (1951-2000), *Toward a General Theory of Action*, New Brunswick, Transaction Publishers.

PATERNOTTE C., GROSE J. (2012), « Social norms and Game Theory : Harmony or discord ? », *British Journal for the Philosophy of Science*, p. 1-37.

PAULUS F. M., MÜLLER-PINZLER L., WESTERMANN S., KRACH S. (2013), « On the distinction of empathic and vicarious emotions », *Frontiers in Human Neuroscience*, 7 (196), p. 1-5.

PELZMANN L., HUDNIK U., MIKLAUTZ M. (2005), « Reasoning or reacting to others ? How consumers use the rationality of other consumers », *Brain Research Bulletin*, 67, p. 438-442.

PFEIFFER U. J. (2013), « Oxytocin-not always a moral moelcule », *Frontiers in Human Neuroscience*, janvier, 7 (10).

PFEIFER J. H., IACOBONI M., MAZZIOTTA J. C., DAPRETTO M. (2008), « Mirroring others' emotions relates to empathy and interpersonal. Competence », *Children NeuroImage*, 39, p. 2076-2085.

PINKER S. , BLOOM P. (1990), « Natural language and natural selection », *Behavioral and Brain Sciences*, 13, p. 707-784.

PINKER S. (1999) [1994], *L'Instinct du langage*, Paris, Odile Jacob.

PINKER S., JACKENDOFF R. (2005), « The faculty of language : What's special about it ? », *Cognition*, 95, p. 201-236.

POULIN-DUBOIS D., SODIAN B., METZ U., TILDEN J., SCHOEPPNER B., (2007), « Out of sight is not out of mind : Developmental changes in infants' understanding of visual perception during the second year », *Journal of Cognition and Development*, 8, p. 401-421.

POWERS S. T. , TAYLOR D. J. , BRYSON J. J. (2012), « Punishment can promote defection in group-structured populations », *Journal of Theoretical Biology*, 311, p. 107-116.

PROUST J. (2013), *The Philosophy of Metacognition Mental Agency and Self-Awareness*, Oxford University Press.

PROUST J., PACHERIE É. (2008), « Neurosciences et compréhension d'autrui », *in* Poirier Pierre, Faucher Luc (éds.), *Des neurosciences à la philosophie*, Paris, Éditions Syllepse.

RAAFAT R. M., CHATER N., FRITH C. (2009), « Herding in humans », *Trends in Cognitive Sciences*, 13 (10), p. 420-428.

RABIN M. (1993), « Incorporating fairness into game theory and economics », *in* Camerer C. F., Lowenstein G., Rabin M. (éds.), *Advances in Behavioral Economics*, Princeton, Princeton University Press.

RACHLIN M. , RAINERI A. , CROSS D. (1991), « Subjective probability and delay », *Journal of Experimental Analysis of Behavior*, 55 (2), p. 233-244.

RACHLIN M., JONES B. A. (2008), « Social discounting and delay discounting », *Journal of Behavioral Decision Making*, 21 (1), p. 29-43.

RADKE S. , GÜROGLÜ B. , DE BRUIJN E. R. A. (2012), « There's something about a fair split : Intentionality moderates contex-based fairness consideration in social decision-making », *PLoS One*, 17, doi:10.1371.

RAMESON L. T. , MORELLI S. A. , LIEBERMAN M. D. (2011), « The neural correlates of empathy : Experience, automaticity, and prosocial behavior », *Journal of Cognitive Neuroscience*, 24 (1), p. 235-245.

REED K., PESHKIN M., HARTMANN M. J., GRABOWECKY M., PATTON J., VISHTON P. M. (2006), « Haptically linked dyads : Are two motor-control systems better than one ? », *Psychol. Sci.*, 17, p. 365-366.

REID V. M., STRIANO T., IACOBONI M. (2011), « Neural correlates of dyadic interaction during infancy », *Developmental Cognitive Neuroscience*, 1, p. 124-130.

REIMAN M. , BECHARA A. (2010), « The somatic marker framework as a neurological theory of decision-making : Review, conceptual comparisons, and future neuroeconomics research », *Journal of Economic Psychology*, 31, p. 767-776.

RENIERS R. L. E. P. *et al*. (2012), « Moral decision-making, ToM, empathy and the default mode network », *Biological Psychology*, 90, p. 202-210.

REYNOLDS L. E. A. , DAPRETTO M. , IACOBONI M. (2009), « Culture in the mind'smirror : How anthropology and neuroscience can inform a model of the neural substrate for cultural imitative learning », *in* J. Y. Chiao (éd.), *Progress in Brain Research*, Paris, Elsevier.

RICHELL R. A. *et al*. (2003), « Theory of mind and psychopathy : Can psychopathic individuals read the "language of the eyes" ? », *Neuropsychologia*, 41, p. 523-526.

RILLING J. K., GUTMAN D. A., ZEH T. R., PAGNONI G., BERNS G. S., KITS C. D. (2002), « A neural basis for social cooperation », *Neuron*, 35 (2), p. 395-405.

RILLING J. K., SANFEY A. G., ARONSON J. A., NYSTROM L. E., COHEN J. D. (2004), « The neural correlate of theory of mind within interpersonal interactions », *NeuroImage*, 22 (4), p. 1694-1703.

RILLING J. K. , GLEEN M. R. , PAGNONI G. (2007), « Neural correlates of social cooperation and non-cooperation as a function of psychopathy », *Biological Psychiatry*, 61 (11), p. 1260-1271.

RILLING J. K., GOLDSMITH R. D., GLEEN A. L., JAIRAM M. R., ELFENBEIN H. A., DAGENAIS J. E., MURDOCH C. D. , PAGNONI G. (2008), « The neural correlates of the affective response to unreciprocated cooperation », *Neuropsychologia*, 46 (5), p. 1256-1266.

RILLING J. K., KING-CASAS B., SANFEY ALAN G. (2008), « The neurobiology of social decision-making », *Current Opinion in Neurobiology*, 18, p. 159-165.

RILLING J. K. , SANFEY A. C. (2011), « The neurosciences of social decision-making », *The Annual Review of Psychology*, 62, p. 23-48.

RILLING J. K. *et al.* (2012), « Effects of intranasal oxytocin and vasopressin on cooperative behavior and associated brain activity in men », *Psychoneuroendocrinology*, 37, p. 447-461.

RIZZOLATTI G., FABBRI-DESTRO M. (2008), « The mirror system and its role in social cognition », *Current Opinion in Neurobiology*, 18, p. 179-184.

ROELFSEMA P. R., VAN OOYEN A., WATANABE T. (2010), « Perceptual learning rules based on reinforcers and attention », Trends in Cognitive Sciences, 14 (2), p. 64-71.

RUBINSTEIN A. (2003), « "Economics and psychology" ? The case of hyperbolic discounting », *International Economic Review*, novembre, 44 (4).

RUGE H., WOLFENSTELLER U. (2010), « Rapid formation of pragmatic rule representations in the human brain during instruction-based learning », *Cerebral Cortex*, 20 (7), p. 1656-1667.

RUGE H., WOLFENSTELLER U. (2013), « Functional integration process underlying the instruction-based learning of novel goal-directed behaviors », *NeuroImage*, 68, p. 162-172.

RUSTICHINI A. (1999), « Minimizing regret : The general case », *Games and Economic Behavior*, 29, p. 224-243.

RUSTICHINI A. (2008), « Dual or unitary system ? Two alternative models of decision making », *Cognitive, Affective, and Behavioral Neuroscience*, 8 (4), p. 355-362.

RUSTICHINI A. (2009), « Neuroeconomics : What have we found, and what should we search for », *Current Opinion in Neurobiology*, 19, p. 672-677.

SAMAYA H., PESTOV I, SCHMIDT J., BUSH B. J. *et al.* (2013), « Modeling complex systems with adaptive networks », *Computer and Mathematics Applications*, 65 (10), p. 1645-1664.

SAMUELSON L. (1997), *Evolutionary Games and Equilibrium selection*, Cambridge, The MIT Press.

SAMUELSON P. A. (1947), *Foundations of Economic Analysis*, Cambridge, Harvard University Press.

SAMUELSON P. A. (1948), « Consumption theory in terms of revealed preference », *Economica*, New Series, 15 (60), p. 243-253

SANFEY A. G., RILLING J. K., ARONSON J. A., NYSTROM L. E., COHEN J. D. (2003), « The neural basis of economic decision-making in the ultimatum game », *Science*, 300 (5626), p. 1755-1758.

SANFEY A. G., LOEWENSTEIN G., MCCLURE S. M., COHEN J. D. (2006), « Neuroeconomics : Cross-currents in research on decision-making », *Trends in Cognitive Sciences*, 10 (3), p. 108-116.

SANFEY A. G. (2009), « Expectations and social decision-making : Biasing effects of prior knowledge on ultimatum responses », *Mind and Society*, 8, p. 93-107.

SAVAGE L. J. (1954), *The Foundations of Statistics*, New York, John Wiley and Sons, Inc.

SCHELLING T. C. (1960), *The Strategy of Conflict*, Cambridge, Harvard University Press.

SCHMIDT C. (1996), « Paradoxes of rationality in decision-making theory », *in* Arrow K. J., Colombatto E., Perlman M., Schmidt C. (éds.), *The Rational Foundations of Economic Behaviour*, Londres, McMillan.

SCHMIDT C. (2001), *La Théorie des Jeux. Essai d'interprétation*, Paris, PUF.

SCHMIDT C. (2010), *Neuroéconomie*, Paris, Odile Jacob.

SCHÜTZ A. (1953), « Common-sense and scientific interpretation of human action », *Philosophy and Phenomenological Research*, 14 (1), p. 1-38

SCHÜTZ A. (1996) [1935], « Political economy : Human conduct in social life », *in* Wagner H., Psathas G., Kerstein F., *Alfred Schütz Collected Papers*, Kluwer, Dordrecht.

Sciences et société. Les normes en question, collectif (2014), Paris, IHEST/Actes Sud, « Questions vives ».

SEARLE J. (1995, 1998), *La Construction de la réalité sociale*, Paris, Gallimard.

SELTEN R. (1975), « Reexamination of the perfectness concept for equilibrium points in extensive games », *International Journal of Game Theory*, 4 (1), p. 25-55.

SEN A. K. (1970), *Collective Choice and Social Welfare*, San Francisco, Holden-Day.

SHENHAV A., BOTVINICK M. M., COHEN J. D. (2013), « The expected value of control : An integrative theory of anterior cingulate cortex function », *Neuron*, 24 juillet, 79, p. 217-240.

SHAMAY-TSOORY S. G. *et al*. (2013), « Giving peace a chance : Oxytocin increases empathy to pain in the context of the Israeli-Palestinian conflict », *Psychoneuroendocrinology*, 38 (12), p. 3139-3144.

SHAPLEY L. S. (1953), « A value for a n-person game », *in* Kuhn W. H., Tucker W. W., « Contributions of the theory of games », Princeton, Princeton University Press, « Annals of Mathematics ».

SHENHAV A., BOTVINICK M. M., COHEN J. D. (2013), « The expected value of control : An integrative theory of anterior cingulate cortex function », *Neuron*, 79 (2), p. 217-240.

SIEGEL J. Z., CROCKETT M. J. (2013), « How serotonin shapes moral behavior », *Annals of The New York Academy of Sciences*, 1299 (1), p. 42-51.

SIMMEL G. (1908, 1999), *Sociologie*, Paris, PUF.

SIMONELLI E. P. (2009), « Optimal estimation in sensory systems », *in* Gazzaniga M. S. (éd.), *The Cognitive Neurosciences*, Cambridge, The MIT Press, 4e éd., p. 525-538

SINIGAGLIA C., RIZZOLATTI G., « Through the looking glass : Self and others », *Consciousness and Cognition*, 20 (2011), p. 64-74.

SINGER T., CRITCHLEY H. D., PREUSCHOFF K. (2009), « A common role of insula in feelings, empathy and uncertainty », *Trends in Cognitive Sciences*, 13 (8), p. 334-340.

SIP K. E., LYNGED M., WALLENTINA M., MCGREGOR W. B., FRITH C. D., ROEPSTORFFA A. (2010), « The production and detection of deception in an interactive game », *Neuropsychologia*, 48 (12), p. 3619-3626.

SMERILLI A., (2012), « We-thinking and vacillation between frames : Filling a gap in Bacharach's theory », *Theory and Decision*, 73, p. 539-560.

SMITH A. (1937) [1759], *An Inquiry into the Nature and the Causes of the Wealth of Nations*, New York, Cannan.

SMITH A. (1974) [1772], *The Theory of Moral Sentiments*, Oxford, Clarendon Press.

« Social epistemology », *Stanford Encyclopedia of Philosophy*, 2001, nouvelle édition 2006.

SPERBER D. (1996), *La Contagion des idées*, Paris, Odile Jacob.

SPITZER M., FISCHBACHER U., HERRNBERGER B., GRÖN G., FEHR E. (2007), « The neural signature of social norm compliance », *Neuron*, 4 octobre, 56, p. 185-196.

STAHL D. O., WILSON P. W. (1995), « On players' models of other player. Theory and experimental evidence », *Games and Economic Behavior*, 10, p. 218-254.

STANTON S. J., LIENING S. J., SCHULTHEISS O. C. (2011), « Testosterone is positively associated with risk taking in the Iowa Gambling Task », *Hormones and Behavior*, 59, p. 252-256.

STROBEL A. *et al.* (2011), « Beyond revenge : Neural and genetic bases of altruistic punishment », *NeuroImage,* 54, p. 671-680.

SUGDEN R. (2001), « Review of Ken Binmore's evolutionary Social theory Game theory and the Social contract », *The Economic Journal*, 111 (469), p. F213-F243.

SUGDEN R. (1995), « A theory of focal points », *The Economic Journal*, 105 (430), p. 533-550.

SUZUKI S., NIKI K., FUJISAKI S. (2011), « Neural basis of conditional cooperation », *Social Cognitive and Affective neurosciences*, 6 (3), p. 338-347.

TAKAHASHI T. (2005), « Loss of self-control in intertemporal choice may be attributable to logarithmic time-perception », *Medical Hypotheses*, 65 (4), p. 691-693.

TAKAHASHI T. (2006), « Time-estimation error following Weber-Fechner law may explain subadditive time-discounting », *Medical Hypotheses*, 67 (6), p. 1372-1372.

TAKAHASHI T. (2007), « A probabilistic choice model based on Tsallis' statistics », *Physica A*, 386 (1), p. 335-338.

TAKAHASHI T. (2007), « A probabilistic choice model based on Tsallis' statistics », *Physica A*, 386, p. 335-338.

TAKAHASHI T. (2007), « Non-reciprocal altruism may be attributable to hyperbolicity in social discounting function », *Medical Hypotheses*, 68, p. 184-187.

TAKAHASHI T. (2009), « Tsallis' non-extensive free energy as a subjective value of an uncertain reward », *Physica A*, 388, p. 715-719.

TAKAHASHI T. (2009), « Theoretical frameworks for neuroeconomics of intertemporal choice », *Journal of Neurosciences, Psychology, and Economics*, 2 (2), p. 75-90.

TAKAHASHI T. (2011), « Psychophysics of the probability weighting function », *Physica A*, 390 (5), p. 902-905.

TAKAHASHI T., HAN R., NISHINAKA H., MAKINO T., FUKUI H. (2013), « The q-Exponential probability discounting of gain and loss », *Applied Mathematics*, 4 (6), p. 876.

TANAKA S. C., DOYA K., OKADA G., UEDA K., OKAMOTO Y., YAMAWAKI S. (2004), « Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops », *Nature Neurosciences*, 7 (8), p. 887-893.

TARDE G. (1890, 1993), *Les Lois de l'imitation*, Paris, Éditions Kimé.

TAYLOR K. S., SEMINOWICS D. A., DAVIS K. D. (2009), « Two systems of resting state connectivity

between the insula and the cingulate cortex », *Human Brain Mapping*, 30 (9), p. 2731-2745.

THIBIERGE C. (2008), « Au cœur de la norme : la trace et la mesure », *Archives de philosophie du droit*, 51, p. 341-371.

THIRIOUX B., MERCIER M. R., BLANKE O., BERTHOZ A. (2014), « The cognitive and neural time course of empathy and sympathy : An electrical neuroimaging study on self-other interaction », *Neuroscience*, 267, p. 286-306.

TIERCELIN C. (1993), LA PENSÉE-SIGNE, Paris, Jacqueline Chambon.

TODOROV A. , MENDE-SIEDLECKI P. , DOTSCH R. (2013), « Social judgments from faces », *Current Opinion in Neurobiology*, 23, p. 373-380.

TOMLIN D. , KYALI M. A. , KING-CASAS B. , ANEN C. *et al.* (2006), « Agent-specific responses in the cingulated cortex during economic exchanges », *Science*, 312 (5776), p. 1047-1050.

TODD R. M. , ANDERSON A. K. (2013), « Salience, state, and expression. The influence of specific aspects of emotion and attention and perception », *in The Oxford Handbook of Cognitive Neuroscience*, Oxford, Oxford University Press.

TSALLIS C. (2006), « Thermostatistically approaching living systems : Boltzmann-Gibbs or nonextensive statistical mechanics ? », *Physics of Life Reviews* 3 (1), p. 1-22.

TVERSKY A. , KHANEMAN D. (1992), « Advances in prospect theory. Cumulative representation of uncertainty », *Journal of Risk and Uncertainty*, 5, p. 297-323.

VAN DEN BOS W., MCLURE S. M., HARRIS L. T., FISKE S. T., COHEN J. D. (2007), « Dissociating affective evaluation and social cognitive processes in the ventral medial prefrontal cortex », *Cognitive Affective Behavioral Neurosciences*, 7 (4), p. 337-346.

VA N LANGE P. A. M. , JOIREMAN J. , PARKS C. D. , VA N DIJK E. (2013), « The psychology of social dilemmas : A review », *Organizational Behavior and Human Decision Processes*, 120 p. 125-141.

VA N SEGBROECK S. , SANTOS F. C. , LENAERTS T. , PACHECO J. (2011), « Emergence of cooperation in adaptive social networks with behavioral diversity », *in* M. G. Kampis G., I. Karsai, Szathmáry E. (éds), *ECAL 2009, Part I*, LNCS 5777, Berlin, Springer Verlag, p. 434-441.

VEBLEN T. (1898), *The Theory of Leisure, an Economic Study of Institutions*, Londres, MacMillan.

VINCENT N. A. (2011), « Neuroimaging and responsibility assessments », *Neuroethics*, 4, p. 35-49.

VO N NEUMANN J. , MORGENSTERN O. (1953) [1943], *Theory of Games and Economic Behavior*, Princeton, Princeton University Press.

VROMEN J. (2011), « Neuroeconomics. Two camps gradually converging : What can economics gain from it ? », *Int. Rev. Econ.*, 58, p. 267-285.

WALDMANN M. R. , WIEGMANN A. (2010), « A double causal contrast theory of moral intuitions in trolley dilemmas », *in* Ohlsson S., Catrambone R. (éds.), *Proceedings of the 32nd annual conference of the cognitive science society*, Austin (TX), Cognitive Science Society.

WATTS D. J., STOGATH S. H. (1998), « Collective dynamics of “small-world” networks », *Nature*, 393,

p. 440-442.

WALTER H., ABLER B., CIARAMIDARO A., ERK S. (2005), « Motivating forces of human actions. Neuroimaging reward and social interaction », *Brain Research Bulletin*, 67 (2005), p. 368-381.

WEIBULL J. W. (1995), *Evolutionary Game Theory*, Cambridge, The MIT Press.

WIEGMANN A., WALDMANN M. R. (2014), « Transfer effects between moral dilemmas : A causal model theory », *Cognition*, 131 (1), p. 28-43.

WILDMAN W. J., MCNAMARA P. (2008), « Challenges facing the neurological study of religious behavior, belief, and experience », *Method and Theory in the Study of Religion*, 20 (3), p. 212-242.

WILLARD A. K., NORENZAYAN A. (2013), « Cognitive biases explain religious belief, paranormal belief, and belief in life's purpose », *Cognition*, 129, p. 379-391.

WOMELSDORF T., FRIES P. D. (2007), « The role of neural synchronization in selective attention », *Current Opinion in Neurobiology*, 17 (2), p. 154-160.

WONG S. (1978), *The Foundations of Paul Samuelson's Revealed Preference Theory*, Londres, Routledge.

WU A Y., LELIVELD M. C., ZHOUC X. (2011), « Social distance modulates recipient's fairness consideration in the dictator game : An ERP study », *Biological Psychology*, 88, p. 253-262.

XIANG T., LOHRENZ T., MONTAGUE P. R. (2013), « The computational substrates of norms and their violations during social exchange », *The Journal of Neuroscience*, 33 (3), p. 1099-1108.

XUE S.-W., WANG Y., TANG Y.-Y. (2013), « Personal and impersonal stimuli differentially engage brain networks during moral reasoning », *Brain and Cognition*, 81, p. 24-28.

YOSHIDA W., FRISTON K. J., DOLAN R. Y. (2008), « Game Theory in mind », *PLoS, Computational Biology*, 4 (12).

YOSHIDA W., SEYMOUR B., FRISTON K. J., DOLAN R. J. (2010), « Neural mechanisms of belief inference during Cooperative ames », *The Journal of Neuroscience*, 30 (32), p. 10744-10751.

YOUNG P. (2007), « Social norms », *Oxford Discussion Paper*, 307.

YU A. J., DAYAN P. (2003), « Expected and unexpected uncertainty : Ach and NE in the neocortex », *Advanced in Neural Information Processing Systems*, 15, p. 157-164.

YU A. J., DAYAN P. (2005), « Uncertainty, neuromodulation, and attention », *Neuron*, 46 (4), p. 681-692.

ZAK P. J., NACK S. (2001), « Trust and growth », *The Economic Journal*, 111 (470), p. 295-321.

ZAK P. J., FAKHAR A., (2006), « Neuroactive hormones and interpersonal trust : International evidence », *Economics and Human Biology*, 4, p. 412-429.

ZAK P. J. (2006), « The neuroeconomics of trust », *Hendricks Symposium-Department of Political Science*, Paper 9.

ZAK P. J., (2011), « The physiology of moral sentiments », *Journal of Economic Behavior and Organization*, 77, p. 53-65.

ZAK P. J., BARRAZA J. (2013), « The neurobiology of collective action », *Frontiers in Neuroscience*, 7 (211).

ZAKI J., OSCHNER K. (2012), « The neuroscience of empathy : Progress, pitfalls and promise », *Nature Neuroscience*, 15 (5), p. 675-680.

Index des matières

Index des noms

DES MÊMES AUTEURS

OUVRAGES DE CHRISTIAN SCHMIDT

Neuroéconomie, Odile Jacob, 2010.
La Théorie des jeux. Essai d'interprétation, PUF, 2003.
Penser la guerre, penser l'économie, Odile Jacob, 1991.

OUVRAGES DE PIERRE LIVET

La Communauté virtuelle, Éditions de l'Éclat, 1994.
Émotions et rationalité morale, PUF, 2002.
Qu'est-ce qu'une action ?, Vrin, 2005.
Les Êtres sociaux (avec Frédéric Nef), Hermann, 2009.

T

Éditions Odile Jacob

Des idées qui font avancer les idées

Retrouvez tous nos livres disponibles en numérique
sur odilejacob.fr
Retrouvez-nous sur Facebook
Suivez-nous sur Twitter

économie

Comprendre nos interactions sociales

Une perspective neuroéconomique

Ce livre est le résultat d'un dialogue fécond entre un économiste et un philosophe. Il montre dans quelle mesure et avec quelles limites les apports récents des neurosciences permettent de mieux comprendre aujourd'hui nos interactions sociales. Il revisite ainsi les questions, classiques en économie, de la coordination des actions et de la coopération des agents, en intégrant les différents processus émotionnels et cognitifs qui y contribuent, notamment à travers les réseaux sociaux. Il explore également les conditions d'émergence des conventions sociales et le fonctionnement des normes qui régissent les comportements des individus. Ainsi se dégagent des idées nouvelles, qu'il s'agisse de l'interintentionnalité des sujets, de la dynamique intertemporelle qui guide leurs relations, des diverses modalités que peut prendre la confiance, ainsi que des formes de contrôle explicite sur les comportements sociaux des agents.

Christian Schmidt
Pierre Livet

Christian Schmidt est professeur émérite à l'université Paris-Dauphine et président de l'Association européenne de neuroéconomie, qu'il a créée en 2011. Il est membre du centre de recherche Phare de l'université Paris-I. Ses travaux portent sur l'économie de la défense, la théorie des jeux, l'analyse du risque et la neuroéconomie.

Pierre Livet est professeur émérite à l'université d'Aix-Marseille, membre du CEPERC. Ses travaux portent sur l'épistémologie des sciences sociales, l'ontologie des êtres sociaux, la théorie de l'action et celle des émotions.

Avec le soutien du CNL

www.odilejacob.fr