

# La neuroéconomie peut-elle changer l'économie ?

Depuis quelques années, des dizaines d'expériences menées par des économistes en collaboration avec des neuroscientifiques dans plusieurs laboratoires d'imagerie cérébrale à travers le monde sont regroupées sous le label de neuroéconomie. Cette discipline émergente prolongerait les champs désormais bien installés de l'économie comportementale et de l'économie expérimentale. En quel sens ? Et, plus généralement, la neuroéconomie peut-elle changer la science économique ?

## Les conditions d'un rapprochement entre économie et neurosciences

L'économie comportementale enrichit de manière maximale conservatrice les modèles standards de rationalité économique en ajoutant à leurs paramètres habituels des facteurs psychologiques dans le but de rendre compte de déviations comportementales apparentes vis-à-vis de la rationalité (Rabin M. [1]<sup>2</sup>, Camerer C. [2]). Il nous semble que la neuroéconomie permet de compléter, voire de repenser, cette approche de la manière suivante : sur quels présupposés concernant le développement psychologique des individus (à une échelle onto- ou phylogénétique) reposent les modèles les plus adéquats en économie comportementale ? Nous défendons l'idée qu'une modélisation d'un comportement en économie comportementale est adéquate lorsqu'elle est réaliste d'un point de vue évolutionnaire. Cette idée provient d'une hypothèse de travail selon laquelle les anomalies typiquement modélisées en économie comportementale correspondent le plus souvent à des stratégies adaptatives implicites de la part des individus (Gigerenzer G. [3]) ou bien sont le résultat de contraintes issues de l'évolution dans le traitement d'une tâche ou d'un stimulus économique type.

Mais, même si ce sont les économistes qui ont affiché le plus grand enthousiasme, ce sont les neuroscientifiques qui, jusqu'ici, ont pu tirer le profit scientifique le plus sain de cette récente hybridation disciplinaire. La raison en est simple. Les neuroscientifiques sont parfois heureux d'emprunter aux économistes des idées permettant de raffiner de manière inédite leurs investi-

gations au sujet, de phénomènes biologiques comme par exemple la récompense, l'optimalité comportementale d'un organisme, la contribution des émotions à la prise de décision rationnelle, etc. De leur côté, les économistes semblent être en quête de fondements et de modèles alternatifs, pouvant aller jusqu'à une remise en cause de la domination des mathématiques dans leur discipline au profit d'une approche d'inspiration biologique.

Cette approche alternative viserait à établir conjointement deux choses. La biologie peut d'abord servir à valider des concepts de base de l'économie. On cherchera donc, dans les neurosciences, une confirmation empirique et naturelle des prédictions propres à l'économie. Ensuite, la neuroéconomie peut être perçue comme la seconde génération de l'économie comportementale qui, depuis une dizaine d'années, est devenue une branche non réellement controversée de l'économie. Ce point rejoint le précédent : de science comportementale, déjà imprégnée de psychologie, l'économie tendrait à se transformer, plus profondément, en une science biologique.

Les sciences psychologiques, comportementales, neurobiologiques peuvent pénétrer la science économique, montrant ainsi sa porosité à des concepts et des modèles scientifiquement exogènes. Ariel Rubinstein [4], dans un article récent critique vis-à-vis de la neuroéconomie, présente une modélisation synthétique de la conception de l'acte de décision qui permet d'expliquer cette porosité.

La modélisation standard d'une prise de décision en économie se fait par le moyen d'une fonction qui associe

Sacha  
Bourgeois-Gironde<sup>1</sup>,  
Institut Jean-Nicod  
(CNRS-EHESS-ENS).

1. Auteur de  
*La Neuroéconomie*,  
Paris, Plon, 2008.

2. Les chiffres entre crochets  
renvoient à la bibliographie en  
fin d'article.

à tout problème de choix  $A$ , dans le domaine pertinent, un élément unique  $c(A) = A$ . La formule  $c(A) = a$  signifie que le sujet choisit  $a$  dans l'ensemble d'alternatives  $A$ . On peut, comme le fait Rubinstein lui-même, adopter une structure plus riche de l'acte de décision. La fonction  $f$  prend alors comme argument non seulement l'ensemble d'alternatives  $A$ , mais la manière  $f$  dont un sujet se représente cet ensemble. La formule  $c(A, f) = a$  signifie à présent que le problème de choix  $A$  est présenté au sujet sous la description ou sous le « cadre » (*frame*)  $f$  et que le sujet choisit  $a$ . À de rares exceptions près (De Martino B. [5], Bourgeois-Gironde S. et Payzan É. [6]), ce recadrage psychologique de l'acte de décision n'est pas l'objet prioritaire des expériences de neuroéconomie. A. Rubinstein suggère que la modélisation (le plus souvent implicite) de l'acte de décision en neuroéconomie est de la forme  $c(A) = (a, x)$ . Le choix de  $a$  par le sujet est concomitant de la production d'un vecteur  $x$  composé d'une série de phénomènes comportementaux observables durant le moment que dure la prise de décision.

De quoi est composé  $x$  et quel est son rapport éventuel avec  $f$ ? La neuroéconomie associe le choix observé  $a$  à un ensemble de mesures qui composent  $x$  re étant l'activité cérébrale d'un ensemble plus ou moins ciblé d'aires du cerveau du sujet. Ces activités sont mesurées au moment du choix, mais la fenêtre temporelle d'intérêt peut être élargie en amont et en aval de ce choix. De même,  $x$  n'est pas exclusivement composée d'activités cérébrales. Toute mesure visant à rendre compte indirectement de processus cognitifs (et par conséquent cérébraux) accompagnant le choix, tels que par exemple les temps de réaction ou les mouvements oculaires, peuvent composer le vecteur  $x$ , indépendamment ou en corrélation avec d'autres mesures, et entrer dans le programme de recherche de la neuroéconomie, sous cette version schématique.

Ce que met en évidence cette synthèse minimaliste, par Ariel Rubinstein, de la neuroéconomie, c'est qu'il n'y a pas de corrélation entre  $f$  et  $x$ . Autrement dit, la mise en évidence de processus cognitifs et cérébraux au moment (élargi) du choix ( $x$ ) n'est pas nécessairement une indication de la manière dont le choix est représenté par le sujet ( $f$ ). Les facteurs psychologiques, qui peuvent affecter la représentation de l'ensemble d'alternatives sur lequel s'exerce le choix du sujet, n'ont pas forcément de réalisation immédiatement observable à travers les processus cognitifs ou neurobiologiques accompagnant ce choix. En quoi l'acquisition de données concer-

nant ces processus peut-il alors avoir une quelconque pertinence en économie? La contrainte qui pèse sur une réponse positive à cette question dépend naturellement du choix de modéliser la relation entre la décision et les processus qui la sous-tendent comme nous venons de le faire. On peut voir, cependant, ces données comportementales comme permettant d'opérer des distinctions – par exemple entre des types de sujets – et des corrélations – par exemple entre des types de choix et la présence de tel ou tel processus remarquable.

Mais il est possible de choisir à la base une conception très différente de l'impact potentiel de la neuroéconomie sur l'économie. Une telle conception alternative consiste à adopter une perspective qu'on appellera fonctionnaliste, et non plus réductionniste, sur la relation entre l'étude du comportement économique et la présence de processus cognitifs et d'activités physiologiques. Le simple fait de dire qu'un comportement est corrélé à un type d'activité  $x$  ne constitue pas en soi une information pertinente pour l'économiste. Mais être en mesure d'expliquer comment les objets habituels de la science économique (les comportements et les institutions économiques en particulier) sont dépendants de conditions psychologiques, biologiques et neurobiologiques ponctuelles (au moment des choix), mais aussi à une échelle temporelle beaucoup plus vaste (celle de la vie du sujet ou encore de l'évolution de l'espèce humaine), implique une mise en perspective des conditions naturelles d'émergence des objets et des interactions que modélise la science économique.

La variété des mesures neuroéconomiques, qui peuvent être associées à des choix dans une vaste panoplie de situations expérimentales, et la multiplicité des approches neuroéconomiques elles-mêmes font ressembler cette discipline naissante à un ensemble de programmes de recherche éclaté, voire à une suite de résultats expérimentaux souvent valorisés en isolation d'une vision plus globale. Le but du reste de cet article est de structurer l'apport de quelques expériences marquantes en neuroéconomie durant ces dernières années autour d'une perspective fonctionnaliste et évolutionniste.

### Émotions et rationalité dans la prise de décision individuelle

À travers ses expériences célèbres des années 1990, Antonio Damasio [7] a mis en évidence le rôle des émotions pour la prise de décision optimale. Les patients de A. Damasio, qui présentaient des lésions dans les aires

ventro-médianes du cerveau, zone dédiée au traitement des signaux émotionnels, sont incapables, sur une tâche où il s'agit de se départir de loteries risquées apparemment attractives mais néfastes à long terme, d'adapter leur comportement et de porter leur choix vers des loteries apparemment moins attractives mais rentables à long terme. On peut tirer maintes leçons de cette expérience séminale de A. Damasio, mais nous voudrions souligner ici un point dans la perspective qui nous intéresse. A. Damasio note que ses patients, bien qu'ils ne parviennent pas à mettre en place un comportement adaptatif, développent une conscience, inefficace donc, des choix plus optimaux qu'ils devraient réaliser.

D'un point de vue économique, ce qui compte sont les choix manifestes, observables : le fait que des sujets peuvent anticiper les conséquences de choix qu'ils prennent peut-être en dépit de leur volonté n'est pas intégré dans la modélisation économique standard. Pourtant, pour tout sujet, les décisions que nous n'avons pas prises, les options que nous avons préféré laisser de côté ont, elles aussi, des effets sous la forme d'états mentaux et physiologiques particuliers, à savoir des émotions de regret ou de soulagement. On peut présumer que ces émotions vont déterminer des comportements tout à fait manifestes. Nous pouvons être déçus d'un choix que nous avons fait et nous pouvons aussi regretter un choix que nous n'avons pas fait. Ce choix qui est resté à l'état virtuel devient la source non observable de mon comportement présent et la cause d'un état émotionnel négatif. Quand, dans les années 1970 et 1980, l'économie s'est progressivement intéressée aux soubassements psychologiques des prises de décision, l'attention a été portée, d'un point de vue théorique puis expérimental, aux deux émotions négatives que sont le regret et la déception. Mais ce ne sont pas les états émotionnels en eux-mêmes qui ont d'abord intéressé les économistes, mais leur rôle dans l'explication de certaines anomalies du comportement.

Revenons très rapidement aux fondements du problème économique du regret. Maurice Allais avait proposé, en 1953, le problème suivant à un groupe d'experts en prise de décision.

Choisissez entre les deux options A et B :

- dans l'option A, vous avez 10 chances sur 10 de gagner 10 000 € ;
- dans l'option B, vous avez 9 chances sur 10 de gagner 15 000 € et 1 chance sur 10 de ne rien gagner du tout.

Face à ce premier choix, la plupart des personnes

préfèrent l'option A. C'est l'attrait de la chose sûre. Pourquoi chercher à gagner 15 000 € dans l'option B en prenant le risque – une chance sur dix – de ne rien gagner du tout alors qu'on gagne 10 000 € à coup sûr en se fixant sur l'option A ?

Une fois ce choix réalisé, on présente aux mêmes individus le choix entre les deux options C et D suivantes :

- dans l'option C, vous avez 1 chance sur 10 de gagner 10 000 € et 9 chances sur 10 de ne rien gagner du tout ;
- dans l'option D, vous avez 0,9 chance sur 10 de gagner 15 000 € et 9,1 chances sur 10 de ne rien gagner du tout.

Ici les préférences se portent très généralement sur l'option D. La raison en est qu'il n'y a pas tellement de différence de probabilité de gagner quelque chose entre C et D. On est un petit peu moins sûr de gagner 15 000 € en choisissant D que de gagner 10 000 € en choisissant C, mais ça vaut la peine de prendre ce risque et de se retrouver le cas échéant avec 15 000 € en poche.

Le problème est que ce comportement constitue une violation de ce que prédit la théorie de la décision rationnelle. Pourquoi ? Parce que si l'on y regarde d'un peu plus près, on peut s'apercevoir que, du point de vue des gains qu'on peut espérer, A et C d'un côté et B et D de l'autre sont des options proportionnellement équivalentes. L'espérance de gagner 10 000 € qui est de 10/10 dans A est de 1/10 dans C et rien d'autre ne change entre ces deux options A et C. De même, l'espérance de gagner 15 000 € dans B qui est de 9/10 est devenue de 0,9 sur 10 dans D. Elle est à nouveau divisée par 10 et rien d'autre ne change à part cela entre les deux options B et D. Étant donné que ces paires de choix sont structurellement similaires sous le rapport de leur espérance de gain, il n'est pas rationnel, dans un cas, de préférer A et, dans l'autre, de préférer D. Alors pourquoi la plupart des sujets adoptent-ils un comportement qui présente une violation manifeste des principes qu'ils admettent clairement par ailleurs ?

La réponse tient dans l'anticipation des regrets. Cette réponse n'est pas anodine, car elle représente le point de départ de travaux que l'on peut mener aujourd'hui en neuroéconomie autour de la rationalité des décisions individuelles. À travers cette réponse, on a commencé à entrevoir quel rôle peuvent jouer les émotions dans nos prises de décision, et on s'est mis à construire des modèles de la rationalité humaine qui les prennent en compte. L'étape suivante était simplement de comprendre non plus seulement théoriquement le

rôle des émotions mais d'observer *in vivo* la contribution d'états émotionnels comme le regret dans nos prises de décision. Quand vous devez choisir entre A et B, c'est-à-dire entre un gain certain de 10 000 € et un gain incertain de 15 000 €, vous vous dites très certainement que si, par malchance, vous n'obteniez pas les 15 000 € mais rien du tout (il y a une chance sur dix que cela arrive) vous éprouverez de très vifs regrets de ne pas avoir choisi les 10 000 € que vous pouviez pourtant empocher à coup sûr. En règle générale, votre amour du risque et de l'argent est moins fort que les regrets que vous anticipez si vous vous retrouvez complètement perdant à l'issue de votre choix. Or, sous le rapport de ces regrets anticipés, la situation n'est pas du tout la même quand vous devez choisir entre C et D. Si vous choisissez l'option D et que vous n'obtenez pas les 15 000 € espérés, vous êtes certainement déçu, mais regrettez-vous vraiment de ne pas avoir choisi C où l'espérance d'obtenir 10 000 € était à peine plus élevée ? Probablement pas. Vous vous dites que le jeu en valait la chandelle et que vous auriez été bien content de gagner les 15 000 €.

En introduisant le regret anticipé dans l'analyse des procédures de décision face à des choix qui comportent des risques et des gains variés, nous pouvons expliquer pourquoi des individus, y compris ceux qui ont des connaissances sophistiquées en théorie de la rationalité, ont l'air de violer un principe de base de la prise de la décision. Ce principe stipule que lorsque deux paires d'options sont de fait équivalentes il est absurde de ne pas faire les mêmes choix dans une paire et dans l'autre. Mais la prise en compte du regret explique la nature des décisions que nous prenons lorsque nous sommes confrontés au problème posé par M. Allais et montre que celles-ci n'ont rien d'absurde. Nous nous projetons dans la situation où les choix seront pris et nous imaginons dans quel état émotionnel nous nous trouverons alors. Cette capacité de projection émotionnelle est suffisante pour engendrer des comportements de choix qui ont l'air de ne pas se conformer aux normes de la théorie économique de la rationalité.

Giorgio Coricelli et Angela Sirigu [8], de l'Institut des sciences cognitives de Lyon, ont étudié les bases cérébrales du regret dans des circonstances où les joueurs pouvaient alternativement éprouver, en principe, l'émotion de regret et l'émotion de déception. Leur expérience procède ainsi.

Au début de chaque coup, le sujet a toujours deux roues devant lui et il doit choisir entre l'une et l'autre.

Une fois qu'il a fait son choix, l'aiguille tourne et le résultat s'affiche. Dans certains cas, seul le résultat pour la roue qu'il a choisi s'affiche. Il gagne ou il perd alors de l'argent. Dans les autres cas, le résultat de la roue qu'il a choisie et le résultat de celle qu'il n'a pas choisie s'affichent tous les deux. Il peut donc comparer ce qu'il a obtenu avec ce qu'il aurait pu obtenir s'il avait choisi l'autre roue. Il gagne ou perd plus ou moins que ce qu'il aurait gagné ou perdu s'il avait joué l'autre roue.

Le premier cas, où seul le résultat de la roue jouée s'affiche, peut donc donner lieu à de la satisfaction ou de la déception. L'aiguille s'est arrêtée juste avant le périmètre gagnant. Un centième de seconde de plus et j'étais gagnant. Je suis donc déçu. Dans le second cas, c'est avant tout du regret que l'on peut éprouver. Car si je gagne 50 € sur ma roue et que l'aiguille de la roue que je n'ai pas jouée indique un gain de 200 €, je regrette de ne pas avoir choisi l'autre roue. C'est grâce à ce simple système de roues de la fortune que G. Coricelli et A. Sirigu parviennent à reproduire, avec une relative simplicité, des situations de déception et de regret dans leur expérience.

Pendant que les sujets répètent ce jeu des roues de la fortune des dizaines de fois, A. Sirigu et G. Coricelli s'intéressent à leur activité cérébrale. Ils se concentrent plus spécialement sur les régions du cerveau où l'on peut penser que s'opère une interaction entre les circuits émotionnels et les circuits de la décision et du raisonnement. Leur problématique neuroscientifique est donc très semblable à celle qu'étudiait A. Damasio : comment les processus de décision et les processus émotionnels interagissent entre eux pour générer des comportements optimaux dans un jeu qui se répète pendant près d'une heure. Mais la réponse qu'ils cherchent à donner est un peu différente de la théorie de A. Damasio. Ce dernier montre que des signaux corporels, somatiques, remontent des profondeurs du système limbique vers les régions du cortex préfrontal où se situent les soubassements neuronaux de nos capacités de réflexion et de décision. A. Sirigu et G. Coricelli, dans leur étude neuroéconomique du regret et de la déception, cherchent à mettre l'accent sur un aspect un peu différent de la relation entre émotion et rationalité. Ils se concentrent sur une séquence particulière au sein de leur protocole expérimental. Ils observent l'activité cérébrale du sujet dans le moment spécial où il fait face aux résultats de son choix ainsi que du choix qu'il n'a pas fait. Dans cette situation, avant d'éprouver une quelconque émotion

de soulagement ou de regret, le sujet compare les résultats. Il opère la soustraction entre ce qu'il a obtenu et ce qu'il aurait pu avoir. Cette comparaison est tantôt favorable et provoque du contentement, et même du soulagement si le sujet ressentait un peu d'inquiétude, et tantôt défavorable, provoquant alors chez le sujet le regret de ne pas avoir cliqué sur l'autre roue. Il faut souligner que la comparaison a bien lieu avant le surgissement de l'émotion et que cette dernière est modulée par la comparaison numérique entre le résultat obtenu et le résultat non obtenu.

G. Sirigu et A. Coricelli isolent donc cette séquence à propos de laquelle ils pensent pouvoir dire que des processus cognitifs précèdent et déterminent certains de nos états émotionnels. Ils se concentrent, comme le faisait également A. Damasio, sur les activités du cortex orbitofrontal des sujets durant le temps de cette séquence. Ils constatent que les activités neuronales dans le cortex orbitofrontal sont bien différenciées selon que les sujets sont censés éprouver du regret ou de la déception. Seule l'émotion de regret provoque une activité significative dans cette zone du cerveau. Mais, ce qui est encore plus remarquable, est que l'activité neuronale liée au regret dépend bien de l'amplitude des écarts entre le gain obtenu et le gain non obtenu. On peut donc se sentir autorisé à dire, comme le font A. Coricelli et G. Sirigu, que les comparaisons qu'effectue le sujet l'amènent à calibrer ses états émotionnels. Les activités cérébrales qui sous-tendent nos états émotionnels dépendent bien ici de la comparaison que nous effectuons entre des états de choses obtenus et des états de choses qui auraient pu être le cas.

On peut, à propos des résultats de cette expérience, émettre les doutes que nous avons évoqués plus haut. En quoi la mise en évidence d'interactions cérébrales entre des activités liées conjointement à des processus cognitifs et à des processus émotionnels a-t-elle un quelconque intérêt pour les économistes, y compris pour les adeptes de cette branche de la théorie de la décision qu'est la théorie du regret ? De telles données sont hors du champ de pertinence de la compréhension économique du comportement humain. Assurément. Le seul fait de confirmer, à l'aide de données neurobiologiques, que l'anticipation des regrets joue un rôle dans la prise de décision ne suffit pas. Le type de questions éventuellement intéressantes pour les économistes relève d'un autre ordre. Nous les formulerons ainsi : la capacité d'éprouver et d'anticiper des regrets est-elle indispensable pour l'implémentation

d'un comportement rationnel ? Comment se fait-il qu'au cours de l'évolution certains organismes aient apparemment développé une capacité à anticiper leurs états émotionnels futurs ? Les regrets anticipés se traduisent-ils nécessairement en poids d'utilité dans la prise de décision ? Est-ce que cette conception de l'influence de ce type d'émotions sur la prise de décision rationnelle – qui est celle qui est retenue plus ou moins explicitement par les théoriciens du regret en économie – est la seule et la plus vraisemblable ?

Une réponse à cette question déborde du strict cadre de l'étude de corrélats neuronaux du regret. Elle suppose de mettre en perspective cette investigation des « bases neuronales » d'une émotion au sein d'un questionnement sur l'émergence, au cours de l'évolution, de la capacité d'anticiper que ses choix futurs sont regrettables. Elle suppose de se poser la question de savoir si la genèse d'un signal de regret se transcrit nécessairement en poids d'utilité dans la prise de décision (ce n'est pas le cas chez les fumeurs pathologiques par exemple). Elle suppose de comparer l'optimalité comportementale de différents types d'organismes différant quant à la genèse et au traitement de ces signaux émotionnels. Une telle approche suppose de faire appel à diverses disciplines : ici, par exemple, la psychologie évolutionniste, l'addictologie, la biologie comparative, et donc de déborder la conception étroite, et faisant tendre à adopter des positions réductionnistes, de la neuroéconomie comme « mise en images cérébrales » des décisions des individus.

### **Théorie de l'esprit et interactions stratégiques**

L'expérience qui a popularisé la neuroéconomie en 2003 est celle qu'a menée Alan Sanfey [9] sur le jeu de l'ultimatum. Un premier joueur dispose de 10 € qu'il doit partager comme bon lui semble avec un second joueur, avec la seule contrainte que si le second joueur refuse l'offre du premier joueur ils seront perdants tous les deux. Si l'offre est acceptée, les deux joueurs disposeront du partage prévu par le premier joueur. A. Sanfey a montré, en utilisant la technique de l'imagerie fonctionnelle par résonance magnétique que, dans le cas d'offres perçues par les seconds joueurs comme trop basses et rejetées par ces derniers, une zone du cerveau, nommée l'*insula*, était significativement impliquée dans ce rejet. Ce qui est frappant, c'est que l'activité de l'*insula* est habituellement associée au traitement du dégoût, de la nausée, et des états proprioceptifs.

La sensation de dégoût face à une offre perçue comme injuste dans le jeu de l'ultimatum peut englober, plus généralement, les comportements qui sont interprétés comme des violations de normes sociales de réciprocité. Le jeu de l'ultimatum nous met dans la position de recevoir une offre puis de l'accepter ou de la refuser selon que nous la considérons comme équitable ou pas. Mais nous restons dans une position relativement statique et passive. Nous agissons certes indirectement sur le choix de notre partenaire, car celui-ci doit anticiper notre éventuelle réaction de refus et nous faisons ainsi tendre son choix, par la simple pression silencieuse que nous exerçons, vers un optimum d'équité. En fait, nous effectuons tout de même deux choses distinctes lorsque nous acceptons ou refusons une offre dans le jeu de l'ultimatum. En premier lieu, naturellement, nous recevons ou renonçons à une somme d'argent, et nous augmentons ou n'augmentons donc pas nos gains en conséquence. Mais en second lieu, par là même, nous faisons que notre partenaire recevra ou perdra lui aussi une somme d'argent complémentaire. De fait, notre choix d'accepter ou de refuser son offre implique un gain ou une perte pour notre partenaire, une récompense ou une punition. Mais la seule frustration qui peut découler de ce jeu est qu'il ne nous donne pas le temps de savourer les conséquences de nos choix. Parfois, lorsque le partenaire nous a fait une offre injuste que nous nous sommes sentis obligés de rejeter, nous aimerions observer chez lui la frustration que nous avons nous-mêmes ressentie.

L'observation des effets de la rétribution positive ou négative de nos partenaires, selon la nature de leur attitude préalable à notre égard, est précisément ce qu'a permis une expérience menée en 2004 par Dominique de Quervain et ses collègues à l'Institut de recherches empiriques en économie de Zurich [10]. Ces chercheurs ont procédé à l'imagerie cérébrale de la phase finale d'un jeu consistant en la décision d'une stratégie de rétribution ou de punition faisant suite à la perception du comportement d'un partenaire interprété comme socialement correct ou incorrect en termes d'équité et de réciprocité.

Le jeu s'appelle à présent le « jeu de la confiance ». Dans cette expérience, les sujets interagissent avec d'autres joueurs à qui ils distribuent des sommes d'argent. Le premier joueur A et son partenaire B reçoivent chacun au départ du jeu une somme de 10 €. Ils peuvent augmenter leur dotation de manière très intéressante, si le joueur A décide de faire confiance

au joueur B. Dans ce cas, A peut décider de transférer la totalité de sa dotation de 10 € à B. S'il le fait, cette somme de 10 € est multipliée par quatre et, à cette étape du jeu, A n'a donc plus d'argent et B détient 50 € : ses propres 10 € initiaux et les 40 € qu'il a obtenus grâce à la confiance que A met en lui. B a alors le choix de tout garder pour lui-même ou bien de diviser son pécule en deux et de retourner 25 € à A. Rien, à cette phase du jeu, ne l'oblige à le faire : B peut s'enrichir et A n'a qu'à regretter d'avoir espéré quelque chose en retour de sa part. Ou B peut considérer que son enrichissement était le fait de la confiance, certes peut-être intéressée, de A et il estimera juste de partager des gains dont il n'est pas entièrement responsable. B remettra alors les deux joueurs à égalité et mutuellement enrichis relativement à leur situation initiale. B est, en tout cas, en situation de peser le pour et le contre de sa décision, sous l'œil scrutateur de A.

Après la décision de B entre un comportement égoïste et unilatéral et un comportement de réciprocité en réponse à la confiance que A a mise en lui, le joueur A, c'est l'intérêt de ce jeu, conserve la possibilité de réagir. A, dans le cas où B n'a pas fait preuve de réciprocité en réponse à sa confiance, peut administrer à B des points de pénalité affectant les phases ultérieures de ce jeu qui se répète plusieurs fois.

Dans la condition la plus remarquable de l'expérience de D. de Quervain, ces points de pénalité ont un coût monétaire pour A. A a perdu son argent mais il veut punir, il peut prendre des unités sur sa prochaine dotation dans le jeu pour imposer une sanction monétaire à B. Pour tout point de pénalité attribué à B — qui correspondra à une unité de moins sur la dotation de ce dernier lors du prochain tour —, A perd deux unités. Dans ce cas de figure, au coup suivant, A aura 8 € et B aura 9 € pour recommencer le jeu. Le joueur A se trouve encore désavantagé par rapport au joueur B, mais il aura, du moins peut-on le présumer, la satisfaction d'avoir puni B pour son comportement précédent. C'est au moment du choix de punir ou de ne pas punir un joueur qui a enfreint la règle tacite de réciprocité dans ce jeu et d'encourir un coût monétaire pour imposer la punition que D. de Quervain et ses collègues décident de scanner le cerveau des sujets à l'aide de la technique de tomographie par émission de positrons, une technique d'enregistrement de l'activité cérébrale basée sur le comportement radioactif d'une molécule ingérée par le sujet. On observe alors une activation significative d'une région du cerveau nommée le noyau

caudé, associée habituellement au traitement de la récompense et de la satisfaction, dans la condition dans laquelle les sujets indiquent un fort désir de punir leur partenaire et peuvent effectivement le faire.

La punition, dans le contexte où la confiance que l'on a placée dans un tiers est déçue, devient une source de satisfaction. Nous passons de la frustration au soulagement, et cela modifie du tout au tout la nature des processus neuronaux que notre cerveau engage dans ces circonstances. Si une voie d'action est ouverte, si nous n'assistons plus passivement aux injustices qui sont commises à notre endroit, ou à ce que nous percevons intensément et physiquement comme tel depuis notre état d'inaction, les circuits cérébraux du plaisir prennent le pas sur ceux du dégoût.

Il faut ainsi clairement souligner que ce qui a changé entre le jeu de l'ultimatum et le jeu de la confiance avec punition possible est la possibilité d'une réplique, ce qui signifie, au fond, un engagement et une responsabilité du sujet vis-à-vis de l'instauration de la justice sociale (la société étant ici schématisée par deux joueurs qui échangent de l'argent !) au lieu, peut-être, d'une attente gracieuse spontanément établie. Nous avons, en effet, laissé le sujet du jeu de l'ultimatum dans un état d'expectative quant aux conséquences de ses choix. Nous le prenons, dans le jeu de la confiance, dans une position dans laquelle il peut mesurer les effets de sa décision de signaler sa préférence pour des normes de réciprocité dans l'échange. L'envoi de ce signal est une source de satisfaction, quand bien même il revêt un coût non négligeable. Le constat remarquable de D. de Quervain est que punir un partenaire qui fait défection est une source de satisfaction pour le sujet objectivée par l'activation d'aires du cerveau typiquement associées au traitement des récompenses. L'expérience de D. de Quervain et de ses collègues présente ainsi l'intérêt de relier l'instauration et la régulation de normes sociales de réciprocité et de coopération à l'étude des mécanismes de traitement de la récompense par le cerveau.

Cette expérience jette un éclairage inédit sur l'énigme que pose la coopération dans le cadre d'une explication évolutionniste du comportement humain. La coopération induit un bénéfice global pour le groupe social, mais elle a un coût pour les individus qui coopèrent. Ces derniers doivent sacrifier une partie de leurs gains possibles. La coopération étant coûteuse pour les individus altruistes, la défection est donc d'autant plus profitable pour ceux qui dans

la société décident de faire cavalier seul. Comment, dans ces conditions, la coopération et la réciprocité peuvent-elles s'établir et se maintenir ? Comment peuvent-elles former des traits durables des sociétés humaines ? Une manière simple de résoudre cette difficulté conceptuelle est de considérer que les cavaliers seuls peuvent être punis par les autres membres de la société, et que l'administration des punitions ne représente pas un coût insurmontable pour ces derniers. Et elle sera, de fait, d'autant moins coûteuse qu'elle se trouvera, par le truchement de mécanismes neurobiologiques anciennement implantés, accompagnée d'une expérience de satisfaction et de réconfort. La conclusion de l'expérience est que la corrélation fonctionnelle de la punition avec le circuit cérébral de la récompense en fait un comportement généralement motivant, et non le fruit d'une disposition mentale étrange, d'une pulsion de vengeance qui ne demande qu'à être libérée ou d'un sursaut éthique idiosyncrasique. D'autres répertoires d'actions sont évidemment à notre disposition dans les interactions sociales, à commencer bien sûr par l'égoïsme et aussi l'indifférence face à l'injustice, à l'égard des autres ou de nous-mêmes. Mais l'engagement moral et affectif envers la justice, à travers notamment la décision de punir les individus antisociaux, est un comportement socialement rentable et ancré profondément dans les circuits cérébraux anciens du plaisir. Nous sommes donc, ici, en présence d'une expérience de neuroéconomie que l'on peut immédiatement relier à une explication fonctionnelle et évolutionnaire du comportement.

Peut-on chercher à intégrer plus avant données neuroscientifiques et théorie économique ? La théorie des jeux repose sur l'hypothèse que les individus puissent prédire les actions des autres. Un joueur est en principe capable de considérer le jeu depuis sa propre perspective comme depuis celle de son adversaire. Les découvertes des neurosciences sociales sur nos capacités à lire les intentions des autres et à nous mettre à leur place, jusqu'à pouvoir en partager les émotions, semblent directement pertinentes dans le cadre d'une réflexion sur les fondements naturels de l'économie. Cependant, le décodage des intentions d'autrui et le partage entre individus d'états affectifs – l'empathie – mobilisent, chacun, des bases neuronales distinctes et il est intéressant de noter que les deux contribuent à l'efficacité de nos prédictions lors de nos interactions sociales. Si les notions liées à l'anticipation stratégique des inten-

tions forment finalement déjà le noyau conceptuel de la théorie des jeux, des raffinements de cette dernière liés à la logique des affects qui président à certaines de nos décisions sont certainement désirables.

Plus généralement l'étude des mécanismes neurobiologiques qui sous-tendent notre relation à autrui – qu'il s'agisse d'émotions ou d'anticipations cognitivement explicites – peuvent permettre de resituer des pans entiers de la théorie des jeux dans une perspective évolutionnaire réaliste. Les phénomènes de coordination pure (coordination sans communication entre les joueurs) étudiés initialement par Thomas Schelling [11] ont reçu après lui deux types principaux d'explication cognitive. En termes de hiérarchie cognitive [12] : les joueurs développent des méta-représentations sur les représentations du jeu par les autres joueurs ; ou en termes de raisonnement d'équipe [13] : les joueurs prennent d'emblée un point de vue collectif sur le jeu qui expliquerait la convergence de leurs réponses comportementales. Ces deux hypothèses supposent des ressources différentes en termes de théorie de l'esprit, et plus particulièrement la possibilité de trancher entre l'hypothèse de représentation des intentions d'autrui ou de simulation émotionnelle des états mentaux d'autrui. Sur ce point, des études de neuroéconomie peuvent permettre d'éclairer les conceptualisations adéquates des capacités des agents dans les problèmes de coordination pure.

D'autres problèmes économiques peuvent être abordés du point de vue des ressources cognitives que leur solution présuppose. À titre d'exemple, les problèmes liés à l'asymétrie d'information entre agents économiques ont donné lieu à différentes théories, essentiellement la théorie du signal et la théorie des contrats. Comme dans l'exemple précédent concernant la coordination pure, la question est de savoir quelles hypothèses concernant le développement de l'esprit humain peuvent être incorporées de manière heuristique au sein de ces théories économiques. Des expériences neuroéconomiques sont ainsi envisageables à la jonction de l'économie des contrats et de l'hypothèse d'un module de détection de la fraude (*cheating-detection module*) désormais classique, qui avait été avancée par Leda Cosmides et John Tooby [14]. L'existence d'un tel module constitue en soi un raccourci dans le traitement de situations complexes du point de vue informationnel et peut donner lieu à des traitements rapides et généralement fiables de la situation.

## Comment le cerveau humain s'est-il adapté à un environnement économique artificiel et complexe ?

Les deux exemples d'expériences de neuro-économie que nous avons discutés, l'expérience de G. Coricelli et A. Sirigu sur les regrets anticipés et l'expérience de D. de Quervain sur la punition altruiste, ainsi que les prolongements que nous en avons envisagés, dans une perspective fonctionnaliste et évolutionniste, amènent à préciser ce qui, selon nous, serait la question centrale d'un programme de recherche unifié et cohérent en neuroéconomie. À savoir : comment le cerveau et l'esprit humains se sont-ils adaptés (avec certaines limites) à un environnement complexe et artificiel très récent à l'échelle de leur évolution ?

Il y a plusieurs manières d'aborder cette question. Nous nous permettrons, pour conclure, de faire référence à deux de nos travaux récents qui se situent dans cette perspective. Tout d'abord, une première approche consiste à explorer les conditions d'un élargissement de l'hypothèse de recyclage culturel des cartes corticales, développée par Stanislas Dehaene et Laurent Cohen [15] dans le contexte de la lecture et de l'arithmétique élémentaire, aux cas des artefacts économiques de base comme la monnaie. S. Dehaene et L. Cohen montrent qu'une partie du cortex humain est spécialisée dans des domaines culturels, tels que la lecture et l'arithmétique, alors que ces inventions sont trop récentes pour avoir pu influencer notre architecture fonctionnelle cérébrale. Ce qui est également frappant est que le traitement de ces stimuli culturels par le cerveau est invariant entre les individus, et ceci à travers le monde. Une même architecture cérébrale assure systématiquement notre traitement des lettres et des nombres. S. Dehaene et L. Cohen expliquent ce paradoxe en formulant l'hypothèse que les inventions culturelles récentes pénètrent des circuits cérébraux anciens et héritent au passage de leurs contraintes structurelles.

Peut-on élargir cette hypothèse au traitement par le cerveau de situations et d'artefacts économiques apparus très récemment à l'échelle de l'histoire de l'espèce humaine ? C'est ce que semblent montrer les résultats d'une expérience menée récemment en imagerie cérébrale par l'auteur de cet article avec Catherine Tallon-Baudry et Florent Meyniel à l'hôpital de La Salpêtrière. Le cerveau utilise systématiquement les zones du cerveau dédiées par l'évolution à la reconnaissance des visages pour traiter des

stimuli monétaires qui ont de la valeur, par rapport aux stimuli monétaires qui ont été démonétisés, et ce indépendamment de la familiarité que les individus ont avec l'un ou l'autre type de stimuli monétaires. Cela montre que pour traiter rapidement des pièces de monnaie, dans la vie courante en somme, le cerveau utilise des ressources anciennes, associées à un traitement automatique des stimuli, et de manière fiable. Autrement dit, nous n'avons pas changé biologiquement au cours de notre histoire récente, même si notre environnement, lui, a radicalement changé, en particulier du point de vue de la complexité de la réalité économique. Nous avons mis en usage des mécanismes cérébraux anciens, dédiés à des fonctions pertinentes à l'échelle évolutionnaire, pour le traitement de ces modifications au sein de notre environnement.

Une interprétation de nombreux résultats d'expériences en neuroéconomie peut être donnée dans les termes de cette hypothèse d'un recyclage culturel de circuits corticaux anciens. Elle peut être associée à une approche complémentaire, consistant en l'étude des représentations naïves des situations, des artefacts, des institutions et de l'environnement économique en général (nous résumons cette approche sous le label de « folk economics »). L'hypothèse complémentaire, ici, est que notre représentation spontanée du monde économique, en deçà des effets immédiats liés à l'éducation variable des individus, hérite de contraintes évolutionnaires. Nous sommes généralement sensibles à l'injustice, nous ne sommes pas spontanément

utilitaristes, nous (les sujets naïfs) comprenons mal le concept d'inflation [16], etc. À titre d'exemple, l'auteur de cet article étudie actuellement comment les enfants autour de l'âge de trois ans (l'âge auquel se met en place les ressources cognitives de compréhension des états mentaux d'autrui) sont sensibles à des distributions injustes dans des jeux d'allocation. S'ils montrent un sens de l'injustice avant trois ans, cela confirme l'idée que les enfants sont avant tout sensibles aux distributions inéquitables indépendamment de la compréhension des intentions d'autrui. Si, avant trois ans, le sens de l'injustice n'est pas développé, on montre, différemment, sa dépendance avant tout envers une capacité à reconnaître les intentions d'autrui. Cela permet notamment d'aider à trancher entre deux modèles d'aversion à l'inéquité : celui de Matthew Rabin [1] basé sur la perception des intentions d'autrui et celui de Ernst Fehr et Klaus Schmidt [17] basé sur le traitement de distributions inéquitables.

La neuroéconomie n'est pas un programme de recherche nécessairement focalisé sur la mise en évidence de corrélats neuronaux des processus de décision, comme l'avaient rapidement définie Alan Sanfey et ses collègues [18]. Elle est plutôt, selon nous, la mise en convergence de disciplines expérimentales diverses ne mobilisant pas systématiquement des techniques d'imagerie cérébrale, visant à répondre à la question des limites (et de la source de ces limites) de notre représentation du monde économique et, par conséquent, des modalités de notre participation à ce monde. //

## Bibliographie

- [1] RABIN M., « Incorporating fairness into game-theory and economics », *The American Economic Review*, vol. 83, 1993, p. 1281-1302.
- [2] CAMERER C., *Behavioral Game-Theory – Experiments in Strategic Interaction*, Princeton, Princeton University Press, 2003.
- [3] GIGERENZER G., SELTEN R. (ed.), *Bounded Rationality – The Adaptive Toolbox*, Cambridge, MIT Press, 1999.
- [4] RUBINSTEIN A., « Comments on neuroeconomics », *Economics and Philosophy*, vol. 24, 2008, p. 485-494.
- [5] DE MARTINO B. *et al.*, « Frames, biases, and rational decision-making in the human brain », *Science*, vol. 313, 2006, p. 685-687.
- [6] BOURGEOIS-GIRONDE S., PAYZAN É., « Behavioral and neural foundations of framing effects », [@jeannicod.ccsd.cnrs.fr](http://jeannicod.ccsd.cnrs.fr), 2005.
- [7] DAMASIO A., *L'Erreur de Descartes*, Paris, Odile Jacob, 2001.
- [8] CORICELLI G., SIRIGU A. *et al.*, « Regret and its avoidance : a neuroimaging study of choice behaviour », *Nature Neuroscience*, vol. 9, 2005, p. 1255-1262.
- [9] SANFEY A. *et al.*, « The neural basis of economic decision-making in the ultimatum game », *Science*, vol. 300, 2003, p. 1755-1758.
- [10] DE QUERVAIN D. *et al.*, « The neural basis of altruistic punishment », *Science*, vol. 305, 2004, p. 1254-1258.
- [11] SCHELLING T., *The Strategy of Conflict*, Cambridge, Harvard University Press, 1960.
- [12] CAMERER C., HO TECK H., CHONG J.-K., « A cognitive hierarchy model of games », *The Quarterly Journal of Economics*, vol. 119, 2004, p. 861-898.
- [13] SUGDEN R., « The logic of team-reasoning », *Philosophical Explorations*, vol. 6, 2003, p. 165-181.
- [14] RUBINSTEIN A., « Comments on neuroeconomics », *Economics and Philosophy*, vol. 24, 2008, p. 485-494.
- [15] DEHAENE S., COHEN L., « Cultural recycling of cortical maps », *Neuron*, vol. 56, 2008, p. 384-398.
- [16] LEISER D., DRORI S., « Naive understanding of inflation », *Journal of Socio-Economics*, vol. 34, 2005, p. 179-198.
- [17] FEHR E., SCHMIDT K., « A theory of fairness, competition and cooperation », *Quarterly Journal of Economics*, vol. 114, 1999, p. 817-868.
- [18] SANFEY A. *et al.*, « Neuroeconomics: cross-currents in research on decision-making », *Trends in Cognitive Sciences*, vol. 10, 2006, p. 108-116.