Les émotions économiques

Sacha Bourgeois-Gironde

Revue européenne des sciences sociales, Tome XLVII, 2009, N° 144, pp. 43-56

Sacha BOURGEOIS-GIRONDE

LES ÉMOTIONS ÉCONOMIQUES
Réflexions sur les mécanismes d'adaptation cérébrale à l'environnement socio-économique

I. EMOTIONS ET RATIONALITÉ

Les relations entre émotions et rationalité posent des problèmes de différentes sortes. Il s'agit parfois de savoir si les émotions peuvent être rationnelles (Elster 1996) et parfois de se demander si les émotions peuvent contribuer à la rationalité (Damasio 1996). Evidemment si elles contribuent à la rationalité – si elles sont par exemple indispensables à la possibilité d'optimiser ses gains au cours d'une série de choix – les émotions sont alors rationnelles dans un sens trivial. Mais ce n'est pas cette question en priorité qui est posée quand des philosophes se demandent si les émotions sont rationnelles. Ils veulent plutôt se demander par là si les émotions peuvent nous apprendre quelque chose, et si ce qu'on peut apprendre à travers les émotions est cohérent, en lui-même, ainsi que relativement à ce qu'on peut apprendre par d'autres canaux cognitifs. Si les émotions ont une dimension cognitive, alors elles peuvent contribuer certainement à la rationalité, pour autant que ces enseignements de l'émotion soient compatibles plus généralement avec les enseignements de mes autres canaux cognitifs. Mais les émotions peuvent aussi contribuer à la rationalité sans nécessairement comporter en elles-mêmes cette dimension cognitive. Elles forment alors des mécanismes non cognitifs dont la présence peut être indispensable à l'usage optimal de mes facultés cognitives indépendantes, conformément alors à l'hypothèse des marqueurs somatiques qu'a défendue Damasio.

On peut différencier plus précisément deux approches concernant les relations entre les émotions et la rationalité. D'une part, les émotions peuvent être considérées comme formant des mécanismes somatiques, liés à une phénoménologie particulière (un signal distinct), me révélant quelque chose sur ma relation à mon environnement. D'autre part, sur la base de ces émotions, et indépendamment même du fait d'avoir été en mesure de déchiffrer ou d'interpréter explicitement le signal délivré par ces émotions, mon comportement peut être infléchi, parfois dans un sens optimal. Autrement dit les émotions peuvent être considérées, alternativement ou conjointement, comme des signaux relativement fiables et comme des poids décisionnels possiblement utiles. Ce sont sous ces deux angles que nous les envisagerons ici. Les regrets présentent bien ces deux caractéristiques. Les regrets sont liés à l'existence ou à la vraisemblance, qui me devient à présent saillante, de conséquences d'un choix que j'ai omis de faire, alors que j'étais en position de le faire, qui sont meilleures que les conséquences d'un choix que j'ai effectivement préféré faire (Zeelenberg 1999). Si j'avais pu anticiper ces regrets, l'émotion alors ressentie m'aurait peut-être guidé vers le choix optimal.

Les émotions que nous avons à l'esprit dans la suite de notre propos sont des émotions que nous sommes susceptibles d'éprouver dans le contexte de choix, et plus précisément dans le contexte de choix économiques où des normes de rationalité ou d'optimalité sont en jeu. Outre les regrets qu'on peut apprendre à minimiser dans nos prises de décision, le dégoût forme un autre exemple. Il constitue une réaction viscérale que l'on a pu identifier dans le cas de jeux de partage, comme le jeu de l'ultimatum (Sanfey et al. 2003). Dans le jeu de l'ultimatum un joueur A dispose de 10 euros et peut en conserver une partie s'il fait une offre de partage de cette somme à un joueur B. Si B accepte l'offre, le partage sera effectif ; si B refuse l'offre aucun des deux joueurs ne retirera un gain de cette tentative d'échange. Il est rationnel (au sens de l'équilibre de Nash) que le joueur A donne le plus petit montant et que B l'accepte. Mais les équilibres sociaux constatés dans ce jeu sont distincts de la prédiction de la théorie de la rationalité et les offres modales acceptées se situent systématiquement entre 4 et 5 euros. Sanfey, dans la première étude de neuroéconomie qui a été popularisée, observe les activités cérébrales des sujets (joueurs B) dans l'ultimatum lorsqu'ils refusent une offre qu'ils perçoivent comme injuste. Le cortex insulaire, dont l'activité est typiquement liée aux états proprioceptifs de nausée ou de dégoût, est fortement impliqué dans le rejet de ces offres basses.

Dans ce cas le dégoût est bien corrélé à l'idée qu'un partage s'est effectué à mon détriment. Mais est-il rationnel dans le sens où il contribuerait par ailleurs à l'optimisation de mes gains ? C'est beaucoup moins clair. Si le jeu de l'ultimatum n'est joué qu'une fois, le rejet de l'offre sur la base de mon dégoût ne me permet pas de maximiser mon utilité monétaire dans ce jeu, au contraire. Je lèse l'autre joueur parce que je sens qu'il m'a lésé, mais je continue à me léser davantage en refusant son offre qui m'a paru inique. Il faut alors resituer le dégoût dans une perspective un peu plus vaste. En premier lieu, si l'on regarde d'un peu plus près les résultats de Sanfey, on pourra voir que les activités de l'insula, liées à l'émotion de dégoût, sont corrélées à des activités du cortex préfrontal dorsolatéral, associées de leur côté à un effort cognitif de délibération. Les données de Sanfey montrent que plus on se rapproche d'offres basses susceptibles d'être rejetées, mais qui ont été de fait acceptées, plus on trouvera un haut niveau d'activité conjointe du cortex préfrontal dorsolatéral et de l'insula, comme si les premières activités, liées à la délibération, avaient servi à inhiber une réponse émotionnelle viscérale. Il y a donc une forme d'apprentissage possible de contrôle de sa réponse émotionnelle. Par ailleurs, quand le jeu est répété, il n'est pas nécessairement contraire à l'intérêt du sujet de ne pas inhiber sa réponse émotionnelle. De fait elle correspond à l'envoi d'un signal négatif et d'une punition monétaire envers un partenaire de jeu qui s'est montré peu altruiste. Dans la suite du jeu, le joueur A, tenant compte de cette punition, aura tendance à faire des offres plus élevées et l'intérêt du joueur B s'en trouvera maximisé.

Le dégoût est une réponse adaptative. On peut penser plus généralement que les émotions sont des réponses adaptatives, non nécessairement munies d'un contenu cognitif. Il se peut, comme on vient de le voir pour le dégoût dans le jeu de l'ultimatum à un coup, que ces émotions entrent en conflit avec ce qu'il est ponctuellement rationnel de faire (accepter l'offre car je n'aurai pas la possibilité de rejouer et que le signal punitif que j'envoie n'aura donc aucune portée dans la suite). Mais ressaisies dans un contexte d'interactions répétées, leur rationalité

devient manifeste. L'approche neuroéconomique des jeux expérimentaux se focalise à juste titre sur la manière dont les émotions peuvent faciliter ou entraver l'optimalité des décisions. Beaucoup des études proposées en neuroéconomie depuis 2003, à la suite finalement du programme de recherche de Damasio entamé dix ans plus tôt, ont cherché à rendre compte du conflit ou au contraire de l'intégration entre les contributions des aires cérébrales dédiées respectivement à l'émotion et à la décision et au contrôle cognitif. Au fond la neuroéconomie permet de dresser un portrait de l'homo oeconomicus enrichi par ces mécanismes adaptatifs, éventuellement inconscients, que sont les émotions.

La neuroéconomie repose toutefois sur des prémisses épistémologiques encore fragiles. Il n'est pas clair que l'économie puisse directement en bénéficier. Les arguments le plus souvent présentés par des économistes favorables à la neuroéconomie consistent à imaginer qu'il est possible d'obtenir une nouvelle mesure objective et observable des préférences des individus (Camerer 2005). De l'autre côté, l'apport pour les neuroscientifiques tient à la mise en place de protocoles d'économie expérimentale en imagerie cérébrale, permettant peut-être d'arriver à des observations inédites sur, par exemple, les mécanismes de décision, l'encodage de la valeur monétaire, les modalités de coordination dans des jeux d'échanges économiques. On est loin, dans les faits, d'une intégration des deux disciplines et peut-être que celle-ci n'a pas vraiment de sens. Il n'y a pas de parallèle que l'on pourrait fermement soutenir entre l'articulation entre la psychologie et les neurosciences d'un côté et entre l'économie et les neurosciences de l'autre, dans la mesure où il n'y a pas en économie d'hypothèses sur le fonctionnement de l'esprit, comme il y en a en psychologie, dont les mécanismes biologiques sous-jacents attendent d'être explicités. Pour que la neuroéconomie puisse avoir un sens il faudrait déjà commencer par dire pourquoi l'économie aurait besoin d'hypothèses psychologiques. C'est précisément ce point que contestent Gul et Pesendorfer (2005) dans un article qui passe pour un manifeste hostile aux relations entre l'économie et les sciences de l'esprit ou du cerveau.

Mais c'est aussi un point que l'on peut contester. L'économie a commencé en grande partie comme une science psychologique et morale dont les hypothèses ont tellement bien été absorbées à sa démarche qu'elles ont quasiment cessé de véhiculer un contenu descriptif dont on pourrait se soucier de la validité. Sans entériner l'idée archéologique de Camerer, et d'autres, que les mesures de l'activité cérébrale dans les tâches de décision nous permettent de revivifier la notion d'utilité cardinale, on peut trouver d'autres motifs fondamentaux à la proposition d'un croisement des modèles économiques et des données psychologiques et neurobiologiques. Un exemple, qui est largement exploité par un autre tenant enthousiaste de la neuroéconomie, Ernst Fehr, consiste à explorer systématiquement les bases neuronales de nos comportements altruistes pour découvrir s'ils sont conditionnés plutôt par des aptitudes et des considérations stratégiques ou par des dispositions et des émotions purement pro-sociales (Singer et Fehr 2005). Une autre question intéressante, dans ce contexte, est de savoir si la perception des intentions d'autrui est réellement une ressource indispensable à l'implémentation d'une stratégie rationnelle dans un jeu, ou si au contraire le développement, dans l'enfance, de cette capacité cognitive aura tendance à nous faire dévier des équilibres de Nash. Chercher à comprendre en vertu de quels mécanismes cognitifs et neuronaux nos comportements sont plus ou moins conformes aux prédictions de

la théorie de la rationalité est certainement un effort pertinent en marge de l'économie.

II. LA NEUROÉCONOMIE DANS UNE PERSPECTIVE ÉVOLUTIONNISTE

Une conception alternative possible de la neuroéconomie est de la comprendre comme une méthode d'investigation des modalités d'adaptation de notre cerveau à un environnement économique artefactuel. C'est comme cela que nous la comprenons. La neuroéconomie permet de soulever des questions de nature évolutionniste et éventuellement d'apporter des éclairages inédits sur la manière dont, au cours de l'histoire récente de l'humanité, nous avons pu mobiliser, avec plus ou moins de bonheur, des ressources et des mécanismes cérébraux qui ont été développés sur une longue échelle historique afin de nous adapter à des mutations rapides et drastiques de notre environnement.

Ces mutations, comme le montre très bien par exemple Jean-Paul Demoule, commencent au Néolithique. Sa conception de la protohistoire amène à se demander ce que la transformation cognitive qui a présidé à la révolution néolithique a signifié en termes de décalage entre mécanismes adaptatifs anciens et nouvel environnement économique. Dans certains cas on peut supposer que les hommes ont remobilisé avec succès des mécanismes cérébraux adaptés aux nécessités de leur ancien environnement dans leurs conditions de vie nouvelles. Dans d'autres cas on peut imaginer au contraire que les mutations de l'environnement, et notamment la complexification des rapports économiques, qui requièrent un renouvellement des ressources adaptatives du cerveau humain, ont échoué à susciter des réactions adaptatives du cerveau. Y a-t-il des émotions propres à l'incapacité éventuelle de s'adapter à un environnement inédit, c'est-à-dire non pas des émotions inadéquates en vue de la réalisation d'une tâche, mais des émotions renvoyant à l'inadéquation entre mes capacités et une exigence de l'environnement ?

Certains biais cognitifs seraient dus à un décalage entre notre environnement humain contemporain artefactuel de décision et le niveau d'évolution actuel de notre cerveau. Nous ne serions plus adaptés à notre environnement et des émotions propres à cette inadéquation pourraient exister. En fait, il y aurait, d'un point de vue évolutionniste, deux façons de concevoir les biais cognitifs qui affectent en particulier la prise de la décision, le traitement des probabilités, la formation des croyances et des préférences. En un sens, les biais cognitifs sont des stratégies mentales et comportementales que nous avons précisément développées au cours de l'évolution et qui jouent donc un rôle adaptatif. Etant donné la rapidité avec laquelle il faut traiter parfois l'information, ce qu'on appelle les biais cognitifs forment des raccourcis optimaux. Mais parfois, alors que l'information est parfaitement disponible, les biais cognitifs peuvent révéler de véritables limites et une incapacité de notre part à traiter cette information de manière adéquate.

C'est selon cette division des biais que l'on peut donner du sens à une enquête sur les relations entre émotions et rationalité. Certains états émotionnels, c'est du moins l'hypothèse que nous proposons, signalent notre statut, plus ou moins temporaire, d'individus inadaptés à l'environnement. D'autres émotions, comme on l'a vu, jouent plus directement un rôle adaptatif et infléchissent nos comporte-

ments dans un sens optimal. La neuroéconomie peut se définir comme la mise en place de nouvelles formes d'investigation sur un problème particulier lié à l'évolution : les mutations de l'affectivité, notamment lors du passage, il y a environ 10 000 ans, à un mode de vie sédentaire, à l'accumulation des richesses, à la naissance de la propriété privée, à la division sociale, à l'apparition de la monnaie, etc. La genèse de l'économie moderne à cette période-là était également celle, probablement, d'une néo-affectivité, d'un ensemble d'émotions nouvelles liées au plus au moins grand succès avec lequel nous mobilisions nos ressources cérébrales anciennes en vue du traitement de stimuli socio-économiques inédits.

Nous faisons l'hypothèse, fortement spéculative et qui reste entièrement à étayer, à la fois par une archéologie cognitive des contextes affectifs anciens et par des protocoles neuroéconomiques pertinents, que la mutation de l'environnement économique immédiat de l'homme qui s'est produite au Néolithique a conduit d'une paléo-affectivité, dans laquelle les émotions constituaient des signaux généralement fiables et adaptatifs, à une néo-affectivité, dans laquelle apparurent des émotions spécifiques liées à la perte d'intelligibilité immédiate de l'environnement et au sentiment de ne plus y appartenir.

Un exemple qui vient à l'esprit est celui de l'exclusion sociale. En quoi consiste l'expérience de l'ostracisme ? Une étude de neurosciences sociales, menée par Naomi Eisenberger et Matthew Lieberman, sur les bases neuronales de l'exclusion sociale peut être resituée dans la perspective plus large que nous cherchons à dessiner, entre paléo-affectivité et néo-affectivité (Eisenberger et Lieberman 2003). Ces auteurs ont proposé une situation, basée sur un jeu nommé le Cyberball, qui consiste à placer un sujet d'expérience devant un écran d'ordinateur sur lequel sont figurés trois joueurs, dont l'un représente le sujet lui-même, en train de s'envoyer une balle. Après quelques échanges les deux joueurs virtuels vont exclure le sujet de la partie et celui-ci va ressentir la douleur d'être exclu du jeu, d'autant qu'il est persuadé d'avoir joué contre deux sujets humains situés dans une autre pièce. Ce simple jeu, d'apparence anodine, génère des sentiments extrêmement vifs de frustration sociale, de confiance trahie, de rejet. Einsenberger et Lieberman ont concentré leurs observations des activités cérébrales des sujets sur le cortex cingulaire antérieur. Cette région du cerveau s'active en effet davantage durant les périodes d'exclusion que durant les périodes d'inclusion des sujets dans le jeu. Mais cette région est également habituellement impliquée dans l'expérience de la douleur physique. Pourquoi une même région du cerveau, au cours de son évolution fonctionnelle, en est-elle venue à traiter des stimuli aussi différents que, par exemple, une brûlure et le fait de se sentir rejeté d'une activité humaine collective ? Une réponse tient au fait que le cortex cingulaire antérieur est spécialisé dans le traitement des stimuli ou des situations qui entrent en conflit avec nos attentes. Nous nous brûlons généralement par surprise et, de même, une situation d'exclusion sociale est finalement quelque chose auquel nous sommes peu préparés. Une autre réponse, plus évidente, est que notre survie dépend de notre intégration au sein d'un groupe de congénères. Si nous nous trouvons exclus de ce groupe il est normal qu'un puissant système d'alarme, en l'occurrence des mécanismes neuronaux liés à la douleur, se déclenche. Les circuits cérébraux de la douleur étaient présents dans le monde animal bien avant que des sociétés humaines complexes ne se développent, ils sont disponibles pour générer les signaux d'alarme propres au traitement du rejet social.

A nouveau nous mobilisons de manière automatique des ressources cérébrales anciennes en vue du traitement de situations qui revêtent un certain degré de nouveauté et de complexité. La douleur liée au rejet social est un marqueur somatique puissamment sensible, mais le cerveau n'a pas toujours une réponse adaptative aux sollicitations environnementales. Pensons au cas de l'addiction, sur lequel nous reviendrons. On peut penser que des sollicitations inédites dans l'environnement favorisent les comportements compulsifs et addictifs et que les mécanismes d'alarme générés par le cerveau s'avèrent, en l'occurrence, inefficaces. Les émotions liées à l'inadaptation d'une réponse comportementale, devenues inefficaces, renforcent alors assurément le sentiment d'inadéquation chez l'individu concerné. Ces émotions sont suscitées par des conflits internes. Ces conflits internes ne concernent pas toujours des enjeux aussi vitaux que les addictions et peuvent simplement consister dans la concurrence entre deux réponses comportementales contradictoires entre lesquelles nous oscillons. Il existe des conflits, des points de friction, au sein de notre architecture cognitive modulaire et certains organismes sophistiqués ont pu développer, au cours de leur évolution, une sensibilité à ces conflits internes via des émotions d'un type particulier : des regrets anticipés, des sentiments d'erreurs, ou ce qu'on nomme dans la littérature neuro-computationnelle des signaux d'erreur fictive (Niv et Schoenbaum 2008). Au sein d'un environnement complexe en mutation susceptible de provoquer des comportements erronés, sous la forme de biais comportementaux et cognitifs, ces émotions jouent un rôle essentiel et les deux aspects des relations entre émotions et rationalité que nous avons évoqués plus haut sont rendus saillants : elles peuvent infléchir le comportement dans une direction optimale, elles forment des indications du caractère plus ou moins temporairement inadéquat d'une de mes réponses comportementales.

III. BIAIS COGNITIFS ET CRITÈRE DE RATIONALITÉ

L'étude des biais cognitifs, et des heuristiques sur lesquelles reposent ces biais, s'est étendue sur une trentaine d'années (Kahneman 2003) mais, au sein de cette histoire, a rarement été clairement formulé un critère qui indiquerait que la présence d'un biais cognitif chez un individu serait une marque de son irrationalité. L'idée dominante est sans doute que, même si les biais sont des déviations par rapport à des normes de raisonnement ou de décision, les heuristiques sur lesquels ils reposent sont aussi des moyens rapides et frugaux pour traiter une information et prendre une décision sous contrainte. Une idée encore plus déculpabilisante, du point de vue de l'attribution de l'irrationalité à l'agent présentant un biais, est que les situations dans lesquelles on met en évidence des biais cognitifs sont très peu écologiques et que ces biais seraient en fait des artefacts expérimentaux (Gigerenzer 1999). Selon ce dernier argument, dans notre environnement habituel ou naturel, les réponses que nous produisons spontanément sont adaptées, et les biais ne sont des biais qu'au regard d'un type de situations expérimentales dans lesquelles on cherche à éliciter une réponse non naturelle de l'agent à un type de problèmes qu'il a par ailleurs de bonnes raisons de traiter autrement. Témoigne de ce clivage méthodologique et interprétatif le débat sur les violations des règles

élémentaires du calcul des probabilités. La plupart des biais dans le raisonnement probabiliste qu'ont mis en évidence Kahneman et Tversky supposent chez les sujets la mise en œuvre de règles et d'axiomes qui n'ont été explicités que très récemment d'un point de vue historique. Le Pléistocène est alors souvent mobilisé, par les tenants d'une rationalité écologique, pour rappeler que nos ancêtres estimaient des fréquences d'événements (le passage d'une proie, l'occurrence d'une sorte de baies) et non pas des probabilités a priori d'événements singuliers.

Dans le déroulement du programme de recherche « heuristiques et biais », il y a donc eu des tentatives plus marquantes en vue de la déconstruction des biais cognitifs qu'en vue de la clarification des critères qui permettraient de dire qu'un biais est irrationnel. Lorsque les axiomes de la théorie de la décision de Von Neumann et Morgenstern et de Savage ont commencé, à la fin des années 1960, à être soumis à des tests empiriques, certains critères de rationalité ont été mis en avant. En particulier, Allais, qui avait dès 1952 proposé une anomalie qui semblait remettre en cause l'axiome d'indépendance de Savage, avait indiqué que ces anomalies ou ces violations des axiomes ne prenaient sens que si on avait affaire idéalement à un individu raisonnable, c'est-à-dire à un individu qui sait quel axiome il viole, qui en comprend la portée et qui est capable de justifier son comportement. C'est ce critère que Slovic et Tversky, en 1974, retiennent dans leur discussion des résultats expérimentaux sur le paradoxe d'Allais. Etre biaisé en connaissance de cause, pour ainsi dire, pouvait devenir le critère d'une rationalité préservée.

Ce recours à l'exposition cognitive des agents aux normes de rationalité qu'ils tendent à violer et à l'adhésion cognitive aux principes alternatifs qu'ils paraissent suivre relève d'une approche qui est en réalité peu compatible avec une compréhension des biais cognitifs dans une perspective évolutionniste. Les biais cognitifs en tant que stratégies adaptatives ne dépendent en principe pas de processus psychologiques conscients. Naturellement si un biais cognitif est rendu saillant à l'esprit d'un individu qui vient d'y succomber, celui-ci aura peut-être tendance à le corriger ou au contraire à l'assumer, mais cela n'indiquera rien sur le caractère adaptatif ou nuisible du biais en question. Les biais découlant de réponses automatiques basées sur des heuristiques, leur prise de conscience correspond tout simplement à leur inhibition et à la mise en place de stratégies comportementales contrôlées dont l'issue peut être effectivement compatible ou non avec les réponses spontanées initiales. Quand il y a incompatibilité entre la réponse initiale basée sur une heuristique et une réponse contrôlée réfléchie, il y a certainement lieu de parler de biais dans le premier cas, mais la lucidité récemment acquise n'indique pas forcément que la réponse initiale était dépourvue de rationalité. Ce qui se produit dans la transition d'une réponse automatique biaisée à une réponse réfléchie normative est peut-être tout simplement le changement de la question. C'est sans doute ce point qu'ont en tête les tenants d'une rationalité dite écologique, ou du moins ceux qui disent que les attributions de rationalité ou d'irrationalité doivent s'effectuer dans des conditions écologiques. Les sujets ont tendance à répondre spontanément à une certaine question, celle qui leur vient à l'esprit, au détriment de la question moins habituelle qui lors d'une expérience peut leur être posée.

Ce qui nous importe ici est qu'il y a des réponses spontanées qui constituent des erreurs au regard de certaines normes de rationalité. Ces réponses spontanées

peuvent trouver une justification si on les resitue dans un contexte plus large que le protocole expérimental à travers lequel elles sont élicitées et qui englobe les traits environnementaux auxquels ces réponses sont en réalité adaptées. Le protocole expérimental peut alors être conçu comme un contexte artificiel – et les biais sont pour les tenants de la rationalité écologique comme Gigerenzer des artefacts expérimentaux – dans lequel on sollicite une réponse qui peut de prime abord sembler inadéquate au sujet. A-t-il le sentiment de cette inadéquation ? Ou au contraire finit-il par comprendre qu'il lui faut, étant donné le contexte, fournir une réponse différente de celle vers laquelle il tend spontanément ? Quand on fait réfléchir les individus au sujet de leurs réponses erronées dans les tâches typiques proposées par Kahneman et Tversky les réponses sont variables selon les biais que l'on considère (voir Stanovich et West 2000) mais ils auront tendance à justifier leur comportement. A titre d'exemple les sujets qui sont victimes d'effets de cadrage ne se rangent pas au principe normatif selon laquelle deux descriptions distinctes d'un même état de choses ne doivent pas modifier les préférences vis-à-vis de cet état de choses. Certes ils acceptent que c'est le même état de choses qui est décrit deux fois de façons différentes, mais ils donnent du poids à cette description (voir Frisch 1992, Sher et McKenzie 2006, Bourgeois-Gironde et Giraud 2009). Certains peuvent même penser que ce n'est pas le même état de choses qui est décrit, qu'il n'y a pas d'équivalence extensionnelle de principe entre les états de choses présentées par les descriptions successives dans les scénarios d'effets de cadrage proposés par Kahneman et Tversky (Livet, dans ce volume). La raison qui fait que les effets de cadrage n'apparaissent pas spontanément comme des violations d'un principe de rationalité pour les sujets est que ceux-ci tendent à resituer l'énoncé des problèmes posés dans un contexte conversationnel où le choix d'une description plutôt qu'une autre est en soi un élément qui véhicule une information discriminante.

Pouvoir expliquer les biais cognitifs en réinterprétant les problèmes qui les suscitent de manière conforme à un usage habituel de leurs dispositions à raisonner, calculer ou décider, n'exclut pas que sous l'angle plus particulier de la norme qui était visée par la position de ces problèmes, les sujets commettent une erreur manifeste. Si on fixe l'attention des sujets sur cette norme et qu'on la place en opposition avec les réponses qu'ils ont spontanément tendu à fournir, les justifications qu'ils donnent ensuite pour leurs réponses ne viennent que renforcer l'idée qu'ils ne se sentent pas en adéquation avec les problèmes posés. Autrement dit ces réponses biaisées à ces problèmes ne montrent certainement pas que nous sommes incapables de raisonner, elles n'indiquent certainement pas de véritables limites cognitives de notre part, mais simplement notre inadaptation à ces environnements expérimentaux artificiels. Il y a bien des biais vis-à-vis de ces environnements, révélant notre difficulté à répondre correctement aux questions qui nous sont réellement posées. Quand ces problèmes nous sont imposés, d'une manière bien plus pressante que par un psychologue expérimental, quand ils sont devenus le type de problèmes que nous rencontrons de manière répétitive dans notre environnement ordinaire, les justifications que nous pouvons tenter de fournir de nos réponses ne font que rendre plus saillante notre inadéquation. Si l'environnement économique dans lequel nous évoluons est tel qu'il exige l'usage de ressources cognitives différenciées des mécanismes mentaux adaptatifs qui prévalaient à une époque antérieure, les justifications fournies peuvent

finir par ressembler à des aveux d'impuissance, et laisser place à d'authentiques sentiments d'inadéquation.

IV. LES SENTIMENTS D'ERREUR

Un critère indirect pour juger de la rationalité ou de l'irrationalité d'un biais cognitif serait de pouvoir mesurer directement le sentiment d'inadéquation du sujet vis-à-vis de sa réponse à la question posée dans le temps même où il fournit sa réponse automatique, plutôt que d'éliciter un jugement rétrospectif sur le caractère non-normatif de la réponse biaisée. Naturellement le problème est de savoir si de tels sentiments existent et même s'ils sont possibles a priori. Si un biais est basé sur une heuristique inconsciente il faudrait envisager une réaction préconsciente rapide, voire quasi-immédiate, qui formerait un indicateur fiable du caractère erroné de l'heuristique en jeu. Bien qu'enclins à l'erreur nous aurions des moyens fiables de détecter ces erreurs dans un temps suffisamment court pour permettre un ajustement comportemental optimal.

Ces sentiments d'erreur peuvent être compris dans le cadre de l'hypothèse des marqueurs somatiques de Damasio. Damasio étudie le comportement de sujets sains et de patients présentant des lésions ventro-médiales dans une tâche de décision qui implique de réfréner la tentation de poursuivre une stratégie d'accumulation de gains rapides importants mais finalement nuisibles afin d'implémenter une stratégie moins attractive au départ mais finalement rentable. Le changement de stratégie – qui consiste en l'occurrence, dans la tâche de Damasio, à cesser de tirer des cartes dans un tas de cartes au profit d'un autre tas de cartes – serait médié par des marqueurs somatiques – ou états corporels d'émotion – avant qu'affleure à la conscience le sentiment que la stratégie initialement suivie est néfaste puis qu'il devienne explicite que la stratégie alternative est la meilleure à terme. Les patients, pour qui les zones de traitement des émotions sont lésées, ne parviennent pas à implémenter la stratégie optimale. Damasio pense que les signaux d'erreur relatifs à leur comportement sous-optimal ne sont pas générés, et donc qu'a fortiori ils ne peuvent pas orienter la décision dans un sens optimal.

Ici nous avons affaire à une tâche d'apprentissage et l'hypothèse de Damasio a été critiquée pour deux raisons. Premièrement, il n'est pas clair que les sujets n'aient pas conscience très tôt, au moins au même moment que la mise en place des processus somatiques liés à l'émotion – de la validité d'une stratégie plutôt qu'une autre. Autrement dit cette conscience ne serait pas nécessairement médiatisée par une phase émotionnelle préconsciente (Maia et McClelland 2004). Deuxièmement, il n'est pas clair non plus que ces mécanismes émotionnels soient liés au traitement de la sous-optimalité d'une stratégie et non pas plutôt au traitement des risques induits par les valeurs présentes sur les tas de cartes les plus attractifs au départ dans le jeu proposé par Damasio (Tomb et Caramazza 2002). Toutefois, quelles qu'en soient les modalités effectives, les données de Damasio montrent que des processus émotionnels accompagnent, de façon relativement certaine, les stratégies d'adaptation comportementale quand un risque de sous-optimalité est en jeu et c'est ce qui nous intéresse ici.

Semblablement le dégagement hors de comportements néfastes peut être facilité par des états émotionnels advenant progressivement à la conscience et qui

formeraient des indicateurs fiables de notre sous-optimalité temporaire. Les marqueurs somatiques de Damasio sont conçus comme des signaux d'anticipation, révélant la capacité des sujets de percevoir à l'avance les risques encourus par leurs choix. Damasio dit non seulement que ces signaux sont fiables, parce qu'ils sont corrélés à un risque de sous-optimalité pour le sujet, mais également qu'ils sont efficaces, parce qu'en leur absence des stratégies optimales ne peuvent pas être établies. Il est peut-être simplement curieux de donner de ces signaux l'interprétation privilégiée qu'en donne Damasio, à savoir qu'ils formeraient des indicateurs, d'abord inconscients, sur l'optimalité *à terme* de mes choix. Disons qu'il est plus simple, et moins coûteux philosophiquement, de les considérer comme des indicateurs des risques encourus sur des séquences comportementales relativement brèves. Ce sont ce qu'on peut appeler des signaux d'erreur fictive (*fictive-error signals*), ces signaux pouvant se traduire ou pas en poids décisionnels infléchissant efficacement les choix des sujets. Damasio conclut en effet du comportement et du type de lésion présentée par ses patients que ces derniers ne produisent pas ces signaux d'anticipation.

On peut envisager plusieurs hypothèses sur ce sujet. Il est sans doute avéré que les patients ventro-médiants de Damasio ne puissent générer ces signaux. Mais on peut également considérer que des patients d'un autre type génèrent ces signaux mais sont dans l'incapacité de les interpréter correctement, et que pour d'autres, ces signaux sont générés et correctement interprétés mais ne peuvent se traduire en poids décisionnel, c'est-à-dire ne permettent pas d'infléchir le comportement en un sens optimal. Quoi qu'il en soit, un des points fondamentaux concerne bien la genèse de ces signaux chez les sujets sains (non cérébrolésés). Pourquoi de tels signaux ? En quel sens sont-ils des marqueurs directs ou bien de la sous-optimalité, comme le pense Damasio, ou bien du risque encouru dans une décision ponctuelle, comme semblent plutôt le montrer Tomb et Caramazza ?

L'idée d'un signal fiable en prise directe avec l'évaluation des conséquences de la réalisation future d'un choix présent ne va pas de soi. Le fait que le paradigme expérimental de Damasio consiste en une tâche d'apprentissage rend naturellement vraisemblable cette idée, mais qu'en est-il de l'anticipation d'une erreur hors d'un contexte présent d'apprentissage ? Le sujet peut garder naturellement longtemps en mémoire la trace d'un comportement qui s'est avéré nuisible par le passé et en reconnaître rapidement l'amorce dans une autre occasion éloignée. Mais le signal d'erreur fictive – à savoir l'anticipation correcte que j'emprunte une voie décisionnelle peu favorable – repose-t-il directement sur ces comparaisons entre le passé et le futur proche, ou est-il, différemment, l'effet sensible d'un conflit entre deux automatismes cérébraux ? Les deux hypothèses ne sont peut-être pas dissociées. Il est possible que la sensibilité au conflit interne aille systématiquement de pair avec l'anticipation sensible des conséquences éventuellement néfastes des choix en cours.

Les deux capacités iraient de pair dans le sens précis où un organisme suffisamment complexe, composé de différents modules susceptibles d'entrer en conflit, serait apte à développer des mécanismes internes de régulation et d'harmonisation entre les sorties comportementales liées à ces différents modules. Il y aurait alors un intérêt à ce qu'en cas de conflit entre ces sorties comportementales, le sujet soit alerté de cette incompatibilité. Il se peut également qu'un des conflits les plus typiques que cet organisme ait à traiter soit celui qui s'instaure, disons-le

de façon extrêmement schématique, entre des sorties émotionnelles et des sorties cognitives. Un organisme suffisamment évolué peut faire l'expérience de la dichotomie entre ses anticipations à terme et ses inclinations de court-terme. Mais au lieu de dire que les signaux prédictifs d'une erreur reposent directement sur la comparaison des choix de court terme avec des perspectives profitables à plus long terme, ces signaux formeraient simplement l'issue sensible d'un conflit entre les systèmes cérébraux qui sous-tendent respectivement les processus de décision ou de planification pour le futur et les mécanismes de gratification immédiate (voir McClure 2004). Ces signaux en eux-mêmes peuvent être myopes au sens où ils sont focalisés sur des conflits ponctuels entre différents systèmes cérébraux. Mais leur fonction, en dépit de leur myopie, est bien d'orienter le comportement, quand ils parviennent effectivement à l'orienter, c'est-à-dire quand ils sont intégrés par l'organisme sous forme de poids décisionnels, vers des perspectives optimales à long terme.

V. OPTIMALITÉ INTER-BIAIS

Paradoxalement, bien que les biais puissent être issus du conflit entre différentes réponses automatiques, ressortant du fonctionnement naturel de différents modules, d'un individu, la coexistence de plusieurs biais peut produire globalement un résultat optimal. Nozick a suggéré que des biais se compensent entre eux et résultent en une rationalité globale (Nozick 1993). Il prend l'exemple des *sunk costs* et de la myopie. Les *sunk costs* (coûts déjà écoulés ou engagés) forment une anomalie comportementale dans la mesure où des investissements passés qui ont cessé d'être productifs doivent être, rationnellement, abandonnés. Il est, en l'occurrence, rationnel ne pas être lié par des investissements (en temps, énergie, argent) passés si ces investissements ont fini par me nuire. En réalité une telle conception de l'irrationalité des *sunk costs* repose sur une préférence forte pour le présent : il s'agit de maximiser son utilité présente. Ainsi, si j'ai pris un abonnement à l'opéra pour l'année au mois de septembre et qu'en cette soirée de janvier je me sens peu enclin à me rendre dans le froid jusqu'à l'opéra, il est rationnel de rester chez moi. Or, il peut au contraire être rationnel, dans au moins deux autres sens, de me sentir lié par cet engagement passé. D'abord j'ai pu contracter un tel abonnement précisément pour me sentir lié, ayant anticipé qu'un certain soir d'hiver prochain ma volonté me ferait défaut. Ensuite je peux vouloir me dire à moi-même que j'agis selon certains principes.

Comme le souligne Nozick, se sentir lié par des investissements passés renvoie à une conception symbolique, et pas seulement instrumentale, de la rationalité. On peut chercher à maximiser une image de soi durable, au moins autant qu'une satisfaction présente. L'idée ici est qu'une certaine irrationalité (les sunk costs) génère une réponse globalement optimale quand elle vient compenser la myopie comportementale. L'état de satisfaction émotionnelle d'un individu à long terme, sa tranquillité identitaire, peut être due à une intégration bénéfique de ses différents biais, c'est-à-dire émerger à partir des irrationalités locales produites par ses limites cognitives.

Les biais tout autant que les émotions façonnent mon identité. On peut donner de cette dernière une définition purement comportementale : la régularité de

certaines réponses dans mon environnement. Face à des mutations environnementales, parfois drastiques, les émotions que je ressens peuvent indiquer des menaces face à la préservation de cette identité, face au fait que le type d'individu que je forme se trouve en état, plus ou moins temporaire, d'inadaptation. Il peut être alors intéressant de compenser cette défaillance par une autre. On peut penser à l'évolution identitaire d'un individu comme étant le fruit d'une série de compensations, plus ou moins explicites et volontaires et parfois inconscientes, de biais, d'irrationalités et d'inadéquations entre elles tendant vers un équilibre adaptatif maximal. Une population d'individus sans réponses automatiques biaisées pourrait être moins adaptée à un environnement sollicitant des ressources cognitives parfois hors d'atteinte qu'une population d'individus dont l'équilibre comportemental, d'apparence sous-optimal, est le fruit de compensations entre des imperfections internes.

La neuroéconomie, comme étude des mécanismes d'adaptation à un environnement économique qui a subi une transformation radicale à plusieurs reprises (le Néolithique, puis la Révolution Industrielle), ou comme étude des mécanismes de décision optimaux dans des contextes expérimentaux (simulations de marchés) hautement artificiels, doit être complétées par d'autres méthodes d'investigation : l'archéologie cognitive, la psychologie du développement et l'éthologie. Nous avons déjà fait allusion aux genres d'hypothèses qui pouvaient être testées par les deux premières disciplines, nous terminerons ces réflexions sur le sens d'une étude des mécanismes d'adaptation cérébrales aux mutations des environnements socio-économiques par des données animales intrigantes. Par exemple, certaines abeilles (certains individus de l'espèce *Apis mellifera*) exhibent des préférences intransitives – la transitivité des préférences constituant un pilier de la rationalité – face à des fleurs artificielles que l'on manipulait selon deux dimensions : la quantité de sucrose contenu et la longueur de la corolle (Shafir 1994). Des abeilles, encore, violent d'axiome d'indépendance des préférences vis-à-vis des alternatives non pertinentes (Si A est préféré à B dans l'ensemble de choix {A,B,C}, A doit continuer à être préféré à B dans l'ensemble de choix {A,B}. A nouveau alors que ces abeilles préféraient la fleur X à la fleur Y en présence de Z, elles préféraient la fleur Y à X en l'absence de Z. Des résultats similaires ont été obtenus sur les geais, les colibris, les rats, les pigeons, et aussi sur les hommes.

Ces exemples indiquent que les normes de rationalité sont violées par les organismes biologiques – c'est-à-dire par les produits de la sélection naturelle. Ce fait pose un problème d'ordre général qui nous intéresse au premier chef. Car il peut sembler à première vue paradoxal que les violations de ces règles aillent de pair avec une conception darwinienne de la sélection de ces organismes. Par exemple, dans le cas de préférences intransitives que nous avons rapporté, il s'agit d'une erreur du point de vue de la maximisation de l'utilité de l'abeille. L'un des choix A>B, ou B>C ou C>A est incohérent avec un comportement optimal pour cet individu. La sélection naturelle est censée substituer à ce comportement un comportement plus adaptatif. Les organismes qui font des choix sous-optimaux, entre des propositions de nourriture, ou comme on l'observe également pour certains oiseaux entre des opportunités de nidation, semblent se comporter de manière non-adaptative et on peut s'attendre à ce qu'ils soient remplacés au cours de l'évolution. Aussi, dans cette perspective évolutionnaire, les violations de la rationalité – parce que l'on suppose que certains patterns comportementaux opti-

maux sont bien exprimés formellement par les normes de rationalité – sont inattendues.

On peut chercher à rendre compte de ces violations biologiques des normes de rationalité. Les deux grandes sources possibles d'irrationalité : l'environnement ou l'organisme lui-même. Les théories de la rationalité limitée, par exemple le programme « heuristiques et biais », font porter la responsabilité de l'irrationalité apparente aux limites ou aux biais par lesquels nous appréhendons l'information dans notre environnement. On a vu aussi que certains critiques de cette approche montrent que l'inadéquation entre l'environnement expérimental et le traitement naturel de l'information par un individu pouvait expliquer la mise en évidence de certaines de ces limites ou biais. Mais cette approche en termes de rationalité limitée, tout autant que sa critique « écologiste », laissent inexpliqué un aspect important du problème. On peut certes admettre que des organismes ne peuvent avoir accès à l'ensemble de l'information qui pourrait être pertinente pour qu'ils prennent une décision optimale, mais pourquoi, cependant, s'engagent-ils dans des processus de traitement erroné de l'information dont ils disposent parfaitement ? Dans son article de synthèse sur la rationalité limitée en 1996, Conslisk souligne que dans la modélisation économique des limites de la rationalité, l'absence d'information et les limites d'accès à l'information étaient clairement prises en compte, mais pas le traitement déficient d'une information disponible. Dans des conditions d'information parfaite, les erreurs de traitement montrent des limites de la rationalité d'un ordre particulier qu'on peut attribuer à un dysfonctionnement ou à un fonctionnement inadéquat, étant donné une information parfaite dans un environnement naturel donné, de l'organisme lui-même (voir Livnat et Pippenger 2008). C'est le sens éventuellement adaptatif de ces dysfonctionnements, des biais et des déficiences clairement inhérents à l'organisme lui-même, dans la mesure où ils sont survécus à la sélection naturelle, que l'on peut chercher à mieux comprendre.

Ecole Normale Supérieure des lettres et sciences humaines (Lyon)
Institut Jean-Nicod – CNRS-ENS-EHESS (Paris)

BIBLIOGRAPHIE

Bourgeois-Gironde, S. (2008). *La Neuroéconomie*, Plon, Paris.

Bourgeois-Gironde, S. et Giraud, R. (2009). Framing effects as violations of extensionality, *Theory and Decision*, en cours de publication.

Camerer, C. et al. (2005). Neuroeconomics : How Neuroscience Can Inform Economics, *Journal of Economic Literature*, (43) 9-64.

Conslisk, J. (1996). Why Bounded Rationality ? *Journal of Economic Litterature*, (34), 669-700.

Damasio, A., (1996). The Somatic marker hypothesis and the possible functions of the prefrontal cortex, *Philosophical Transactions of the Royal Society, London B Biological Sciences*, 351(1346):1413-20.

Demoule, J.-P., (2008). *La Révolution néolithique*, Le Pommier, Paris.

Eisenberger, N., Liberman, M. et Williams, K. (2003). Does rejection hurt ? An fMRI study of social exclusion. *Science,* (302) 290-292.

Elster, J., (1996). Rationality and the Emotions, *The Economic Journal*, 106 (438 : 1386-1397.

Frisch, D. (1992). Reasons for Framing-Effects, Organizational Behavior and Human Decision Processes, (54), 399-429.

Gigerenzer, G. et Todd P.M. (1999). *Simple Heuristics that Make us Smart*, Oxford University Press, NY.

Gul, F. et Pesendorfer, W. (2005). The case for mindless economics, working paper.

Kahneman, D. Maps of bounded rationality : psychology for behavioral economics. *American Economic Review*, (93), 1449-1475

Livnat, A. et Pippenger, N. (2008). Systematic mistakes are likely in bounded optimal decision-making systems, *Journal of Theoretical Biology* (250) 410-23.

Maia, T. et McClelland, J. A reexamination of the evidence for the somatic marker hypothesis : what participants really know in the Iowa gambling task. *Proc Natl Acad Sci USA*, (101) 16075-16080.

McClure, S., Laibson, D. et al. (2004). Separate Neural Systems Value Immediate and Delayed Monetary Rewards, *Science*, (306) 503-507.

Niv, Y. et Schoenbaum, G. (2008). Dialogues on prediction. *Trends in Cognitive Sciences*, (7), 265-272

Nozick, R. (1993). *The Nature of Rationality*, Princeton University Press, Princeton.

Sanfey A. et al., (2003). The neural basis of economic decision-making in the Ultimatum Game. *Science*, 300(5626) :1755-8

Shafir, S. (1994). Intransitivity of preferences in honey bees : Support for 'comparative' evaluation of foraging options, *Animal Behavior* (48), 55-67.

Sher, S., et McKenzie, C. (2006). Information leakage from logically equivalent frames, *Cognition*, (101) 467–494.

Singer, T. and Fehr, E., (2005). « The Neuroeconomics of Mind Reading and Empathy.» Neuroscientific Foundations of Economics Decision-Making, 95 (2), 340-345.

Slovic, P. et Tversky A. (1974). « Who accepts Savage's axioms », Behavioral Science, 19, p. 368-37.

Stanovich, K. et West, R. (2000). Individual Differences in Reasoning : Implications for the Rationality Debate ? *Brain and Behavioral Sciences*, (23) 645-665.

Tomb, I., Hauser, M., Deldin, P. et Caramazza, A. Do somatic markers mediate decisions on the gambling task ?. Nature Neuroscience, (5), 1103-4

Zeelenberg, M., (1999). Anticipated regret, expected feedback and behavioral decision-making. *Journal of Behavioral Decision Making*, 12, 93-106.