

Réconcilier le formel et le causal : le rôle de la neuroéconomie

Benoit Hardy-Vallée
Department of Philosophy
University of Toronto
Jackman Humanities Building
170 St. George St., 4th Floor,
Toronto, ON, M5R 2M8
Tel.: (416) 978-3316
Fax.: (416) 978-8703
ben.hardy.vallee@utoronto.ca
<http://www.hardyvallee.net>

Benoit Dubreuil
Aspirant au FNRS
Centre de Théorie Politique
Université Libre de Bruxelles
Correspondence:
Benoît Dubreuil
90 rue Rose-de-Lima
Montréal (Qué)
H4C 2K9, CANADA
Tel.: (514) 935-3435
Fax: (514)343-7348
benoit.dubreuil@umontreal.ca
<http://african.cyberlogic.net/bdubreuil>

Réconcilier le formel et le causal : le rôle de la neuroéconomie

Introduction

De Pappus d'Alexandrie à Crick et Watson, la mise au jour de mécanismes a contribué à l'explication de la nature. Par *mécanisme*, on entend une analyse du fonctionnement d'un système qui décompose celui-ci en éléments, chacun pouvant effectuer certaines opérations, puis qui montre comment les caractéristiques du système résultent de l'activité des éléments. On peut distinguer deux interprétations du mécanisme, l'une causale, l'autre formelle. Dans le premier cas, les éléments du mécanisme ont des contreparties matérielles : pour chaque élément de l'explication ou du modèle, on peut identifier une structure réelle, matérielle, capable d'interaction causale. (Notons que par causalité, nous entendons ici une notion réaliste, telle que le transfert d'énergie). Le diagramme d'un moteur à explosion, par exemple, représente des pièces (piston, bielle, soupape) qui peuvent être identifiées dans un moteur réel.

Dans le second cas, le mécanisme décrit des éléments et des opérations sans pour autant que ceux-ci puissent être appariés à des structures matérielles: on vise plutôt l'adéquation entre les relations fonctionnelles (entrées/sorties) du modèle et celles du système à l'étude. Ainsi la linguistique chomskyenne propose un modèle mécaniste-formel d'explication de la compétence linguistique en supposant des éléments (par exemple, les groupes nominaux ou verbaux) ainsi que des opérations (règle de formation et de transformation des phrases). Les éléments comme les opérations, cependant, ne correspondent pas à des structures neuronales, anatomiques ou moléculaires, mais à un ensemble de processus computationnels réalisés par le cerveau. Les modèles chomskyens rendent compte de la compétence linguistique humaine, mais n'expliquent pas, par exemple, comment nos aires motrices et visuelles interagissent afin de produire une

phrase.

La théorie économique a, en général, opté pour une approche formelle de la rationalité, en particulier depuis l'adoption de la conception *ordinale* de l'utilité. Auparavant, dans la psychologie hédoniste de Bentham et de Mill, l'utilité était un étalon de mesure du bonheur global de l'individu, qui subsumait en une dimension les variétés de plaisirs et de douleurs. On représentait alors l'utilité comme une valeur objective et théoriquement mesurable, une utilité *cardinale*. Elle comportait toutefois une incertitude relative à sa mesure, et c'est pourquoi le premier pas vers la formalisation du choix rationnel fut d'abandonner la conception cardinale (mais aussi causale) de l'utilité pour une logique des préférences comparées, et donc d'abandonner les relations *causales* pour établir des relations strictement *formelles*.

Dans la théorie de l'utilité *ordinale*, la décision économique peut être construite sans hypothèses psychologiques sur l'intensité ou le contenu des sensations ou perceptions. En effet, plutôt que de supposer que les choix de l'agent peuvent être ordonnés en fonction de leur propension objective à maximiser le bien-être, la théorie porte plutôt sur l'ordonnement des préférences de l'agent. L'utilité devient ainsi une notion relationnelle plutôt qu'absolue, qui définit une relation de préférence entre deux alternatives : « A est préféré à B » équivaut à « A est plus utile que B ». Un agent économique n'a pas à être capable d'attribuer une valeur absolue à un choix A, mais il peut certainement déterminer si (1) A est préféré à B, (2) B est préféré à A ou (3) A et B sont équivalents. De la sorte, utilité et préférences deviennent deux notions indissociables, l'utilité référant à l'échelle sur laquelle se mesurent les préférences relatives et subjectives, et les préférences définissant une fonction d'utilité.

1. Le formalisme à l'épreuve de l'expérience

La préférence est donc la notion première de la rationalité économique : la théorie de la décision, la théorie des jeux et la théorie de l'équilibre général modélisent l'agent rationnel comme celui qui sélectionne des décisions ou stratégies optimales, soit les possibilités d'action— parmi celles qui lui sont offertes—qui maximisent son utilité subjective. L'explication et la prédiction du comportement rationnel s'effectuent donc en identifiant des éléments—des états, des actions et des issues, selon le modèle de Leonard Savage (1954)—et leurs relations. La prise de décision est alors représentée par une mécanique formelle et la valeur du modèle se mesure à sa capacité à décrire le comportement du système à l'étude, plutôt qu'à le décomposer en entités causalement interactives. Avec les axiomatisations de von Neumann-Morgenstern, Savage et Arrow-Debreu, la science économique acquiert, selon Pareto, « la rigueur de la mécanique rationnelle » (1966 :160) et réalise donc l'idéal de « newtonisation » des sciences humaines. Les préférences, stratégies et fonctions d'utilité fournissent une explication mécaniste de la prise de décision sans pour autant s'engager sur le plan ontologique.

Aussi la théorie du choix rationnel (TCR), en elle-même, n'est-elle qu'un formalisme mathématique et non une théorie comportementale ou psychologique. On doit lui adjoindre des hypothèses auxiliaires afin d'en faire un discours qui peut être utilisé dans l'analyse empirique— sans pour autant en faire une mécanique *causale*. L'interprétation standard de la TCR s'est appuyée sur deux hypothèses: (1) le postulat de rationalité et (2) l'attribution d'une fonction d'utilité égoïste. Selon la première, on tient pour acquis que les agents sont rationnels, au sens où l'on présuppose une cohérence entre leurs croyances, désirs et intentions : par exemple l'intention de faire *A* doit pouvoir, du moins en théorie, être logiquement déduite des croyances et désirs de l'agent à propos de *A*. Selon la seconde, les individus possèdent une fonction d'utilité qui les

amène à préférer leur intérêt individuel, défini généralement en termes monétaires.

L'ajout de ces hypothèses auxiliaires permet à la TCR de quitter le simple formalisme mathématique et de réaliser des prédictions empiriques. On peut désormais déduire les préférences d'un agent de sa fonction d'utilité et déterminer l'action qu'il devrait sélectionner dans tel ou tel contexte. La TCR peut ainsi devenir scientifiquement utile et générer un programme de recherche en économie, en sociologie ou en science politique. L'économie expérimentale, dès ses débuts dans les années 1950, indique toutefois que la TCR, ou du moins les hypothèses auxiliaires qu'elles s'étaient adjointes, ne sont pas vérifiées. Aversion au risque, à l'incertitude, violation de l'axiome d'indépendance, préférences pour des solutions moralement préférables mais économiquement suboptimales, la liste des dissonances est longue.

Plutôt que de considérer que la TCR avait été falsifiée, la théorie économique, l'épistémologie des sciences sociales et la philosophie de l'action ont généralement opté, relativement au fossé entre la théorie et les faits, pour une des trois attitudes suivantes, que nous nommerons *téléologisme*, *instrumentalisme*, et *interprétativisme*. (Cette nomenclature ne vise pas à départager les auteurs, mais plutôt les divers arguments face à aux incongruences entre la TCR et les données expérimentales.)

Selon la première (téléologiste), des agents situés suffisamment longtemps dans une dynamique de marché finiront par adopter les prescriptions de la TCR. S'ils ne le font pas, ils seront éliminés. En d'autres mots, le marché finira par sélectionner les individus et firmes rationnelles. La TCR décrit donc un ensemble de règles que les agents tendront à adopter. L'apprentissage et l'évolution du marché interagissent de façon à maintenir les agents sur la voie de la rationalité. Les expériences qui mettent à jour l'irrationalité des agents ne font que souligner

les imperfections propres aux débutants ou non-experts.

Selon la seconde interprétation (instrumentaliste), la TCR ne décrit pas le comportement d'agents réels, mais d'agents idéaux. Elle construit des modèles mathématiques de la prise de décision, des modèles « sans frictions », dans lequel on ne tient pas compte des erreurs (de calcul, de temps, d'information) et des limitations des agents. L'agent idéal occupe, épistémologiquement parlant, le même rôle que le gaz idéal ou la lentille parfaite. Comme Friedman et Savage (1948: 298) le soulignent, on pourrait prédire adéquatement le comportement d'un joueur professionnel de billard si on pouvait calculer, à l'aide de formules physiques, les différentes trajectoires possibles et évaluer les plus profitables. On fait l'hypothèse que le joueur se comporte « comme si » il connaissait les formules et pouvait estimer les angles et calculer les trajectoires. La science économique, selon Friedman (1953), n'a pas d'engagement ontologique et ses entités ne sont que des paramètres dans le calcul économique. Les expériences montrent certes que des agents réels ne sont pas des *Homo oeconomicus*, mais cela n'est pas plus surprenant que d'apprendre qu'un gaz réel ne se comporte pas complètement comme un gaz idéal. L'important est, en bout de ligne, la prédictivité du modèle.

Selon la troisième attitude, (*interprétativiste*), la rationalité est une norme d'interprétation du comportement, qui permet la compréhension d'autrui; la TCR y joue un rôle périphérique (voir par exemple Davidson 1993). Même si les agents ne se conforment pas à la TCR, ils sont toujours interprétables comme étant motivés par des raisons. La rationalité des agents est constitutive d'un ensemble de principes interprétatifs plutôt que de règles explicites. Bien qu'on puisse codifier des normes d'inférence et de décision (logique, théorie bayésienne du choix rationnel, théorie des jeux, etc.), il n'y a pas de principes universels qui nous permettent de répondre univoquement à des questions comme « que croire ? », « quelle croyance réviser ? » ou «

que faire ? », ou encore des principes qui nous permettent d'interpréter univoquement le comportement (linguistique ou autre) d'une personne. Un agent qui agit ou qui interprète a toujours le choix entre réviser une croyance locale ou une théorie générale, de sorte que la bonne croyance ou la bonne action ne peut pas être déduite des modèles théoriques. Le holisme des croyances rend caduque toute tentative d'explicitation des normes d'interprétation du comportement. Ne demeurent alors que des principes généraux et constitutifs : attribuer une activité mentale, une rationalité, des croyances, des désirs, des connaissances, etc. Les théories de la rationalité sont des connaissances d'arrière-plan qui nous servent à interpréter le comportement plutôt qu'à le produire. Elles proposent des normes épistémiques et pratiques d'utilisation des concepts de croyances, désirs, action et rationalité. Les expériences démontrent non pas l'irrationalité des agents, mais plutôt que les raisons qui motivent ceux-ci ne sont pas conformes aux théories codifiées. Bien que techniquement irrationnels, les sujets sont toujours interprétables comme des agents rationnels.

2. Le formalisme est un normativisme

Derrière ces trois interprétations se cache le même arrière-plan épistémologique que nous nommerons le normativisme, à savoir l'idée selon laquelle une théorie de la rationalité est essentiellement une théorie normative. Cet arrière-plan module à la fois la méthodologie modélisatrice—le formalisme mécanistique—et les interprétations de la TCR et des résultats expérimentaux. Nous illustrerons ces propos avec un exemple en particulier, soit le jeu de l'Ultimatum (cf. Camerer 2003).

Dans l'Ultimatum, un agent A doit proposer une fraction $f > 0$, de son choix, d'un montant m d'argent à un agent B. Si B accepte, B reçoit f , et A empêche $m - f$. Si B refuse, personne ne

touche un sou. L'interprétation standard de la théorie des jeux prédit que, en bons maximisateurs d'utilité, les agents se comporteront comme suit : A devrait proposer la plus petite fraction possible, et B accepter n'importe quel montant. Or dans les faits, les données expérimentales montrent que les agents A proposent des offres représentant en moyenne 20 à 50% du butin, et les agents B refusent généralement les offres inférieures à 20 à 30%. Le résultat est robuste peu importe le montant d'argent, la culture, le degré d'anonymat, l'expérience, etc. Les seuls cas où les agents se comportent approximativement de façon conforme à l'interprétation standard sont les cas où B (mais non A) est un ordinateur ou lorsque A et B sont des groupes qui prennent des décisions collectives.

On peut donc appliquer les trois interprétations de la TCR à ces résultats :

- Téléologisme : si les agents jouent de façon répétée à l'Ultimatum, ils finiront par adopter la stratégie optimale; les offres seront aussi minimales que possible, et toute offre sera acceptée.
- Instrumentalisme : si les agents ne suivent pas une stratégie optimale, cela est dû à leur imperfection; généralement, la théorie fournit néanmoins des prédictions utiles.
- Normativisme : on peut interpréter les sujets comme ayant des raisons de sélectionner les stratégies suboptimales; par exemple, ils valorisent l'équité ou encore l'absence de risque.

Implicitement, chacune de ces interprétations tend vers la même idée : une théorie de la rationalité est essentiellement une théorie des normes de rationalité, lesquelles sont soit des règles qu'un agent rationnel finit par suivre (téléologisme), soit celles que suivrait un agent idéal

(instrumentalisme), soit celles par lesquelles nous interprétons un agent rationnel (interprétativisme). L'idée qu'une théorie de la rationalité est une théorie normative est ancrée dans la pratique philosophique et économique au point où, lorsque des expériences ont montré que les agents ne se comportaient pas comme la théorie le prédisait, on a préféré parler de « paradoxes » (d'Allais, d'Ellsberg, etc.) plutôt que de réfutations empiriques ou de contre-exemples, marquant par là le caractère logique, formel et normatif de ces théories.

Les perspectives normativistes mettent cependant la TCR dans une position épistémologiquement problématique. En effet, le téléologisme et l'instrumentalisme tiennent pour acquis l'utilité de la TCR : que ce soit en tant que finalité ou en tant qu'abstraction pratique, la TCR décrit toujours un état idéal des agents rationnels, et cet idéal n'est pas sujet à la réfutation. Des théorèmes et des axiomatiques démontrent formellement l'issue rationnelle des jeux : la validité de la TCR ne s'évalue pas à l'aune de ses prédictions, mais de ses vertus formelles (complétude, consistance, etc.). Si, dans l'Ultimatum, les sujets font des offres irrationnelles (plus que le minimum) et refusent des offres non nulles, ce n'est pas que la théorie prédit mal, mais que les agents démontrent des biais irrationnels. *Augmentez les montants, et vous verrez que les agents adopteront des stratégies optimales*, affirmaient les économistes sceptiques devant les résultats expérimentaux de l'Ultimatum. Or, dans une version où les joueurs devaient se diviser un montant de 100\$, même des offres de 30\$ faisaient l'objet d'un refus (Camerer et Thaler, 1995). Si la TCR décrit une finalité ou un idéal de rationalité, on peut certainement exiger que cette finalité et cet idéal soient justifiés.

Lorsqu'on tente de fournir une justification du téléologisme et de l'instrumentalisme, on constate le support mutuel que s'apportent le formalisme et le normativisme. L'argumentaire justifiant les normes de rationalité va généralement comme suit : tout comme la grammaire est

constituée de normes linguistiques, les théories de la rationalité sont constituées de normes de décision qui recommandent des cours d'action. La TCR est « un ensemble consistant de normes, et non une étude empirique » (Marschak, 1951: 13). Si une personne ne choisit pas ce que la TCR recommande, elle commet une erreur. De même, si une personne se trompe en multipliant 234 par 92, ce n'est pas l'arithmétique qui est en faute, mais la personne. Cette dernière *devrait* trouver 21 528. L'économie théorique n'est pas alors une science susceptible d'entretenir des propositions falsifiables, mais de décrire comment des agents économiques se comporteraient s'ils étaient des individus complètement rationnels. Les proposeurs *devraient* offrir des montants minimaux et les receveurs *devraient* accepter toute proposition. Une pratique est donc rationnelle si elle peut être rationalisée à l'intérieur d'un modèle théorique. La théorie n'est alors qu'une tautologie complexe, justifiée par sa rigueur formelle. Ce faisant, cependant, la théorie s'exclut elle-même, en grande partie, du champ scientifique puisque la science s'efforce de procéder à une tentative systématique de vérification, falsification et comparaison des théories.

L'interprétativisme, au fond, ne fait qu'explicitier ce point de vue : la rationalité ne s'analyse pas en règles et en codifications, mais en principes d'interprétation. L'agent idéal est tout simplement représenté autrement que par la TCR. Popper disait ainsi du principe de rationalité qu'il s'agissait d'un principe « quasiment vide » (1985 :359). Donc la théorie de la rationalité oscille entre tautologie et banalité.

La justification des théories économiques possède donc un mode opératoire assez particulier. Comme le résume Binmore (1994: 150), la pratique standard en économie est de considérer les théories de la rationalité comme des exercices formels:

Des axiomes sont proposés et les propriétés des individus rationnels en sont alors déduites mathématiquement. Si les mathématiques nécessaires sont suffisamment évidentes, l'attention

se concentre alors sur la vérité des théorèmes de l'auteur plutôt que [...] sur leur aptitude à formaliser les concepts qu'ils prétendent capturer.

En d'autres termes, les théories de la rationalité sont justifiées *a priori*. Kant (CRP, III, 28) considérait que des connaissances étaient justifiées *a priori* lorsque, dans la justification, on ne faisait intervenir aucune information qui vienne de l'expérience introspective ou empirique. Outre cette propriété négative, les connaissances *a priori* ont aussi une définition positive: à l'instar du concept de causalité, ce sont avec elles qu'on aborde l'expérience empirique, voire par elles qu'on donne sens ou rend possible cette expérience. Selon Kant, les propositions qui seront nécessaires, fondationnelles (dérivée d'aucune autre connaissance) et universelles (sans exception) seront absolument *a priori* et donc à l'abri de la réfutation. Or les propositions de la TCR partagent bel et bien ces propriétés. Elles sont :

- *Nécessaires*, car en tant que connaissances logico-déductives leur vérité n'est pas faillible ou contingente : elles sont aussi nécessaires que 2 et 2 font 4.
- *Fondationnelles*, car elles ne sont pas dérivées de connaissances ou de généralisations empiriques : ce sont des axiomes à partir desquels on peut déduire d'autres propositions, par exemple l'issue d'un jeu.
- *Universelles*, car elles ne décrivent pas ce qui est le cas, mais ce qui *doit* être le cas; aucune exception n'est même pensable. Un agent qui ne se conforme pas aux axiomes est tout simplement non-rationnel ou irrationnel.

L'économie, cependant, n'est pas une science qui traitent de formes *a priori*, mais des comportements humains : des achats, des investissements, des ventes, des mises en marchés, *etc.* Dans la prochaine section, nous nous intéressons à deux manières de contourner les difficultés

relatives au normativisme : la première est proposée par l'économie expérimentale et la seconde par la neuroéconomie.

3. Réconcilier le formel et le causal : une vision mécaniste de l'utilité

Une manière de contourner les difficultés du normativisme est d'abandonner les hypothèses auxiliaires mentionnées plus haut : le postulat de rationalité et l'attribution d'une fonction d'utilité égoïste. C'est généralement ce qu'a choisi de faire l'économie expérimentale. Ce faisant, elle renverse cependant l'objet de la recherche économique : il ne s'agit plus de prédire le comportement des agents, mais bien de partir de leurs actions pour découvrir des biais cognitifs ou pour explorer leur fonction d'utilité. Le programme de recherche de Tversky et Kahneman, par exemple, ne porte plus sur la prédiction de l'action, mais sur la manière dont l'action révèle des biais cognitifs (Kahneman et al. 1982). Les recherches conduites par Werner Güth, Ernst Fehr, Frans van Winden, George Ainslie et d'autres économistes expérimentaux, quant à elles, visent à utiliser les décisions réelles des acteurs pour décrire d'une manière plus précise leur fonction d'utilité (en intégrant par exemples des préférences pro-sociales ou l'aspect dynamique des préférences). Ce renversement est semblable à celui mis de l'avant dans la méthode des préférences révélées, utilisée pour interpréter le comportement des consommateurs (Samuelson 1938). Celle-ci ne cherchait plus à prédire le comportement des agents à partir d'une fonction d'utilité hypothétique, mais à explorer les préférences des agents et, incidemment, leur fonction d'utilité en observant leurs choix réels..

Grâce à l'économie expérimentale, la TCR peut s'adjoindre des hypothèses auxiliaires plus précises que celles associées à son interprétation standard. Dans le jeu d'Ultimatum, par exemple, le rejet des offres jugées insuffisantes justifie l'ajout de préférences sociales à la

fonction d'utilité égoïste. L'économiste expérimental peut ainsi mesurer l'utilité de la préférence sociale pour le rejet, qui s'élèvera à 20%-30% du total pour la plupart des joueurs. Même avec l'ajout de l'économie expérimentale, la TCR demeure toutefois une théorie formelle où les entrées (perceptions) sont reliées aux sorties (actions) par une fonction d'utilité (qui attribue une valeur aux états du monde) et par des mécanismes formels qui ordonnent les préférences. La TCR propose donc toujours un modèle normatif qui détermine ce que les agents doivent faire en fonction de leurs ressources cognitives et de leur fonction d'utilité.

La TCR peut cependant éviter les problèmes reliés au normativisme en cherchant à relier les mécanismes formels à des processus et des structures causalement impliqués dans le contrôle de l'action. Les neurosciences et la neuroéconomie (les neurosciences de la décision) ont ainsi montré que la prise de décision est le produit de plusieurs mécanismes distinct d'évaluation, de motivation et de sélection de l'action. Bien qu'encore à leurs balbutiements, ces domaines sont déjà à même de nous fournir des modèles causaux permettant de préciser la nature de la décision économique.

Une des premières trouvailles apte à suggérer une révision importante de la TCR est la découverte d'une dissociation entre l'appréciation ('liking') et la motivation ('wanting'). Les recherches de Berridge et de ses collègues, par exemple, montrent que les systèmes dopaminergiques—un ensemble de structures sous-corticales, majoritairement situées dans l'aire tegmentale ventrale—sont en grande partie responsables de la motivation. L'activité de ces neurones est nécessaire pour que des indices reliés à des récompenses acquièrent une signification et une saillance motivationnelle. Des souris et des rats génétiquement modifiés qui ne produisent plus de dopamine perdent la *motivation* à acquérir de la nourriture sans pour autant cesser de *l'apprécier*. Ils démontrent par exemple tous les signes externes typiques de

l'appréciation du sucre ou de la nourriture (passer la langue sur les lèvres), mais ne démontrent aucune motivation à ingérer du sucre ou toute autre nourriture auquel l'animal est sensible normalement (Berridge et Robinson, 1998). Le même phénomène a été constaté chez des humains. Par exemple des fumeurs à qui on administre un antagoniste de la dopamine éprouvent le même plaisir à fumer sans pourtant en avoir le désir (Berridge et Robinson, 2003). Ces recherches suggèrent que la motivation n'est pas en premier lieu la recherche d'un stimulus agréable, mais la recherche de ressources, et que cette motivation est le produit des systèmes dopaminergiques. La distinction entre la motivation et l'appréciation au niveau neuropsychologique nous rapproche d'une explication causale de la relation l'utilité et l'action.

Une autre des trouvailles de la neuroéconomie est que le concept d'appréciation ne correspond pas à un seul mécanisme, mais à plusieurs. Dans un premier temps, on peut distinguer l'appréciation *primaire* de l'appréciation *secondaire*. L'appréciation primaire est une évaluation du caractère hédonique (l'intensité et la valence) de certains stimuli. Elle s'incarne elle-même dans plus d'une structure. On sait par exemple que le nucleus accumbens est impliqué dans la *récompense* (appréciation primaire positive) alors que l'insula est impliquée dans la *punition* (appréciation primaire négative). Des stimuli plaisants (nourriture, drogue, etc.) suscite l'activité du premier, alors que des stimuli répulsifs ou dégoûtants (nourriture avariée) suscite celle du deuxième (Sprengelmeyer, 2007; Berridge et Robinson, 2003). De même, l'amygdale est principalement impliquée dans les réponses au danger (peur, panique, etc.).

Apprécier un stimulus comme étant plaisant ou déplaisant est donc distinct de la motivation, mais distinct aussi de ce qu'on peut appeler les préférences (ou l'*appréciation secondaire*). En effet, préférer A à B implique que, si on fait la somme des pour et contre, des coûts et bénéfiques, A se révèle supérieur. Des aires préfrontales, comme le cortex orbitofrontal et

le cortex préfrontal ventromédian, intègrent les inputs en provenance des aires d'appréciation primaires (Wallis, 2007; Rushworth et al, 2007). Dans des situations de prises de décision, on considère que ces aires encodent la valeur économique, la valeur des buts ou encore la disposition à acheter (*willingness to pay*). Dans une expérience de Plassmann et ses collègues (2007), où des sujets affamés doivent décider de la valeur économique d'un plat, l'activation du cortex orbitofrontal est proportionnelle à la valeur que les sujets attribuent au stimulus. Les études de lésions cérébrales par Damasio et ses collaborateurs illustrent la différence—et les relations complexes—entre l'appréciation primaire et secondaire. Au Jeu de l'Iowa, des participants doivent piger des cartes parmi un choix de 4 paquets, dont les taux de paiement et de punition varient considérablement (les sujets sont informés de la valeur de leur gain ou de leur perte immédiatement après la sélection d'une carte). Des sujets normaux arrivent rapidement à identifier les paquets profitables et à éviter les autres. Damasio (2000) a montré que le comportement des sujets se basent sur des « marqueurs somatiques », soit des réactions physiologiques—telles qu'un accroissement de la conductance de la peau— qui surviennent à la fois lorsqu'ils apprennent la valeur de leur carte et lorsqu'ils anticipent cette valeur. Des patients victimes de lésions au cortex orbitofrontal ou au cortex préfrontal ventromédian manifestent des réactions physiologiques uniquement lorsqu'ils apprennent la valeur de leur carte mais non pas lorsqu'ils l'anticipent. Ils sont donc incapables de développer une appréciation secondaire, c'est-à-dire une préférence pour une option, et sont condamnés à expérimenter, après coup, la valeur de leur choix. Une préférence, en termes neuronaux, équivaut donc à l'encodage d'une appréciation secondaire par l'intégration d'appréciations primaires.

Enfin, outre la motivation, l'appréciation primaire et l'appréciation secondaire, on peut identifier une troisième forme d'appréciation dite *cognitive*. Des aires préfrontales (en

particulier la partie gauche du cortex dorsolatéral préfrontal, CDLPF) permettent de maintenir l'attention sur des buts non hédoniques et non émotifs, mais davantage intellectuels. Il peut s'agir, par exemple, d'acquérir de l'argent, de poursuivre un but pré-sélectionné ou de se conformer à des normes sociales. Dans chaque cas, l'issue de la décision est déterminée par une valeur qui n'est pas (ou pas nécessairement) associés à des stimuli agréables. Cette caractéristique structurelle du système de prise de décision permet d'expliquer la distinction—absente de l'appareil conceptuel de la TCR classique mais omniprésente dans le langage courant—entre l'utile et l'agréable (Jackendoff 2007). L'agréable réfère avant tout aux situations ou événements qui activent les aires d'appréciation primaire et secondaire, alors que l'utile désigne les situations ou événements qui mènent à des appréciations cognitives. Si nous apprécions l'argent, par exemple, c'est parce que nous la trouvons utile. Or nous la trouvons utile parce que nous en faisons un moyen pour acquérir des biens ou des services qui permettront à leur tour d'activer les aires d'appréciation primaire et secondaire.

Les structures sous-jacentes au concept d'appréciation cognitive nous permettent par ailleurs d'expliquer de manière causale ce que plusieurs théories de l'utilité ont déjà modélisé en économie expérimentale. La picoéconomie de George Ainslie (1992), par exemple, a bien mis en évidence comment nos fonctions d'utilité variaient avec le temps et comment notre capacité à différer une récompense était déterminante dans le processus de prise de décision. Cela coïncide avec le fait que le développement des aires préfrontales chez l'enfant lui permet de poursuivre des buts plus abstraits en inhibant, par exemple, une récompense immédiate pour obtenir une récompense à long terme. Un des principaux défis de la neuroéconomie au cours des années à venir consistera à décrire de manière plus détaillée les mécanismes impliqués dans la poursuite des buts cognitifs. Même si les aires préfrontales sont généralement reconnues comme le siège

des processus de contrôle cognitif de haut niveau, il n'existe malheureusement pas une théorie unique quant à leur fonctionnement (Jurado et Rosselli 2007). Au minimum, on doit supposer l'existence de mécanismes permettant 1) le contrôle de l'attention, 2) la planification des actions complexes et 3) l'inhibition des actions ou des stimuli non pertinents, mais nous sommes encore dans l'attente d'un modèle qui intégrera d'une manière cohérente les données psychologiques et neurologiques. Or un tel modèle permettra de décrire avec beaucoup plus de précision le processus de prise de décision.

La neuroéconomie permet donc d'atteindre une conception plus raffinée de l'appréciation que celle généralement favorisée dans les théories normatives sous le concept d'utilité. Les distinctions que nous avons esquissées peuvent être rapprochées de celles proposées par Kahneman et ses collègues (1997) entre les différents types d'utilité. *L'utilité de la décision*, au sens classique, est assimilée à l'activité motivationnelle des neurones dopaminergique : c'est elle qui est révélée par la décision de faire X. Elle peut toutefois être distinguée de l'utilité *expérimentée* (l'appréciation de X pendant la décision), de l'utilité *prédite* (l'appréciation de X avant la décision) et de l'utilité *remémorée* (l'appréciation de X après la décision). Cette typologie évite de faire de la maximisation de l'utilité une thèse normative. Elle fournit un schéma qui fait le pont entre les données neuronales et les données comportementales et qui permet de réconcilier les aspects formels de la théorie de la rationalité avec des mécanismes causaux.

Ce schéma permet d'expliquer, par exemple, comment le comportement peut révéler des préférences (motivation/utilité classique) sans pour autant que ces préférences ne soit reflétées dans l'utilité prédite (l'appréciation avant la décision). Ou encore comment l'utilité prédite peut être cognitive sans être hédonique, c'est-à-dire comment une chose peut-être utile sans être

agréable. De même, la préférence—en tant qu’appréciation secondaire survenant *avant* la décision—peut être cohérente ou non avec l’utilité de la décision : la cohérence en question sera une question de fait plutôt que de définition. Sans se substituer aux mécanismes formels décrits par l’économie expérimentale, la neuroéconomie laisse donc entrevoir la possibilité d’une description mécaniste et causale de la prise de décision rationnelle.

4. La mécanique de la décision : deux exemples

Nous avons soutenu jusqu’à présent que la neuroéconomie permettait de transformer une analyse normative de la décision en une analyse causale. Les modèles neuroéconomiques permettent ainsi de conserver le formalisme de la TCR, mais en ancrant le processus de décision dans des structures et des processus neuronaux. Les concepts d’utilité, de sélection de l’action ou de préférence acquièrent ainsi une définition plus précise et une signification causale. Donnons-en une illustration à partir de deux exemples concrets : (1) la décision d’acheter et (2) le jeu de l’ultimatum.

Knutson et ses collègues (2007) ont présenté à des sujets l’image d’un produit attrayant (du chocolat Godiva) en leur laissant 4 secondes pour décider s’ils souhaitaient acquérir ce produit. Ensuite, ils affichaient le prix du produit (parfois bas, parfois élevé) et demandaient aux sujets de décider s’ils voulaient ou non acheter le produit à ce prix. Évidemment, les sujets refusaient d’acheter les produits dont le prix était trop élevé (80\$ pour une boîte de chocolat) et préféraient les produits dont le prix était généralement bas (7\$ pour la même boîte). L’intérêt de l’expérience se situe sur le plan neuropsychologique. Lorsqu’on demande aux sujets s’ils souhaitent acheter, les produits désirables provoquent une activation du *nucleus accumbens*, cette région du cerveau associée au plaisir et à l’anticipation des récompenses. En revanche, lorsque le

prix est affiché, on détecte une activation de l'insula, une région associée à la douleur, au dégoût et aux autres émotions négatives. L'activation est moindre lorsque le prix est jugé acceptable, auquel cas les structures préfrontales, impliquées dans la planification et le contrôle, sont également sollicitées. Sur le plan causal, l'activation de ces régions permet de prédire de manière efficace si le sujet achètera ou non le produit: l'activation préfrontale permet de prédire l'achat, alors qu'une forte activation insulaire permet de prédire le refus du produit (Knutson et al. 2007). En somme, on explique la décision comme un compromis, où les aires préfrontales jouent le rôle de médiateur, entre le plaisir d'acquérir, manifeste dans l'activation du nucleus accumbens, et le désagrément relié à l'achat qui se manifeste dans l'insula. L'analyse permet donc d'expliquer comment la fonction d'utilité de l'acheteur est en fait déterminée par l'activation de structures neuronales concrètes. Elle permet également d'expliquer des caractéristiques bien connues du processus de décision. Par exemple, le fait que le temps de décision soit plus rapide pour les prix très élevés et très bas que pour les prix intermédiaires. La différence dans le temps de décision s'explique en effet par la nécessité de faire intervenir les aires préfrontales pour résoudre le conflit entre l'insula (déplaisir) et le nucleus accumbens (le plaisir).

Une deuxième série d'études digne de mention est celle reliée au jeu d'ultimatum. Les études d'imageries fonctionnelles révèlent que l'insula antérieure, le cortex préfrontal dorsolatéral—lié au contrôle cognitif et à l'attention—et le cortex antérieur cingulé—lié au conflit cognitif, la détection des erreurs et la modulation émotionnelle—sont activés au moment d'accepter ou de rejeter une offre (Sanfey et al., 2003). Une offre perçue comme insuffisante déclenche, dans le cerveau des répondants, un sentiment de « dégoût » : plus les offres sont jugées insuffisantes, plus l'activation de l'insula antérieure est intense. Cette réaction affective n'est pas uniquement une réponse à une récompense monétaire jugée insuffisante, puisque l'activation est sensiblement

inférieure quand le proposeur est un ordinateur. Ainsi, le fait d'être déçu par un humain nous dégoûte davantage que par un ordinateur.

Quand une offre est jugée suffisante, il semble normal de l'accepter : il y a un gain monétaire et aucun sentiment négatif. Quand l'offre est insuffisante, cependant, le cerveau fait face à un dilemme : punir le proposeur injuste, ou obtenir un peu d'argent ? La décision finale dépend du poids respectif de l'activation du CDLPF et de l'insula antérieure. L'activation de l'insula antérieure est donc corrélée à la fois au degré d'insuffisance et à la décision de rejeter des offres insuffisantes (Sanfey et al. 2003:1756). Le cortex antérieur cingulé, plus actif quand les offres sont insuffisantes, se comporte comme un modérateur entre le but cognitif (le désir d'obtenir plus d'argent) et le but émotif (le désir de punir). Ainsi, les sujets réagissent de manière viscérale à l'incapacité des autres de répondre à leurs attentes. Les données neuroéconomiques sont également confirmées par les données physiologiques et celles tirées de l'introspection. L'activation de l'insula antérieure, par exemple, est corrélée avec les changements physiologiques associés à l'émotion de colère, comme une plus grande conductance de la peau (van't Wout et al., 2006), et les sujets interprètent leur réaction émotionnelle comme telle (Ben-Shakhar 2007).

Les offres insuffisantes suscitent donc une réponse émotive forte. Mais le comportement ne se réduit pas à une simple réponse affective primaire, puisque des mécanismes cognitifs de haut niveau interviennent également dans le contrôle de l'action. Par exemple, d'autres études ont montré comment la partie droite du CDLPF (et non la gauche), lorsqu'on perturbe artificiellement son fonctionnement l'aide de stimulation magnétique transcranienne (SMT), parvient moins efficacement à moduler le comportement : les sujets soumis à la SMT, plutôt que de rejeter les offres insuffisantes comme à l'habitude, acceptent toutes les offres possibles

(Knoch et al. 2006). Curieusement, ils acceptent ces offres bien qu'ils les considèrent toujours insuffisantes. Or on sait que le CDLPF droit est très largement associé à l'inhibition des réponses prédominantes. Lorsqu'on trouble son fonctionnement, il devient plus difficile pour les sujets de résister à la tentation. Les sujets deviennent également plus susceptibles de s'engager dans des comportements risqués (Knoch et Fehr 2007).

Une autre aire essentielle pour comprendre le fonctionnement de la décision dans les jeux d'ultimatum est le cortex préfrontal ventromédian (CPFVM). Lorsqu'il affectée (chez les patients cérébro-lésés), les sujets rejettent plus facilement les offres insuffisantes (Koenigs et Tranel, 2007). La compréhension du fonctionnement du CPFVM demeure l'un des grands défis de la neuroéconomie. Par exemple, les mêmes patients cérébraux-lésés, lorsqu'ils font face à des dilemmes moraux complexes, ont tendance à adopter davantage des raisonnements utilitaristes que les patients normaux, ce qui laisse croire qu'ils contrôlent davantage leurs émotions ou qu'ils sont moins émotifs (Koenigs et al. 2007). Comment une même lésion peut elle rendre les gens plus émotifs dans un cas (le jeu d'ultimatum) et moins dans un autre (les dilemmes moraux complexes)? Une réponse possible est que le CPFVM permet à la fois prise en considération des gains futurs et la modulation des émotions typiquement sociales (la sympathie et la détresse face au malheur d'autrui) mais pas d'autres émotions négatives comme la colère (Moll and de Oliveira-Souza 2007). Ainsi, le dysfonctionnement du CPFVM rendrait plus difficile la prise en considération des gains futurs (dans le jeu d'Ultimatum), tout en facilitant les jugements utilitaristes. Bien qu'incomplètes, les recherches sur le CPFVM démontrent l'intérêt de la neuroéconomie pour l'étude de la rationalité, que ce soit sur le plan formel ou causal. Plutôt que d'opposer trop simplement les mécanismes d'appréciation et de contrôle (les émotions et la cognition), elles nous rappellent l'importance dans la prise de décision des mécanismes

d'intégration des émotions et des risques, de même que leur interaction avec les stimuli sociaux.

[INSÉRER FIGURE 1 ICI]

Conclusion

Selon la définition classique de Lionel Robbins, l'économie est la « science qui étudie les comportements humains en tant que relation entre les fins et les moyens rares à usages alternatifs » (Robbins, 1932: 13). On peut comprendre cette relation entre les moyens et les fins comme une relation formelle, ou une relation causale. Dans le présent article, nous avons soutenu que la théorie du choix rationnel, dans son interprétation standard, ne se présentait pas comme une théorie causale, mais bien normative. L'économie expérimentale permet de résoudre une partie du problème en remplaçant les hypothèses sous-jacentes à l'interprétation standard de la TCR par des hypothèses plus réalistes. L'attribution d'une fonction d'utilité qui inclut des préférences sociales, par exemple, permet d'améliorer la prédictivité de la théorie dans des jeux comme celui d'ultimatum. L'économie expérimentale ne parvient pas cependant à surmonter une autre difficulté théorique à laquelle fait face la TCR, à savoir la nature formelle des entités postulées. Nous avons montré comment la neuroéconomie, parce que son étude ne se limite pas à ses manifestations comportementales, laisse entrevoir le développement d'une véritable science naturelle du choix rationnel. De la sorte, le rapprochement entre biologie et économie, inauguré entre autre par Darwin, pourra trouver son expression dans une science générale de « l'économie de la nature ».

Références

- Ainslie, G. 1992. *Picoeconomics: The Strategic Interaction of Successive Motivational States Within the Person*, Cambridge University Press.
- Ben-Shakar, G., Bornstein, G., Hopfensitz, A. et van Winden, F. 2007 Reciprocity and Emotions in Bargaining: Using Physiological and Self-Report Measures. *Journal of Economic Psychology*, 28(3), 314-323.
- Berridge, K. C., et Robinson, T. E. 1998. What Is the Role of Dopamine in Reward: Hedonic Impact, Reward Learning, or Incentive Saliency? *Brain Research Reviews*, 28(3), 309-369.
- Berridge, K. C., et Robinson, T. E. 2003. Parsing Reward. *Trends in Neuroscience*, 26(9), 507-513.
- Camerer, C. 2003. *Behavioral Game Theory : Experiments in Strategic Interaction*. Princeton, Princeton University Press.
- Damasio, A. 2000. *L'Erreur de Descartes*. Paris, Odile Jacob.
- Davidson, D. 1993. *Actions et événements*. Paris, Presses universitaires de France.
- Kahneman, D., Slovic, P., et Tversky, A. 1982. *Judgment under uncertainty: Heuristics and biases*. New York, Cambridge University Press.
- Kahneman, D., Wakker, P. P., et Sarin, R. 1997. Back to Bentham? Explorations of Experienced Utility. *The Quarterly Journal of Economics*, 112(2), 375-405.
- Knoch D et Fehr E. 2007. Resisting the power of temptations: The right prefrontal cortex and

- self-control. *Annals of the New York Academy of Sciences*, 1104, 123-134.
- Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, V., et Fehr, E. 2006. Diminishing Reciprocal Fairness by Disrupting the Right Prefrontal Cortex. *Science*, 314(5800), 829-832.
- Knutson, B., Rick, S., Wimmer, G. E., Prelec, D., et Loewenstein, G. 2007. Neural Predictors of Purchases. *Neuron*, 53(1), 147-156.
- Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M. et Damasio, A. 2007. Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature* 446, 908–911.
- Koenigs, M. et Tranel, D. 2007. Irrational Economic Decision-Making after Ventromedial Prefrontal Damage: Evidence from the Ultimatum Game. *Journal of Neuroscience*, 27(4), 951-956.
- Moll, J. et de Oliveira-Souza, R. 2007. Moral judgments, emotions and the utilitarian brain, *Trends in Cognitive Sciences*, 11(8), Pages 319-321.
- Pareto, V. 1966 [1909]. *Manuel D'économie Politique* (2e éd.). Genève, Librairie Droz.
- Plassmann, H., O'Doherty, J., et Rangel, A. 2007. Orbitofrontal Cortex Encodes Willingness to Pay in Everyday Economic Transactions. *Journal of Neuroscience*, 27(37), 9984-9988.
- Robbins, L. 1932. *An Essay on the Nature and Significance of Economic Science*. London, Macmillan.
- Rushworth, M., Behrens, T., Rudebeck, P., et Walton, M. 2007. Contrasting roles for cingulate and orbitofrontal cortex in decisions and social behaviour. *Trends in Cognitive Sciences*, 11(4), 168-176.

Samuelson, P. 1938. A Note on the Pure Theory of Consumers' Behaviour. *Economica* 5:61-71.

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., et Cohen, J. D. 2003. The Neural Basis of Economic Decision-Making in the Ultimatum Game. *Science*, 300(5626), 1755-1758.

Savage, L. J. 1954. *The Foundations of Statistics*. New York, J. Wiley.

Sprengelmeyer, R. 2007. The neurology of disgust. *Brain*, 130(7), 1715-1717.

van 't Wout, M., Kahn, R. S., Sanfey, A. G., et Aleman, A. 2006. Affective State and Decision-Making in the Ultimatum Game. *Experimental Brain Research*, 169(4), 564-568.

Wallis, J. D. 2007. Orbitofrontal Cortex and Its Contribution to Decision-Making, *Annual Review of Neuroscience*, 30, 31-56.

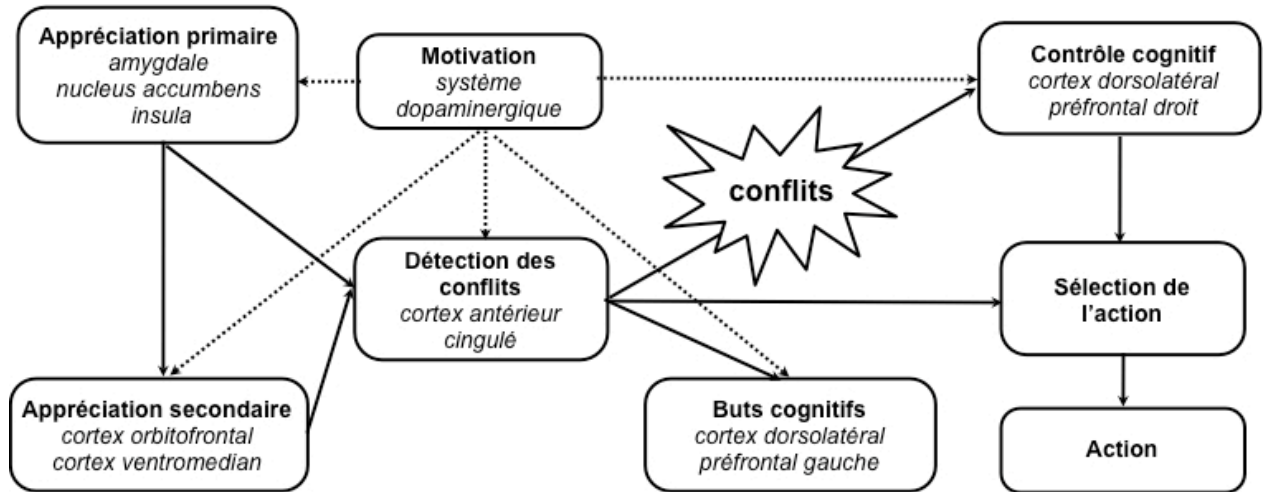


Figure 1: Mécanismes neuroéconomiques de décision

Résumé: La théorie du choix rationnel (TCR), dans son interprétation standard, ne se présente pas comme une théorie causale, mais bien normative. L'économie expérimentale, en remplaçant les hypothèses sous-jacentes à l'interprétation standard de la TCR par des hypothèses plus réalistes – une fonction d'utilité qui inclut des préférences sociales, par exemple – a permis d'améliorer la prédictivité de la TCR. Elle n'est pas parvenu cependant à surmonter une autre difficulté théorique à laquelle fait face la TCR, à savoir la nature formelle des entités postulées. Dans cet article, nous montrons comment la neuroéconomie, parce que son étude ne se limite pas à des manifestations comportementales, laisse entrevoir le développement d'une véritable science naturelle du choix rationnel.

Abstract: Rational choice theory (RCT) is standardly interpreted not as a causal theory, but rather as a normative one. Experimental economics, replacing underlying assumptions of the standard interpretation of RCT by more realistic ones—such as a pro-social utility function, for instance—has improved the predictability of the RCT. That was not sufficient, however, to overcome another theoretical difficulty RCT faces, which is the formal nature of its postulated entities. In this paper, we show how neuroeconomics, because it is not limited to behavioral manifestations, may pave the way to the development of a genuine natural science of rational choice.

Mots-clés: théorie du choix rationnel, neuroéconomie, économie expérimentale, jeu d'ultimatum

Key words: Rational choice theory, neuroeconomics, experimental economics, ultimatum game

Classification JEL: A12, B13, C79, , D01, D87