Rationalité et sciences sociales

M. Jon Elster, professeur

L'irrationalité

Ce cours sur l'irrationalité fait suite au cours de l'année précédente sur le désintéressement. Comme dans ce dernier cours, j'ai fait appel à deux corps théoriques distincts, d'un côté les classiques de la pensée, de Montaigne à Tocqueville, et d'un autre la psychologie et l'économie récentes, cherchant les bonnes questions chez les classiques pour essayer ensuite de trouver les bonnes réponses chez les modernes. Ensemble, les deux cours constituent les deux volets d'une critique de la théorie de l'homme économique, ou *homo economicus*. Dans ce qui suit, je propose non pas tant un résumé du cours en tant que tel qu'une explicitation du concept d'irrationalité tel que le cours l'a déroulé.

Sans trop caricaturer, on peut dire qu'un grand nombre d'économistes continuent d'utiliser les deux hypothèses du choix rationnel et des motivations intéressées, qui ont en leur faveur la simplicité et la parcimonie. Dans la mesure où la recherche de vérité doit l'emporter sur la quête de simplicité et dans la mesure où l'hypothèse de l'homme économique est réfutée par les observations empiriques, je défends dans le cours l'idée qu'il faut y renoncer. J'ajoute qu'en faisant appel à d'autres hypothèses (par exemple la théorie des perspectives proposée par Daniel Kahneman et Amos Tversky ou la théorie de l'escompte hyperbolique proposée par R.H. Strotz et élaborée par George Ainslie), on peut éviter l'arbitraire en déduisant de celles-ci des faits nouveaux (les *novel facts* de Imre Lakatos). Le fait que ces hypothèses alternatives ne se laissent pas intégrer dans une théorie unifiée, semblable à cet égard à la théorie du choix rationnel, ne saurait constituer une objection décisive.

Concernant les deux composantes de la théorie économique que je viens de mentionner, la rationalité et les motivations intéressées, il n'est pas vrai de dire que la première implique la seconde. Il s'agit là d'une vue simpliste et fausse, qui comporte souvent un brin de mauvaise foi. Pour attaquer la théorie du choix

rationnel il est sans doute commode de s'en prendre à ses avocats les plus grossiers, pour lesquels il semble en effet aller de soi que l'agent rationnel ne poursuit que son intérêt propre. Or c'est une victoire trop facile, car c'est une position qui n'est défendue par aucun économiste sérieux. En réalité, une motivation désintéressée comme l'altruisme est non seulement compatible avec la rationalité, mais elle l'exige. Si j'alloue une partie de mon revenu à la réduction de la pauvreté dans le tiers-monde, le même souci désintéressé qui m'y conduit doit aussi me faire rechercher la fondation philanthropique qui en fasse le meilleur usage. Si mon argent finit par profiter plus aux fonctionnaires de la fondation — ou aux dictateurs — qu'aux pauvres, on pourra mettre en question non seulement mon altruisme mais également ma rationalité. Le cours explore à cet égard ce que les économistes appellent le « *warm glow* » (dans ma terminologie « l'effet Valmont ») qui correspond au plaisir ressenti dans toute action altruiste.

Il est tout à fait possible que certains comportements d'apparence altruiste soient en effet motivés par un désir d'autosatisfaction. Or il faut ajouter que dans cette hypothèse il faut aussi présupposer l'irrationalité, sous la forme de la duperie de soi-même. Afin d'obtenir la satisfaction intime d'avoir fait le bien, il faut penser avoir agi pour le bien d'autrui. En de tels cas, on ne peut pas garder à la fois l'hypothèse de motivations égocentriques et l'hypothèse de rationalité.

La théorie de l'homme économique comporte aussi une troisième composante : l'hypothèse que chaque agent est parfaitement informé de la situation de tous les autres acteurs. Il sait en particulier qu'ils sont rationnels, leur motivation intéressée, et qu'ils sont eux aussi parfaitement informés. Cette composante est surtout importante dans les jeux stratégiques. Elle est un peu moins essentielle que les deux autres composantes, en ce sens qu'on peut souvent obtenir des prédictions précises même dans le cas d'information imparfaite. Elle a pourtant une place dans le type idéal de l'homme économique.

Une quatrième composante, dont le statut est assez différent, est celle de l'individualisme méthodologique, selon lequel, en gros, l'élucidation complète de la psychologie individuelle ferait disparaître la sociologie. Il n'entre pas dans mon propos ici de défendre cette doctrine, à laquelle je souscris profondément. Il convient néanmoins d'observer que, contrairement à ce que l'on peut lire chez les durkheimiens, l'individualisme méthodologique n'implique ni la rationalité des agents ni leur motivation intéressée. En fait, tout ce que je dis dans ce cours sur les comportements irrationnels présuppose un cadre individualiste. On peut donc retenir cette dernière composante de la théorie de l'homme économique tout en rejetant ou en critiquant les trois autres.

Trois des idées dont je viens de parler — la rationalité, les motivations intéressées, et l'individualisme méthodologique — sont strictement indépendantes les unes des autres. Il existe pourtant une idéologie bien-pensante selon laquelle elles sont étroitement solidaires et, bien entendu, sont toutes à rejeter. Qui défend l'une d'entre elles est accusé, par réflexe, d'accepter les autres. Le présent cours est

largement une critique de la force explicative de la notion de rationalité. Je tiens à souligner que du point de vue normatif, la rationalité est une idée incontournable — transculturelle et transhistorique — au contraire de l'idée de motivations intéressées.

Ce cours opère ainsi un va-et-vient constant entre le normatif et l'explicatif. La force normative de la rationalité sert de correctif permanent aux tendances irrationnelles spontanées. C'est ainsi que l'on peut expliquer tous les dispositifs que nous construisons — ou que la société met à notre disposition — pour combattre la faiblesse de volonté, ainsi que je les ai décrits dans un livre récent, Agir contre soi. Or comme toujours lorsqu'il y a des moyens de correction, il peut aussi y avoir hypercorrection. C'est ce que j'appellerai « surrationalité ». La force normative de la rationalité est si puissante que nous sommes souvent tentés de l'appliquer hors de son domaine naturel. Ainsi la « critique » de la rationalité que je proposerai est en partie une critique au sens kantien, ou peut-être pascalien : « Il n'y a rien de si conforme à la raison que ce désaveu de la raison ».

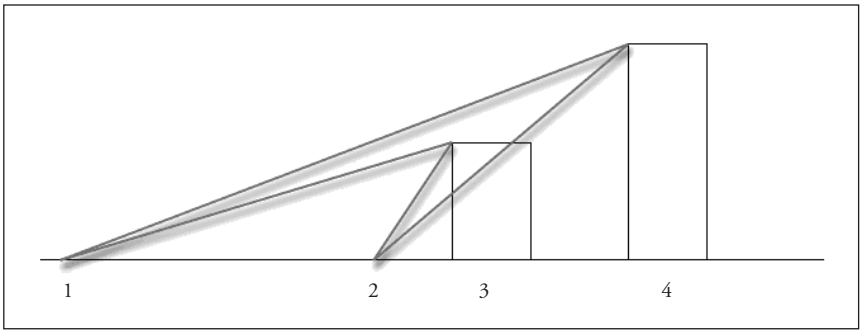
Le cours a suivi en gros le plan suivant :
10 janvier. Introduction générale
17 janvier. Les structures élémentaires de la rationalité
24 janvier. Indétermination et irrationalité
31 janvier. La surrationalité
7 février. Réduction et production de dissonance cognitive
14 février. La faiblesse de volonté
21 février. Les croyances motivées
13 mars. L'escompte du futur
20 mars. Les passions
27 mars. Les passions (suite)
3 avril. Biais et heuristiques
10 avril. La théorie des perspectives
17 avril. Vue d'ensemble

Introduction

L'idée d'irrationalité est d'origine relativement récente, comme l'est aussi celle de rationalité par contraste avec laquelle elle se définit. On affirme souvent que la notion de rationalité instrumentale dans son sens contemporain date des années 1860-1870, quand eut lieu la révolution marginaliste en économie. Il me semble plus exact de remonter à Leibniz, qui concevait le choix par Dieu du meilleur des mondes possibles par analogie avec l'entrepreneur rationnel.

Même si l'idée de la rationalité instrumentale est de date assez récente, le comportement rationnel est de tous les temps. Il en va de même de l'irrationalité. Pour les anciens, cependant, les phénomènes irrationnels étaient liés aux perturbations physiologiques et viscérales. L'irrationalité était essentiellement

« chaude ». Or nous savons aujourd'hui qu'il existe également une irrationalité « froide » qui n'est accompagnée d'aucune perturbation de l'organisme. Pour en donner un exemple, considérons une personne qui marche vers deux bâtiments, qu'on va supposer transparents.



Lorsque la personne se trouve au point 1, à quelque distance des deux bâtiments, leurs grandeurs apparentes relatives et leurs grandeurs réelles relatives sont les mêmes. Le bâtiment le plus grand apparaît comme le plus grand. Or quand la personne arrive au point 2, le bâtiment plus petit domine le plus grand. Il s'agit là d'une simple illusion que le cerveau corrige automatiquement sans que nous y fassions attention.

On peut aussi lire le diagramme de manière différente, en interprétant l'axe horizontal comme une dimension temporelle plutôt que spatiale et les hauteurs des bâtiments comme des biens qui deviennent accessibles aux moments 3 et 4 respectivement. Les lignes obliques représentent maintenant la valeur présente des deux biens aux divers moments du temps. Lorsque la personne contemple le choix entre les biens au moment 1, le bien plus grand lui semble préférable, mais au moment 2 la préférence s'est renversée. Elle va donc choisir le bien moins grand.

De manière qualitative, le diagramme suggère pourtant ce qui sera confirmé plus loin, à savoir qu'il peut y avoir un changement de préférences sans que rien ne se passe, sauf le passage du temps. C'est un cas paradigmatique de l'irrationalité froide. A ce point de l'exposé, l'essentiel est de comprendre que le présupposé traditionnel et implicite selon lequel l'irrationnel découle toujours des passions, au sens large des anciens, n'est plus tenable aujourd'hui.

Dans les trente dernières années, les sciences sociales ont découvert un grand nombre de mécanismes précis qui sont générateurs de comportements irrationnels et dont il est longuement question dans ce cours. Pour pouvoir affirmer l'irrationalité de tel ou tel comportement, il faut *a priori* définir la notion de rationalité dont on se sert. Prenons le comportement de quelqu'un qui se soucie peu des conséquences éloignées dans le temps de ses actions présentes, avec des conséquences négatives pour sa santé, ses finances et ses relations personnelles. D'un point de vue intuitif,

on dirait sans doute que cet individu constitue le paradigme même de l'irrationalité. Du point de vue que j'adopte dans ce cours, il n'en est rien. Je serais prêt à dire qu'il se comporte bêtement, mais la bêtise n'est pas la même chose que l'irrationalité. D'un point de vue objectif, cet individu souffre des effets de ses actions, ce qui n'exclut pas que de son point de vue subjectif il fasse ce qui lui semble le mieux. Dans ce cours, j'adopte une définition résolument subjective des notions du rationnel et de l'irrationnel, non pas que cette définition soit plus « correcte » qu'une autre, idée qui d'ailleurs n'a pas de sens, mais simplement parce qu'elle me semble la plus utile à mes fins.

En ce qui concerne ces fins, elles sont surtout explicatives. La théorie du choix rationnel suggère des hypothèses dont on peut se servir pour rendre compte des comportements observés. De même, l'intérêt des mécanismes de l'irrationalité est de fournir des outils pour l'explication de l'action. Cela dit, la théorie du choix rationnel est aussi, et même d'abord, une théorie normative. Elle dicte à l'agent ce qu'il doit faire afin de réaliser ses projets au mieux possible. Une fois établie cette prescription, l'observateur peut la transformer en prédiction. En posant comme hypothèse explicative que l'individu dont il s'agit est en effet rationnel, on vérifie celle-ci en comparant son comportement observé avec le comportement que recommande la théorie.

Il importe de voir que cette vérification fournit un critère nécessaire mais non suffisant de la rationalité du comportement. Autrement dit, même un comportement qui est conforme aux prescriptions de la théorie du choix rationnel pourrait être le résultat d'un mécanisme irrationnel. Ou bien, pour le dire encore autrement, la rationalité ou l'irrationalité d'une action n'est pas un attribut de l'action elle-même mais du processus qui l'engendre. Le cours explore également les mécanismes susceptibles de « mimer » la rationalité (c'est-à-dire où tout se passe « comme si » l'agent était animé par la rationalité subjective, même lorsque l'on peut démontrer que tel n'est pas le cas), tels que la sélection naturelle. On rencontre parfois l'idée de la rationalité des émotions, fondée sur l'efficacité des réactions émotionnelles et quasi automatiques aux situations dangereuses. Or à mon avis il ne faut pas confondre le caractère adaptatif de ces réactions et leur prétendu caractère rationnel. Le fait que la sélection naturelle ait produit des comportements qui sont souvent les mêmes que ceux qu'aurait choisis un agent rationnel ne prouve en rien qu'il s'agit d'actions rationnelles. De manière plus importante, je montre dans les deux conférences sur les passions que la simulation de la rationalité par la sélection naturelle risque d'être imparfaite, et que les réactions émotionnelles au danger enfreignent souvent les normes de la rationalité. Le comportement des gouvernements occidentaux à la suite du 11 septembre 2001 en offre sans doute un exemple.

A mon avis, l'influence de la sélection naturelle sur la capacité à faire des choix rationnels délibérés est beaucoup plus importante que l'existence en nous, produite par l'évolution, de réactions automatiques capables de simuler la rationalité. Il s'agit en quelque sorte d'une distinction entre la vente en gros et la vente au détail.

Dans un environnement complexe, la capacité à former des croyances bien fondées et à opérer des arbitrages cohérents entre les diverses fins qui s'imposent est essentielle. Le cours traite à cet égard du problème des poids relatifs qu'il convient d'accorder aux biens proches dans le temps et aux biens plus éloignés. Il existe maintenant une littérature considérable suggérant fortement que cet arbitrage est typiquement incohérent, comme je l'ai indiqué dans le diagramme de tout à l'heure. On pourrait penser que cette incohérence constitue un handicap sévère pour l'organisme, mais apparemment il n'en est rien.

En second lieu, on peut faire appel à une analogie sociale de la sélection naturelle, notamment à la concurrence du marché. Ainsi, l'on pourrait réconcilier l'hypothèse de l'agent rationnel avec certaines modélisations de la rationalité qui lui imposent un fardeau cognitif très lourd. Admettons que des femmes et des hommes de chair et de sang soient incapables d'accomplir les calculs pour lesquels les chercheurs ont besoin de plusieurs pages de mathématiques avancées. Dans la réalité, les agents sociaux utilisent des critères de décisions souvent très grossiers. Ils coupent la poire en deux, imitent le voisin, recherchent un seuil de satisfaction plutôt qu'un maximum, etc. On peut néanmoins affirmer qu'ils se comportent comme s'ils étaient capables de calculs sophistiqués, puisque ceux qui s'écartent du comportement optimal sont éliminés par le marché. L'irrationalité, dans cette perspective, ne serait qu'un phénomène passager et éphémère.

Ce raisonnement, qui est au fondement de l'hypothèse de la rationalité « comme si » adoptée par l'économie moderne, est pourtant extrêmement faible. Il y a de nombreuses disanalogies entre la sélection naturelle et la concurrence du marché, dont la plus importante découle peut-être du fait que l'environnement économique change trop vite pour que la concurrence ait le temps d'éliminer les agents qui auraient choisi des critères de décision sous-optimaux. De plus, la plupart des choix des agents sociaux ne se font pas dans le cadre d'une situation concurrentielle dont les perdants seraient éliminés. Finalement, dans l'histoire du monde, les sociétés de marché constituent un phénomène exceptionnel.

Il convient ainsi de bien distinguer les deux propositions suivantes. D'une part, une proposition empirique : ni les émotions ni le marché ne tendent en général à produire, de manière systématique, des comportements adaptatifs. D'autre part, une proposition conceptuelle : l'adaptation, qui est un fait objectif, n'a rien à voir avec la rationalité, qui est un fait subjectif. Dans la suite de ce cours, j'insisterai toujours sur le caractère radicalement subjectif de la rationalité, même si de temps en temps il me faudra faire face aux implications contre-intuitives de cette approche. N'est-il pas absurde, par exemple, d'affirmer que le toxicomane soit rationnel ? Je répondrai que ce n'est pas absurde, même si, le plus souvent, c'est faux.

Supposons maintenant que nous nous proposions de vérifier l'hypothèse de l'homme économique dans une situation précise, et qu'elle s'avère fausse. Puisque l'hypothèse comporte les diverses composantes qu'on a vues, il est difficile de savoir ce qu'il faut en conclure. Selon la thèse dite de Duhem-Quine, nos

hypothèses ne confrontent pas le monde une par une, mais en bloc et de manière simultanée. Même lorsqu'une expérience est conçue afin de sonder une hypothèse précise, un résultat négatif n'infirme pas forcément celle-ci, car il se peut que le coupable soit l'une des hypothèses auxiliaires adoptées implicitement ou explicitement par le chercheur. Dans le cas qui nous concerne ici, il est parfois difficile de savoir si les résultats d'une expérience donnée réfutent l'hypothèse de la rationalité ou celle d'une motivation intéressée. Dans mon cours de l'année dernière, j'ai cité le comportement électoral comme un exemple possible. Pour expliquer pourquoi les électeurs se donnent la peine de voter, on peut comprendre leur vote comme un don à la société. Ce serait l'explication par le désintéressement. On pourrait également interpréter leur décision de se déplacer pour voter comme l'effet d'une sorte de pensée magique. Chaque individu se dirait, de manière plus ou moins consciente, que s'il vote, d'autres avec les mêmes caractéristiques que lui le feront également. « Si je vote, ceux qui sont comme moi voteront aussi. » Ce serait l'explication par l'irrationalité.

Je voudrais dire deux mots sur le rôle de l'inconscient dans les phénomènes irrationnels. Il me semble évident que les processus inconscients y jouent un rôle important. La réduction de la dissonance cognitive et la formation de croyances motivées sont des processus qui se déroulent « dans le dos » de l'agent, sans qu'il en soit conscient. Il ne s'agit pas là d'un constat empirique, mais d'une vérité conceptuelle. La nature exacte de ces processus nous est largement inconnue. On peut déduire leur existence à partir de leurs effets, un peu comme on a déduit l'existence de la matière noire dans l'univers.

Il serait tentant de conclure, avec le premier Freud, que les processus de l'inconscient sont sujets au Principe de Plaisir. Ainsi l'on adopte parfois une opinion ou une croyance non pas parce qu'elle est appuyée par les observations ou les expériences, mais pour le plaisir qu'on en tire. Le cours en examine de nombreux exemples. La propriété fondamentale de ces processus inconscients consiste en ce qu'ils sont dirigés vers la satisfaction immédiate. L'inconscient est incapable de reculer pour mieux sauter. Agir en vue d'une fin éloignée présuppose que celle-ci soit représentée sur l'écran mental de l'agent, ce qui justement est un trait constitutif de la conscience. Attribuer cette capacité à l'inconscient serait donc en faire une conscience. Je citerai pourtant aussi des cas dans lesquels un agent semble adopter une croyance qui ne correspond ni à ce qu'il a de bonnes raisons de croire ni à ce qu'il désire être le cas. Qu'on pense par exemple à la jalousie d'Othello ou à celle du Narrateur chez Proust. Pour comprendre ces comportements, peut-on faire appel au second Freud, celui de Au-delà du principe de plaisir ? A mon avis, l'idée de la pulsion de mort est trop spéculative pour être d'une utilité quelconque. Le cours n'offre pas de meilleure alternative, malheureusement.

Une autre question très difficile concerne l'existence de croyances et d'émotions inconscientes. Dans les phénomènes de mauvaise foi ou de duperie de soi-même, il semble y avoir non seulement une croyance motivée, mais également la

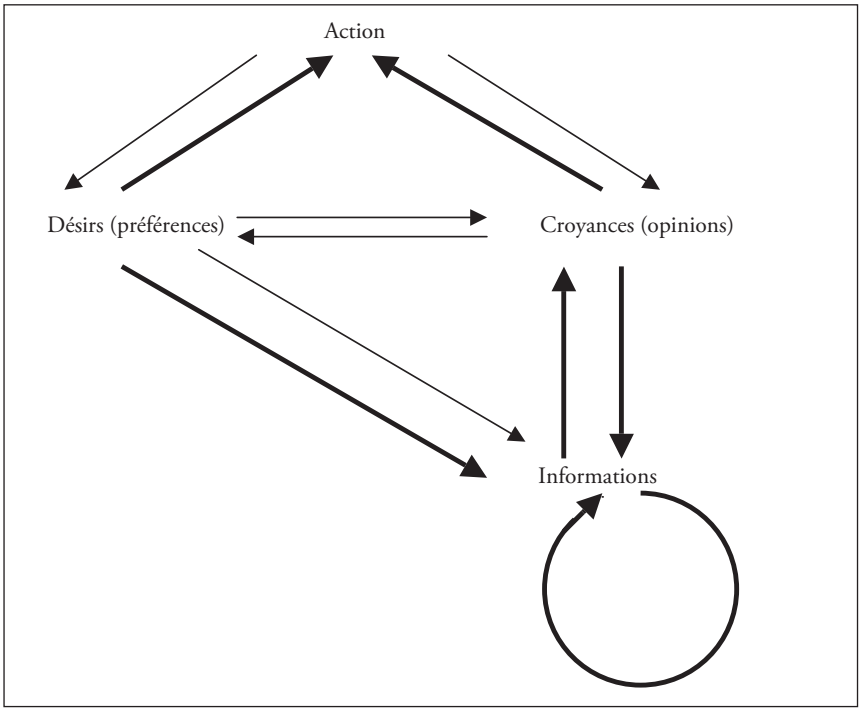
suppression d'une croyance initiale plus pénible encore que mieux fondée. Or comme on le sait depuis Sartre, nommer l'instance mentale qui effectue cette suppression cause des problèmes formidables.

De plus, il faut se demander si les croyances inconscientes ont en commun avec les croyances conscientes de pouvoir servir de prémisses à l'action. Considérons l'exemple hypothétique suivant. Un homme se ment à lui-même sur la fidélité de son épouse, ayant supprimé la conscience du fait qu'elle le trompe avec son meilleur ami. Afin qu'il puisse rester ignorant au niveau de la conscience, son inconscient l'empêche de se promener dans les parties de la ville où il risquerait de rencontrer son épouse avec son amant. En principe, on pourrait tester l'hypothèse d'une croyance inconsciente en annonçant au sujet que l'amant de sa femme va se trouver en tel lieu, où le sujet se rend régulièrement, un jour donné, pour voir s'il évite d'y aller. Dans la conférence sur les croyances motivées on voit que certaines expériences psychologiques suggèrent la possibilité d'une telle manipulation des sujets par leur inconscient. Tant que le problème de l'instance de la suppression n'est pas résolu, l'interprétation de ces résultats reste pourtant fragile. L'idée d'émotion inconsciente est ambiguë. Il y a des émotions qui s'ignorent, et des émotions qu'on supprime. Un observateur peut constater la colère ou l'amour chez une personne qui n'a pas elle-même conscience de ressentir ces émotions. Dans une culture qui n'a pas conceptualisé la notion de dépression, comme c'est apparemment le cas à Tahiti, un jeune homme dont l'amie l'a quitté pour un autre et qui exhibe toutes les signes cliniques de la dépression, dira simplement qu'il est « fatigué ».

Les émotions qu'on supprime mais qui persistent dans l'inconscient présentent un problème plus aigu. Nous avons tous, sans doute, observé la transmutation de l'émotion de l'envie en indignation. Un observateur constate sans difficulté la persistance de l'envie, par le ton des remarques dérogatoires que fait le sujet envieux sur l'objet de son émotion, mais le sujet lui-même se voit dans un état de juste colère. Nous avons affaire dans ce cas non pas à une simple ignorance, mais à une ignorance motivée. La nature hideuse de l'envie induit un désir de la supprimer ou de la transmuer, mais même reléguée à l'inconscient, elle continue d'exercer une influence causale sur le comportement. On ne comprend pas très bien comment cela se fait, mails il est difficile de nier l'existence du phénomène.

Voici le schéma de base qui sert de cadre conceptuel pour le cours tout entier.

Les flèches épaisses ont une double interprétation, puisqu'elles représentent à la fois des relations de causalité et des relations d'optimalité. Considérons les rapports entre action, désirs et croyances. D'une part, les flèches indiquent que l'action choisie est le meilleur moyen de réaliser les désirs de l'agent, étant donné ses croyances. D'autre part, elles indiquent que ces désirs et ces croyances constituent les causes de l'action. Nous avons vu que Max Weber, en soulignant uniquement l'optimalité de l'action rationnelle, a sous-estimé l'importance des relations causales.



Les flèches minces, en revanche, représentent des relations causales qui ne sont pas en même temps des rapports d'optimalité. Dans le cours, ces relations causales ont évidemment une grande importance. Sans entrer dans tous les détails, considérons simplement la flèche qui va des désirs vers les croyances. Cette influence causale équivaut en gros à prendre ses désirs pour des réalités. On forme la croyance que le monde est tel qu'on voudrait qu'il soit. Je dis « en gros », puisqu'on a déjà vu des cas, comme celui de la jalousie d'Othello, dans lesquels l'agent tend à former les croyances qu'il a intérêt à trouver fausses.

Le schéma représente l'explication non seulement d'une action, mais également des croyances de l'agent et de sa recherche d'information. De manière plus précise, ce dernier explanandum comprend la quantité de ressources — que ce soit en temps ou en argent — que l'agent consacre à l'acquisition de nouvelles informations, en sus de celles qu'il possède déjà. Cette variable, souvent négligée dans l'analyse du choix rationnel, est d'une importance fondamentale. On verra notamment que plusieurs formes d'irrationalité ont leurs origines dans un investissement soit insuffisant soit excessif dans l'acquisition d'information.

L'acquisition d'information est une action ou un ensemble d'actions. Donc de manière générale, l'action principale se double d'une action secondaire ou préalable, sauf si l'agent décide de ne recueillir aucune information supplémentaire. Dans

certains cas, l'action principale et l'action secondaire coïncident. Dans la guerre, dit Napoléon, « on s'engage et puis on voit ». Supposons qu'un général, avant de livrer une bataille, cherche à déterminer l'esprit combatif des troupes adverses. Pour y parvenir, la méthode la plus sûre est souvent d'engager le combat. Dans le cas typique, il s'agit pourtant d'actions distinctes. L'achat d'une voiture est autre chose que les visites chez les vendeurs d'automobiles et la lecture des brochures.

En revanche, la formation de croyances ne constitue pas une action. On ne peut pas se décider à croire, même dans le cas où avoir une certaine croyance serait avantageuse. Considérons par exemple le cas du fumeur qui voudrait arrêter de fumer mais qui n'y arrive pas. Il sait que s'il croyait que le risque de développer un cancer du poumon était certain, il arrêterait. Il a donc intérêt à y croire. Or les croyances se commandent en amont, par les raisons qui les justifient, non pas d'en aval, par les conséquences qui en découlent. Comme l'a démontré Bernard Williams dans un article justement célèbre,

> On ne peut pas à la fois croire que **p** et croire que la croyance que **p** est une conséquence de la décision de croire que **p** (Bernard Williams, « Deciding to believe », in *Problems of the Self).*

Cela dit, comment comprendre la formation des croyances ? Au fond, c'est un processus passif. Sartre dit quelque part qu'on tombe en mauvaise foi comme on tombe en sommeil, et c'est vrai aussi, je pense, pour les croyances ordinaires. On se trouve avoir telle ou telle opinion.

Dans cette perspective, l'idée de croyances rationnelles pourrait sembler mystérieuse. A mon avis, elle est étroitement liée aux qualités de jugement et de bon sens. Chez celui qui possède ces qualités, la synthèse spontanée des diverses éléments d'information, de pertinence et de fiabilité souvent très variables, se fait d'une manière qui accorde à chacune d'entre elles son poids approprié. On peut citer ici les observations d'un économiste éminent, Paul Krugman, sur l'ancien directeur de la Réserve Fédérale aux Etats-Unis Alan Greenspan. Plutôt que de s'appuyer sur des modèles formels de l'économie,

> Greenspan avait la capacité de deviner, à partir de données fragmentaires et parfois contradictoires, la direction du vent économique (Paul Krugman, *New York Times,* octobre 28 2005).

La formation de croyances rationnelles est ainsi une question de capacités personnelles et intimes, dont le possesseur lui-même ignore le mode d'opération, plutôt que de procédures mécaniques susceptibles d'être enseignées et transmises.

Certes, cette proposition est controversée. Les spécialistes des sciences de la décision proposent toute une gamme de techniques qui sont supposées permettre la formation et la mise à jour de croyances rationnelles. A mon avis, ces idées n'ont pourtant aucune réalité psychologique. Comme je ne suis pas moi-même spécialiste en la matière, ce jugement pourrait sembler téméraire. Le cours essaie dans une certaine mesure de le justifier.

Pour revenir au schéma, les antécédents directs de l'action sont les désirs (ou préférences) et les croyances (ou opinions) de l'agent. Dans ma conception, les croyances et les opinions sont de nature exclusivement positive. Même si l'on dit couramment, « A mon avis l'avortement est inacceptable », je compte de telles propositions comme l'expression d'une préférence plutôt que d'une opinion. Pour éviter tout malentendu, il convient aussi de préciser que je n'utilise pas le mot « préférence » au sens d'un simple goût, ni le mot « désir » au sens d'une impulsion plus ou moins violente. Ce sont des termes techniques qui couvrent toutes sortes de motivations, hédoniques, esthétiques, éthiques ou autres.

Il vaut peut-être la peine de s'attarder un instant sur deux différences entre la notion de désir et celle de préférence. Les préférences mettent nécessairement en jeu deux objets ou plusieurs, pour les comparer, tandis qu'un désir porte sur un seul objet et ne comporte en lui-même aucun élément comparatif. Ainsi on peut parler d'un renversement de préférences, mais seulement d'un changement de l'objet du désir. Cette distinction va s'avérer importante dans les analyses de ce que l'on peut appeler l'irrationalité diachronique.

Un système de préférences est susceptible d'être incohérent, si par exemple on préfère un objet X à un autre objet Y, l'objet Y à l'objet Z, et enfin Z à X. Un désir est incohérent si la description de son objet comporte une contradiction, comme c'est le cas du désir d'être présent à ses propres funérailles pour y entendre son oraison funèbre. Cette distinction est pertinente pour les analyses de l'irrationalité synchronique.

J'opère également une distinction dans le cours entre préférences substantielles et préférences formelles. Les premières expriment l'attitude de l'agent envers des objets spécifiques, comme une préférence pour les oranges sur les pommes ou la préférence pour un candidat politique sur un autre. Les dernières expriment l'attitude envers le temps et envers le risque. On peut ainsi préférer un bien moindre immédiat à un bien plus important mais futur. Je parlerai alors d'impatience. Un agent peut aussi avoir une préférence pour l'action immédiate par rapport à une action différée. Dans ce cas, je parlerai d'urgence. Enfin, on observe souvent l'aversion pour le risque, quand un agent préfère un bien sûr à un bien incertain ayant une valeur attendue plus élevée.

Pour illustrer :

L'impatience : l'agent préfère 100 euros aujourd'hui à 200 euros dans un an.
L'urgence : l'agent préfère agir aujourd'hui pour obtenir 100 euros après-demain plutôt qu'agir demain pour obtenir 200 euros après-demain.
Le risque : l'agent préfère 100 euros à une loterie qui lui donne ou bien 50 euros avec une probabilité de 50 % ou bien 200 euros avec une probabilité de 50 %.

Tandis que l'impatience et le risque sont des phénomènes bien connus, l'urgence l'est moins. Dans le cours, je défends néanmoins l'idée que dans les choix faits sous l'impulsion de l'émotion, l'urgence est susceptible de prendre une importance considérable.

Les croyances sont ou bien factuelles ou bien causales. Autrement dit, elles portent sur l'existence des diverses actions qu'on pourrait choisir ainsi que sur les conséquences du choix de l'une d'entre elles. On risque de mal choisir faute d'avoir assez réfléchi aux conséquences à long terme de chacune des actions possibles, mais aussi faute d'avoir parcouru une gamme d'options suffisamment large. La distinction est importante surtout en ce qui concerne la recherche d'informations supplémentaires. Il y a souvent un arbitrage entre l'exploration en profondeur des conséquences du choix de l'une des options connues et l'exploration en extension du champ des options. Cet arbitrage est pourtant sujet à une incertitude profonde.

Les croyances qui portent sur les conséquences de l'action sont susceptibles d'avoir deux composantes. D'une part, l'agent peut croire que s'il fait A, une des conséquences X, Y ou Z va se produire, tandis qu'il peut exclure les conséquences V et W. D'autre part, il peut assigner une probabilité numérique précise à chacune des conséquences. Comme toute probabilité, il s'agit d'une évaluation subjective, même si elle peut s'appuyer en partie sur des fréquences objectives. Si la croyance comporte la première composante mais non pas la seconde, nous avons une situation d'incertitude, tandis que la présence des deux composantes définit une situation de risque.

Pourtant il convient de nuancer un peu. Dans la définition technique de l'incertitude, on suppose qu'exactement une des conséquences possibles va se produire. Elles sont mutuellement exclusives et conjointement exhaustives. Pour que l'agent puisse faire cette appréciation très précise, il faut évidemment qu'il ait une connaissance très approfondie de la situation. Dans la pratique, on imagine mal qu'il ne puisse pas s'appuyer sur cette connaissance afin de former une opinion sur la probabilité relative des diverses conséquences. Même s'il est incapable d'assigner des probabilités quantitatives précises, il peut du moins conclure que telle conséquence est plus probable que telle autre.

Par contraposition, si l'agent est vraiment incapable de dire quoi que ce soit sur la probabilité relative des conséquences, on ne peut pas lui imputer une appréciation précise des conséquences possibles. Selon la formule désormais célèbre de Donald Rumsfeld, la situation peut comporter des « inconnus inconnus », unknown unknowns, qui viennent en sus des « inconnus connus» dont on connaît la nature tout en ignorant leur probabilité. Dans la présence d'inconnus inconnus, il convient de parler d'ignorance plutôt que d'incertitude. Le réchauffement climatique en est sans doute un bon exemple. Les effets lointains et indirects du réchauffement sont susceptibles, et même presque certains, de prendre des formes dont nous n'avons aujourd'hui aucune idée.

La pertinence de ces questions pour le thème du cours est double. D'une part, les phénomènes d'incertitude et d'ignorance affaiblissent la force normative et prédictive de la théorie du choix rationnel. Puisque l'agent fait son choix en fonction des conséquences probables des diverses options, une connaissance moins complète de celles-ci limite sa capacité à faire un choix rationnel. Même s'il est parfois possible

d'exclure certaines actions comme étant manifestement irrationnelles, il aura souvent l'embarras du choix parmi celles qui restent. Ainsi il faudrait substituer à la notion de choix rationnel celle, plus faible, de choix non irrationnel.

D'autre part, et il s'agit là d'une implication de plus grande portée, l'incertitude et l'ignorance sont des sources profondes d'irrationalité. L'esprit humain a horreur du vide. Il nous est très difficile d'accepter le fait que nous n'avons pas suffisamment d'information pour avoir une opinion sur un sujet donné. Dans une boutade amusante, Albert Hirschman caractérise une certaine culture latino-américaine par le besoin d'avoir « une opinion ferme et instantanée sur n'importe quel sujet ». Même si ce trait est en l'occurrence culturel, il s'agit aussi d'une tendance tout à fait universelle. Selon Montaigne,

> Il s'engendre beaucoup d'abus au monde : ou pour dire plus hardiment, tous les abus du monde s'engendrent, de ce, qu'on nous apprend à craindre de faire profession de nostre ignorance ; et sommes tenus d'accepter, tout ce que nous ne pouvons refuter (Montaigne, *Essais* III. XI.).

Le cours se penche longuement sur les manifestations de cette crainte d'admettre notre ignorance ou, dans le langage des psychologues, du manque de tolérance de l'ambiguïté.

Dans le schéma de l'action rationnelle, les croyances et les opinions ont uniquement une valeur instrumentale. Elles servent à rendre plus probable ou moins coûteuse la réalisation des fins de l'agent. Sans doute sont-elles aussi parfois sources de satisfaction intrinsèque, comme lorsqu'on a une bonne opinion de soi, mais cet avantage n'entre pas parmi les éléments du schéma.

Il est vrai que certains chercheurs ont proposé l'idée selon laquelle un agent rationnel cherche l'arbitrage optimal entre une opinion plaisante ou agréable et une opinion bien fondée. Puisque le plaisir constitue la fin ultime de toute action, il serait irrationnel, selon eux, de négliger les plaisirs que l'on peut tirer de la croyance que le monde est tel qu'on voudrait qu'il le soit. Celui qui ignore les signes d'un cancer naissant aura peut-être une espérance de vie plus courte, mais en revanche il aura vécu, supposons le, avec moins d'angoisse.

Les multiples absurdités de ce raisonnement sont sans doute évidentes, mais il sera néanmoins utile de les épeler. En premier lieu, il faudrait évidemment que le choix d'une opinion agréable, plutôt que d'une opinion bien fondée, soit un choix inconscient. Afin de tirer du plaisir d'une croyance agréable (le « warm glow ou effet Valmont »), il faut croire qu'elle est bien fondée. En deuxième lieu, il n'y a aucune raison — empirique ou théorique — de penser que l'inconscient soit capable de faire les arbitrages qui s'imposent. Comme j'en ai fait la remarque plus haut, imputer à l'inconscient la capacité de calculer, c'est le faire trop semblable à la conscience. En dernier lieu, la prémisse selon laquelle le plaisir est la fin ultime de toute action est indéfendable. Celui qui se bat pour une cause ne le fait pas pour son plaisir personnel, même si celui-ci est susceptible d'être augmenté par la surestimation des chances de victoire.

L'acquisition d'information est guidée par ce qu'on appelle « la règle d'arrêt rationnelle ». Avant de commencer la collecte d'information, on doit définir les conditions dans lesquelles on arrêtera de chercher pour passer à l'action, conditions qui dépendent à la fois des désirs de l'agent et de ses croyances.

Considérons d'abord comment l'investissement dépend des bénéfices et des coûts attendus de l'information. En ce qui concerne les bénéfices, on frôle le paradoxe, car comment peut-on déterminer la valeur d'informations supplémentaires sans déjà les posséder ? Dans les situations qui se répètent régulièrement, l'expérience peut nous guider. Ainsi les médecins ont des connaissances très précises de l'accroissement de la probabilité de détection d'un cancer qui se produit avec chaque test supplémentaire. Dans les situations sans précédent, ou ayant seulement des précédents partiels, il est plus difficile et parfois même impossible de déterminer la valeur attendue de la recherche. Je reviendrai sur ce point la semaine prochaine.

Il est parfois possible de résoudre cette difficulté par une sorte de *tâtonnement*, représente par la boucle du diagramme. Supposons que je sois parti cueillir des champignons dans une région que je connais mal. Je sais qu'en général les champignons poussent en groupes, mais j'ignore la distribution de ceux-ci sur le terrain. La question qui se pose est la suivante: quand dois-je arrêter de chercher et commencer, tant bien que mal, à cueillir ? Même s'il n'y a pas de réponse générale, la solution peut se présenter d'elle-même si je tombe sur une concentration de champignons tellement dense que j'en aurai assez pour remplir mon panier. De manière semblable, on met parfois fin à des expériences médicales avant le temps prévu si le traitement expérimental s'avère rapidement si efficace qu'il serait contraire à l'éthique de le refuser au groupe de contrôle.

Les coûts d'acquisition de l'information se divisent en coûts directs et en coûts d'opportunité. Si l'on va de magasin en magasin pour acheter un produit donné le moins cher possible, il faut tenir compte du prix du taxi ou du ticket de métro. Il y a ensuite le coût d'opportunité, qui est la valeur de la meilleure utilisation alternative du temps consacré à la collecte d'information. Même si le prix du ticket de métro est inférieur au gain brut escompté par l'achat au prix le plus bas, la demi-heure de voyage aurait pu être consacrée à des activités qui, pour l'agent, ont plus de valeur que le gain net. Comme on le voit dans le cours, la négligence des coûts d'opportunité est susceptible d'être source d'irrationalité.

Si nous passons à l'impact des préférences sur l'acquisition d'information, considérons d'abord l'impact des préférences formelles. Supposons que nous ayons affaire à un agent « myope », mot que j'utilise systématiquement pour nommer un agent qui attache peu d'importance aux conséquences éloignées dans le temps de son choix présent. Dans l'achat d'un bien de consommation durable comme une voiture ou une machine à laver, cette personne n'a pas intérêt à passer beaucoup de temps à comparer la durée de vie des diverses marques. De même, l'attitude envers le risque peut influer sur les ressources qu'on investit pour déterminer les taux d'accident des différentes compagnies aériennes.

Les préférences substantielles façonneront également l'acquisition d'informations supplémentaires. Considérons deux personnes qui ont été licenciées et qui cherchent un nouvel emploi, dans le cadre d'un régime d'assurances-chômages généreuses qui leur permettent de maintenir un niveau de vie assez élevé. L'une d'entre elles ne s'intéresse au travail que pour le revenu qu'il lui procure, tandis que pour l'autre avoir un emploi est une condition essentielle du respect de soi, sans lequel elle tire peu de plaisir de ses activités de loisir. Cette deuxième personne investira certainement un plus grand effort que la première dans sa quête d'un emploi, en cherchant des informations précises et détaillées sur le marché du travail qui lui permettent de former des croyances bien fondées sur ses chances d'être embauché.

Comme j'en ai déjà fait la remarque, un impact *direct* des préférences sur les croyances n'est pas compatible avec les principes de la rationalité. On vient de voir qu'un impact *indirect*, par l'intermédiaire de la collecte d'information, n'a en soi rien d'irrationnel. Si vous regardez encore une fois le diagramme, vous constatez pourtant qu'il y a aussi une ligne mince — donc un mécanisme causal non rationnel — qui part des désirs pour arriver à l'information. Parfois, la formation de croyances motivées s'opère en effet par un mécanisme plus subtil que la simple tendance à prendre ses désirs pour des réalités. Au lieu de la règle d'arrêt rationnelle, on peut adopter une « règle d'arrêt hédonique » et cesser la collecte d'information lorsque la croyance justifiée par les données accumulées est celle qu'on aimerait croire vraie. Il semble par exemple que Gregor Mendel, le père de la génétique quantitative, ait utilisé ce principe dans ses recherches statistiques. On observe un comportement apparenté chez les médecins qui arrêtent leur examen quand ils ont identifié une cause, sans se demander s'il pourrait y en avoir d'autres.

J'en arrive maintenant à une question que vous vous êtes sans doute posée, à savoir le rôle unique et spécial des désirs dans l'analyse de l'action rationnelle. Comme vous le constatez, il n'y a aucune flèche épaisse qui aboutisse aux désirs. Les désirs constituent les données primitives à partir desquelles se construisent les diverses optimisations dont je viens de parler. Puisqu'ils servent de critères de rationalité, ils ne sont pas susceptibles d'être évalués comme étant eux-mêmes plus ou moins rationnels.

Certains trouveront sans doute bizarre l'affirmation qu'on peut être à la fois bête et rationnel. Le paradoxe disparaît pourtant dans le contexte explicatif, dans lequel il s'agit uniquement de comprendre le comportement de l'agent à partir de la seule hypothèse de la rationalité subjective. Un agent rationnel chercherait si nécessaire à ajuster ses croyances, en acquérant des informations supplémentaires, mais il n'a aucune incitation d'ajuster ses désirs.

On peut néanmoins qualifier un désir ou un système de préférences d'irrationnel s'il est incohérent. J'ai déjà donné quelques illustrations de cette idée. Un exemple plus complexe, brièvement mentionné ici et dont je reparle longuement dans le cours, concerne les renversements de préférence à la suite du passage du temps. Or dans de tels cas, il ne s'agit pas d'évaluer un désir ou un ordre de préférence d'un point de vue externe, mais simplement de constater leur incohérence interne.

Le cours couvre également les cas de figure dans lesquels l'agent est piégé dans et par ses croyances, semblable à l'agent myope qui est piégé dans et par son horizon temporel court, ainsi que de « l'ignorance pluraliste », c'est-à-dire les comportements collectifs qui se maintiennent par les croyances fausses, stables et s'auto-justifiant qu'ont les agents sociaux les uns des autres. Pour simplifier il s'agit d'une situation où on suppose que la croyance ou le désir en question est peu répandu mais qu'il y une croyance très répandue qu'ils sont très répandus.

J'explique dans le cours la logique de l'ignorance pluraliste par des exemples qui permettent aussi d'introduire une application importante de la théorie du choix rationnel, à savoir la théorie des jeux. A travers les exemples du Dilemme du Prisonnier ou du Jeu de l'Assurance, je propose de distinguer l'irrationalité non seulement de la bêtise, mais également de la malchance, qui s'avère en fait subsumer la bêtise. Il serait facile, par exemple, de taxer de bêtise le toxicomane qui meurt d'une surdose à vingt ans. Dans certains cas, l'accusation est sans doute justifiée, mais elle ne l'est pas si la seule et unique cause de son addiction se trouve dans le fait d'avoir un taux d'escompte du futur élevé. Bien que la psychologie ne soit pas encore en mesure d'expliquer la myopie de certains individus, il est certain que celle-ci n'est jamais choisie. L'individu myope est piégé.

Publications 2007-2008

« The night of August 4 1789 : A study in collective decision making », *Revue Européenne des sciences sociales*, 2007.