L'émergence d'un nouveau domaine de savoir: la neuroéconomie

Nicolas Vallois

UNIVERSITÉ PARIS I PANTHÉON-SORBONNE
U.F.R. ECONOMIE

THÈSE

pour obtenir le grade de
DOCTEUR ÈS-SCIENCES ECONOMIQUES

présentée et soutenue publiquement par
Nicolas VALLOIS
le 05/12/2012

sous le titre

L'émergence d'un nouveau domaine de savoir: la neuroéconomie

Directeur de Thèse

Madame Annie L. Cot, Professeur à l'Université Paris I Panthéon-Sorbonne

Membres du Jury :

Monsieur Giorgio Coricelli, Chargé de recherches au CNRS

Monsieur Richard Arena, Professeur à l'Université de Nice – Sophia Antipolis

Monsieur Xavier Guchet, Maître de conférences à l'Université Paris I Panthéon-Sorbonne

Monsieur Guillaume Hollard, Directeur de recherches au CNRS.

Monsieur Mathias Pessiglione, Chargé de recherches à l'INSERM.

Remerciements

Je tiens à remercier :

Annie L. Cot, qui a dirigé, corrigé et relu mes travaux au cours de ces quatre années,

Mathias Pessiglione, Xavier Guchet, Guillaume Hollard, qui ont encouragé mes débuts dans la recherche, m'ont fourni de nombreux conseils et des réflexions stimulantes, et ont accepté d'être membres du jury,

Giorgio Coricelli et Richard Arena, qui ont bien voulu être les rapporteurs du présent travail,

Jean-Sebastien Lenfant, Don Ross et pour des suggestions et des discussions fructueuses,

Les membres du séminaire AOH (Jérôme Lallement, Sophie Pellé, Cléo Chassonnery-Zaïgouche, Judith Favereau, Dorian Jullien, Nadeera Rajapakse, Isselmou Ould Boye, Mathieu Renault, Agnès Gramain, Nicolas Brisset, Maxime Desmarais-Tremblay, Pierrick Dechaux, Raphael Fevre), pour leurs relectures et leurs commentaires,

Axelle Neyrinck, qui m'a éveillé au sens de l'histoire,

Luc Mallet et Sacha Bourgeois Gironde, qui ont eu la gentillesse de m'accorder des entretiens.

Table des matières

Introduction générale.....9

(i) Enjeux et objectifs d'une étude historique sur la neuroéconomie.....13
(ii) Hypothèse interprétative:l'approche pathologique du comportement comme apport spécifique de la neuroéconomie.....18
(iii). Une révolution technologique? 22
(iv) Nouvelle approche ou nouvelle discipline? Les rapports de la neuroéconomie à l'économie comportementale.....30

Chapitre liminaire. Questions de méthode : neuroéconomie et idéologie scientifique.....42

I. Le point de vue externe : la neuroéconomie et son contexte.....46

A. Une perspective intermédiaire: la rhétorique de la neuroéconomie.....46

1. L'art de la persuasion des neuroéconomistes : les modèles de la neurobiologie comme métaphores élégantes.....47
2. Une visée prescriptive: « la bonne science est une bonne conversation ».....50

B. Les interprétations psychosociales.....52

1. Biographies et communautés intellectuelles: la vie des auteurs comme matière pour l'histoire de la pensée.....53
a. Vernon Smith et les neurosciences, Paul Glimcher et l'économie: le parcours intellectuel de la neuroéconomie.....53
b. Les limites d'une lecture biographique: la notion contestable de « vision du monde ».....55
1. Une science intéressée: le rôle des débouchés et des applications pratiques.....59

II. La notion d'idéologie scientifique chez Georges Canguilhem : une source d'inspiration pour l'historien de la pensée économique.....66

A. L'originalité du regard historique: l'attention aux ruptures.....67

B. Peut on parler d'idéologie dans les sciences ? Canguilhem et la sociologie des sciences.....69

C. Idéologie scientifique et neuroéconomie: esquisse d'un récit historique.....75

Conclusion du chapitre 1.....79

PREMIERE PARTIE – PSYCHIATRIE ECONOMIQUE ET NEO-COMPORTEMENTALISME: LA PREHISTOIRE THEORIQUE DE LA NEUROECONOMIE (1955-1999) 81

(i) un nouvel agent économique: l'animal de laboratoire 82
(ii) une approche évolutionniste et séquentielle 83
(iii) la notion de pathologie.....87

Chapitre 2. Impulsivité et choix inter-temporel : la naissance d'une idéologie scientifique en sciences comportementales.....89

I. La psychiatrie économique comme idéologie scientifique.....92

A. La naissance d'une médecine mentale à vocation économique: la science quantitative de la motivation.....92

1. Quantification des récompenses et utilité espérée.....93
2. Les rapports ambivalents du néo-comportementalisme à l'économie comportementale.....94

B. Science du diagnostic et maximisation de la normalité: les ambitions théoriques de la psychiatrie économique.....97

II. Choix inter-temporel et impulsivité: la reprise d'un problème théorique en économie par la psychologie néo-comportementale.....102

A. Robert Strotz et la myopie temporelle (1955) : une première approche économique de l'impulsivité.....102

B. Un problème théorique abandonné par les économistes? Incohérence séquentielle et incohérence dynamique.....107

C. Les apports de la psychologie expérimentale : l'impulsivité comme constante comportementale.....109

III. Néo-comportementalisme et économie : un projet interdisciplinaire contesté.....115

A. Melioration versus Maximisation : le débat interne sur la signification économique des expériences néo-comportementales.....116

1. La séquentialité comme obstacle à la rationalité : Richard Herrnstein et la théorie de la mélioration.....116
2. L'adoption d'un formalisme économique en psychologie comportementale et ses critiques.....120
a. La controverse Rachlin / Herrnstein.....121
b. Une vision fonctionnelle de la maximisation.....124
c. Vers une rationalité au deuxième degré : l'intelligence des séquences.....125

B. L'intégration de la science quantitative de la motivation à l'économie.....128

1. La reprise du problème de l'impulsivité par l'économie comportementale: mise en évidence et mesure d'une actualisation hyperbolique chez l'homme 129
2. Le modèle d'actualisation quasi-hyperbolique: l'incohérence temporelle comme anomalie.....131

IV. La disparition de la norme économique: la pathologie comme évidence fondatrice de

l'analyse....134

A. L'impulsivité, une anomalie d'un genre problématique pour l'économie comportementale: la remise en cause du partage normatif/descriptif....134

B. Justification biologique de l'actualisation hyperbolique: l'impulsivité comme dérèglement sans norme....137

C. La pathologie comme solution à l'absence de norme....140

V. Impulsivité et maximisation de la normalité – Les objets hyperboliques du néo-comportementalisme....143

Conclusion du chapitre 2....146

Chapitre 3. Les années 1990 : La vocation économique de la neurobiologie – Paul Glimcher et l'« utilité espérée physiologique »...148

I. Nouveaux instruments, nouveaux modèles : les micro-électrodes au service d'une théorie de l'utilité espérée physiologique....152

A. Les débuts de l'électrophysiologie : l'influence déterminante du modèle stimulus-réponse....152

B. Platt et Glimcher, 1999 : De la réflexologie à la théorie de l'utilité espérée....157

II. Vers une théorie fréquentielle de la décision : la prédiction de la récompense comme processus stochastique....164

A. Les aléas de la prédiction : une limite à la maximisation du choix discret....164

B. De la maximisation du choix discret à la théorie des jeux évolutionnaires: la stochasticité comme réponse optimale....167

1. La variabilité des décisions comme moyen d'apprentissage: un premier sens de la stochasticité....168

2. La variabilité des décisions comme stratégie mixte (optimale): un second sens de la stochasticité....169

III. De l'utilité espérée à l'apprentissage de la récompense....175

A. les neurones dopaminergiques et l'erreur de prédiction de la récompense: un circuit neuronal spécialisé dans l'apprentissage hédonique....175

B. La formalisation des processus d'apprentissage: de Pavlov aux algorithmes du *machine learning*....178

1. Apports et limites des théories pavloviennes de l'apprentissage 179

2. L'apprentissage par différence temporelle (*Temporal Difference Learning* ou *TD learning*)....180

C. Approfondissement et remise en question de la théorie économique....184

IV. Glimcher et l'économie: des rapports ambivalents....186

A. Utilité espérée physiologique, apprentissage de la récompense: un vocabulaire théorique confus pour l'économiste 186

B. Un rapport ambigu à l'économie....189

Conclusion du chapitre 3.....191

Conclusion de la Première Partie.....192

DEUXIEME PARTIE – D'UNE INNOVATION TECHNOLOGIQUE A LA CONSTRUCTION D'UNE NOUVELLE DISCIPLINE: LA NEUROECONOMIE DES ANNEES 2000 ET L'IMAGERIE FONCTIONNELLE PAR RESONANCE MAGNETIQUE 196

Chapitre 4. Une discipline sous la tutelle de l'économie comportementale et des neurosciences (2000-2005).....199

I. Les apports de l'imagerie fonctionnelle par résonance magnétique : extension et approfondissement d'un paradigme expérimental conçu initialement sur l'animal.....201

A. L'erreur de prédiction de la récompense: un paradigme expérimental prometteur pour les neurosciences en 2001.....201

B. La notion de monnaie neuronale commune :un résultat fondateur pour l'« économie neuronale ».....206

C. Apprentissage de la récompense et choix inter-temporel : le rôle spécifique du cortex pré-frontal.....209

II. La tutelle de l'économie comportementale ou « l'économie comportementale dans le scanner ».....213

A. L'imagerie fonctionnelle comme instrument de vérification de l'économie comportementale.....213

1. L'actualisation quasi-hyperbolique dans le cerveau.....214
b. La théorie de la punition altruiste dans le cerveau.....217
3. La théorie des regrets dans le cerveau.....219
4. La théorie des *prospects* dans le cerveau.....220

2. Les modèles duaux et les premiers manifestes de la neuroéconomie, entre simplification descriptive et rhétorique anti-économique.....228

III. La tutelle des neurosciences : Antonio Damasio et la théorie des marqueurs somatiques 232

A. De la neuropsychiatrie à la théorie des jeux : les lésions du cortex préfrontal et l'*Iowa Gambling Task* (IGT).....232

B. Émotions versus évaluation économique : les rapports ambivalents de Damasio à la neuroéconomie.....238

Conclusion du chapitre 4.....240

Chapitre 5. De l'économie comportementale dans le scanner à la neuro-psychiatrie computationnelle: la constitution d'une discipline autonome (2005-2010).....241

I. Les critiques externes comme moteurs théoriques.....243

A. Les critiques de l'économie comportementale dans le scanner.....243

1. Le traditionnel argument des préférences révélées et de l'« économie sans pensée » (*mindless economics*) (Gul et Pesendorfer. 2005).....244

2. Les neuroéconomistes sont-ils des économistes comportementalistes « gâtés »? La neuroimagerie comme jouet coûteux (Rubinstein, 2008) 248

3. La neuroéconomie comme néo-phrénologie: une attaque directe contre les neurosciences (Harrison, 2008-a et 2008-b).....251

4. Le problème de l'inférence inverse.....254

B. Les critiques de la théorie des marqueurs somatiques et de l'*Iowa Gambling Task*. 257

II. La construction d'un paradigme autonome (1): la référence à la pathologie.....263

A. Le paradigme du *reward learning* comme outil de diagnostic 264

B. La connaissance clinique des pathologies comme « ontologie cognitive » au service de la neuroimagerie.....267

1. Du cognitif au cérébral, et du cérébral au cognitif: le double niveau de lecture des études de neuropsychiatrie computationnelle.....268

2. La connaissance clinique des pathologies comme outil de définition des fonctions cognitives.....270

3. La distinction clinique du normal et du pathologique: une référence implicite et sous-estimée des travaux de neuroimagerie.....273

4. La neuroéconomie comme compromis entre le langage clinique et l'approche fonctionnelle 276

III. La construction d'un paradigme autonome (2): la mise en avant d'une monnaie neuronale commune à l'homme et à l'animal.....279

A. L'ambivalence des émotions.....279

B. Abandon des modèles duaux et retour des interprétations biologiques et évolutionnistes: le cas du choix inter-temporel.....284

IV. La neuroéconomie comme projet de « psychiatrie économique »: le cas des comportements addictifs.....293

A. Une explication biologique et évolutionniste: la notion d'environnement addictif. .294

B. Une définition pathologique de la rationalité économique: le rôle des critères cliniques.....297

C. L'affirmation d'une identité propre: distanciation critique et réappropriation des figures historiques de la discipline.....302

Conclusion du chapitre 5.....307

Conclusion de la deuxième partie.....308

TROISIEME PARTIE – LA CONQUETE DE NOUVEAUX OBJETS : LE CHOIX INTER-PERSONNEL ET L'ANALYSE DU BIEN-ETRE.....310

Chapitre 6. L'analyse du choix inter-personnel: la neuroéconomie entre neurosciences sociales, neuro-éthique et économie comportementale.....312

I. L'étude des interactions sociales: un sous-domaine particulier de la neuroéconomie....314

II. Émotions et Préférences sociales: le choix inter-personnel vu par l'économie comportementale dans le scanner.....319

A. Trois études fondatrices pour la neuroéconomie sociale (2001-2003).....319

B. La théorie de la punition altruiste dans le cerveau: le programme de recherche de Ernst Fehr.....322

C. Apports et limites des modèles duaux.....327

1. Une approche fédératrice pour la neuroéconomie sociale.....328

2. Émotions et préférences sociales: le problème de la répétition des interactions et de l'apprentissage.....334

III. De l'apprentissage de la récompense au choix inter-personnel: la notion d'intelligence sociale.....338

A. *Reward Learning* et choix inter-personnel: une perspective dynamique sur les jeux inter-personnels.....338

B. Modèles duaux versus monnaie neuronale commune.....341

C. L'intelligence sociale comme capacité à former des émotions sociales complexes. 344

IV. Les implications prescriptives. Troubles et pathologies de l'intelligence sociale.....348

Conclusion du chapitre 6.....353

Chapitre 7. Paternalisme Libertarien et Psychiatrie Économique. L'apport de la neuroéconomie à l'analyse comportementale du bien-être. 355

I. L'économie comportementale du bien-être : la portée normative des *behavioral economics* 358

A. De la description des erreurs à la prescription des normes.....358

B. Limites normatives.....365

II. Une perspective intermédiaire: la collaboration entre le neurobiologiste Antonio Rangel et l'économiste Douglas Bernheim.....372

A. Un fondement neurobiologique à l'économie comportementale du bien-être.....372

B. Émotions et processus réflexes – Un approfondissement insuffisant des concepts neurobiologiques....377
III. Neuroéconomie et psychiatrie économique....384
Conclusion du chapitre 7....391

Conclusion générale....393

Annexes....403

Compte-rendus d'entretiens....403
-Entretien avec Sacha Bourgeois-Gironde (12/02/2009)....403
-Entretien avec Mathias Pessiglione (12/11/2009)....414
-Entretien avec Luc Mallet, 14/09/2010....426

Bibliographie....432

Introduction générale

L'expression de neuroéconomie est apparue dans littérature scientifique il y a maintenant dix ans. La première occurrence écrite de ce terme se trouve dans un article publié en 2002 dans *The Flame*, magazine des étudiants en troisième cycle de l'université de Claremont (États Unis). Le professeur Paul Zak y affirme que ses recherches, qui se situent aux frontières de l'économie et des neurosciences, relèvent d'une nouvelle discipline, qu'il appelle neuroéconomie (cité par diPetrio, 2002). La même année, une revue académique de sciences de gestion -*Management Science*- fait paraître le compte-rendu d'une expérience de neuroéconomie, réalisée par Kevin McCabe et ses coauteurs, dans lequel le terme de neuroéconomie est utilisé à nouveau: « *la démonstration d'une relation entre l'activité du cerveau et le choix économique observé atteste de la faisabilité d'une science neuroéconomique de la décision* »[1] (Smith *et al.*, 2002, p.712).

Neuf étudiants en médecine, recrutés par les expérimentateurs, participent à cette étude. Les sujets doivent choisir successivement entre deux loteries financières. A chaque fois, l'un des deux paris implique un risque simple, et l'autre de l'ambiguïté, c'est-à-dire une incertitude sur la distribution des gains et des pertes possibles. Les réponses fournies par les participants permettent ainsi de révéler, ou non, et de mesurer ce que les économistes appellent l'aversion à l'ambiguïté. Au moment où les choix sont effectués, les auteurs enregistrent, grâce à une technique d'imagerie cérébrale[2], l'activité dans le cerveau de chacun des sujets (Smith *et al.*, 2002, p.713).

Les auteurs inscrivent leur approche au sein de cette « *science neuroéconomique de la décision* », qui consisterait donc à établir des corrélations entre des variables neuronales ou cérébrales -l'activité dans certaines zones du cerveau- et des variables comportementales -ici, le degré d'aversion à l'ambiguïté- déduites du choix observé. Le terme de « *décision* » est cependant ambigu. Il peut être compris d'abord au sens matériel de « comportement économique », c'est-à-dire de choix individuels ayant un rapport, même lointain, avec les activités marchandes: achat, vente, investissement, coopération, négociation, *etc*. Ces choix seraient simulés en laboratoire, et la neuroéconomie représenterait donc une tentative pour rendre compte de « ce qui se passe dans le cerveau des individus » lorsque ceux-ci prennent

1 J'ai réalisé moi-même la traduction, de l'anglais vers le français, de cet extrait. Il en va de même pour tous les articles et ouvrages en langue anglaise cités dans la suite du texte. La fidélité de la traduction a été privilégiée au détriment de l'expression, ce qui explique la style parfois peu élégant de certaines citations.

2 La méthode d'imagerie utilisée dans cette expérience est la tomographie par émission de positons, connue en anglais sous le nom de « PET scan ».

des décisions à caractère économique. Paul Zak souscrit ainsi à cette définition lorsqu'il affirme, dans son entretien au journal *The Flame*, que « *dans la neuroéconomie, notre but est d'observer et de mesurer ce qui se produit dans le cerveau lorsque les gens prennent des décisions* » (cité par diPetrio, 2002, p.20). Ici, le champ des décisions « à caractère économique » est très large, puisque Zak y inclut notamment les comportements altruistes, et, plus généralement, tous les choix ayant un rapport avec les interactions sociales.

Cette manière de comprendre la neuroéconomie laisse à penser, par ailleurs, que des expériences pouvant, rétrospectivement, être qualifiées de neuroéconomiques, avaient déjà été réalisées avant 2002. Par exemple, un an auparavant, Kevin McCabe avait déjà proposé une étude de neuroimagerie portant sur les dynamiques de coopération entre les individus dans un protocole connu sous le nom de jeu de l'investisseur[3] (McCabe *et al.*, 2001). Antérieurement à sa première occurrence écrite, le mot de neuroéconomie était déjà utilisé de façon informelle dès la fin des années 1990. Le véritable inventeur (et promoteur) de la neuroéconomie est en fait Kevin McCabe, qui déclare « *avoir inventé le terme de neuroéconomie pour expliquer ce que nous faisions à l'époque. Vernon* [Smith], *qui était un grand défenseur de cette approche, et moi-même avions donné de nombreuses conférences sur le sujet à la fin des années 1990 et au début des années 2000, jusqu'à la publication dans PNAS*[4] ».

Le terme de décision peut cependant aussi être entendu de manière formelle, au sens de la « théorie de la décision », celle-ci désignant l'application de théories économiques et mathématiques à des problèmes de choix abstraits (Kast, 2002, p.8). La neuroéconomie consisterait alors à emprunter à l'économie des procédures formelles de prise de décision pour modéliser le fonctionnement du cerveau. Pour le neurobiologiste (et neuroéconomiste) Paul Glimcher par exemple, la neuroéconomie viserait à appliquer des modèles d'utilité espérée au système nerveux, élaborant ainsi une « *théorie économique du cerveau* » (Glimcher, 2003, p.322). D'une façon relativement proche, l'hypothèse dite du « cerveau bayésien » proposée notamment par Stanislas Dehaene[5] suppose que le cerveau construit, à partir des entrées

3 Ce jeu met en relation deux joueurs. Le premier choisit de transférer, ou non, une part de sa dotation financière initiale. S'il y a transfert, le second joueur peut décider de « trahir la confiance » qui a été accordée et il empoche alors la totalité de la cagnotte commune. Mais il peut aussi choisir de se comporter de manière altruiste, et de reverser une part de ses gains au premier joueur. Si le jeu n'est pas répété (entre les mêmes individus), et si, par conséquent, aucun effet de réputation n'est envisageable, un joueur cherchant à maximiser son propre gain financier choisira toujours de trahir la confiance qui lui est accordée; par suite, ce même individu, dans la position du premier joueur, s'il anticipe que son partenaire est lui aussi égoïste et forme les mêmes anticipations, choisira de ne pas faire confiance au second joueur. Or, les expériences réalisées ne laboratoire montrent que cette prédiction n'est pas vérifiée pour la plupart des joueurs (*cf.* McCabe, Rassenti et Smith, 1996*;* McCabe *et al.*, 2001).

4 Communication personnelle avec l'auteur (21 avril 2012). L'article en question, intitulé « une étude par neuroimagerie de la coopération dans l'échange réciproque entre deux personnes, a été publié en 2001 dans les *Proceedings of the National Academy of Sciences* (McCabe *et* al., 2001).

5 Pour une introduction générale à la notion de cerveau bayésien, on se référera au cours donné par Stanislas

sensorielles, un modèle du monde extérieur, et que ce processus peut être compris comme une inférence bayésienne (Wacongne, Changeux et Dehaene, 2012; Moreno-Bote *et al.*, 2011).

L'objet exact de cette « *science neuroéconomique de la décision* » (Smith *et al.*, 2002, p.712) peut donc être défini de deux manières différentes. Dans l'expérience de McCabe précédemment citée (Smith *et al.*, 2002), le choix en laboratoire, entre des loteries financières, peut être compris comme la simulation d'une décision réelle d'investissement financier. Mais ce même choix fait aussi référence à un problème classique de la théorie de la décision, lié à l'aversion à l'ambiguïté et connu sous le nom de paradoxe d'Ellsberg (Ellsberg, 1971). La nature de l'échange disciplinaire entre économie et neurosciences est complètement différente selon ces deux acceptions. Dans le premier cas, un comportement à caractère économique -ici, une décision d'investissement- est expliquée à partir du fonctionnement du cerveau: les neurosciences s'emparent d'un objet de la théorie économique, opérant ainsi ce que Mäki appelle un « *impérialisme inversé* » (Mäki, 2009, p.4). Dans le second cas, ce sont au contraire des modèles économiques formels qui viennent s'appliquer à des objets non-économiques (variables neuronales ou cérébrales).

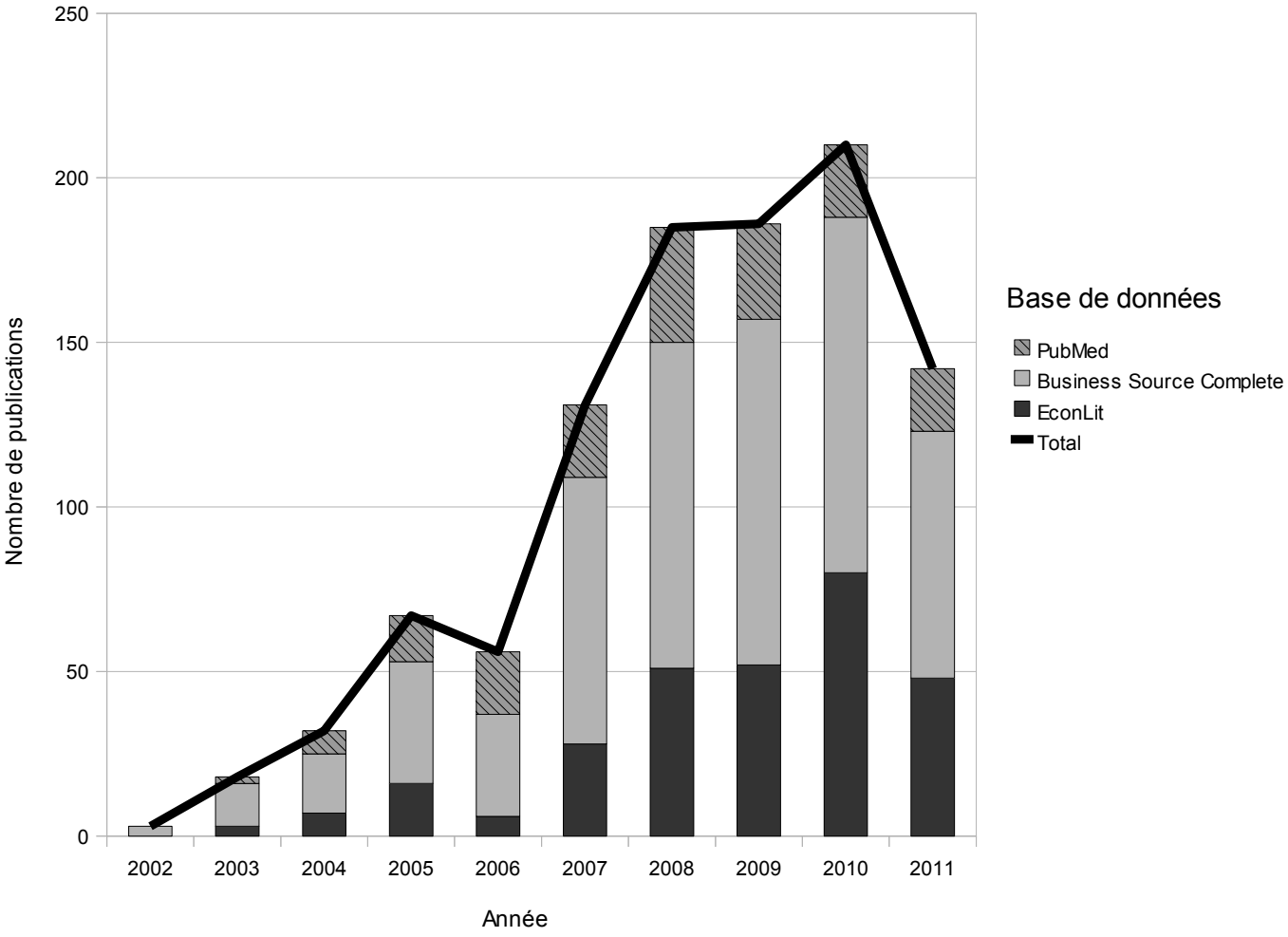
Les définitions de la neuroéconomie mobilisant l'expression de « décision » sont donc assez équivoques. Le *Journal of Economic Literature* a ainsi introduit en 2006 au sein de son système de classification une rubrique « neuroéconomie » (code D87) qui « *regroupe des études utilisant des techniques neuroscientifiques pour analyser les méthodes de prise de décision* »[6]. Or, cette analyse peut porter aussi bien sur des méthodes formelles ou réelles de décision. Le sens devant être donné à l'échange disciplinaire entre économie et neurosciences n'est donc souvent pas bien fixé, ce qui est à l'origine, nous le verrons, de nombreuses confusions.

Le très fort développement académique de la neuroéconomie est en revanche clair et incontestable. A partir des premières occurrences écrites en 2002-2003, le nombre de publications contenant le terme de neuroéconomie explose à partir de la deuxième moitié des années 2000. Cette expansion gagne trois domaines de la littérature scientifique : la neurobiologie et les sciences médicales (base de données *PubMed*), la théorie économique (base de données *Econ Lit*), et les sciences de gestion et la presse économique (base de données *Business Source Premier*)[7] :

Dehaene au Collège de France le 31 janvier 2012, disponible en ligne à cette adresse: http://www.college-de-france.fr/media/stanislas-dehaene/UPL4454455001590428634_Cours2012_CerveauStatisticien_7.pdf

6 *Cf.* le système de classification du *Journal of Economic Literature*, disponible en ligne: http://www.aeaweb.org/jel/jel_class_system.php

7 *PubMed*, *EconLit* et *Business Source Premier* sont les bases de donnée de référence dans ces trois domaines. *PubMed* et *EconLit* recensent uniquement des publications académiques, dans des revues avec comité de



Au cours de la décennie 2002-2012, un peu plus de mille publications contiennent le mot de neuroéconomie. La délimitation du corpus théorique de la neuroéconomie suppose néanmoins de se restreindre uniquement aux revues académiques, ce qui représente tout de même environ 800 publications. En outre, avant l'apparition du terme de neuroéconomie, certains chercheurs faisaient pour ainsi dire « de la neuroéconomie sans le savoir »[8] et réalisaient des travaux qui, rétrospectivement, peuvent être qualifiés de neuroéconomiques bien qu'ils n'y fassent pas référence (*cf. supra*). C'est le cas notamment de l'expérience célèbre de McCabe *et al.* (2001), et, surtout, de tout un courant de recherche qui utilise, dans les années 1990, les microelectrodes pour étudier, chez le primate, l'apprentissage de la

lecture. *Business Source Premier* a un champ d'investigation plus large et inclut des publications économiques en économie et en gestion, mais aussi des articles de presse.

8 Cette expression est empruntée à Mathias Pessiglione, voir entretien du 21 novembre 2009 en annexe.

récompense. Par exemple, les expériences importantes de Glimcher (Platt et Glimcher, 1999) ou de Wolfram Schultz (Dayan, Schultz, et Montague, 1997) n'apparaissent pas dans ce relevé. Cependant, dans notre perspective, la neuroéconomie ne débute véritablement que dans les années 2000. Pour des raisons qu'il conviendra de justifier ultérieurement, nous proposons de limiter le corpus théorique de la neuroéconomie aux travaux publiés à partir de 2000. Les recherches antérieures, quoique proches dans leur principe, seront traitées comme des antécédents théoriques de la neuroéconomie.

Parallèlement à cette diffusion dans la littérature académique, la neuroéconomie a poursuivi un mouvement rapide d'institutionnalisation. A partir des premières conférences réunissant à Martha Vineyard une trentaine de chercheurs, la neuroéconomie s'est dotée d'une association officielle[9]. Celle-ci sert de plateforme d'intégration académique, favorisant la rencontre entre économistes et neuroscientifiques, par l'organisation notamment des conférences annuelles. Cet effort de promotion a eu un certain succès, puisque plusieurs centres de recherches, équipés d'IRM_f, spécifiquement dédiés à la neuroéconomie ont été ouverts aux États-Unis (universités de Duke, de New York, de Claremont, université George Mason) et même en Europe (à l'université de Zurich). Un diplôme de troisième cycle en neuroéconomie a même été crée à l'université de Claremont. Par ailleurs, la neuroéconomie est désormais reconnue comme discipline ou sous-domaine de l'analyse économique. Elle dispose d'une rubrique dédiée (D87) dans le *Journal of Economic Literature* (*cf. supra*). En outre, beaucoup de neuroéconomistes s'attendent aujourd'hui à ce que le Prix de la Banque de Suède en sciences économiques en mémoire d'Alfred Nobel soit attribué à l'un d'entre eux[10]. La neuroéconomie pénètre aussi les milieux non-académiques, et notamment celui de la presse économique, comme le signale le nombre important de publications y faisant référence au sein de *Business Source Premier* (*cf. supra*). La neuroéconomie entre également dans l'usage courant, puisqu'elle dispose depuis 2007 d'une entrée officielle dans le dictionnaire anglais[11].

(i) Enjeux et objectifs d'une étude historique sur la neuroéconomie

Notre étude vise à expliquer l'apparition et le développement de la neuroéconomie:

9 Cette association s'appelle la *Society for Neuroeconomics* (http://www.neuroeconomics.org/)

10 Dès 2002, Kevin McCabe estimait qu' « *il ne fait pas de doute que si le champ se développe de la manière avec laquelle nous le souhaitons, il débouchera sur un Nobel à un moment ou un autre* » (diPetrio, 2002). Chaque année, les spéculations concernant l'attribution du prix à l'un des neuroéconomistes sont relancées, les chercheurs ayant les faveurs des pronostics étant Ernst Fehr et Colin Camerer.

11 Le *Oxford Dictionnary* définit la neuroéconomie comme une « combinaison d'économie, de neurosciences et de psychologie visant à déterminer la manière avec laquelle les individus prennent des décisions ». Le terme de neuroéconomie n'est cependant pas encore entré dans le dictionnaire français.

pour quelles raisons des économistes et des neurobiologistes se mettent-ils à travailler ensemble au début des années 2000? Comment ce projet de recherche a-t-il pu donner naissance à un nouveau domaine de savoir? Ce questionnement historique demeure donc strictement extérieur au champ dont il entreprend l'analyse: il s'agit d'identifier les conditions théoriques de possibilité de la neuroéconomie. En cela, notre travail se distingue de la littérature secondaire existante, qui reste le plus souvent dans une perspective interne à la discipline.

Les travaux prenant la neuroéconomie comme objet d'étude cherchent généralement soit à défendre, soit à critiquer cette dernière. Les partisans de la neuroéconomie veulent promouvoir son approche et vulgariser ses concepts. Les justifications théoriques avancées incluent d'ailleurs assez souvent des analyses historiques portant sur le développement de la discipline. Les neuroéconomistes ont ainsi été les plus prompts à écrire leur propre histoire : assez rapidement, les articles et livres de synthèse consacrent quelques pages à la genèse historique de la discipline et aux travaux pionniers en la matière. Dès 2003, Paul Glimcher, l'une des figures centrales du domaine, publie un ouvrage de référence, dans lequel il replace la neuroéconomie au sein de l'histoire des connaissances relatives au système nerveux depuis l'Antiquité (Glimcher, 2003).

Bien évidemment, ces récits d'« auto-histoire » se caractérisent par un point de vue très particulier. Pour Glimcher par exemple, il s'agit de montrer l'intérêt des approches économiques en neurosciences. Selon cet auteur, la neuroéconomie représente une avancée théorique majeure pour la neurophysiologie, dont le paradigme est amené à se substituer au paradigme classique du réflexe (Glimcher, 2003, *xix*). Dans les travaux destinés plutôt à un public d'économistes (Camerer, Loewenstein et Prelec, 2005 ; Glimcher *et al*., 2008 ; Bourgeois-Gironde, 2008 ; Schmidt, 2010), le récit chronologique est mis au service d'une défense de l'utilisation des neurosciences en économie, pour montrer, pour reprendre le sous titre de l'ouvrage de Christian Schmidt, « *comment les neurosciences transforment l'analyse économique* » (Schmidt, 2010).

Notre analyse de la neuroéconomie n'a nullement pour objectif de défendre son potentiel théorique. Au contraire, notre approche, inspirée de Georges Canguilhem, propose une contribution critique à la discipline. Pour Canguilhem, l'histoire des sciences a une « *tâche critique* », qui « *consiste à annuler les discours intériorisants et reproducteurs* » (Canguilhem, 1977, p.38). Il s'agit de remettre en cause les discours tenus par les scientifiques sur leur propre science. Les histoires de la neuroéconomie écrites par des neuroéconomistes

pour les besoins de promotion et de vulgarisation de leur approche appellent ainsi la production de récits historiques alternatifs.

Il convient cependant de bien préciser en quoi consiste la tâche critique que Canguilhem assigne à l'historien. Il existe en effet déjà au sein de la littérature secondaire de très nombreuses études méthodologique à vocation critique. Dès le début des années 2000, l'approche proposée par les neuroéconomistes a soulevé de nombreuses objections, aussi bien en neurosciences qu'en économie. Sans rentrer dans les détails de l'argumentation, les auteurs remettent en question la compatibilité des approches neurobiologiques et économiques de la décision (*cf.* Gul et Pesendorfer, 2005 ; Harrison, 2008-a et 2008-b ; Rubinstein, 2008).

Cependant, même si elles sont souvent redoutables pour les neuroéconomistes, ces analyses critiques ont joué un rôle moteur dans la maturation de la discipline. Cet apport est tout d'abord visible en termes bibliométriques (*cf. supra*): la publication d'une étude méthodologique en économie par exemple, même très critique, accroît l'impact et la visibilité académique de la discipline. L'expansion de la neuroéconomie dans le langage académique (*cf. supra*) est à cet égard assez trompeur, notamment en économie, car la très grande majorité des articles de neuroéconomie publiés dans des revues d'économie sont en fait des études purement méthodologiques, et ne relèvent pas directement de la neuroéconomie à proprement parler, en tant que discipline expérimentale. Les numéros spéciaux des revues *Games and Economic Behavior, Economics and Philosophy* et du *Journal of Economic Methodology* consacrés à la neuroéconomie en fournissent une bonne illustration. Ces publications ont eu un impact académique important, en produisant respectivement en 2005, 2008 et 2010 douze, quinze et onze articles de recherche classés dans la catégorie « neuroéconomie » de l'index du *Journal of Economic Literature*. Pourtant, seuls cinq de ces trente-huit articles (dans le numéro du *Games and Economic Behavior*) rendent compte d'études expérimentales originales, les trente-trois autres travaux se contentant de commenter, citer et critiquer des recherches de neurobiologie déjà publiées [12].

Surtout, la neuroéconomie, comme toutes les disciplines expérimentales, suppose non seulement la production mais aussi l'interprétation de données empiriques. Le versant critique et méthodologique de la littérature ne constitue donc pas un à-côté du corpus de la neuroéconomie, mais y appartient de plein droit. Comme le souligne Francesco Guala, le fait que certains résultats expérimentaux soient acceptés et d'autres rejetés fait partie du processus

12 Cet effet bibliométrique ne doit cependant pas être interprété comme la cause de l'apparition de la neuroéconomie. L'« impact académique » d'un programme de recherche est la conséquence de sa scientificité; la mesure de cette variable ne saurait fournir une explication de celle-ci.

de régularisation normale des sciences de laboratoire, et ceci est aussi valable pour l'économie expérimentale (Guala, 2005, p.123). L'existence d'un tel rejet est plutôt gage de scientificité.

La critique méthodologique ou épistémologique a donc participé directement au développement de la discipline. Sur le fond, les critiques méthodologiques formulées par les économistes ont notamment permis de corriger certaines confusions relatives aux concepts économiques mobilisés par les neurobiologistes[13]. C'est la raison pour laquelle nous adoptons une définition large du corpus de la neuroéconomie, en y incluant à la fois la littérature primaire (compte-rendus d'expériences) et secondaire (articles de synthèse, analyses méthodologiques).

Les études réalisées sur la neuroéconomie dans le domaine de la méthodologie et de la philosophie économique (*economic methodology*, *philosophy of economics*) ne sont donc pas, à proprement parler, extérieures au champ scientifique qu'elles entendent critiquer. C'est précisément parce qu'elles doivent être pleinement et positivement considérées comme partie intégrante de cette nouvelle discipline que ces analyses ne parviennent pas à expliquer la logique historique de son développement. L'objectif visé n'est pas en effet de décrire le processus de construction de la discipline, mais de juger de ses résultats. Il s'agit toujours d'évaluer les apports possibles des neurosciences à l'économie, ou de l'économie aux neurosciences, et donc de se prononcer sur le contenu scientifique de la neuroéconomie.

A l'inverse, notre étude se veut strictement historique et descriptive. Cela entraîne deux conséquences importantes. Tout d'abord, nous ne cherchons pas à statuer sur la valeur scientifique à accorder à la neuroéconomie. La scientificité de cette dernière est prise comme un fait, qu'atteste son inscription dans le champ académique. En second lieu, notre interprétation n'a vocation à fournir aucune contribution théorique, aussi bien pour l'économie que pour les neurosciences. La compréhension des conditions historiques de possibilité de la discipline n'a aucun intérêt direct pour le théoricien, ce qui est pleinement assumé. En cela, la critique historique se distingue de la critique dite méthodologique ou épistémologique.

Cette perspective extérieure au champ se distingue donc de la littérature secondaire existante, qui elle aussi prend de manière réflexive les travaux expérimentaux comme objets

13 C'est notamment le cas, nous le verrons, pour le concept d'utilité. Les travaux des années 1990 sur le primate ont souvent été compris au départ comme des tentatives de réduction directe du concept d'utilité au niveau neuronal (voir par exemple Camerer, 2005, p.449). Cette interprétation a été favorisée par un usage assez indéterminé du concept par les neurobiologistes. En 2005, Paul Glimcher propose notamment la notion d'« *utilité espérée physiologique* » pour désigner les signaux d'activité neuronale encodant l'anticipation d'une récompense (Dorris, Bayer et Glimcher, 2005). Les accusations de réductionnisme avancées par un auteur comme Glenn Harrison par exemple (Harrison, 2008) ont obligé Glimcher à expliciter davantage le sens du concept d'utilité qu'il utilise dans ses travaux. Dans son ouvrage de 2010, Glimcher se montre plus précautionneux dans son usage de la notion, en prenant soin de distinguer sa compréhension en neurosciences de son élaboration en économie (Glimcher, 2010).

d'étude, mais dans une visée d'évaluation et, éventuellement, de correction. Les auteurs endossent à la fois le rôle de juge et celui d'éducateur. En effet, les analyses méthodologiques critiquent sévèrement l'état actuel de la neuroéconomie mais proposent des perspectives d'approfondissement pour l'avenir. Elles considèrent ainsi qu'il existe des mauvaises, mais aussi des bonnes façons de faire de la neuroéconomie. Don Ross, économiste et philosophe des sciences, fournit sans doute la meilleure illustration de cette attitude de conseiller critique. Dans un article de 2008 ayant eu un large impact dans la littérature spécialisée, Ross distingue deux voies de recherches en en neuroéconomie, l'une étant jugée prometteuse, l'autre étant largement rejetée. La première, baptisée du nom d' « *économie neurocellulaire* » (*neurocellular economics*), renvoie à un courant de recherches très spécialisées, associées pour l'essentiel à l'équipe de Paul Glimcher, et donc très nettement localisées. Or, ce courant est sous-estimé car « *il est inconnu de la plupart des économistes* ». A l'inverse, le courant dominant en neuroéconomie, qualifié d' « *économie comportementale dans le scanner* » (*behavioral economics in the scanner*) bénéficie d'une audience beaucoup plus importante, du fait d'une attitude faussement provocatrice qui se propose d'expliquer les phénomènes sociaux et individuels à l'échelle neuronale (Ross, 2008, p.473-474).

Cet exemple est particulièrement révélateur du rôle joué par cette littérature dite méthodologique, de méta-réflexion, dans la constitution de la neuroéconomie. Don Ross a ultérieurement publié un ouvrage intitulé *La mutinerie de la mésencéphale: la Pico-économie et la neuroéconomie de l'addiction aux jeux financiers* (Ross *et al*., 2008) qui peut se comprendre à la fois comme une réflexion sur la portée économique de la neurobiologie mais aussi et surtout comme un manuel de neuroéconomie à proprement parler. Francesco Guala et Geoffrey Hodgson assument quant à eux pleinement ce rôle actif du philosophe et de l'économiste dans la direction des recherches expérimentales: dans un article intitulé « Le philosophe dans le scanner », les deux auteurs affirment que les spécialistes de méthodologie doivent participer à la pratique de la recherche, en critiquant les interprétations des résultats, en proposant des tâches expérimentales, *etc.* (Guala et Hodgson, 2010).

Qu'elles visent à critiquer ou à promouvoir la neuroéconomie, les études méthodologiques sont adossées dans les deux cas à une certaine conception de ce que doit être une bonne science. Il s'agit de mieux fonder une discipline encore jeune, ce qui explique cette vision téléologique de l'histoire. Or, si cette tâche a un intérêt scientifique, elle demeure en revanche très éloignée des préoccupations qui orientent une étude historique. Au mieux s'agit-il d'« histoire-fiction » : on juge la portée future de l'hypothèse d'un rapprochement entre économie et neurosciences. Harrison et Ross écrivent par exemple : « *la viabilité du*

programme et son importance potentielle reposent sur l'hypothèse empirique selon laquelle les signaux dopaminergiques dans le striatum ventral et le cortex médial préfrontal constituent une «monnaie commune », qui a de nombreuses propriétés communes avec le concept d'utilité en économie. Si cette hypothèse est correcte, alors les neuroscientifiques pourraient exploiter utilement un siècle de progrès réalisés par les économistes pour modéliser les processus de valuation dans le cerveau » (Harrison et Ross, 2010, p.187)[14].

Cette dimension spéculative ou d'histoire-fiction apparaît aussi dans les analyses visant non pas à critiquer mais à promouvoir et diffuser la neuroéconomie, qui se caractérisent par une certaine tonalité « eschatologique »: le renouvellement de la théorie économique par les neurosciences doit avoir lieu, mais il est encore à venir. Christian Schmidt en appelle à la fin de son ouvrage à une ré-interprétation de la pensée économique par la neurobiologie, et souhaite l'émergence d'une « *neuroéconomie politique* » (Schmidt, 2010, p.251). Paul Glimcher conclut son ouvrage sur une déclaration programmatique, inclinant à la modestie quant à l'état des avancées actuelles: « *la construction d'une théorie économique du cerveau est une gigantesque tâche, largement empirique. Ce livre sert plus à décrire comment l'on pourrait commencer à construire une telle théorie plutôt qu'à indiquer à quoi cette théorie pourrait ressembler, à terme* » (Glimcher, 2003, p.322).

(ii) Hypothèse interprétative:l'approche pathologique du comportement comme apport spécifique de la neuroéconomie

Qu'il soit envisagé comme le signe d'une promesse ou d'une insuffisance, le caractère non-achevé de la neuroéconomie incite à juger de son contenu scientifique. Plutôt que de prendre part au débat entre promoteurs et critiques méthodologiques de la neuroéconomie, notre étude vise à comprendre, en amont, comment la collaboration entre neurosciences et économie a pu faire l'objet d'un débat au sein de la communauté des chercheurs. Notre hypothèse est la suivante: la condition historique de possibilité de ce projet théorique (ainsi que de la critique qui lui est associée) est l'apparition, en économie, d'un questionnement théorique relatif à la pathologie et aux troubles du comportement. C'est parce que l'analyse économique se transforme, pour reprendre les termes de Gul et Pesendorfer, en une approche « *thérapeutique* » du comportement individuel (Gul et Pesendorfer, 2005, p.8), que le recours aux neurosciences devient pertinent dans ce domaine. Dans cette visée thérapeutique, il ne s'agit plus d'analyser « *la manière avec laquelle les choix individuels interagissent au sein*

14

d'un ensemble institutionnel, étant donnés les objectifs de ces individus », mais d'« *améliorer les objectifs de l'individu »*(Gul et Pesendorfer, 2005, p.8), en distinguant ce qui relève du normal et du pathologique. La collaboration entre neurosciences et économie témoignerait ainsi de l'avènement d'un paradigme « *pathologique* » (Vallois, 2011).

Cette hypothèse interprétative situe l'apparition de la neuroéconomie dans le prolongement direct des discussions concernant le paternalisme libertarien (Thaler et Sunstein, 2003, 2008; Sunstein et Thaler, 2003; Jolls et Sunstein, 2006; Gruber et Koszegi, 2001, O'Donoghue et Rabin, 2003, Ariely; 2008), également appelé paternalisme asymétrique par Camerer *et al.* (2003). En effet, il est significatif de constater que le terme de neuroéconomie n'apparaît et ne se développe dans le langage théorique qu'à partir de 2002-2003. Pourtant, les techniques de neuroimagerie sont inventées et utilisées dès les années 1990. Surtout, des travaux portant sur l'activité neuronale chez le singe, comme ceux de Paul Glimcher notamment, qui par la suite ont été intégrés à la neuroéconomie, sont publiés également dans la décennie précédant les années 2000. En particulier, l'étude célèbre de Platt et Glimcher, parfois considérée comme l'expérience fondatrice de la neuroéconomie (bien qu'elle n'en utilise pas encore le terme), a été réalisée en 1999 (Platt et Glimcher, 1999).

Ce délai d'incorporation, de quelques années, des avancées des neurosciences au sein de l'économie peut apparaître relativement bref. Cependant, au vu de la rapidité avec laquelle l'utilisation des techniques neuroscientifiques s'est développée par la suite, il est légitime de se demander ce qui a pu provoquer cet intérêt si soudain pour la neurobiologie autour des années 2002-2003. Quoique pleinement convaincus, par la suite, de l'intérêt des techniques neuroscientifiques, les économistes travaillant et réalisant des expériences en laboratoire n'étaient sans doute pas, avant les années 2000, enclins à considérer ces mêmes techniques comme étant utiles à leurs recherches. Contrairement à ce qui est souvent défendu par les neuroéconomistes (voir par exemple Camerer, Loewenstein et Prelec, 2005), le mariage entre économie comportementale et/ou expérimentale[15] et neurosciences ne n'est pas d'emblée imposé comme une évidence.

Bien au contraire, l'économie comportementale n'était pas au départ, selon nous, compatible avec la neurobiologie. Les recherches dites comportementales en économie sont en effet largement été influencées par les travaux du psychologue Daniel Kahneman (Heukelom, 2009, p.123); or celui-ci a toujours été assez hostile aux approches évolutionniste dont sont nourries les neurosciences[16]. Une rupture théorique a donc été

15 La distinction entre économie expérimentale et comportementale est précisée plus bas.

16 Une importante controverse a en particulier opposé Kahneman et Tversky au psychologue Gigerenzer, plus favorable aux schémas d'explication évolutionnistes (Kahneman et Tversky, 1996; Gigerenzer, 1996). Pour

nécessaire pour que certains économistes comportementalistes se transforment en neurobiologistes. Notre hypothèse est que cette transformation est liée à l'émergence d'un débat autour de la notion de paternalisme libertarien. Ces discussions, lancées par les premières contributions de Cass Sunstein et Richard Thaler (Sunstein et Thaler, 2003 ; Thaler et Sunstein, 2003), concernent le versant dit normatif de l'économie comportementale. Il s'agit en effet d'élaborer des recommandations en matière de régulation et de politique publique, de manière à améliorer le bien-être des agents. Le paternalisme libertarien repose notamment sur l'idée -supposée être corroborée par des expériences de laboratoire- selon laquelle les individus prennent de mauvaises décisions et sont souvent incapables d'optimiser leurs propre bien-être. L'enjeu consiste alors à restaurer une norme de rationalité individuelle (Thaler et Sunstein, 2003, 2008; Sunstein et Thaler, 2003; Jolls et Sunstein, 2006; Gruber et Koszegi, 2001, O'Donoghue et Rabin, 2003, Ariely; 2008).

Ces analyses normatives[17] ont abouti à remettre en cause la stricte séparation, défendue par Kahneman[18], entre ce qui relève de l'économie normative et de l'économie descriptive. Il s'agit en effet de se prononcer sur des questions de régulation et d'intervention publique, sur des comportements qualifiés d'« erreurs » ou de « pathologies ». C'est précisément à ce moment que s'est révélée pertinent pour les économistes comportementalistes traitant de ces problèmes théoriques, tout un champ de réflexion en neurosciences et en psychologie évolutionniste, portant sur les conduites pathologiques. C'est la raison pour laquelle nous considérons que la neuroéconomie débute en 2002-2003, à partir des premières contributions de Sunstein et Thaler (*cf. supra*).

La condition de possibilité de la neuroéconomie est donc l'apparition d'une approche médicale ou « *thérapeutique* » du choix individuel (*cf.* Gul et Pesendorfer, 2005, p.8). Il y a là quelque chose de tout à fait inédit pour l'économie. Celle-ci a pu peut-être, dans le passé, s'inspirer ou être influencée par la médecine ; en revanche, l'idée selon laquelle les choix observés ne permettraient pas d'être utilisés directement pour l'analyse du bien-être, parce que

une analyse de cette controverse, voir Jullien et Vallois, 2012.

17 Conformément aux usages en vigueur dans la littérature (Thaler et Sunstein, 2003, 2008; Sunstein et Thaler, 2003; Jolls et Sunstein, 2006; Gruber et Koszegi, 2001, O'Donoghue et Rabin, 2003, Ariely; 2008 ; Bernheim, 2008), nous utiliserons l'expression d'économie normative au sens défini ci-dessus, c'est-à-dire comme l'élaboration théorique des moyens permettant d'améliorer le bien-être des individus. Elle est dite normative car il s'agit de restaurer une norme de rationalité individuelle. L'expression d'économie normative est prise comme un synonyme d'économie du bien-être ; Douglas Bernheim par exemple parle indifféremment d'analyse comportementale normative ou du bien-être (*behavioral welfare analysis* ou *normative welfare analysis*) (Bernheim, 2008). Celle-ci se distingue de l'analyse descriptive ou positive -ces deux expressions étant elles aussi utilisées comme équivalentes- qui vise simplement à rendre compte des comportements, sans statuer ou évaluer leur rationalité.

18 Cette thèse constitue le principe directeur des travaux de Kahneman et apparaît, implicitement ou explicitement dans la quasi-totalité de ses travaux; voir par exemple Tversky et Kahneman, 1986, p.8252.

les individus sont incapables de maximiser correctement leurs fonctions d'utilité est complètement nouvelle en économie. Cette vision thérapeutique va en effet complètement à l'encontre du principe des préférences révélées, comme le soulignent Gul et Pesendorfer (Gul et Pesendorfer, 2005, p.10) ; Wade Hands parle d'un « *tournant normatif* » de l'analyse économique contemporaine, initiée par ces discussions autour du paternalisme libertarien (Hands, 2009, p.1).

Notre interprétation de la neuroéconomie n'a pas l'ambition, on l'a vu, de fournir une contribution directe à la théorie économique et aux neurosciences. Il ne s'agit pas de promouvoir ce paradigme pathologique, ou d'en rejeter sa validité, ni non plus de mieux démontrer ce qui est déjà démontré dans les expériences. Cela étant, notre description historique remplit bien une tâche critique (*cf. supra*), à deux niveaux différents. Tout d'abord, en associant neuroéconomie, médecine et paternalisme libertarien, nous montrons que ce nouveau domaine de savoir est inséparable d'un projet politique de régulation des comportements. Par delà la neutralité revendiquée des chercheurs, la neuroéconomie est animée par un projet que nous proposons d'appeler « psychiatrie économique », et de comprendre comme une « idéologie scientifique » au sens de Canguilhem (*cf.* chapitre 1; Canguilhem, 1977).

Par ailleurs, la mise en avant du rôle joué par la pathologie permet de remettre en question les analyses internes de la discipline. Notre étude débouche en effet sur deux propositions principales. D'une part, la neuroéconomie se distingue radicalement de l'économie comportementale influencée par Daniel Kahneman (*cf. supra*). Plus généralement, le rôle théorique joué par les économistes dans ce domaine de recherches expérimentales très spécialisées est quasi-nul. D'autre part, les nouveaux instruments de neuroimagerie, apparus dans les années 2000, n'ont eu qu'une importance secondaire dans le développement de la discipline. Le véritable apport des neurosciences à l'économie concerne selon nous l'analyse normative, et porte plus précisément sur la définition de la norme de rationalité. Les théories du paternalisme libertarien manquent, nous le verrons, de fondement normatif solide, car elles qualifient des comportements d'irrationnels sans avoir de définition assurée de la norme de rationalité. Les neurosciences offrent une solution à cette difficulté, non pas grâce à une définition de cette norme, mais plutôt en désignant, à l'aide des critères cliniques, un ensemble de comportements et d'individus dits « pathologiques », qui peuvent par la suite être raisonnablement considérés comme non-rationnels.

Même si, encore une fois, il ne s'agit pas de rejeter ou de défendre cette approche, notre hypothèse interprétative permet néanmoins de nous situer au sein de la littérature

secondaire. Les auteurs y considèrent au contraire, on va le voir, que la neuroéconomie représente une ouverture de l'économie comportementale aux neurosciences, ce qui amène à poser le problème de l'interdisciplinarité. Par ailleurs, cet échange est envisagé essentiellement comme un transfert de techniques (expérimentales), ce qui conduit à développer une réflexion sur le rôle des instruments. Notre hypothèse interprétative tend à remettre en cause ces analyses méthodologiques.

(iii). Une révolution technologique?

La mise en avant du rôle joué par la pathologie tend tout d'abord à diminuer l'importance des techniques expérimentales, celles-ci étant souvent considérées comme l'innovation principale de la neuroéconomie. Les critiques de la neuroéconomie ciblent souvent les difficultés liées à l'utilisation de ces instruments en économie (voir par exemple Harrison, 2008-a ; Rubinstein, 2008). Pour les promoteurs de la neuroéconomie, les nouveaux instruments d'observation de neuroimagerie seraient porteurs d'une révolution théorique à laquelle l'économie ne saurait échapper. La neuroéconomie, surtout à ses débuts, a ainsi été portée par un enthousiasme technologique certain. Colin Camerer, Georges Loewenstein et Drazen Prelec -trois économistes récemment convertis aux neurosciences- justifiaient ainsi l'utilisation d'un appareillage expérimental emprunté aux neurosciences en affirmant, dans un article célèbre paru en 2005 que « *les nouveaux outils définissent des nouveaux champs scientifiques et transgressent les anciennes frontières entre disciplines. Le télescope créa l'astronomie en sortant la science de la pure spéculation cosmologique. Le microscope rendit possible des avancées similaires en biologie. La même chose est vraie de l'économie* » (Camerer, Loewenstein et Prelec, 2005, p.10).

Les « *nouveaux outils* » ne se limitent pas à la seule imagerie par résonance magnétique fonctionnelle (IRM_f). Ils mobilisent plusieurs techniques différentes (Camerer, Loewenstein, Prelec, 2005, p.11-14). Les plus connues sont les techniques d'imagerie cérébrale ou de neuroimagerie, qui offrent la possibilité d'observer le fonctionnement du cerveau *in vivo*: imagerie par résonance magnétique fonctionnelle (IRM_f), magnétoencéphalographie (MEG), électroencéphalographie (EEG), tomographie par émission de positons (TEP). Parmi ces instruments, l'IRM_f dispose de nombreux avantages et reste la technique la plus utilisée dans les expériences de neuroéconomie[19]. Cette méthode d'imagerie

19 L'IRM_f a deux principaux avantages: elle est une méthode non-invasive, c'est-à-dire qu'elle n'implique pas d'effraction de la peau, et elle offre une bonne résolution spatiale, ce qui signifie que la localisation de l'activité dans les différentes régions du cerveau est assez précise . A l'inverse, la tomographie par émission

a cependant des inconvénients. Sa résolution temporelle est relativement faible, et son utilisation demeure soumise à d'importantes contraintes[20]. Au regard des avantages et des inconvénients de chacune de ces techniques, les chercheurs sélectionnent l'instrument le plus adapté aux besoins de l'expérience (Camerer, Loewenstein, et Prelec, 2005, p.11).

La neuroimagerie inclut également les micro-électrodes. Ces dernières sont des techniques dites d'électrophysiologie, qui mesurent l'activité électrique directement au niveau du neurone. Les micro-électrodes ont beaucoup été utilisées dans les années 1990, pour les recherches sur le primate. En dehors de la neuroimagerie au sens large (IRM_f, TEP, MEG, EEG ou micro-électrodes), les neuroscientifiques se servent aussi, chez l'animal, de techniques de stimulation électrique (*electric brain stimulation*), qui permettent d'activer un groupe de neurones ou une région du cerveau. Chez l'homme, la stimulation magnétique transcrânienne répétitive (SMTr) permet à l'inverse d'inhiber l'activité dans certains régions du cerveau. En observant les effets sur le comportement d'une stimulation par SMTr, il est ainsi possible d'inférer la fonction cognitive ou comportementale d'une région (*cf.* Knoch *et al*., 2006).

La révolution technologique consisterait donc à associer l'ensemble de ces instruments plus ou moins récents (Camerer, Loewenstein et Prelec, 2005, p.10). Pour les défenseurs de la neuroéconomie comme Camerer, Loewenstein et Prelec, cette convergence technologique dans les sciences du cerveau serait susceptible de transformer radicalement la science de son époque, et donc l'économie. Ces propositions, en retour, nourrissent des critiques relatives à la fiabilité de ces nouveaux instruments (Harrison, 2008-a ; Rubinstein, 2008), et qui remettent en question l'idée selon laquelle le progrès technique dans les neurosciences serait à l'origine

---

de positons (TEP) suppose l'injection, dans le sang, d'un marqueur (faiblement) radioactif. Apparues dans les années 1990, l'IRM_f et la TEP améliorent toutes deux sensiblement la résolution spatiale des méthodes traditionnelles d'imagerie. Avant l'apparition de l'IRM_f et de la TEP, il existait deux principales techniques. La plus ancienne de ces méthodes est l'électro-encéphalographie (EEG), mise au point en 1929 par le neurologue Hans Berger. Elle consiste dans une mesure directe de l'activité électrique du cerveau. Relativement proche dans son principe, la magnéto-encéphalographie (MEG) s'appuie sur une mesure des champs magnétiques induits par l'activité du cerveau. EEG et MEG disposent d'une bonne résolution temporelle, mais leur résolution spatiale est beaucoup moins bonne que celles de l'IRM_f et la TEP.

20 Ces contraintes sont financières tout d'abord, puisque cet équipement coute très cher, et son accès demeure en règle générale réservé en priorité aux recherches médicales. Aux États Unis, quelques rares centres de recherches, comme le *Center for Neuroeconomic Studies* de la *Claremont University* ou le *Center for the Study of Neuroconomics* de la *George Mason University*, disposent de leur propre IRM_f, dont l'usage est spécifiquement dédié à la recherche en neuroéconomie. Le plus souvent, comme c'est le cas notamment en France, ces installations sont destinées à la recherche médicale, et leur utilisation pour des buts non-thérapeutiques est limitée à des plages horaires relativement restreintes, en fonction des besoins de l'activité hospitalière. A ces contraintes d'accès s'ajoutent des contraintes pratiques et logistiques. Bien que non-invasive, l'IRM_f est une technique qui suppose la mise en place d'un examen assez lourd, source d'inconfort et de gêne pour le patient. Celui-ci est obligé de se tenir immobile, allongé, dans un scanner étroit, dont le fonctionnement est par ailleurs extrêmement bruyant. Ces conditions compliquent donc fortement la tâche des expérimentateurs qui souhaitent reproduire de la façon la plus fidèle possible le prise de décision économique en situation « réelle ».

d'un important progrès théorique en économie.

Ce débat méthodologique surestime selon nous l'importance des techniques de neuroimagerie et de stimulation. Notre étude historique montre en effet que la plupart des concepts qui forment le noyau dur théorique de la neuroéconomie ont été élaborés bien avant l'apparition de ces nouveaux instruments dans les années 1990 et 2000. D'ailleurs, certaines expériences récentes de neuroéconomie sont réalisées sans ces techniques, en comparant par exemple, comme c'est l'usage en psychopathologie, les comportements de sujets sains et « pathologiques » (sujets lésés, suivant un traitement neuropharmacologique, *etc.*).

Surtout, le principal défaut de ces analyses méthodologiques portant sur le rôle des techniques est, de notre point de vue, de prendre les objets respectifs de l'économie et des neurosciences comme donnés, sans étudier la manière avec laquelle ces objets ont été construits, pour ensuite défendre ou critiquer l'existence d'intérêts théoriques communs. La défense « technologique » de la neuroéconomie par Camerer, Loewenstein et Prelec (*cf. supra*) s'appuie sur le présupposé selon lequel neurosciences et économie partageraient un ensemble d'objets d'études communs. Colin Camerer, en collaboration avec un autre économiste -Ernst Fehr- a proposé par la suite une synthèse d'un sous-domaine de la neuroéconomie, qualifié de « *neuroéconomie sociale* ». Cet article s'ouvre sur l'affirmation selon laquelle les « *instruments des neurosciences [...] génèrent des outils puissants pour étudier les processus cérébraux à l'œuvre dans les interactions humaines* » (Fehr et Camerer, 2007, p.419). Ce qui est admis sans autre justification ici, est que l'économie et les neurosciences, aient naturellement vocation à traiter d'un même objet, qui serait les « *interactions humaines* ».

D'une façon similaire, pour Christian Schmidt, l'« *étude des décisions raisonnées* » fournit un point d'ancrage à la rencontre entre neurosciences et économie (Schmidt, 2010, p.21). De leur côté, des auteurs critiques comme Gul et Pesendorfer contestent cette possibilité, en faisant valoir que la modélisation de la décision en économie ne porte pas sur ses processus neuronaux sous-jacents (Gul et Pesendorfer, 2005). Une part importante de l'opposition entre les auteurs est liée ici à l'ambiguïté du terme de « décision », signalée plus haut : la neuroéconomie peut en effet faire référence à une approche neurobiologique des comportements économiques « réels », comme l'entendent Gul et Pesendorfer, mais aussi à une application des théories économiques au fonctionnement du cerveau (*cf. supra*), comme le défendent notamment Ross et Harrison (2008). Quoiqu'il en soit, l'idée selon laquelle l'économie et les neurosciences auraient par nature vocation à partager ou non un ensemble d'objets d'étude constitue ici un présupposé commun aux promoteurs de la neuroéconomie et

à ses critiques.

Cette perspective conduit à des propositions anachroniques. En particulier, les neuroéconomistes justifient souvent leur approche en tirant parti de l'existence d'un intérêt théorique pour le fonctionnement du cerveau chez les économistes, intérêt qui n'aurait pu être assouvi jusqu'à l'apparition des techniques modernes des neurosciences. William Jevons, économiste marginaliste du XIXe siècle, qui voulait trouver un moyen de mesurer quantitativement, sur l'homme, l'utilité ordinale, est régulièrement cité à ce titre comme un précurseur de la neuroéconomie. Camerer, Loewenstein et Prelec considèrent que la neuroimagerie accomplit l'espoir déçu de Jevons, en fournissant les variables neuronales du plaisir : « *La théorie économique a été construite en admettant que les détails concernant la boite noire du cerveau humain resteraient inconnus. Ce pessimisme était exprimé par Jevons en 1871. Mais aujourd'hui les neurosciences ont montré que la prédiction pessimiste de Jevons était fausse : l'étude du cerveau et du système nerveux humain autorise une mesure directe des pensées et des sentiments* » (Camerer, Loewenstein et Prelec, 2005, p.10).

Pour Camerer, Loewenstein et Prelec, les neurosciences contemporaines viendraient donc combler une lacune importante de la théorie économique, en offrant les outils nécessaires pour ouvrir la « boite noire » du cerveau humain. Cette avancée technique répondrait à un besoin théorique -rendre compte des processus neurophysiologiques- qui, selon Camerer, Loewenstein et Prelec, a toujours existé. Les neurosciences rempliraient un trou ou un manque de la théorie économique, et il y aurait une coïncidence parfaite entre les interrogations formelles des économistes, et les recherches empiriques des neurophysiologistes: « *les économistes [...] fournissent de riches outils conceptuels pour comprendre et modéliser le comportement, alors que les neurobiologistes fournissent des outils pour l'étude du mécanisme* [neuronal] » (Glimcher et Rustichini, 2004, p.447).

Pourtant, il n'est pas du tout sûr que les économistes aient toujours été disposés à considérer la physiologie du cerveau comme un possible instrument de réflexion théorique. Si Jevons avait eu à sa disposition une IRM fonctionnelle, il est très peu probable qu'il en ait fait le même usage que celui des neuroéconomistes aujourd'hui, indépendamment de toutes considération technique. Dans notre perspective, l'émergence de la neuroéconomie n'a été possible qu'à la faveur d'une double transformation, expérimentale et médicale, de la théorie économique, qui ne s'était pas encore produite au XIXème siècle. En mettant l'accent sur ses conditions historiques de possibilité, notre analyse tend donc à nuancer le rôle joué par les innovations technologiques très récentes en neurosciences.

Les connaissances produites aujourd'hui par la neuroéconomie n'auraient pas pu valoir comme des éléments de théorie économique à l'époque de Jevons. Les neuroéconomistes cherchent en effet à établir des corrélations entre variables neuronales et comportementales dans des jeux économiques ou des protocoles standardisés (choix entre des loteries, jeu de l'investisseur par exemple). Par exemple, McCabe *et al.* (2001) observent que la coopération est associée à une activité des régions préfectorales lorsqu'elle a lieu spécifiquement entre sujets humains (McCabe *et al.*, 2001). Avant ces premiers travaux sur l'homme, des études sur le singe, qui peuvent aussi être rattachées à la neuroéconomie, portaient sur des tâches de conditionnement très frustres, dans lesquelles l'animal associe à un stimulus l'obtention d'une récompense alimentaire (jus de fruit) (voir par exemple, Dayan, Schultz et Montague, 1997).

De telles expériences semblent avoir une portée très restreinte. La réaction spontanée des économistes face à ce genre de travaux consiste en général à souligner les difficultés liées à la généralisation des résultats obtenus: quel est le point commun entre la consommation de jus de fruit par un singe et des décisions d'investissement prises par un agent économique? Chez l'homme, l'environnement aseptisé du laboratoire, et l'inconfort important auquel les techniques d'imagerie soumettent les sujets, ne représentent-ils pas une déformation importante des conditions « réelles » du choix économique et n'empêchent-ils pas de simuler les décisions telles qu'elles sont prises sur le marché? Les résultats, obtenus sur un petit nombre de participants, soigneusement sélectionnés par les expérimentateurs, valent-ils pour l'ensemble de la population?

Ces remarques ne signifient pas nécessairement que la neuroéconomie représente un projet théorique vain. Cependant, il y a indiscutablement dans la démarche proposée par les neuroéconomistes quelque chose qui apparaît comme très déroutant pour beaucoup d'économistes, et qui est lié à la nature inductive des recherches dans le domaine des sciences biologiques. Les neuroéconomistes ne cherchent pas en effet à répondre à tester des modèles économiques de l'investissement, ou à mesurer l'utilité dans le cerveau: il s'agit toujours, dans le cadre de protocoles fortement codifiés, d'identifier des activités neuronales et/ou cérébrales, pouvant éventuellement être associées à une fonction cognitive bien délimitée.

La démarche des neuroéconomistes apparaît donc comme très parcellaire, et très limitée pour les économistes, parce qu'elle procède, de proche en proche, par une généralisation progressive et prudente de résultats à la portée très restreinte. Inversement, du point de vue des neuroéconomistes, les modèles des neuroéconomistes sont beaucoup trop généraux. Ils sont de ce point de vue critiquables non pas nécessairement parce qu'ils sont

invalidés par les données empiriques, mais précisément parce que la confrontation aux faits n'intervient qu'en dernière analyse, pour vérifier si la conjoncture initiale était fondée. Paul Zak, par exemple, constate que les conclusions auxquels il aboutit dans ses recherches expérimentales sur l'altruisme et la confiance convergent avec celles de modèles proposés par des économistes (Andreoni 1990; Rabin 1993; Fehr et Gächter, 2000; Sally 2000, 2001; Levitt and List 2007). Quoique proches dans leur esprit, ces modèles sont cependant rejetés par Zak: « *mais nous soulignons que ces modèles demeurent déductifs; ils utilisent une approche qui s'appuie sur des conjectures et des vérifications plutôt que sur une expérimentation systématique* » (Vercoe et Zak, 2010, p.128).

Évidemment, la défense d'une approche purement inductive est assez naïve, dans la mesure où toute induction doit nécessairement s'appuyer à un moment ou un autre sur des considérations théoriques afin de généraliser les résultats obtenus. Par exemple, McCabe *et al.* (2001) associent à l'activité des régions pré-frontales l'exercice de facultés de « lecture des pensées » (*mind-reading*) qui, selon les auteurs, joue un rôle important dans la confiance interpersonnelle. Cette localisation ne reste bien sûr qu'une hypothèse, dans la mesure où elle suppose d'importants postulats théoriques: en particulier, les auteurs supposent que l'exercice de ces facultés est nécessaire, dans le jeu considéré, pour adopter des comportements coopératifs; ils supposent par ailleurs que ces facultés sont à l'œuvre spécifiquement entre sujets humains, et non face à un ordinateur; l'activité dans les régions préfrontales est distinguée d'autres fonctions qui seraient susceptibles d'être engagées, comme la simple réflexion stratégique; sans compter sur les nombreuses simplifications statistiques sur lesquelles repose le traitement des données de la neuroimagerie.

Cependant, même si, sur le plan strictement méthodologique, toute induction dans les disciplines expérimentales n'est jamais pure et repose sur une part de raisonnement déductif, la neuroéconomie demeure un domaine de recherche étrange pour beaucoup d'économistes. Les neuroéconomistes observent des champs de phénomènes très restreints, et n'aboutissent jamais à proposer de véritable modèle global du comportement. La neuroéconomie n'a de ce point de vue pas du tout vocation à répondre à des questions très générales sur la conscience, la pensée, l'inconscient, *etc.*: le corpus de la neuroéconomie, même dans ses parties les plus spéculatives, se distingue nettement de ce qui est connu sous le nom de « neurophilosophie », c'est-à-dire de toute la littérature qui vise à évaluer la portée ou la signification philosophique des neurosciences (*cf.* Andrieu, 2007).

Pour Camerer, Loewenstein et Prelec, Jevons aurait été intéressé par un appareil capable de mesurer dans le cerveau l'intensité du plaisir et des satisfactions. Les techniques d'imagerie ne remplissent cependant pas du tout cette fonction. Pour que ces nouveaux outils commencent à être envisagés comme de possibles instruments d'observation en économie, il a fallu, au préalable, que l'économie se transforme science expérimentale du comportement humain. La neuroéconomie participe donc d'une évolution de la théorie économique, qui a vu celle-ci s'ouvrir aux techniques d'expérimentation en laboratoire.

Afin de bien comprendre les domaines de recherche dans lesquels s'inscrit la neuroéconomie, il est donc nécessaire de décrire rapidement les grandes lignes de cette évolution. Les économistes admettent généralement (voir par exemple Willinger et Eber, 2005) que la première expérimentation en économie a été réalisée en 1948 par Edward Chamberlin, qui simulé, à l'aide de ses étudiants le fonctionnement d'un marché concurrentiel (Chamberlin, 1948). Assez rapidement, le recours aux simulations expérimentales se développe en économie, sous l'impulsion notamment de Herbert Simon, fondateur de l'économie comportementale. Au départ, les simulations expérimentales font figure d'hétérodoxie, ou plutôt s'affirment comme telles, et Herbert Simon a conçu son approche comme une alternative au programme de cherche dit « dominant ». Pour les historiens (Heukelom 2009; Sent, 2004, 2005), le programme de recherche mené par Simon s'est ainsi avéré être un relatif échec, dans la mesure où celui-ci n'est jamais parvenu à intégrer le courant « dominant », « standard », ou « orthodoxe » de l'analyse économique[21].

Selon Ester Mirjam Sent, ce n'est que dans les années 1980, à la suite des travaux de Kahneman et Tversky, que l'économie comportementale parvient à s'affirmer comme un nouveau courant de la théorie économique orthodoxe. Sent distingue ainsi l'« ancienne économie comportementale » (*old behavioral economics*) de Simon de la « nouvelle

21 Bien que les expressions d'économie « orthodoxe », « standard », ou « courant dominant » apparaissent fréquemment dans la littérature, y compris en histoire de la pensée économique, et soit ici utilisée par Heukelom (2009) ou Sent (2004, 2005), nous n'adopterons pas dans ce travail cette terminologie. Il apparait en effet assez contestable, pour un historien de la pensée économique, de considérer que l'ensemble des travaux relevant de ce courant soi-disant dominant relève de la même approche. Cela tend à nier d'une part l'importance des controverses et d'autre part la diversité pourtant bien réelle des approches au sein de ce courant dominant. Si l'économie « orthodoxe » représente un seul et unique programme de recherche, comment définir son « noyau dur théorique »? Le plus souvent, l'économie dominante est renvoyée à l'« économie néo-classique », ce qui, encore une fois, n'est pas satisfaisant, puisque l'économie néoclassique ne représente pas une seule et même école de pensée du XIXe siècle à nos jours, et que d'autres branches de l'économie, comme la théorie des jeux par exemple, semblent mériter tout autant le qualificatif de courant dominant. Nous adopterons donc des définitions de l'économie comportementale et de l'économie expérimentale sans faire référence à leur positionnement à l'égard de ce courant dominant (*cf. infra*), en s'appuyant seulement sur leur objet d'étude (le choix individuel pour l'économie comportementale, le fonctionnement des mécanismes de marché pour l'économie expérimentale), et leur méthode (recours à des simulations en laboratoire).

économie comportementale » (*new behavioral economics)* de Kahneman et Tversky (Sent, 2004). Cette dernière admettrait une position plus conciliante vis à vis de l'économie orthodoxe, en tant qu'elle accepterait, on y reviendra plus en détails, de reconnaître à la théorie du choix standard une validité normative. Dans le même temps, l'économie expérimentale développée par Vernon Smith, ancien étudiant de Chamberlin, propose une perspective de réconciliation similaire dans le domaine des simulations de marché, là où l'économie comportementale se concentre plutôt le choix individuel. Smith en effet postule que les mécanismes de marché et de régulation collective permettent de contre-balancer les limites individuelles à la rationalité et offrent ainsi un moyen, là où la rationalité des acteurs est déficiente, pour approcher l'équilibre concurrentiel (Smith, 2007).

Cette histoire de l'expérimentation en économie est évidemment trop sommaire, mais notre objectif ici est simplement de présenter et définir les prédécesseurs théoriques immédiats de la neuroéconomie. Nous retiendrons donc l'expression d'économie expérimentale ou de *market experiments* pour désigner le programme de recherche, qui, sous l'impulsion de Vernon Smith notamment, traite en économie des problèmes d'équilibre de marché en s'appuyant sur des simulations expérimentales de marché. Le terme d'économie comportementale ou *behavioral economics* sera spécifiquement réservé à ce que Sent appelle la « *nouvelle économie comportementale* », associée aux travaux de Kahneman et Tversky, et qui porte sur la rationalité individuelle (Sent, 2004).

L'émergence de la neuroéconomie a donc eu pour condition de possibilité historique l'émergence elle-même antérieure de l'économie expérimentale et comportementale. John Davis voit dans cette ouverture progressive de l'économie, dans la deuxième moitié du XXe siècle, aux disciplines expérimentales et, plus généralement, à toute une série d'autres disciplines (science sociales, humaines, littérature, physique, *etc.*) un mouvement de particularisation et d'atomisation de courant dominant. Pour Davis, l'orthodoxie ne représente plus aujourd'hui, en économie, un courant unitaire et fortement structuré, mais a tendance à se scinder en de multiples sous-domaines, spécialisés vers le traitements d'objets spécifiques, et placés sous l'influence de disciplines diverses. L'économie, de plus en plus, subit l'influence d'autres sciences (Davis, 2008), et la neuroéconomie participe évidemment de cet impérialisme inversé (Davis, 2010).

Effectivement, la neuroéconomie semble bien participer de ce mouvement de particularisation et d'éclatement de la théorie économique, puisque, d'une part, la portée de ses résultats expérimentaux est limitée à des comportements ou des individus particuliers. D'autre

part, ces mêmes résultats fournissent des contributions théoriques aux domaine restreint de la théorie de la décision individuelle, et, plus précisément, à la théorie du choix inter-personnel et inter-temporel. Il y a donc tout un pan de l'analyse économique qui n'est pas exploré, et qui n'a pas vocation à l'être, par la neuroéconomie, celle-ci étant avant tout une entreprise d'analyse du choix individuel. Les phénomènes macroéconomiques notamment, ou le fonctionnement des institutions, ne constituent notamment pas des objets d'étude pour les neuroéconomistes. Il n'y a donc pas de neuro-macro-économie ou de neuroéconomie expérimentale, pour la simple raison que les marchés et les phénomènes macroéconomiques n'ont pas de cerveau. Bien sûr, cela n'empêche nullement les neuroéconomistes de fournir des explications microéconomiques, en termes de rationalité individuelle, à ces variables macro-économiques: par exemple, les travaux sur le mimétisme sont susceptibles d'éclairer les mécanismes sous-jacents à l'euphorie financière, des mécanismes psychologiques comme l'aversion aux pertes peuvent rendre compte de processus de formation ds prix dans les enchères étudiés par Smith (voir par exemple Delgado, 2008). Cependant, comme le souligne Christian Schmidt, ce type de travaux reste assez rare, et les implications collectives de la neuroéconomie sont assez largement ignorées (Schmidt, 2010, p.289).

Le questionnement méthodologique portant sur le rôle des techniques d'observation en neuroéconomie est très éloigné des préoccupations qui animent une étude historique. Les auteurs qui participent à ce débat prennent les objets et les méthodes respectifs des neurosciences et de l'économie comme donnés, et cherchent à évaluer leur compatibilité, sans interroger leur construction théorique. Or, on l'a vu, avant l'apparition des premières simulations économiques en laboratoire, les économistes n'avaient pas cet intérêt théorique pour la psychologie expérimentale, qui, aujourd'hui, est à l'origine directe de la neuroéconomie.

(iv) Nouvelle approche ou nouvelle discipline? Les rapports de la neuroéconomie à l'économie comportementale

Le traitement du problème de l'interdisciplinarité dans la littérature secondaire est également caractérisé par une absence de mise en perspective historique. Les disciplines limitrophes de la neuroéconomie (économie, économie comportementale, neurosciences, psychologie) sont considérées comme ayant des frontières stables et préexistantes, dont l'origine n'est pas questionnée. Le problème méthodologique qui se pose alors est de

déterminer dans quelle mesure ces frontières peuvent ou non être franchies (voir par exemple Mäki, 2009). Dans cette perspective, la neuroéconomie est conçue comme une tentative d'approfondissement de l'économie comportementale par les neurosciences. Cela conduit à faire de la neuroéconomie une nouvelle approche -neuroscientifique- en économie comportementale, plutôt qu'une nouvelle discipline à part entière (*cf. infra* ; Camerer, Loewenstein et Prelec, 2005; Sanfey, Loewenstein et Cohen, 2006). A l'inverse, notre interprétation dissocie radicalement *behavioral economics* et neuroéconomie. Cette dernière est située dans le prolongement d'un domaine de recherches expérimentales spécialisées portant sur le *reward learning*: d'un point de vue historique, la question pertinente n'est pas de savoir si des frontières disciplinaires considérées comme données peuvent être ou non franchies, mais de décrire la manière avec laquelle un domaine de savoir autonome s'est progressivement doté de ses propres frontières, pour se démarquer des programmes de recherche pré-existants.

A ses débuts, la neuroéconomie a souvent été présentée comme un approfondissement direct de l'économie comportementale. Pour Camerer, Loewenstein et Prelec par exemple, les *behavioral economics* représentent un premier élargissement de l'économie à la psychologie, qui a débouché naturellement sur un second élargissement aux neurosciences (Camerer, Loewenstein et Prelec, 2005, p.9). L'économie comportementale a jusqu'alors été principalement influencée par la *« recherche comportementale sur la décision »*, qui correspond au programme de recherche de Daniel Kahneman en psychologie (*cf.* Heukelom, 2009). Cette *« recherche comportementale sur la décision »* aurait elle-même vocation à s'ouvrir aux neurosciences. Ce point de vue est également partagé par Sanfey, Loewenstein et Cohen, qui, dans un article ultérieur intitulé « *le croisement des courants de recherches sur la théorie de la décision* », considèrent que la neuroéconomie représente une superposition des méthodes de l'économie, de la psychologie et des neurosciences dans l'étude du comportement (Sanfey, Loewenstein et Cohen, 2006).

Dans cette perspective, la neuroéconomie ne serait pas à proprement parler une nouvelle discipline, mais plus modestement une nouvelle approche au sein de l'économie comportementale. Elle se limiterait à étudier les même problèmes théoriques que ceux de l'économie comportementale. Effectivement, les premières expériences de neuroéconomie reproduisent des protocoles classiques des *behavioral economics*: jeu de l'investisseur (McCabe *et al.*, 2001), paradoxe de Ellsberg (Smith *et al.*, 2002), *etc.* La seule nouveauté, finalement, consiste dans l'utilisation d'un appareillage neuroscientifique. Ce type de travaux

expérimentaux a ainsi été qualifié dans la littérature d' « *économie comportementale dans le scanner* » (*cf.* Ross, 2008, p.473).

Pourtant, il n'est pas du tout sûr que cet approfondissement de l'économie comportementale par les outils des neurosciences laisse celle-ci inchangée. Il est paradoxal d'affirmer, comme le font Camerer, Loewenstein et Prelec, que l'ouverture de l'économie à la psychologie de Kahneman a profondément affecté l'économie et a été à l'origine d'un nouveau domaine de recherche, tout en refusant cette qualité à la plus récente ouverture de cette même économie comportementale aux neurosciences. Il n'y a pas de raison de supposer *a priori* que la neurobiologie confirme les explications proposées par les *behavioral economics*, ou se contente d'utiliser les mêmes concepts théoriques.

La possibilité d'une transformation radicale de l'économie par les neurosciences est bien, cependant, assumée par Camerer, Loewenstein et Prelec, puisque ceux-ci distinguent entre « *deux types de contributions* » des neurosciences à l'économie, « *incrémentales* » et « *radicales* ». Dans le premier cas, « l*es neurosciences ajoutent des variables au modèle conventionnel de prise de décision ou suggèrent des formes fonctionnelles spécifiques pour remplacer les hypothèses abstraites* [des économistes] *qui n'ont jamais été bien justifiées empiriquement* ». En revanche « *l'approche radicale implique [...] de se demander comment l'économie aurait pu avoir évolué si elle avait depuis le départ été influencée par des idées et des résultats neuroscientifiques désormais disponibles* » (Camerer, Loewenstein et Prelec, 2005, p.12). Selon Camerer, Loewenstein et Prelec, la neuroéconomie fournit à la fois des contributions incrémentales et radicales. Les modèles neuroéconomiques de la décision peuvent dans certains cas simplement ajouter des variables explicatives, ou calibrer des paramètres non précisés dans les modèles économiques, et dans d'autres cas remettre en question complètement ces mêmes modèles. Glimcher souligne également que les avancées des neurosciences sont susceptibles d'être interprétées soit comme des confirmations, soit comme des remises en cause profondes de la théorie économique (*cf.* Glimcher, 2008, p.32). Cependant, il ne fait pas de doute pour Camerer, Loewenstein et Prelec que les neurosciences sont, dans l'ensemble, porteuses d'une évolution théorique radicale: « *les neurosciences fournissent selon nous un ensemble entièrement nouveau de concepts explicatifs pour expliquer la prise de décision* » (Camerer, Loewenstein et Prelec, 2005, p.12). Or, si cet ensemble de concepts explicatifs est entièrement nouveau, il est logique de considérer que l'économie comportementale ne saurait demeurer inchangée par son ouverture aux neurosciences.

L'économie comportementale est donc susceptible d'être remise en question dans ses fondements théoriques. Les neurobiologistes peuvent reprocher aux économistes comportementalistes de formuler des hypothèses comportementales non-justifiées sur le plan neuronal ou cérébral, là où, précisément, les économistes comportementalistes reprochaient aux économistes qu'ils qualifient de « standards » de proposer des hypothèses logico-déductives non-justifiées sur le plan comportemental (voir par exemple Tversky et Kahneman, 1986, p.8251).

Si les neurosciences et l'économie comportementale peuvent donc étudier les mêmes protocoles expérimentaux, la neuroéconomie ne se réduit pas à une simple superposition théorique, comme le postulent Sanfey, Loewenstein et Cohen, pour qui « *l'intégration d'approches théoriques et de méthodologies disparates offre un potentiel stimulant pour la construction de modèles de prise de décision plus adéquats* » (Sanfey, Loewenstein, et Cohen, 2006, p.108). Cette intégration bouleverse, à chaque niveau d'analyse, les schémas interprétatifs utilisés.

Encore s'agit-il ici de prospective : il est question de savoir si l'approfondissement de l'économie comportementale donnera lieu ou non à une nouvelle discipline. De façon plus décisive, notre hypothèse d'un paradigme pathologique implique une distinction, dès le départ, entre le programme de recherche de l'économie comportementale, que nous associons aux travaux de Daniel Kahneman en psychologie[22], et celui de la psychologie évolutionniste, dont est issue la neuroéconomie (*cf. supra*). L'approche proposée par les neuroéconomistes s'inscrit en effet dans une tradition d'étude du comportement animal qui remonte aux années 1960. Ce courant de recherche, souvent méconnu des économistes, regroupe des études assez disparates et ne constitue pas au départ un programme de recherche à proprement parler (ou en tout cas n'est pas désigné comme tel dans la littérature). Ces travaux ont cependant progressivement constitué une problématique de recherche commune à l'économie, à la psychologie évolutionniste puis aux neurosciences, autour de l'étude des comportements dits d'« égalisation des rendements » (*matching behavior*) et des dynamiques dites d' « apprentissage de la récompense » (*reward learning*). Nous avons donc décidé d'appeler ce programme de recherche « néo-comportementalisme » ou « science quantitative de la

22 L'influence, supposée ici prédominante, de Daniel Kahneman sur l'économie comportementale, pourrait être discutée. Cette interprétation correspond néanmoins au point de vue de Floris Heukelom, historien spécialiste de l'économie comportementale (Heukelom, 2009, p.123). Il existe bien sûr des économistes comportementalistes qui ne sont pas nécessairement hostiles, comme Kahneman (*cf. supra*), aux références évolutionnistes en psychologie, et qui, précisément, ont pu se « convertir » aux neurosciences, comme Colin Camerer. Néanmoins, le programme de recherche des *behavioral economics* doit selon nous, dans l'ensemble, être distingué de celui de la neuroéconomie.

motivation »[23]. Le terme de néo-comportementalisme vise à souligner que cette tradition théorique s'inscrit en rupture à la fois avec le comportementalisme en psychologie, mais aussi avec l'économie comportementale influencée par la recherche comportementale sur la décision de Kahneman[24].

La neuroéconomie doit selon nous être comprise comme le prolongement direct de cette science quantitative de la motivation et non comme celui de l'économie comportementale. Les expériences de neuroéconomie portaient au départ principalement sur des protocoles classiques des *behavioral economics*, ce qui a été critiqué comme une forme d'« *économie comportementale dans le scanner* » (Ross, 2008, p.473). Cependant, ces travaux du début des années 2000 représentent selon nous une tentative d'annexion de la neuroéconomie par l'économie comportementale, qui a produit des résultats théoriques assez mitigés. En effet, les comportements et les processus neurophysiologiques étudiés en neuroéconomie ne prennent sens selon nous, qu'à l'intérieur d'un cadre théorique bien précis, portant sur les mécanismes d'apprentissage de la récompense.

Le concept d'émotions fournit ici un bon exemple. Il existe en effet, avant l'apparition de la neuroéconomie, une littérature théorique spécialisée en économie sur le rôle des émotions dans la prise de décision, en particulier chez Jon Elster (1999, 2000, 2009). En économie comportementale, Georges Loewenstein est le principal promoteur de ce concept théorique, auquel il a consacré un article intitulé « les émotions dans la théorie économique et dans le comportement économique » (Loewenstein 1999; voir aussi Loewenstein, 1996). La mise en avant des émotions, des affects ou de ce que Loewentein appelle des « *facteurs viscéraux* » (*visceral factors*, Loewwenstein, 1996), implique une perspective dualiste sur la cognition, dans la mesure où les processus émotionnels sont opposés à la raison. Ce dualisme a été assez influent au sein des *behavioral economics*, et a été notamment repris par Kahneman (2003) sous la forme d'une distinction entre « *intuition* » et « *raisonnement* » (Kahneman, 2003, p.1450).

Lorsque les économistes comportementalistes ont commencé à s'intéresser aux neurosciences, les résultats expérimentaux ont très souvent été interprétés comme confirmant, ou apportant une preuve supplémentaire de l'importance des émotions dans le comportement

23 L'expression de « science quantitative de la motivation » s'inspire de la *Society for Quantitative Analysis of Behavior*, à laquelle ont participé la plupart des chercheurs que nous avons choisi d'appeler néo-comportementalistes.

24 Il importe de souligner que l'économie comportementale ou *behavioral economics* n'a aucun rapport avec le comportementalisme en psychologie. L'expression d'économie comportementale est en fait assez trompeuse, car elle ne renvoie pas du tout aux travaux comportementalistes sur le conditionnement, mais à la psychologie kahnemanienne.

humain. Camerer, par exemple, dans une large revue de la littérature, considère que l'économie comportementale et les neurosciences convergent sur le thème des émotions (Camerer, 2003, p.1673). Cette lecture des neurosciences, associée souvent à une référence aux travaux d'Antonio Damasio (1994), a été influente, et apparaît dans les ouvrages de vulgarisation de la neuroéconomie (voir par exemple Bourgeois-Gironde, 2008, p.10). Les neurosciences ont ainsi assuré au début des années 2000 une grande popularité à un concept qui, jusqu'alors, avait été largement ignoré des économistes non-comportementaux. Or, l'appréhension des résultats expérimentaux en terme d'émotion a conduit à de nombreuses difficultés théoriques, liés notamment au cadre dualiste, et a dû par la suite être abandonnée. Cela peut s'expliquer, selon notre interprétation, par le fait que la neuroéconomie soit inscrite dans un courant de recherche portant spécifiquement sur l'apprentissage de la récompense, qui implique au contraire une approche unitaire des processus cognitifs d'évaluation[25].

L'exemple des émotions montre que les correspondances établies par les neuroéconomistes entre activités neuronales et cérébrales et fonctions cognitives n'ont de validité que dans les protocoles de *reward learning*. Par conséquent, la neuroéconomie n'est pas née de l'application de techniques neuroscientifiques à des problèmes préexistants de l'économie comportementale, ou de l'interprétation d'expériences neuroscientifiques à partir de concepts empruntés aux *behavioral economics* (effet de cadrage, biais d'ancrage, de *statu quo*, émotions ou facteurs viscéraux, *etc*), mais plutôt de l'extension progressive de cette approche du comportement individuel en terme d'apprentissage de la récompense.

La genèse théorique de la neuroéconomie montre donc que celle-ci n'a désigné au départ qu'un ensemble de recherches expérimentales très spécialisées, qui ont progressivement donné naissance à un nouveau sous-domaine de recherche autonome de l'économie, autour de l'étude du *reward learning*. Cependant, que ces réflexions ne valent que sur le plan théorique. Sur le plan institutionnel et académique, un effort important a été poursuivi pour faire de la neuroéconomie une nouvelle discipline à part entière. Néanmoins, les neuroéconomistes sont toujours au départ soit des neurobiologistes qui, à ce tire, publient principalement dans des revues de neurosciences, et sont soumis à des impératifs de recherche propres aux neurosciences; soit des économistes de formation, le plus souvent issus de

25 Les réflexions proposées par Pierre Livet (voir notamment Livet, 2009) représentent cependant un cas particulier, dans la mesure où cet auteur appréhende les émotions comme des signaux d'alarme permettant de réviser nos croyance (Livet, 2009). Cette perspective est très proche de celle du *reward learning.* Pierre Livet a proposé un modèle dit des « émotions mixtes » pour représenter les processus d'évaluation étudiés en neuroéconomie, qui demeure compatible avec la notion d'apprentissage de la récompense (Livet, 2009). Le rôle des émotions dans la neuroéconomie, et en particulier l'influence des travaux d'Antonio Damasio, seront approfondis dans les chapitres 5 et 6.

l'économie comportementale. La neuroéconomie ne dispose pas de sa propre revue académique et, lorsqu'elle est enseignée à l'université, elle est abordée en règle générale dans des cursus associant de façon large « économie et psychologie », ce qui inclut aussi l'économie comportementale. Il serait possible de discuter très longuement sur le fait de savoir si, institutionnellement, la neuroéconomie constitue une nouvelle discipline ou un simple approfondissement des *behavioral economics*. Si la neuroéconomie jouit d'un prestige académique indéniable, la pratique concrète de la recherche dans ce domaine s'opère le plus souvent à l'intérieur de programmes d'étude relevant de l'économie comportementale et/ou des neurosciences.

Sur le strict plan théorique, le partage est plus net : la neuroéconomie ne s'inscrit dans aucune tradition d'analyse en économie, parce qu'il n'y existait pas, auparavant, de conceptualisation des problèmes d'apprentissage de la récompense. D'où une implication, peut être surprenante, de notre hypothèse interprétative :la modélisation économique n'a pas joué de rôle moteur dans la construction théorique de la neuroéconomie. Le problème du *reward learning* a été inventé par les neurobiologistes, et s'est transformé *de facto* et *a posteriori* en un problème d'analyse économique.

Cette proposition va à l'encontre des analyses qui suggèrent que la neuroéconomie représente une tentative pour appliquer aux neurosciences des modèles formels de l'économie. Cette vision a notamment été promue par Glimcher, qui a défendu son approche à ses débuts comme une « *théorie économique dans le cerveau* » (Glimcher, 2003, p.322). Selon Glimcher, avant l'apparition de la neuroéconomie, la neurophysiologie dépendait d'un paradigme qu'il qualifie de « réflexologie », hérité de William Sherrington. Dans cette perspective, le cadre de référence pour penser le fonctionnement du système nerveux est le réflexe, ce qui signifie que le système nerveux central a pour fonction de coder des informations sensorielles et de transmettre des commandes motrices (Glimcher, 2003, p.253). Or, les réflexes n'ont que peu d'intérêt pour les économistes, dans la mesure où ils ne constituent pas ce que Schmidt considère comme l'objet commun aux disciplines, c'est-à-dire les « *décisions raisonnées* » (Schmidt, 2010, p.21).

Pour Glimcher, les neurosciences auraient progressivement adopté, dans les années 1990, un paradigme « économique », dans lequel le système nerveux serait caractérisé par une fonction de prédiction de la récompense, indépendante du sensori-moteur (Glimcher, 2003, p.253). La théorie économique de l'utilité espérée aurait joué selon Glimcher un rôle central au cours de cette évolution, en constituant un nouveau cadre théorique pour penser le

fonctionnement du système nerveux (Glimcher, Dorris et Bayer, 2005, p.213).

Or l'« utilité espérée physiologique » dont parle Glimcher (Glimcher, Dorris et Bayer, 2005) est une expression trompeuse car elle ne correspond pas du tout à la notion d'utilité espérée en économie. D'où deux confusions très fréquentes dans la littérature : la neuroéconomie est parfois comprise comme une confirmation, au niveau neuronal, de la théorie de l'utilité espérée[26], ou bien comme une tentative pour mesurer, dans le cerveau, le phénomène de l'utilité (Camerer, Loewenstein et Prelec, 2005, p.10). Or les neuroéconomistes sont très loin de pouvoir mesurer la quantité de plaisir dans le cerveau. La proposition selon laquelle les neurosciences auraient « résolu » le problème de Jevons -mesurer l'utilité- (Camerer, Loewenstein et Prelec, 2005, p.10) s'appuie en fait sur une confusion entre la notion d'utilité en économie et ce que les neurobiologistes appellent des « processus d'évaluation des récompenses » ou de « processus d'évaluation dans la prise de décision » (*value based decision making*)[27].

26 C'est souvent en ce sens qu'ont été compris les travaux de Glimcher par les économistes. Camerer a par exemple écrit :« *l'ironie de la neuroéconomie est que les neuroscientifiques trouvent souvent les principes les plus élémentaires de la rationalité [économique] utiles pour expliquer le choix des êtres humains. Glimcher (2003) montre avec élégance comment le modèle simple de l'utilité espérée clarifie le type d'encodage réalisé par les neurones pariétaux* » (Camerer, 2005, p.449).

27 Pour Andrew Landreth et John Bickle, les neurosciences « *fournissent de nouvelles informations sur la structure des processus d'évaluation* » (Landreth et Bickle, 2008, p.425). Ces processus sont assimilés, à tort, par Camerer, Loewensetin et Prelec (2005) avec la notion d'utilité en économie. Cette confusion est très courante en neurobiologie, et en neuroéconomie, car pour les neurobiologistes, l'évaluation dans la prise de décision renvoie, dans la nature, à tous les mécanismes d'évaluation de n'importe quel type de « récompense » (*reward*), celle-ci pouvant prendre les formes les plus diverses. Par conséquent, ce concept d'évaluation en biologie est censé s'appliquer également à l'homme, et, par suite à l'économie: « *l'évaluation dans la prise de décision est omniprésente dans la nature. Elle se produit à chaque fois qu'un animal fait un choix entre plusieurs alternatives à partir d'une valeur subjective qu'il a attribué à chacune d'entre elles. Des illustrations de ces processus peuvent être trouvées dans des comportements animaux élémentaires, comme le butinage des abeilles, mais aussi dans des décisions complexes prises par l'être humain, comme l'échange sur un marché financier. La neuroéconomie […] étudie ces processus cérébraux nécessaires pour adopter des évaluations dans la prise de décision […] Elle cherche à établir une théorie expliquant la manière avec laquelle les êtres humains prennent des décisions, qui peut s'appliquer à la fois aux sciences naturelles et sociales* » (Rangel, Camerer, et Montague, 2008, p.545). De nombreux auteurs, comme Camerer, Loewenstein et Prelec (2005, *cf.supra*) ont donc soutenu, en particulier aux débuts de la neuroéconomie, que les processus d'évaluation étudiés par les neurobiologistes pouvaient servir de « substitut » (*proxy*) pour l'utilité ordinale que les économistes ne peuvent mesurer, et qu'ils sont contraints de déduire des choix observés. Or, d'une part, ces processus d'évaluation de distinguent de l'utilité car ils ne portent pas sur une quantité de plaisir ou de satisfaction effectivement ressentie, mais sur leur anticipation au moment du choix. L'évaluation de la récompense remplit toujours, en biologie, une fonction de prédiction. Des études de neuroimagerie indiquent précisément que ces fonctions d'estimation sont localisées dans différentes zones du cerveau et suivent des dynamiques indépendantes de celles liées au plaisir (*cf.* Berns *et al.*, 2001). Paul Glimcher a ainsi proposé la notion d'« utilité espérée physiologique » (*physiological expected utility*) pour désigner les signaux d'activité neuronales observés dans ces processus d'évaluations, qui se produisent au moment du choix. Cette notion, quoique plus précise, reste trompeuse, parce que le système nerveux ne procède pas, comme en économie, par une estimation statique, à partir de probabilités et de préférences données. Ces mécanismes de prédiction de la récompense s'appuient en effet sur des processus d'essais et d'erreurs. Ces dynamiques d'apprentissage portent à la fois sur la valeur et sur le risque et l'incertitude, qui constituent deux dimensions non-séparables du choix. Par conséquent, ce que Glimcher appelle « utilité espérée physiologique » constitue un concept distinct de l'utilité espérée des économistes.

Il est vrai que les psychologues et neurobiologistes ont souvent, comme Glimcher, défendu une influence de la théorie économique sur leurs travaux (Glimcher, 2003; Dorris, Glimcher et Bayer, 2005). Or cette référence à l'économie renvoie simplement à l'utilisation de processus d'optimisation et de maximisation sous contrainte, qui peuvent effectivement mobiliser un vocabulaire économique. Mais la formalisation de ces processus a bien été réalisée par les psychologues et les neurobiologistes, et non pas par des économistes. L'inspiration économique revendiquée du néo-comportementalisme doit donc être comprise comme un effort pour formaliser et quantifier davantage les résultats des neurosciences, sur le modèle de l'économie, mais non pas nécessairement en adoptant des modèles de l'économie théorique. Cette orientation « économique » correspond selon cette interprétation simplement à approfondissement des méthodes quantitatives en neurosciences. Tout laisse à penser que cette évolution n'a au fond rien à voir avec l'économie : pour Steven Quartz par exemple, la neuroéconomie représente le stade supérieur des neurosciences cognitives, qui seraient elles-mêmes nées d'un dépassement des sciences cognitive (Quartz, 2008).

Si ce nouveau domaine de recherches expérimentales constitue finalement une branche de l'analyse économique, c'est parce qu'il a apporté, ultérieurement, des solutions théoriques à un problème soulevé par l'économie comportementale, à propos du paternalisme libertarien et des comportements pathologiques, à partir notamment d'une approche originale des conduites impulsives et addictives. L'intérêt théorique ayant donné naissance à la neuroéconomie est bien venu des économistes, mais ceux-ci se sont largement contentés d'importer des concepts plutôt que d'exporter les leurs.

La neuroéconomie apparaît au départ comme une simple application, à l'économie, de plusieurs techniques expérimentales fournies par les neurosciences contemporaines (neuroimagerie, SMTr). Cette innovation technologique n'a cependant été possible qu'à la faveur de deux transformations théoriques majeures: transformation de l'économie en discipline expérimentale avec l'apparition de l'économie comportementale; transformation de l'économie comportementale elle-même, avec la montée en puissance d'un questionnement normatif, étranger au cadre kahnemanien de référence, lié au problème de la pathologie.

Une histoire critique de la discipline, au sens de Canguilhem, ne vise pas à effectuer une critique épistémologique, c'est-à-dire à examiner la recevabilité scientifique des explications proposées par les neuroéconomistes. La tâche critique consiste à expliquer la

naissance d'un débat concernant les rapports de l'économie aux neurosciences, et non de prendre part à ce débat. La très grande réserve et le fort scepticisme qui apparaît dans la « méta-réflexion » des économistes sur la neuroéconomie ne sauraient dissimuler ce fait théorique indubitable : au début des années 2000, la signification, pour l'analyse économique, des résultats expérimentaux fournis par les neurosciences devient un enjeu central en économie. Les défenseurs les plus ardents de la neuroéconomie, qui sont persuadés du potentiel théorique des techniques d'imagerie en économie, comme leurs adversaires les plus véhéments représentent selon nous les deux faces d'une même transformation, récente, de la théorie économique contemporaine.

Selon notre hypothèse, cette transformation est liée à l'apparition, au début des années 2000, d'une réflexion théorique en économie sur le traitement des conduites pathologiques. La neuroéconomie ne débute donc selon nous véritablement qu'à partir des années 2002-2003 (*cf. supra*), mais elle dispose néanmoins d'antécédents théoriques. Dès les années 1960, des chercheurs appartenant au courant que nous avons appelé néo-comportementalisme essayent de modéliser certains troubles du comportement dans des termes inspirés de l'analyse économique, jetant ainsi les prémisses théoriques de la neuroéconomie.

Une étude historique de la neuroéconomie appelle d'abord des précisions méthodologiques. L'exploration du passé théorique de la discipline doit être menée à partir d'une certaine conception de la science et de son progrès. Le premier chapitre explicite et justifie le choix d'une perspective historiographique inspirée par Georges Canguilhem (chapitre liminaire. Questions de méthode : neuroéconomie et idéologie scientifique). Celle-ci conduit à accorder une stricte autonomie de la théorie à l'égard de son contexte d'apparition et de diffusion. Cela apparaît d'autant plus justifié à propos de la neuroéconomie que l'environnement académique et social ne joue qu'un rôle tout à fait secondaire dans l'émergence de la discipline. De notre point de vue, la neuroéconomie est née d'un projet purement théorique, associant à l'économie une réflexion sur la maladie mentale. Notre interprétation de cette psychiatrie économique à partir du concept d'« idéologie scientifique » empruntée à Canguilhem (1977) se distingue ainsi de la sociologie des sciences.

Le récit historique proprement dit est scindé en trois parties. La première est consacrée aux travaux néo-comportementalistes antérieurs aux années 2000, qui forment la préhistoire théorique de la neuroéconomie. La science quantitative de la motivation débute au début des années 1960 par une première série d'expériences sur le pigeon (chapitre 2. Impulsivité et choix inter-temporel : la naissance d'une idéologie scientifique en sciences comportementales). Ces travaux visent à établir, chez l'animal, des relations quantitatives

entre le comportement conditionné et les récompenses attendues. La loi dite d'égalisation des rendements, proposée par Richard Herrnstein (1961), fournit un principe général du comportement, permettant d'expliquer des troubles et dérèglements de la motivation. Cette approche quantitative débouche par ailleurs sur un rapprochement avec l'économie, via une référence à la notion de maximisation. La psychiatrie économique trouve ainsi son origine dans un questionnement théorique, apparu dans les années 1960 chez les psychologues, associant pathologies mentales d'un côté, rationalité et théorie de la décision de l'autre. Les années 1990 voient l'incorporation de techniques neurophysiologiques à ce courant de recherches. Chez le singe, les micro-électrodes sont utilisées pour étudier les dynamiques d'apprentissage de la récompense (chapitre 3. Les années 1990 : La vocation économique de la neurobiologie – Paul Glimcher et l'« utilité espérée physiologique »). Cette innovation technologique permet ainsi d'approfondir à l'échelle neuronale les résultats acquis sur le pigeon à l'échelle comportementale.

La deuxième partie est consacrée à la neuroéconomie telle qu'elle a été définie ici, c'est-à-dire aux travaux des années 2000. La neuroéconomie apparaît d'abord soit comme un prolongement neurobiologique de l'économie comportementale, soit comme un prolongement économique des neurosciences (chapitre 4. Une discipline sous la tutelle de l'économie comportementale et des neurosciences, 2000-2005). D'un côté, les économistes comportementalistes se servent au départ de l'IRMf comme d'un instrument permettant de vérifier, au niveau du cerveau, leurs propres explications du comportement. Cette approche s'est largement appuyée sur une représentation dualiste de la cognition, sous la forme d'une opposition entre la raison et les émotions. Les travaux du neurobiologiste Antonio Damasio sur les émotions font ainsi figure de référence incontournable dans le domaine. Ce n'est que dans la deuxième partie des années 2000 que la neuroéconomie parvient à se constituer comme une discipline autonome. Le réductionnisme dont est porteuse l'« économie comportementale dans le scanner » (Ross, 2008) a soulevé d'importantes critiques méthodologiques. Les neuroéconomistes ont alors progressivement pris leurs distances avec les interprétations dualistes inspirées de Kahneman. Dans le domaine du choix inter-temporel, le retour à la notion de monnaie neuronale commune marque ainsi une rupture avec l'économie comportementale et le programme kahnemanien (chapitre 5. De l'économie comportementale dans le scanner à la neuro-psychiatrie computationnelle : la constitution d'une discipline autonome).

La troisième et dernière partie vise à dégager le portée de ce paradigme néo-comportementaliste en économie. La psychiatrie économique fait l'objet de deux

approfondissements théoriques, en dehors du choix inter-temporel dont il aura jusqu'alors été question. La notion d'apprentissage de la récompense ou *reward learning* peut d'abord s'appliquer au choix inter-personnel (chapitre 6. L'analyse du choix inter-personnel : la neuroéconomie entre neurosciences sociales, neuro-éthique et économie comportementale). Comme pour le choix inter-temporel, la neuroéconomie peut se comprendre comme une approche « pathologique » du comportement. La neuroéconomie du choix inter-personnel étend ainsi aux troubles de la cognition sociale les analyses portant sur les addictions. Enfin, la neuroéconomie peut valoir comme une contribution à l'analyse du bien-être (chapitre 7. Paternalisme libertarien et Psychiatrie économique. L'apport de la neuroéconomie à l'analyse comportementale du bien-être). Le traitement des questions normatives fait figure à la fois d'illustration et d'aboutissement théorique à notre étude, puisqu'il met en évidence l'écart théorique entre les propositions normatives de la neuroéconomie et celles de l'économie comportementale.

Chapitre liminaire. Questions de méthode : neuroéconomie et idéologie scientifique

> « *Ceux qui affirment qu'il n'y a point de philosophie sans choix politique, que toute pensée est « progressiste » ou réactionnaire […], leur sottise est de croire que la pensée exprime l'idéologie d'une classe* »
> Michel Foucault, *Les Mots et les Choses*, 1966, p.339

L'histoire de la neuroéconomie peut s'écrire selon des points de vue très différents. Le mode d'analyse retenu est notamment fonction de l'appartenance (ou non) de l'auteur à la discipline. Les intentions sont alors variables, selon qu'elles visent à promouvoir, vulgariser la neuroéconomie, ou bien critiquer sa méthodologie, son influence sur les pratiques sociales, *etc.* Ce chapitre a pour objectif de justifier la mise en perspective adoptée ici, qui s'inspire de l'histoire des sciences pratiquée par Georges Canguilhem.

Il peut apparaître d'abord surprenant d'écrire l'histoire d'un objet si récent. Certains pourraient arguer que ce travail relève plus du journalisme que de l'histoire. Par ailleurs, notre approche emprunte à Georges Canguilhem la notion d'idéologie scientifique. Or, les travaux de cet auteur portent essentiellement sur la biologie et les sciences médicales au XIX^{e} siècle. Dans le livre *Idéologie et Rationalité dans l'histoire des sciences de la vie* (Canguilhem, 1977), l'idéologie scientifique fait référence principalement au darwinisme social de Herbert Spencer. En quoi un concept utilisé par Canguilhem à propos de la pensée biologique du XIX^{e} siècle peut-il être pertinent pour caractériser les développements contemporains de la neuroéconomie ?

Le véritable problème ne réside pas dans l'impossibilité présumée d'écrire l'histoire du temps présent. Nombreux sont les historiens qui, au contraire, revendiquent aujourd'hui avec vigueur de se saisir de l'extrême contemporanéité en écrivant l'« histoire du temps présent »[28].

28 L'expression d'« *histoire du temps présent* »fait son apparition en 1978 avec la création par le CNRS d'un « *Institut d'Histoire du Temps Présent* ». Également connu sous le nom d'« *histoire immédiate* », ce courant historiographique connaît depuis les années 1970 un important renouveau, en revendiquant notamment la spécificité (orale) de ses sources. En dépits de la vogue récente dont elle fait l'objet, l'histoire immédiate n'est cependant pas entièrement nouvelle. Les historiens ont en effet toujours affirmé pouvoir se saisir du temps présent comme d'un objet historique. Cette possibilité est même justifiée méthodologiquement par les fondateurs de l'École des Annales, puisque ceux-ci définissent l'histoire comme une relation du passé au présent. Un numéro de la revue a ainsi été consacré à l'Allemagne nazie en 1937 (*cf.* Dosse, 2010). Pour certains historiens, l'histoire du temps présent n'a donc au fond aucune spécificité en tant que telle. Elle est selon Antoine Prost par exemple une « *histoire comme les autres* », car « *l'histoire du passé proche revient à faire de l'histoire tout court* » (Prost, 2007)

Du point de vue de Canguilhem, c'est au contraire l'éloignement dans le temps qui complique la tâche de l'historien. En effet, l'histoire des sciences est essentiellement compréhension, elle « *exige une installation dans le contenu des énoncés scientifiques* ». La spécificité du point de vue de l'historien ne réside donc pas dans un prétendu recul réflexif mais bien plutôt dans la recherche d'une certaine familiarité acquise par la pratique: cette « *installation ne peut être qu'une pratique* » (Canguilhem, 1977, p.17). Or, cette exigence compréhensive semble d'autant plus difficile qu'elle s'applique à des énoncés théoriques appartenant à des époques plus anciennes. Inversement, la compréhension semble plus aisée *a priori* pour des objets contemporains. Dans cette perspective, le projet d'écrire une histoire de la neuroéconomie est ainsi sans doute plus facile et plus légitime que l'écriture, par exemple, d'une histoire de la pensée économique au XVI^{e} siècle

L'écriture d'une histoire de la neuroéconomie n'est donc pas difficile parce que celle-ci est une science récente, ou contemporaine, mais plutôt parce qu'elle se donne comme une discipline *encore en construction*. En effet, ce qui est attendu d'une étude prenant pour objet la neuroéconomie est une évaluation de son contenu scientifique. Cela correspond notamment aux objectifs que se donnent les analyses réalisées à partir d'un point de vue interne à la discipline (*cf.* Introduction). La neuroéconomie est comprise comme une discipline jeune, au statut scientifique encore relativement indéterminé. Il s'agit alors soit de critiquer, soit de promouvoir, l'approche neuroscientifique en économie. L'observateur veut donc se faire juge, et séparer, au sein de ce champ de recherche émergent, ce qui ressort du vrai, du faux, ou du simplement probable[29].

Ici, la scientificité de la neuroéconomie ne sera pas discutée ni remise en question. Les multiples manifestations de son existence académique (publications d'articles, de livres, créations de *cursus* et de laboratoires spécifiques dans les universités, organisation de colloques, de conférences, *etc.*) sont autant de *faits* qui n'ont nullement besoin d'être prouvés, et qui confirment son appartenance légitime à la science moderne. Notre questionnement se veut strictement historique et descriptif. Il s'agit de rendre compte de l'émergence d'un nouveau domaine de savoir, et non d'évaluer son contenu théorique.

Mais cette approche historique soulève à son tour une difficulté, en risquant de réduire

29 La catégorie du probable est en fait l'une des plus importantes : souvent, les études sur la neuroéconomie aboutissent à des conclusions mitigées et consensuelles, entre rejet radical et promotion, dans lesquelles on considère que les modèles de la neuroéconomie ont encore un statut incertain, mais qu'ils ont indiscutablement un « *potentiel* » (voir par exemple, Mäki, 2010). C'est le cas par exemple de Emrah Aydinonat qui affirme que la neuroéconomie est « *plus qu'une inspiration, moins qu'une révolution* » (Aydinonat, 2010).

la science à des facteurs académiques ou sociaux purement contingents. Or l'histoire des sciences de Georges Canguilhem maintient une stricte autonomie du savoir scientifique à l'égard de son contexte d'apparition et de diffusion. Cette perspective nous semble particulièrement pertinente pour la neuroéconomie car l'examen des éléments non-théoriques ne suffit pas, on va le voir, à expliquer l'apparition de la discipline. Notre enquête historique vise donc à mettre en évidence une causalité purement théorique, et non contextuelle ou sociale.

Pour autant, l'histoire des sciences ne saurait se limiter à restituer le progrès des connaissances. Comme le souligne Paul Veyne, « *l'histoire [...] demeure fondamentalement un récit et ce qu'on nomme explication n'est guère que la manière qu'a le récit de s'organiser en une intrigue compréhensible* » (Paul Veyne, 1970, p.111). L'historien des sciences doit donc « raconter » ou mettre en scène les théories qu'il étudie. Le recul qu'il adopte par rapport à la science fait toujours dépendre son récit théorique de son « style » de narration[30] .

L'apparition d'un nouveau domaine de savoir comme la neuroéconomie pose ainsi à l'historien la question de savoir comment le progrès, dans l'histoire des sciences, doit être conçu ou « raconté ». La neuroéconomie, telle qu'elle apparaît et se développe au début des années 2000, dispose on l'a vu d'antécédents théoriques (*cf.* Introduction). Décrire l'émergence de la discipline implique donc de mettre en rapport celle-ci avec sa préhistoire théorique. Toute la difficulté consiste alors à comprendre comment ces deux éléments hétérogènes peuvent entrer en liaison, sans se confondre: comment des travaux théoriques réalisés dans les années 1960 peuvent-ils à la fois préfigurer et se distinguer de recherches entreprises quarante ans plus tard?

Le problème peut se comprendre comme le choix d'un mode de narration, et d'un recul interprétatif. La notion d'« idéologie scientifique » avancée par Canguilhem est particulièrement adéquate, de notre point de vue pour « raconter » l'histoire de la neuroéconomie, en rompant avec le progrès linéaire des connaissances supposé par les analyses internes à la discipline (*cf.* introduction), tout en se limitant à des éléments strictement théoriques. L'approche de Canguilhem conduit en effet à adopter une conception large -mais non-relative- de la scientificité, qui inclut la possibilité de l'erreur, de

30 Plus généralement, de nombreux courants de l'historiographie contemporaine défendent ainsi une proximité assumée avec la littérature et soulignent l'importance du style (voir par exemple Prost, 1996). Pourtant, si cette mise en intrigue engage même, selon White, un « *acte poétique* » (White, 1973), le style de l'historien ne repose nullement sur l'imagination. L'histoire n'est pas fiction, et si elle peut être comprise comme un roman, encore faut-il rappeler, comme le fait Paul Veyne, qu'il s'agit d'un « *roman vrai* ». Par delà ses inévitables lacunes, le récit doit en effet répondre aux exigences d'une question initiale posée à l'historien (Paul Veyne, 1970).

l'approximation, et surtout d'un certain excès de prétention du scientifique (Canguilhem, 1977).

Si la perspective interne se traduit par une certaine cécité historique, le point de vue externe, pouvant être adopté par des historiens ou des sociologues des sciences engage une attitude relativiste. Celle-ci, dans sa version faible ou forte, soulève le problème de savoir ce qui distingue les théories économiques et/ou neuroscientifiques de simples croyances collectives. Surtout, le contexte non-théorique semble jouer un rôle mineur dans l'apparition et la diffusion de la neuroéconomie. Au final, une histoire non-théorique ou une sociologie de la neuroéconomie peine à expliquer les raisons de son développement théorique (I. Le point de vue externe : la neuroéconomie et son contexte). La notion d'« idéologie scientifique » proposée par Canguilhem vise à résoudre ces difficultés, et il faudra montrer en quoi la perspective du livre *Idéologie et Rationalité dans l'histoire des sciences de la vie* (Canguilhem, 1977) peut se révéler pertinente à propos de la neuroéconomie (II. La notion d'idéologie scientifique chez Georges Canguilhem : une source d'inspiration pour l'historien de la pensée économique).

I. Le point de vue externe : la neuroéconomie et son contexte

L'adoption d'un point de vue interne à la neuroéconomie conduit soit à justifier, soit à critiquer les principes méthodologiques sous-jacents au rapprochement entre économie et neurosciences (*cf.* introduction). Dans une perspective externe à la discipline, il est possible en revanche non pas d'interroger la validité de l'approche défendue par les neuroéconomistes, mais d'expliquer l'origine et les motivations d'une telle orientation méthodologique. La question ne concerne plus la validité scientifique de la neuroéconomie : il s'agit d'expliquer son émergence, donc de la rapprocher de son contexte d'apparition et de déploiement.

Cette démarche qualifiée usuellement d'« externaliste » en histoire de la pensée tend ainsi à contourner le problème de la scientificité, en montrant que la validité des théories économiques est relative à une époque, un environnement intellectuel particulier, à des données psycho-sociales, *etc.* Ce relativisme peut s'exprimer de façon faible ou forte, selon que l'on considère une relation de dépendance au contexte soit pour une partie (A), soit pour la totalité du discours théorique (B). Pourtant, dans les deux cas, l'historien ou le sociologue externaliste est reconduit à la visée évaluatrice caractéristique des analyses « internalistes ». En effet, dans la version faible, il faut finalement admettre l'existence d'un noyau pur de « bonne théorie » derrière les ruses de la rhétorique. Dans la version forte du relativisme, le contenu de la neuroéconomie est dépouillé de toute prétention scientifique, car les théories sont réduites à des croyances psycho-sociales justifiées par des besoins pragmatiques et contingents. Dans les deux cas, l'historien exprime bien toujours, qu'il le veuille ou non, un jugement de valeur quant à la validité scientifique des théories de la neuroéconomie.

A. Une perspective intermédiaire: la rhétorique de la neuroéconomie

Une perspective intermédiaire entre internalisme et externalisme consiste à envisager un lien de dépendance faible entre les théories économiques et leur contexte : certains éléments du discours théorique, et non la totalité de celui-ci, se voient attribuer une valeur relative. Cette version modérée du relativisme s'exprime notamment dans les travaux célèbres

de Deirdre McCloskey (1985) sur la rhétorique de l'économie. Une telle perspective peut être appliquée à l'étude de la neuroéconomie (*cf.* Mäki, 2010). Les emprunts de modèles à la neurobiologie apparaissent comme autant de métaphores, qui enrichissent l'« art de la persuasion » des économistes (A). Ce type d'études sémantiques, à partir de la notion de rhétorique empruntée à McCloskey, invite cependant implicitement à évaluer le contenu du savoir scientifique, car la « *bonne science* » y est caractérisée, pour reprendre les termes de McCloskey, comme une « *bonne conversation* » (McCloskey, 1985, p.27)

1. L'art de la persuasion des neuroéconomistes : les modèles de la neurobiologie comme métaphores élégantes

Le livre de McCloskey, *The Rhetorics of Economics*, publié en 1985, a entraîné un renouvellement important des méthodes en histoire de la pensée économique. Le rôle primordial, souligné par McCloskey, des dispositifs rhétoriques dans les théories économiques a en effet conduit l'historiographie contemporaine à s'intéresser davantage au style du discours tenu par les économistes, et, plus généralement, aux effets pratiques recherchés par ces derniers. La perspective de McCloskey a été appliquée récemment à la neuroéconomie par Mäki (2010). Les arguments utilisés par les neuroéconomistes, et par leurs adversaires, apparaissent ainsi comme un ensemble de figures de style utilisées pour convaincre dans l'affrontement théorique. La neurobiologie est dès lors caractérisée comme un réservoir d'outils de persuasion rhétorique disponibles pour l'analyse économique.

Pour McCloskey, « *la rhétorique est l'art de parler. Plus généralement, elle est l'étude de la manière avec laquelle les gens persuadent* » (McCloskey, 1985, p.133). Déterminer la nouveauté de la neuroéconomie reviendrait donc à identifier les modifications non pas du contenu mais du « *style* » des théories économiques par la neurobiologie: « *les changements ne sont pas principalement des changements de modèles explicatifs, mais des changements dans les manières de parler* » (*ways of talking*). Or, cet art de la persuasion dépend non pas du sujet de la discussion, mais de l'audience à laquelle s'adresse l'orateur (McCloskey, 1985, p.133-135).

Effectivement, la neuroéconomie s'appuie sur un dispositif rhétorique d'autant plus élaboré que l'audience à laquelle s'adressent les neuroéconomistes est hétérogène. Ces derniers doivent non seulement convaincre les neurobiologistes du bien-fondé de leur

approche, mais aussi les économistes appartenant à des courants de recherches assez divers (théorie de la décision, théorie des jeux, économie comportementale, expérimentale, *etc.*), les philosophes, des spécialistes de méthodologie économique, *etc.* La diversité de l'audience explique ainsi, selon Mäki, la grande richesse rhétorique de la neuroéconomie. Cet auteur a tenté d'identifier dans un article récent (Mäki, 2010) les différents procédés rhétoriques sur lesquels s'appuient les neuroéconomistes pour convaincre.

Selon Mäki, la neuroéconomie s'appuie d'abord sur la « *rhétorique de la scientificité* », qui invoque l'autorité et le prestige des sciences dites naturelles (biologie, neurosciences). La rhétorique de la « *profondeur* » et de la « *micro-fondation* » affirme que la recherche de causes « *plus profondes* » permet de révéler des mécanismes cachés. La rhétorique de la « *bonne nouvelle et du progrès* » suggère que les économistes devraient tirer parti de manière opportune des avancées récentes des neurosciences. Cela s'accompagne de plaidoiries en faveur d'un plus grand degré de réalisme des théories économiques. La neuroéconomie rejoint ici un procédé rhétorique courant en économie comportementale, en se proposant de donner des fondations psychologiques plus réalistes aux modèles économiques. La vertu de la neuroéconomie résiderait en outre dans son caractère interdisciplinaire : elle poursuivrait ainsi le processus d'unification des sciences particulières. Enfin, et de manière décisive, l'audience de la neuroéconomie inclut également un public de décideurs politiques et de néophytes. Par conséquent, les neuroéconomistes développent souvent une rhétorique de la « *pop science* », c'est-à-dire de la science qui a vocation à intéresser le grand public. D'une part, tout le monde a un intérêt à comprendre le fonctionnement du cerveau dans la vie quotidienne. D'autre part, la possibilité d'utiliser la neuroéconomie pour contrôler et manipuler les comportements peut apparaître séduisante pour des décideurs (Mäki, 2010, p.110-111).

A ces procédés « *positifs* » visant à promouvoir directement l'usage des neurosciences s'ajoutent des éléments de « *prudence* » destinés à assurer néanmoins un terrain d'entente avec l'économie. La rhétorique du progrès est par exemple nuancée par une rhétorique du compromis : alors que la nécessité d'une révision théorique est admise, l'abandon pur et simple du modèle néoclassique n'est pas pour autant requis. Les neuroéconomistes affirment par exemple que l'économie a jusqu'alors réalisé d'importants progrès, qui doivent être approfondis. La rhétorique de la modestie est une autre stratégie de compromis, qui vise à éviter des réactions trop hostiles chez les économistes : les connaissances de la neurobiologie pourraient irriguer, ou enrichir, (*inform*) la théorie économique, sans nécessairement remettre celle-ci radicalement en question. La rhétorique de la nouveauté et de l'« *excitation* » promet de rendre l'analyse économique plus attrayante et plus séduisante. Les images du cerveau

obtenues par la neuroimagerie jouent encore une fois un rôle de premier ordre.

Ce travail d'analyse rhétorique de la neuroéconomie soulève plusieurs remarques. Tout d'abord, les procédés rhétoriques identifiés par l'auteur ressemblent davantage à des arguments d'autorité qu'à des figures de style à proprement parler. Mäki considère néanmoins que les images du cerveau « *habillent* » l'argumentation d'une manière élégante, ce qui peut suggérer un enrichissement métaphorique du discours économique, mais cette idée n'est pas approfondie[31]. D'une manière générale, l'étude faite ici de la rhétorique n'est pas spécifiquement langagière. Elle porte plutôt sur des raccourcis théoriques et des schémas d'explication insuffisamment justifiés. Le terme de rhétorique est à comprendre dans un sens très large, qui inclut les artifices les plus divers utilisés pour convaincre. Par ailleurs, comme le souligne Mäki, les adversaires de la neuroéconomie eux-mêmes sont contraints de recourir à une rhétorique, car « *la critique libre de toute rhétorique n'est pas possible. En critiquant une certaine rhétorique, l'on est obligé d'exercer soi-même sa propre rhétorique: rhétorique 2 à propos de la rhétorique 1* » (Mäki, 2010, p.114)

En outre, il y a une incertitude quant au fait de savoir si cette rhétorique est l'effet nécessaire de l'utilisation des neurosciences, ou si ces dernières sont contraintes d'utiliser de tels procédés pour convaincre de leur intérêt pour l'économie. Dans le premier cas, le dispositif rhétorique s'explique par les neurosciences et leur est inséparable; dans le second cas, il serait possible d'envisager, dans l'absolu, une neuroéconomie débarrassée de ses artifices rhétoriques. Mäki penche en faveur de la deuxième solution. Il est en cela, on va le voir, fidèle à McCloskey. En effet, selon Mäki, le recours à de tels procédés est rendu nécessaire pour promouvoir la neuroéconomie dans le monde académique, pour convaincre et trouver des ressources: « *une grande part de cette rhétorique peut être excusée simplement comme dispositif d'une campagne, à une échelle plus large, pour créer un espace social et épistémique pour cette nouvelle initiative* » (Mäki, 2010, p.113). La neuroéconomie ne serait pas complètement réductible à un art de la persuasion, et pourrait prétendre, par delà ces procédés oratoires, à des fondements objectifs.

31 Il y aurait pourtant incontestablement ici matière à réaliser une analyse proprement (et véritablement) rhétorique, c'est à dire portant sur les procédés de langage, car, comme le souligne Pierre-Henri Castel, la neuroimagerie fonctionnelle introduit bien « *une nouvelle façon de parler* » (Castel, 2009, p.15). Ingvar (1977) a réalisé une étude intéressante quoique relativement ancienne sur la question. A propos de la corrélation des images du cerveau avec des états mentaux comme « nouveau langage », cet auteur propose la notion d' « *idéogramme cérébral* » (Ingvar, 1977).

2. Une visée prescriptive: « la bonne science est une bonne conversation »

McCloskey ne nie nullement la possibilité d'élaborer une théorie économique « *scientifique* ». Le recours à des procédés rhétoriques par les divers courants de pensée n'empêche pas, sur le long terme, l'élaboration de compromis objectifs entre les parties, dès lors que certains « *standards rhétoriques* » sont respectés (McCloskey, 1985, p.141). En effet pour McCloskey, « *la bonne science est une bonne conversation* » (McCloskey, 1985, p.27): la non-violation des règles de la discussion scientifique est garante de la scientificité du discours. Les analyses de McCloskey conservent dès lors une visée prescriptive ou évaluatrice, car il s'agit toujours de restaurer une certaine vérité, qui peut ou qui a pu être déformée par un usage illégitime de la rhétorique.

En appliquant à la neuroéconomie la perspective de McCloskey, Mäki maintient également la possibilité d'établir objectivement la validité des théories, car, on l'a vu, à court terme, les excès rhétoriques sont excusés comme « *dispositif d'une campagne, à une échelle plus large, pour créer un espace social et épistémique pour cette nouvelle initiative* ». Or, sur le long terme, le potentiel scientifique « réel » de la neurobiologie sera révélé, par delà la rhétorique des débuts: « *naturellement, si cette campagne n'a que peu de substance en dehors de cette rhétorique, alors cette rhétorique est difficile à justifier. Dans un temps suffisamment long, la rhétorique doit être alignée avec la performance substantielle réelle* [...] *Même un écart très important entre la rhétorique et la performance réelle n'est pas nécessairement fatal aux premiers stades de la recherche. Ceci semble aujourd'hui compris par les neuroéconomistes eux même* » (Mäki, 2010, p.113).

Ce type d'analyse se comprend donc, dans une tradition au fond très positiviste, comme une entreprise d'analyse et de restauration du sens des discours théoriques. La mise en évidence d'artifices rhétoriques permet de faire apparaître le résidu objectif latent des arguments utilisés, et leur validité. Mäki se montre ainsi confiant, sans être exagérément optimiste, en ce qui concerne le potentiel théorique de la neuroéconomie: « *l'utilisation de multiples sources de preuves et leur triangulation sont des manières d'améliorer la qualité des données ainsi que la qualité de l'évaluation des modèles. On peut au moins l'espérer* ». Par conséquent, il n'y a pas de raison de rejeter en tant que telles les données neuronales: « *la question de savoir si les données neuronales sont inférieures aux tests économétriques réalisés avec des données de marché [...] ne devraient donc pas se poser, ni jouer un rôle important* » (Mäki, 2010, p.116).

La réflexion sur les procédés rhétoriques doit donc bien conduire à la formulation d'un jugement de valeur sur la validité, ou au moins sur le « potentiel » de courants de recherche. Pour McCloskey, le philosophe est ainsi « *guide de la science* » (McCloskey, 1985, p.183). La cible réelle des critiques de McCloskey concerne moins au fond la rhétorique elle-même, que les économistes qui ne se préoccupent pas d'analyser leur propre discours. McCloskey fustige notamment l'économie dite « *moderniste* », incarnée par Milton Friedman et sa méthodologie du « comme si », qui manifeste « *cynisme et mépris à l'égard des normes de la conversation académique* » (McCloskey, 1985, p.25). La philosophie a des vertus positives pour la théorie que, malheureusement, la plupart des économistes ne soupçonnent pas. Cette invitation à l'auto-critique n'est pas sans rappeler la psychanalyse : « *une critique rhétorique, comme un cours de psychanalyse pourrait rendre les économistes plus conscients d'eux-mêmes, modestes et tolérants, meilleurs sur le plan personnel et professionnel* » (McCloskey, 1985, p.175).

L'historien, ou le philosophe critique, affiche donc l'ambition d'apporter une « valeur ajoutée à la théorie ». McCloskey et Mäki adoptent un point de vue quasi-interne à la discipline. Il s'agit toujours d'améliorer la théorie, et non d'expliquer son apparition ou son développement. L'histoire a un intérêt théorique en tant qu'elle met à jour certains effets de distorsions de la théorie liés au contexte. Ce relativisme moyen, qui comprend l'histoire ou la méthodologie comme une entreprise de restauration du sens des énoncés, n'est pas exclusivement attaché aux travaux de McCloskey, mais représente une posture mitigée et consensuelle adoptée assez largement par les historiens de la pensée économique. Par exemple, dans un article récent, Robert Dimand attribue à l'histoire des théories une fonction positive pour l'analyse économique elle-même. Il constate que l'absence de réflexion historique des économistes conduit ceux-ci à construire des visions exagérément critiques ou élogieuses des différents courants de pensée. Cette tendance à créer des « *gentils et des méchants* » justifie selon l'auteur le travail de l'historien, qui, en établissant la valeur *véritable* des théories, produit un bénéfice *théorique*: « *une telle célébration et dénonciation déforment la compréhension et font dévier l'orientation des recherches [...] Elles contribuent pourtant aussi à faire avancer la connaissance, en stimulant une investigation, une réflexion et un débat critique qui n'auraient pas eu lieu* [en l'absence de ces distorsions]. *Ce processus à la fois aide et complique l'histoire intellectuelle, en créant des difficultés pour une recherche historique qui n'aurait cependant pas pu être entreprise, soutenue ou remarquée d'une autre manière* » (Dimand, 2007, p.77).

Dimand remarque ainsi justement que, si l'on considère que la théorie économique est

déformée par les économistes eux-mêmes pour les besoins de la persuasion, cette déformation a au bout du compte un effet positif, car elle rend possible un travail historique de restauration du sens. La réflexion historique et méthodologique est fondamentalement légitimée par le besoin de clarification de la théorie. L'histoire de la pensée, comme toute activité qui « *continue à être poursuivie », « doit avoir certains bénéfices, et non seulement des coûts* » (Dimand, 2007, p.93). C'est à cette condition qu'elle peut présenter un intérêt pour l'économiste. Le spécialiste de méthodologie, d'histoire ou de « *rhétorique* » conserve une responsabilité vis à vis de la théorie. Il a peut être même, selon McCloskey, la plus grande responsabilité, en tant que « guide » de la recherche. Une attitude plus radicale consiste à prendre un recul plus tranché, pour réaliser un pur travail d'historien, détaché de toute obligation à l'égard de l'analyse économique.

B. Les interprétations psychosociales

La version forte du relativiste rejette la possibilité d'établir, ni même de restaurer, une valeur objective à la théorie indépendamment de son contexte. Elle s'exprime notamment dans le « programme fort » de la sociologie des sciences anglo-saxonne des années 1970 (*cf.* Bloor, 1976), dans les travaux de Bruno Latour[32] ou à travers certaines interprétations de Thomas Kuhn. L'analyse anthropologique de la vie de laboratoire réalisée par Latour, Woolgar et Biezunski dans un institut de neuroendocrinologie (Latour, Woolgar et Biezunski, 1979) suggère un rapprochement possible entre sociologie des sciences et neuroéconomie.

Aucune étude sociologique de ce type n'a été réalisée jusqu'à présent à propos de la neuroéconomie. Néanmoins, la neuroéconomie est souvent suspectée de n'être qu'un avatar de la forte vogue dont jouissent actuellement les neurosciences (Rubinstein, 2008 ; Harrison 2008). Ses critiques les plus fréquentes sont ainsi liées le plus souvent à des considérations concernant la sociologie des neurosciences. Par ailleurs, les travaux de McCloskey sur la rhétorique (McCloskey, 1985, 1994-a et 1994-b) ont incontestablement permis d'ouvrir de nouvelles perspectives historiographiques, soulignant davantage l'influence du contexte sur les théories économiques. En particulier, les historiens se sont beaucoup intéressés au cours des dernières années aux biographies des auteurs et au rôle joué par les communautés intellectuelles. Il serait possible de réaliser une analyse de la neuroéconomie selon une

32 *cf.* Latour, 1989

perspective similaire, en s'intéressant notamment à la personnalité de Vernon Smith. Il s'agira donc ici non pas de passer en revue, comme cela a été fait précédemment, des études réalisées sur la neuroéconomie, mais d'envisager, à partir de ce qui a été fait ailleurs, à propos d'autres objets, des modes d'analyse possibles, centrés sur le parcours personnel et collectif des auteurs (A). La sociologie des sciences est également susceptible de fournir un éclairage plus large du contexte, en s'intéressant aux débouchés et aux applications pratiques recherchés par les scientifiques. Elle offre ainsi la représentation d'une science « intéressée », dont le contenu théorique est dépendant de demandes, de contraintes, et d'impératifs sociaux (B).

1. Biographies et communautés intellectuelles: la vie des auteurs comme matière pour l'histoire de la pensée

Une tendance récente de l'historiographie a remis au goût du jour les études biographiques. L'émergence de la neuroéconomie pourrait ainsi être racontée à travers le prisme de personnalités comme Paul Glimcher et/ou de Vernon Smith (a). Cette perspective centrée sur la vie des auteurs soulève néanmoins de nombreuses difficultés d'un point de vue historique, qu'illustre assez bien le recours fréquent à la notion de « vision du monde » (*worldview*) (b).

a. Vernon Smith et les neurosciences, Paul Glimcher et l'économie: le parcours intellectuel de la neuroéconomie

La vie des économistes semble intéresser de manière croissante les historiens de la pensée économique. En 2007, la revue *History of Political Economy* consacrait ainsi un numéro spécial aux biographies et en 2010 un numéro sur les « communautés intellectuelles ». Le parcours intellectuel des auteurs fournit en effet un schéma d'explication commode pour décrire la formation des théories. Pour Malachi Hacohen, les données biographiques sont indispensables. En l'absence de celles-ci, l'histoire de la pensée n'est qu'une abstraction, et sa construction risque de n'être qu'une re-construction artificielle, éloignée des enjeux théoriques réels : « *la situation logique de la vie d'un individu, dont la recherche caractérise toute bonne biographie, autorise l'historien à évaluer*

l'accomplissement et la cohérence théorique de l'individu, ainsi que les limites et les conditions de possibilité de la créativité. En l'absence d'une telle précision, l'histoire des idées risque l'usage factice du contexte et des discours faussement cohérents ; des transgressions que l'histoire des discours a largement commises en critiquant l'usage des biographies » (Hacohen, 2007, p.10).

Plus généralement, la biographie fait l'objet d'un intérêt renouvelé par l'ensemble de l'historiographie contemporaine[33]. Ce type d'analyse semble légitime à propos de la neuroéconomie pour plusieurs raisons. Il existe tout d'abord une tradition assez importante en la matière, chez les historiens de la psychologie. Les ouvrages de James Capshew[34], Ellen Herman[35] et de Nikolas Rose[36] ont notamment mis en évidence le rôle joué par la professionnalisation des psychologues dans la construction de la psychologie comme discipline académique après 1945. La neuroéconomie apparaît comme un prolongement naturel de ce type d'études, à la fois sur le plan chronologique et thématique.

En outre, certains neuroéconomistes jouent un rôle privilégié dans la discipline, et pourraient faire l'objet d'études biographiques pertinentes. Le parcours de Paul Glimcher, par exemple, rend compte des débuts de la neuroéconomie. Ses travaux en neurobiologie invoquent initialement un certain argumentaire à tonalité économique, sans maîtriser complètement le sens des concepts économiques. Les carrières personnelles de George Ainslie, Richard Herrnstein ou de Howard Rachlin sont également particulièrement intéressantes pour comprendre l'apparition d'une biologie et d'une psychiatrie de laboratoire à vocation économique (*cf.* chapitre 2).

C'est surtout la personnalité de Vernon Smith qui invite naturellement à une enquête biographique. Cette figure incontournable de l'économie expérimentale, qui a reçu, en 2002, le Prix de la Banque de Suède en sciences économiques en mémoire d'Alfred Nobel, s'est beaucoup intéressée aux neurosciences. Il a largement contribué à la création d'un centre de recherches spécialisé en neuroéconomie dans son université (le *Centre for the Study of Neuroeconomics* à la *George Mason University*). Le positionnement théorique de Vernon Smith, économiste converti à la neurobiologie, offre donc la possibilité de comprendre l'apparition d'un intérêt théorique des économistes pour les neurosciences. Son cas est d'autant

33 Voir notamment l'article « Biographies et prosopographies » dans lequel François Dosse parle de « *fièvre biographique* » (Dosse, 2010-a).

34 James H. Capshew, 1999. *Psychologists on the March: Science, Practice, and Professional Identity in America, 1929-1969*. Cambridge University Press.

35 Ellen Herman, 1996. *The Romance of American Psychology: Political Culture in the Age of Experts*. University of California Press.

36 Nikolas Rose, 1999. *Governing the Soul: The Shaping of the Private Self*. Free Association Books.

plus intéressant qu'il semble apporter un appui fort à l'argument proposé ici selon lequel la neuroéconomie doit être comprise comme un paradigme « pathologique » en théorie de la décision[37]. Vernon Smith est en effet atteint du syndrome d'Arsperger[38], ce qui l'a conduit à s'intéresser de près à l'autisme, et, plus généralement aux troubles de la cognition sociale. Ici, la situation logique de l'individu apparaît comme la plus pure expression d'un courant de recherches dont l'apport principal repose sur la notion de pathologie économique.

b. Les limites d'une lecture biographique: la notion contestable de « vision du monde »

Il convient cependant d'apporter une nuance forte à l'idée selon laquelle le parcours personnel pourrait servir de révélateur, ou d'instrument explicatif de la théorie pour l'historien. Suite aux travaux de Michel Foucault et de Roland Barthes notamment, il paraît en effet aujourd'hui difficilement tenable d'analyser l'œuvre comme miroir de la vie de son auteur [3940]. Ces critiques ont été prises en compte par les historiens de la pensée économique (*cf.* Forget et Goodwin, 2011). La remise au goût du jour récente des études biographiques en économie s'appuie en fait sur une causalité qui se veut plus élaborée, non-déterministe, entre le texte et la vie individuelle de son auteur. Forget et Weintraub veulent par exemple « *problématiser le concept d'auteur et reconnaître qu'aucun texte et aucune idée n'a émergé d'une conscience unique* » (Weintraub et Forget, 2007, p.21). Comme le souligne François Dosse, le renouveau de la biographie n'a été possible qu'en modifiant les règles de l'ancienne prosopographie : « **«** *la figure du biographié n'est plus envisagée à partir d'une totalité uniforme postulée mais tout au contraire interrogée dans ses tensions, contradictions, ses cités diverses*

37 *cf.* Introduction. Pour une présentation synthétique de cet argument, voir Vallois, 2011.

38 Vernon Smith est passé à la télévision, dans un reportage sur l'autisme, dans lequel il décrit sa maladie. La vidéo est disponible à l'adresse suivante : http://www.youtube.com/watch?v=w5bYbpdMy2c

39 Voir notamment Michel Foucault, 1969; Roland Barthes, 1968.

40 Comme le souligne Hacohen, le retour en force récent des études biographiques en histoire des idées semble étonnant : « *la biographie est de retour. Ceux qui ont vécu les transformations de l'histoire, de la sociologie des sciences, et de la critique littéraire dans les années 1970 et 1980 pourraient être surpris, comme je le suis. Pendant mon parcours universitaire, dans les années 1980, la biographie a pratiquement disparu comme genre d'écriture en histoire. L'histoire culturelle et sociale, déployant des méthodes empruntées à la littérature, à l'anthropologie et aux études culturelles, était résolument hostile à la biographie, en particulier lorsque ses sujets étaient des penseurs de l'élite. Il y eut un tournant culturel également en histoire et en philosophie des sciences, et, en dépits du caractère varié des analyses portant sur la formation du discours, la biographie n'en faisait pas partie. Même en histoire intellectuelle, l'approche dominante dissous à la fois le texte et les auteurs dans le discours et le langage, plutôt que de considérer la vie comme sujet problématique approprié* » (Hacohen, 2007, p.9).

d'appartenance. D'où une attention très forte aux interactions, au tissu même de l'enchevêtrement des vies […] En ce moment herméneutique, le genre biographique est devenu plus réflexif et ne prétend plus faire parler directement le réel mais en saturer le sens » (Dosse, 2010-a, p.85).

Si les idées ne peuvent émerger librement de la conscience individuelle d'un auteur, une solution possible consiste alors à élargir l'enquête vers la vies des groupes, et la formation d'« identités théoriques collectives ». C'est la voie suivie notamment par Mata et Lee, qui affirment : « *nous ne sommes pas des biographes, puisque nous ne nous servons pas de ces matériaux pour construire un compte-rendu de la vie. Nous utilisons plutôt les histoires personnelles comme traces de l'élaboration de relations d'identité et de différences dans une communauté* » (Mata et Lee, 2007, p.156). Dans cette perspective, la vie de groupe apparaît comme une ressource privilégiée pour comprendre l'élaboration des paradigmes théoriques, qui seraient le reflet d'« identités intellectuelles collectives », de croyances partagées et élaborées en commun. Étant une science de laboratoire, donc une discipline dont les modalités d'organisation pratiques sont par essence collectives, la neuroéconomie semble particulièrement prédisposée à ce type d'enquête de terrain, comme le suggère l'étude précédemment citée, par Latour, Woolgar et Biezunski, d'un institut de recherche expérimentale en biologie (Latour, Woolgar et Biezunski, 1979).

L'éclairage de la théorie par le contexte biographique, qu'il soit individuel ou de groupe, a néanmoins toujours pour objectif de montrer que les économistes possèdent en amont certaines croyances qui ont forcément une influence, même faible et contingente, sur leur travail théorique. Cette découverte historique d'une croyance collective participant à la fois à la théorie et à la vie des théoriciens est généralement jugée d'autant plus pertinente qu'elle porte sur des économistes qui prétendent être débarrassés de tout jugement de valeur et faire de la science « objective ». La valeur ajoutée du travail historique se mesure alors à l'aune de l'écart entre la prétention à l'objectivité du théoricien d'une part, et l'importance théorique plus ou moins grande de ses croyances personnelles d'autre part. Backhouse et Bateman veulent montrer par exemple que la théorie de Keynes *contrairement à l'opinion de celui-ci et de la plupart de ses commentateurs* est « *imprégnée de jugements moraux* » (Backhouse et Bateman, 2009).

Dans cette perspective, l'histoire de la pensée économique a pour tâche d'étudier et déconstruire ce travail de « déni » des économistes, qui serait caractéristique de la genèse de l'économie contemporaine. Fontaine et Marciano soulignent ainsi l'importance de la « *distinction entre le fait que la vision du monde des économistes (et des autres scientifiques)*

soit inévitablement, quoique partiellement, modelée par leurs croyances idéologiques; et le fait que la plupart des économistes depuis la Seconde guerre mondiale ont considéré leurs théories comme étant indépendantes de ces croyances. Une telle distinction est cruciale pour comprendre l'histoire de l'économie de l'après-guerre. Cela nous aide à mieux comprendre le processus par lequel la mise sous silence de certaines voix a permis aux économistes de construire une image d'eux mêmes comme constructeurs d'une science a-politique » (Fontaine et Marciano, 2007, p.569).

L'éclairage historique de la sociologie des sciences, appliquée à la vie des auteurs, rencontre ici cependant une limite importante, et ce même dans une version plus sophistiquée insistant sur la notion d' identité collective. En effet, ce type d'analyse psychosociale, s'il peut éventuellement aider à comprendre la formation de certaines idées, concepts ou théories, ne peut nullement fournir les motifs de leur succès et de la durabilité plus ou moins grande de leur diffusion. Admettons que les croyances idéologiques des économistes de l'après guerre soient explicatifs de la naissance d'un discours économique à vocation a-politique. Mais pourquoi ce discours se maintient-il? Qu'est-ce qui explique sa durabilité? Pourquoi certaines idée demeurent de simples croyances collectives, alors que d'autres accèdent au statut de théorie scientifique? Comme le souligne Hacohen, « *la biographie est un genre essentiel et viable en histoire des idées économiques et peut au mieux répondre des questions concernant la formation de la théorie économique […] Elle demeure une composante méthodologique essentielle, une étape, dans l'histoire du discours économique, mais elle ne peut répondre la question plus large et plus délicate liée à cette dernière, qui concerne le triomphe des paradigmes économiques et leur influence historique* »(Hacohen, 2007, p.24).

Pour Hacohen, l'historien de la pensée est contraint dès lors à perdre en précision, et à palier l'insuffisance des données strictement biographiques par une interprétation plus libre du contexte : « *pour répondre à de telles questions, nous devons souvent abandonner le détail et la précision de la biographie […] La question de la perméabilité du discours est un paramètre important, et les sources sont ici faibles et peu nombreuses. Étudier le discours implique alors hasard et réussite, et, parfois, la spéculation* » (Hacohen, 2007, p.25). En effet, la question de la permanence et de la stabilité des énoncés théoriques nécessite la prise en compte non seulement d'éléments biographiques, mais d'abord et surtout du contexte plus large. L'ensemble des impératifs et des contraintes sociales de l'époque expliquent à la fois le succès et l'échec des théories. Pour rendre compte de manière contextuelle de la formation et de la durabilité des savoirs théoriques, les historiens de la pensée économique se contentent alors en général d'évoquer une « *mentalité collective* », partagée par les économistes mais

aussi par l'ensemble du corps social. La « *vision du monde* »[41] de l'auteur (*worldview*) est ainsi considérée comme l'élément explicatif déterminant, à l'interstice entre la simple biographie et l'histoire sociale (voir par exemple Fontaine et Backhouse, 2007).

Comme le souligne Hacohen, cette solution a pour inconvénient de rendre l'explication plus confuse et moins précise. Le terme de « vision du monde » exprime ainsi assez bien les problèmes soulevés par ces analyses biographiques élargies. Le terme fait en effet plus ou moins référence à un « air du temps », à des croyances collectives latentes dans une société et un temps donnés. Le degré d'imprécision varie selon les études, mais en général l'expression n'est jamais explicitement définie, et renvoie à des catégories beaucoup trop lâches et incertaines. Cette indétermination offre bien sûr l'avantage de pouvoir passer de manière commode du texte au contexte et vice-versa, pour mieux brouiller la frontière et donner l'impression d'une relation «complexe» de l'un à l'autre. Par exemple, à propos de l'émergence de la notion d'altruisme en économie dans les années 1960, Philippe Fontaine écrit: « *les conséquences malheureuses des changements affectant les sociétés occidentales rappelèrent aux économistes que l'égoïsme et la confiance dans l'individu seul ne suffisent pas à atteindre la cohésion sociale […]. Ceci était significatif d'un souci plus général pour la question de l'atteinte de résultats sociaux favorables à partie d'une multitude d'actions égoïstes, un souci qui amorça un tournant vers la considération de ce qui apparaît comme comportement non-égoïste* » (Fontaine, 2007, p.331-32). Il est difficile d'admettre qu'un phénomène historico-

41 L'expression de « vision du monde » et de « mentalités » collectives sont ici considérées comme équivalentes. Pour être tout à fait précis, le terme de « mentalité collective » renvoie plus spécifiquement à un courant historiographique associé à l'École des Annales. Les représentants majeurs de l'histoire des mentalités sont notamment Marc Bloch, qui étudie les phénomènes de croyance collective dans son ouvrage de 1924 *Les rois thaumaturges*, Lucien Febvre et sa notion d' « outillage mental », Jacques le Goff, et surtout Georges Duby et son livre célèbre sur *Les trois ordre ou l'imaginaire du féodalisme*. Sur l'histoire des mentalités, on se réfèrera plus spécifiquement à Mandrou, 1968; Dosse, 2010-b. Ce courant historiographique connait son apogée dans les années 1970, avant d'être progressivement remis en question dans les années 1980. Les critiques principales (voir notamment de Certeau, Julia et Revel, 1970; Lloyd, 1993) portent sur la dichotomie établie entre culture populaire et culture d'élite -le terme de mentalité étant emprunté à l'ethnologue Lucien Lévy-Bruhl, qui l'utilise pour désigner les comportements pré-logique précédant la modernité occidentale. Aujourd'hui, la notion supposée plus complexe et moins binaire de représentation a pris le relais dans les recherches historiennes (Dosse, 2010-b). Le terme de « mentalité », après avoir été très en vogue il y a quelques années en histoire, apparaît donc aujourd'hui assez désuet. Ceci explique peut être le retour d'expressions naguère rejetées, comme celle de « vision du monde » ou d'« air du temps ». Ces controverses peuvent cependant être ignorées pour notre propos, et nous utilisons ici les termes de mentalité, représentation et vision du monde comme synonymes en tant qu'elles font toutes référence à une croyance collective, ce qui autorise ainsi l'historien à faire le lien avec une étude du contexte social. Comme le souligne Dosse, le concept de représentation joue aujourd'hui un rôle similaire à celui de mentalité dans les années 1970: « *il habille large et permet à l'historien de butiner dans les divers champs d'investigation que se sont appropriées les sciences sociales sœurs* » (Dosse, 2010-b, p.222). Par delà les querelles historiographiques, la qualité des études historiques ne dépend pas, à l'évidence, d'un choix de vocabulaire mais bien plutôt de la finesse de l'analyse. Or, force est de constater qu'en histoire de la pensée économique, ce genre d'enquêtes est en règle générale peu convaincant, les auteurs se contentant le plus souvent de faire appel à une causalité sociale assez grossière (*cf. supra*).

social aussi indéterminé que « *les conséquences malheureuses des changements affectant les sociétés occidentales* » après 1945 soit explicatif de quoi que ce soit. Quels sont ces changements? En quoi influent-ils les économistes dans leur travail théorique? L'auteur ne répond pas à ces questions préalables. Il considère comme acquis que l'altruisme appartient à la vision du monde des années 1960.

Comme le souligne Hacohen, l'imprécision est le prix à payer pour élargir l'explication au delà de la simple formation des énoncés théoriques. Ce type d'interprétation sociale des biographies n'en est pas pour autant illégitime. Certes, le recours à des notions telles que celle de « vision du monde » apparaît le plus souvent comme un échappatoire commode pour l'historien, mais la difficulté à définir le contenu de ces mentalités collectives témoigne aussi de l'intérêt historique de la démarche. Les critiques formulées plus haut suggèrent néanmoins qu'un travail de distinction entre deux types de contexte, l'un de formation, l'autre de diffusion, est nécessaire, préalablement à leur articulation, pour éviter toute confusion. Le parcours intellectuel des économistes, même élargi à une communauté intellectuelle, ne donne pas lieu en effet à la même interprétation des théories que celle qui pourrait être élaborée à partir de l'étude des pratiques économiques et sociales. A propos de la neuroéconomie, l'évolution personnelle de certains chercheurs rend par exemple compte, on l'a vu, des stratégies de rapprochement entre disciplines qui ont donné progressivement naissance à la neurobiologie de la décision dans sa forme actuelle. Cette description n'a *a priori* que peu d'éléments communs avec une analyse des débouchés et des applications pratiques de la neuroéconomie, et tout l'enjeu -mais aussi toute la difficulté- consistera à associer ces deux schémas explicatifs. Le deuxième volet de l'enquête sociologique conduit donc à chercher dans des impératifs technico-économiques la cause de la durabilité et de la permanence des énoncés théoriques. Le succès de la neuroéconomie en tant que nouvelle discipline serait ainsi lié à une demande sociale qu'elle parviendrait à satisfaire.

1. Une science intéressée: le rôle des débouchés et des applications pratiques

Le regard « externaliste » porté sur la science peut conduire à deux types d'explication différents. Le premier porte sur le contexte de formation des théories; le second sur le contexte de diffusion. Cependant, les deux modes d'analyse ne sont pas mutuellement

exclusifs. Ils ne sont en général pas distingués dans la plupart des études[42]. Une vision élargie du contexte de formation conduit en effet, on l'a vu, à rattacher des éléments biographiques à une « vision du monde » caractéristique de la société dans son ensemble. L'individu est à l'intersection de multiples lignes d'influences sociales, sa vision du monde offre un condensé de l' « esprit d'une époque »[43]. La figure de l'auteur est davantage un préalable à l'analyse du contexte de diffusion. En outre, envisagée depuis le point de vue de ses applications immédiates ou même seulement possibles, la science engage la résolution de problèmes pratiques qui sont donnés comme antérieurs au savoir lui-même. Ces problèmes se posent aux scientifiques mais aussi à leur environnement social au sens large. C'est l'autre raison pour laquelle l'analyse des débouchés techniques et de la diffusion des théories scientifiques peut se concevoir comme un élargissement des perspectives biographiques au sens strict.

Ces explications associant contexte de formation et de diffusion sont assez fréquentes en histoire de la pensée économique. Il doit permettre de mettre à jour une cohérence d'impératifs techniques, d'intérêts sociaux qui entretiennent un rapport nécessaire avec cette vision du monde auquel participe la science économique: de quels types de croyances collectives les économistes sont-ils représentatifs? Comment envisagent-ils la société et ses problèmes? Quelles solutions prétendent-ils y apporter? Les réponses à ces questions sont plus ou moins évidentes selon les théories économiques. Une ligne idéologique clairement et explicitement articulée permet par exemple de saisir assez aisément les ambitions pratiques et politiques d'une école de pensée.

En revanche, les choses sont moins évidentes *a priori* pour la neuroéconomie. Les neuroéconomistes, notamment ceux issus de la neurobiologie considèrent en effet que leur travail relève relève de la recherche pure. Celle-ci pourrait avoir éventuellement des retombées pratiques, mais de manière non-directe, sur un horizon très long et prospectif. Par conséquent, la neuroéconomie ne saurait de leur point de vue ni transformer la société, ni même proposer des possibilités de transformation[44].

42 Sur l'inséparabilité du contexte de formation et de diffusion, voir par exemple Terrall, 2006, p.73 « *écrire sur la vie des individus, extraite et reconstruite à partir de preuves matérielles, oblige aussi à écrire sur la science elle-même, à travers l'expérience vécue de ses praticiens. En dévoilant ces lettres oubliées et en cherchant ces «forces volatiles et sensibles» qui dirigent nos sujets, nous sommes entraînés dans le monde des lecteurs, des spectateurs, institutions, collaborations, disputes et de toutes les autres interactions qui forment le quotidien de la science* » .

43 Ainsi, pour Hacohen, les biographies « *ne sont rien d'autre que l'étude de l'intersection locale de discours qui construisent le sujet* » . Elles « *fournissent de superbes points de vue pour l'analyse de l'intersection des discours. Les rumeurs sur la mort du sujet et l'éclipse de l'expérience sont largement exagérées* » (Hacohen, 2007, p.22)

44 Sur les faibles retombées pratiques de la neuroéconomie, voir par exemple l'entretien avec Mathias Pessiglione du 22 novembre 2009 en annexe. Il y a bien sûr des exceptions à ce scepticisme quant au potentiel pratique de la neuroéconomie; Olivier Oullier, chercheur en neurosciences à l'université d'Aix-Marseilles, a par exemple participé à un rapport du Conseil d'Analyse stratégique sur les « impacts éthiques

Pourtant, cette affirmation selon laquelle le savoir sur le cerveau n'aurait pas ou peu d'utilité pratique directe semble aller contre l'intuition commune, qui associe spontanément neurosciences et transformation, voire carrément manipulation des cerveaux. Cette revendication d'une science pure, dénuée de toute ambition réformatrice, par les neurobiologistes, apparaît ici comme l'écho du travail de déni des économistes de l'après guerre évoqué précédemment (*cf.* Fontaine et Marcicano, 2007, cité p.23), qui construisent leur discipline comme a-politique, alors même que, dans les faits, leurs théories ont imprégné, et continuent d'imprégner nos croyances, nos comportements et nos pratiques concrètes. Le refus ostensible d'une logique réformatrice ne doit donc pas être pris pour argent comptant: doivent être prises en compte les transformations sociales concrètes induites ou rendues possibles par un courant théorique.

Quels changements pratiques la neuroéconomie peut-elle apporter? La neurobiologie dispose de deux segments d'application en sciences de gestion, l'un en marketing et l'autre en finance (plus exactement en théorie de la décision financière). Le « neuromarketing » utilise les résultats théoriques des neurosciences pour améliorer les stratégies de communication et de vente des entreprises. Les connaissances relatives à la perception des stimuli environnementaux, ou à la mémorisation des expériences perceptives peuvent notamment avoir un intérêt immédiat pour les chercheurs en marketing (Droulers et Roullet, 2007 ; Droulers et Roullet, 2010). Dans ce domaine, une étude fréquemment citée est l'expérience réalisée en 2004 par McClure *et al.* sur les préférences relatives aux marques de soda , qui montre que le cerveau prend en charge dans des régions distinctes l'encodage de valeurs « culturelles », comme une marque de soda, et « sensorielles », liées à la simple consommation de la boisson (McClure *et al.*, 2004). La neurofinance vise quant à elle à améliorer les aptitudes décisionnelles des investisseurs financiers. Ce sont ici les travaux neuroscientifiques sur le comportement face au risque et le choix intertemporel qui font l'objet d'une application pratique.

Neuromarketing et neurofinance sont deux formes d'instrumentalisation distinctes. Dans un cas, il s'agit d'apprendre à « tromper » (dans une certaine mesure) le cerveau des consommateurs ; dans l'autre, l'enjeu consiste à aider les décideurs. Néanmoins, au delà de cette différence de forme, le contenu théorique de ces applications pratiques est sensiblement le même. Les conclusions des travaux expérimentaux sont utilisées comme autant de recettes psychologiques destinées à améliorer soit les stratégies de vente des entreprises, soit les

des neurosciences ». Néanmoins, ce genre de prise de positions sur des « sujets de société » est assez rare et la plupart des chercheurs affichent la plus grande prudence et refusent de parler d'application en l'état actuel des connaissances.

décisions prises par les investisseurs. Les neurosciences ne constituent pas, à proprement parler, un nouvel outil pour le marketing et la finance. Le contenu scientifique des arguments mobilisés par le neuromarketing et la neurofinance joue en fait un rôle secondaire; c'est probablement la raison pour laquelle les neurobiologistes refusent en général d'être associés à ces démarches. Les recommandations pratiques qui sont formulées ne sont en général pas spécifiques aux neurosciences ; il s'agit, au mieux, d'une confirmation de résultats déjà connus en psychologie. Par exemple, l'étude citée précédemment de McClure *et al.* (2004) n'apporte aucun élément décisif pour le marketing, puisque l'on y apprend seulement que les individus sont sensibles aux aspects culturels des marchandises et aux effets de marque. La différence entre deux systèmes d'évaluation s'illustre par des différences d'activation entre des zones du cerveau. Néanmoins, l'idée selon laquelle les marques ont une influence sur les jugements de valeur n'est pas nouvelle : il s'agit simplement d'une confirmation, au niveau neuronal, d'une hypothèse largement admise en marketing.

En ce qui concerne la neurofinance, l'étude de Sokol-Hessner (2009), intitulée « Penser comme un trader réduit l'aversion au risque », montre que les individus améliorent leurs performances dans un jeu d'investissement financier lorsque l'on leur demande de s'« imaginer être » dans la peau d'investisseurs professionnels. Là encore, la nouveauté ne provient pas tant du résultat comportemental en tant que tel, mais de la corroboration de cette conclusion par des données neuronales. Invité à participer à un séminaire de management destiné à des investisseurs financiers, Paul Zak a quant à lui élaboré « *huit leçons de la neuroéconomie pour les managers financiers* » (Sapra et Zak, 2010). Encore une fois, le contenu des recommandations est assez décevant, et il semble difficile d'y voir de véritables découvertes des neurosciences: nécessité de contrôler notre goût démesuré pour le risque qui résulte de notre impulsion biologique à chercher la nouveauté (Sapra et Zak, 2010, p.67), de vérifier les « points de référence » dans les jugements probabilistes (cette notion faisant référence directement à la théorie des perspectives de Kahneman plutôt qu'aux neurosciences (Sapra et Zak, 2010, p.72), tendance au mimétisme (Sapra et Zak, 2010, p.70), *etc.*

Si la neuroéconomie ne débouche que sur des conseils de psychologie pratique, il semble effectivement difficile de voir dans les neurosciences une nouvelle technologie économique d'importance. La neuroéconomie ne se borne cependant pas à fournir des schémas d'explication psychologiques du comportement à partir de l'interprétation des activations du cerveau, mais offre la possibilité de contrôler et de manipuler ces zones, dans le but de modifier les comportements. Ceci soulève la difficile question de savoir si les neurosciences peuvent donner lieu à des applications réellement et directement coercitives et

contraignantes. D'un côté, les neurophysiologistes sont effectivement capables de contrôler certaines structures cérébrales par voie médicamenteuse. C'est notamment le cas du système dopaminergique, qui joue un rôle central en neuroéconomie en tant qu'il est responsable de l'évaluation des récompenses espérées. Par exemple, Palminteri *et al.* ont montré que l'administration de L-Dopa et/ou de halopéridol[45] permettait de restaurer un comportement « sain » face au risque pour des sujets atteints de la maladie de Parkinson et du syndrome Gilles de la Tourette (Palminteri *et al.*, 2009).

Toutefois, ce type de modification a des effets globaux sur le comportement. Il est très délicat de soigner les individus malades en neuropsychiatrie à l'aide des médicaments, car il est difficile d'élaborer des actions réellement ciblées. La neuropharmacologie transforme radicalement la personnalité. Par conséquent, il est plus qu'improbable d'espérer, en l'état actuel des recherches, pouvoir agir sur le cerveau des individus afin de les forcer à consommer un type de produit ou d'améliorer leurs stratégies d'investissement. Même en psychiatrie, le recours aux médicaments sur la base des informations fournies par la neuroimagerie n'est pour l'instant pas très développé. Les outils de neuroimagerie qui ont servi de support à l'essor des neurosciences contemporaines sont en effet des outils de mesure et de détection. Ils permettent d'enregistrer l'activité de neurones individuels (microélectrodes) ou de zones du cerveau (via la mesure de leur taux d'oxygénation -appelé signal BOLD- par l'IRM fonctionnelle). Ces instruments sont donc distincts des techniques de la neuropharmacologie. Les progrès dans la connaissance du fonctionnement du cerveau ne sont donc pas nécessairement parallèles à ceux réalisés dans le domaine du traitement médicamenteux. La neuroimagerie peut néanmoins avoir une utilité médicale (*cf* chapitre 5 et 6), mais plus en tant qu'instrument de diagnostic que de technique thérapeutique à proprement parler.

Au delà de la fonction thérapeutique actuellement limitée des instruments utilisés en neurosciences, il y a peut être aussi en amont, de manière encore plus décisive, une incompatibilité de principe entre neurobiologie et psychopharmacologie. En effet, comme le souligne Pierre-Henri Castel, la première évolue dans un cadre conceptuel évolutionniste: « *la psychopathologie* [dérivée des neurosciences cognitives] *bouscule la vieille psychiatrie sur un autre plan, car elle est de plus en plus évolutionniste* ». La référence à l'évolution a d'abord une conséquence théorique, car « *la contextualisation des phénomènes restreint la généralité des explications. Il n'y a pas de désordre biologique « en soi », mais en présence de certains*

45 Le L-Dopa est une substance augmentant le taux de dopamine. Le halopéridol est un médicament neuroleptique qui agit sur les récepteurs de la dopamine en les inhibant.

contraintes du milieu. » Ainsi, contrairement à l'idée reçue d'un réductionnisme neuronal dont seraient porteuses les neurosciences, la psychiatrie évolutionniste « *s'efforce de tenir, les deux bouts de la chaîne, entre le déterminisme moléculaire et génétique et la vie de relation des êtres humains* »; ainsi, elle « *transcende les anciens clivages de la médecine mentale traditionnelle* » et « *laisse derrière elle les thérapies béhavioristes des troubles mentaux (qui ignorent l'histoire du développement)* » (Pierre-Henri Castel, 2010, p.26). Les neurosciences ne sont pas porteuses d'une médicamentation forcenée des troubles psychiques mais au contraire d'une régulation raisonnée du milieu de vie du patient. En outre, les neurosciences n'ayant pas à ce jour inventé de thérapies nouvelles ni de nouveaux médicaments, les psychiatres évolutionnistes s'en remettent le plus souvent aux techniques de soin ayant fait leurs preuves dans le passé: « *la problématique « évo/dévo » ...remet en selle les prémisses des bonnes vieilles psychothérapies de la relation parents-enfants -voire, à la surprise générale, des versions cognitives « naturalisées » de certains idées psychanalytiques!* » (Pierre-Henri Castel, 2010, p.26)

Ceci soulève incontestablement des problèmes majeurs pour une interprétation sociologique de la neuroéconomie. Il est délicat d'identifier des intérêts clairs et évidents derrière l'utilisation des neurosciences en économie. En particulier, il apparaît douteux d'attribuer à un « lobby pharmaceutique », soucieux de généraliser l'utilisation des psychotropes dans la société, un rôle direct et influant dans l'émergence de la discipline. Les craintes d'une « neuromédicalisation » de la société sont largement infondées en l'état actuel des connaissances, et relèvent plus au fond de ce que Ruwen Ogien appelle une « *panique morale* » (Ogien, 2004). Les préoccupations éthiques sont en effet disproportionnées au regard des menaces réelles. Les autorités sont très vigilantes en ce qui concerne l'utilisation de la neuroimagerie : la législation française sur la bioéthique fait peser par exemple de nombreuses contraintes sur l'utilisation de ces techniques à des fins non-médicales[46].

Les enquêtes sociologiques concernant les neurosciences insistent en général sur les liens de dépendance entre la recherche de laboratoire et les grandes firmes pharmaceutiques. Les travaux de David Healy en particulier ont montré de manière convaincante que la méthodologie et l'orientation des recherches en neurosciences étaient largement dépendantes des préférences financières de l'industrie pharmaceutique (Healy, 1998 ; Healy, 2002). Dans un article récent intitulé « D*evenir accroc aux médicaments: l'école de Chicago, le projet pharmaceutique et le néolibéralisme médical* », Edward Nik-Khah explore les liens existant

46 Voir notamment la loi n°2004-800 du 6 août 2004 , et dans les annexes, les commentaires de Luc Mallet et Mathias Pessiglione sur le difficile accès aux IRMf à l'hôpital de la Pitié-Salpêtrière, pourtant l'un des centres hospitaliers parmi les mieux dotés en la matière en France.

entre certains économistes de l'école de Chicago (George Stigler notamment), leurs travaux sur le capital humain et l'économie de la santé, et le lobby pharmaceutique (Nik-Khah, 2009). Du fait du rôle joué par les neurosciences, la neuroéconomie semble au premier abord pouvoir faire l'objet d'une enquête socio-historique du même type. Or, les neuroéconomistes ne plaident nullement en faveur d'une médicalisation des troubles comportementaux, en particulier pour les comportements addictifs (voir notamment Ross *et al.*, 2008). La neuroéconomie est porteuse d'une nouvelle conception de la santé mentale[47]dans laquelle la pharmacologie semble jouer un rôle secondaire.

La tâche de l'historien « externaliste » se complique donc fortement. Les liens entre neuroéconomie et industrie pharmaceutique semblent ténus. Cela ne signifie pas pour autant que la neuroéconomie ne réponde pas à une certaine demande sociale. Toutefois, cette demande ne doit sûrement pas être comprise de manière trop simple comme un impératif ou une pression émanant d'un groupe d'intérêt commercial ou économique. Les choses sont plus complexes, car la neuroéconomie n'a *a priori* aucun débouché pratique immédiat. Dans la perspective de Canguilhem, les données du problème sont mal posées. Il faut en effet non pas partir d'impératifs socio-économiques mais chercher au contraire l'origine de la science dans des limites rencontrées par la technique: « *l'essor de la pensée scientifique a pour essor l'échec de la technique* » (Canguilhem, 1977, p.23). Cette affirmation signifie que l'échec de la technique est donc aussi, en amont, un échec de la pensée.

Un problème pratique est donc l'occasion, non la cause, d'une révolution scientifique. A défaut d'expliquer, les facteurs contextuels peuvent ainsi, *a minima*, servir à repérer les ruptures théoriques. De notre point de vue, c'est l'échec d'une pensée « asilaire » de la maladie mentale, dont le fonctionnement concret a été étudié notamment par Goffman[48], qui explique l'essor de cette nouvelle forme de réflexion théorique de la santé mentale, inspirée par l'économie. Il faudra donc montrer en quoi la neuroéconomie participe de ce projet théorique, lié à l'ouverture des asiles sur le monde social, plus vaste et non-réductible à la simple satisfaction d'intérêts économiques et financiers.

47 L'un des objectifs de ce travail consistera à insister sur les re-configurations générales du soin psychiatrique autorisées et rendues possibles par une vision « économique » des troubles mentaux. De ce point de vue, il n'est pas illégitime de se préoccuper de l'utilisation possible des neurosciences. Toutefois, ces préoccupations devraient se porter non pas sur les possibilités de « manipulation » des cerveaux, mais plutôt sur l'essor de ce qui sera caractérisé ici comme néo-comportementalisme économique, et notamment sur sa déclinaison politique, en économie du bien-être, sous la forme d'un paternalisme libertarien (*cf.* introduction et chapitre 7).

48 *cf.* Erving Goffman, 1968. *Asiles; études sur la condition sociale des malades mentaux et autres reclus*. Les Editions de Minuit.

II. La notion d'idéologie scientifique chez Georges Canguilhem : une source d'inspiration pour l'historien de la pensée économique

Les études sur la neuroéconomie peuvent être réalisées selon deux perspectives, soulevant chacune des problèmes spécifiques. Le regard interne à la discipline elle-même n'est pas à proprement parler historique, ou du moins n'en nourrit pas l'ambition. L'approche dite externaliste rencontre un obstacle particulier dans la mise en rapport des contextes de formation et de diffusion de la neuroéconomie. En outre, même si ce problème est résolu, il apparaît délicat de distinguer, du point de vue de la sociologie des sciences, les théories scientifiques (neuroéconomiques par exemple) de simples croyances collectives.

L'enjeu consiste ici à essayer de résoudre ces difficultés à partir de la notion d'« idéologie scientifique » proposée par Georges Canguilhem (Canguilhem, 1977). Selon ce dernier, la position de l'historien ne saurait être interne à la science elle-même. L'histoire de la pensée n'est pas une simple restitution des théories, et se distingue par une inventivité propre (A). Toute la difficulté consiste alors à conserver précision et rigueur analytique dans le récit historique. L'idéologie désigne un élément non-théorique, mais qui se donne paradoxalement à voir uniquement dans les théories elles-mêmes. Parler d'idéologie dans les sciences, pour Canguilhem, ne revient donc pas à élargir de manière plus ou moins arbitraire au contexte d'élaboration ou de diffusion, mais au contraire à établir un réseau de nécessités dans les textes théoriques eux-mêmes. A propos de la neuroéconomie, l'idéologie scientifique désignera l'ambition (théorique) nourrie par les neuroscientifiques et les psychologues de traiter la maladie mentale dans des termes économiques, que nous désignerons par l'expression de « psychiatrie économique ». Il sera donc possible *in fine* d'esquisser la forme que prendra cette histoire de la neuroéconomie, écrite dans une perspective largement inspirée par Georges Canguilhem (C).

A. L'originalité du regard historique: l'attention aux ruptures

Pour Canguilhem, l'histoire des sciences doit aller plus loin que la simple reconstruction rétrospective du discours scientifique. L'histoire des sciences, en effet, n'est pas seulement la mémoire de la science mais aussi son « *laboratoire* », c'est-à-dire qu'elle doit permettre de mimer l'expérience vécue de la découverte scientifique. L'historien ne se borne pas, par conséquent, à juxtaposer chronologiquement des contenus théoriques : « *du fait qu'une élaboration n'est pas une restitution, on peut conclure que la prétention de l'épistémologie à rendre plus qu'elle n'a reçu est légitime* » (Canguilhem, 1977, p.13) L'histoire des sciences n'a pour autant nullement besoin de recourir à l'imagination ou à l'invention pour apporter une compréhension subjective de la construction du savoir scientifique. La description historique n'est ni spéculative, ni fictive. L'historien signale donc son originalité en s'intéressant non pas directement à la science mais plutôt à son processus d'élaboration. Or, l'étude de ce processus nécessite, selon Canguilhem, « *une éducation de l'attention aux ruptures* » (Canguilhem, 1977, p.24).

Canguilhem se montre ici fidèle à la notion de « rupture épistémologique » proposée par Gaston Bachelard : le progrès scientifique résulte d'abord de l'élimination de connaissances antérieures (Canguilhem, 1977, p.22). Si la science procède ainsi par « rature » et approfondissement, cela suggère paradoxalement que l'esprit scientifique n'est jamais « jeune »: en d'autres termes, celui-ci ne progresse que sur la base de connaissances antérieures inadéquates. C'est la raison pour laquelle, selon Canguilhem, ce ne sont pas tant les « corrections » ultérieures, mais plutôt les conjectures hasardeuses, les anticipations spéculatives qui doivent être considérés comme les moments importants de l'histoire des sciences. Par conséquent, les ruptures décisives de la pensée scientifique doivent être recherchés dans des « *dépassements présomptueux* » de la science par les scientifiques eux-mêmes, par lesquels les savants tentent d'élargir la connaissance sans attendre l'aboutissement des recherches: « *la production progressive de connaissances scientifiques nouvelles requiert, à l'avenir comme dans le passé, une certaine antériorité de l'aventure intellectuelle sur la rationalisation, un dépassement présomptueux, par les exigences de la vie et de l'action, de ce qu'il faudrait connaître et avoir vérifié, avec prudence et méfiance, pour que les hommes se rapportent à la nature selon de nouveaux rapports et en toute sécurité* » (Canguilhem, 1977, p 38).

Cette « *impatience* » du scientifique s'explique donc, en amont, par la prise en compte

d'un certain nombre de besoins pratiques, liés selon Canguihem à « *la vie et à l'action* ». L'idéologie scientifique désigne cette convergence d'intérêts à la fois pratiques et contingents d'un côté, et scientifiques de l'autre. La science entretient ainsi des relations de voisinage avec des éléments non-scientifiques, comme des croyances politiques par exemple. Canguilhem illustre notamment la notion d'idéologie scientifique avec la théorie évolutionniste au XIX^{e} siècle. Au début, les théories de Darwin peinaient à s'imposer, car elles passaient pour une vulgaire importation, en biologie, de la pratique des éleveurs. L'évolutionnisme social de Herbert Spencer, qui peut être considéré comme une idéologie politique au sens commun donné à cette expression, joua alors un rôle central dans l'acceptation des thèses de Darwin, avant que, ultérieurement, la génétique moderne n'apporte une validation définitive à celles-ci.

Précisément, cette boucle ultérieure de consolidation de la théorie génère et explique l'illusion rétrospective qui donne à la reconstruction historique l'aspect de la linéarité: l'évolutionnisme est scientifique, parce que la génétique le confirme. Le travail de l'historien consiste à revenir sur ces « *dépassements présomptueux* » -par exemple, les doctrines de Spencer- pour montrer que ceux-ci ne sont pas des freins au progrès de la science, mais en sont au contraire des conditions nécessaires, en tant qu'ils motivent l'intérêt des théoriciens, soulèvent et indiquent des problèmes à résoudre, *etc.*

Cette tendance à aplanir le déroulement de la pensée scientifique est visible dans les histoires de la neuroéconomie écrites depuis un point de vue interne. Dans cette perspective, la neuroéconomie peut se comprendre comme une tentative de réponse à des problèmes posés par l'économie comportementale, à partir des outils des neurosciences, notamment de la neuroimagerie. Pourtant, il doit bien y avoir au départ une motivation particulière, de nature non-théorique, à utiliser ces instruments en économie. L'élaboration théorique de la neuroéconomie dans les années 2000 doit bien répondre à des changements pratiques. Il semble en effet que la neuroéconomie corresponde initialement à un effort de codification en termes économiques de problèmes techniques en médecine, concernant la santé mentale. La rupture décisive qui devra être étudiée concernera les changements de conception et de traitement dans la maladie mentale dans les pays occidentaux.

B. Peut on parler d'idéologie dans les sciences ? Canguilhem et la sociologie des sciences

Le progrès de la science engage, on l'a vu, un « *dépassement présomptueux* » de l'état actuel des connaissances de la part du scientifique (Canguilhem, 1977, p.23). Cette anticipation spéculative de l'aboutissement des recherches doit être mise en rapport avec ce que Canguilhem appelle une « *idéologie scientifique* ». Or l'application du concept d'idéologie dans le domaine scientifique soulève des difficultés. Cette expression évoque en effet au premier abord un rapprochement avec la sociologie des sciences; mais, pour Canguilhem, l'histoire des sciences n'a aucune vocation sociologique. Faut-il voir l'indice d'une aporie? L'enjeu ici consiste à préciser le sens de cette expression d'« idéologie scientifique », afin de comprendre la tâche assignée par Canguilhem à l'historien de la pensée.

Le terme d'idéologie, observe Canguilhem, désigne initialement le fait qu'un jugement puisse être orienté, ou dénaturé, par l'intérêt. Cette définition minimale, qui s'accorde aussi avec le sens qu'en donne Marx[49], invite naturellement à critiquer les prétentions à l'objectivité du sujet connaissant. Ce dernier est en effet supposé, consciemment ou inconsciemment, déformer et orienter le contenu de ses croyances en fonction de son intérêt. L'idéologie met donc en évidence deux illusions : celle portant sur le contenu des représentations d'une part, celle concernant la prétention à l'objectivité de ces mêmes représentations : « *toute idéologie est un écart, au double sens de distance et de décalage, distance de la réalité, décalage relativement au centre d'investigation à partir duquel elle s'imagine procéder* » (Canguilhem, 1977, p.36).

Reconnaître l'existence d'idéologies dans les sciences revient ainsi à admettre que les jugements scientifiques soient influencés par l'intérêt. Nous sommes donc ici reconduits au problème précédemment évoqué à propos de la sociologie des sciences. Comment expliquer l'autonomie et la durabilité des théories scientifiques, une fois admis que la science est dépendante de son contexte de formation et/ou de diffusion?[50]. Canguilhem montre ainsi par

49 En effet, chez Marx, l'idéologie est une croyance illusoire, qui a néanmoins une certaine utilité pour le sujet: en tant que que « *fabulation rassurante, complaisance inconsciente à un jugement orienté par l'intérêt* », elle exerce une « *fonction de compensation* » (Canguilhem, 1977, p.37)

50 Ces difficultés ont été semble-t-il comprises par Marx lui-même. En effet, ce dernier, dans *L'Idéologie Allemande*, ne mentionne pas -ou, peut être, évite de mentionner- la science comme possible idéologie. La raison se trouve formulée dans sa *Contribution à la Critique de l'Économie Politique* : si l'on admet que l'idéologie est relative à un état social, comment celle-ci pourrait elle (dans le cas de la science) conserver une valeur permanente?En outre, la société communiste promet abolition de l'idéologie: mais celle-ci n'est-elle pas condition du progrès scientifique? Ceci impliquerait qu'il puisse y avoir une science débarrassée de toute idéologie.

exemple qu'au XIX^{e} siècle, les sciences biologiques sont largement guidées dans leur développement par des impératifs commerciaux. Néanmoins, ces impératifs, qui évoluent et divergent au cours du temps, n'expliquent nullement la convergence théorique qui s'établit à la fin du siècle autour de la notion d'évolution.

La solution à cette aporie se trouve peut être le plus explicitement formulée chez Canguilhem dans sa critique de la sociologie des sciences, qui vise notamment Thomas Kuhn :

> « *Kuhn parvient mal à répudier l'héritage de la tradition logico-empiriste et à s'installer décidément sur le terrain de la rationalité, de laquelle semblent pourtant relever les concepts-clefs de cette épistémologie, ceux de paradigme et de science normale. Car paradigme et normal supposent une intention et des actes de régulation, ce sont des concepts qui impliquent la possibilité d'un décalage ou d'un décollage à l'égard de ce qu'ils régularisent. Or Kuhn leur fait jouer cette fonction sans leur accorder les moyens, en ne leur reconnaissant qu'un mode d'existence empirique comme faits de culture. Le paradigme c'est le résultat d'un choix d'usager. Le normal c'est le commun, sur une période donnée, à une collectivité de spécialistes dans une institution universitaire ou académique. On croit avoir affaire à des concepts de critique philosophique, alors qu'on se trouve au niveau de la psychologie sociale. D'où l'embarras dont témoigne la Postface de la deuxième édition de Structure des Révolutions scientifiques lorsqu'il s'agit de savoir ce qu'il convient d'entendre par vérité d'une théorie* » (Canguilhem, 1977, p.23)

Le reproche porte sur la réduction « *psychosociale* » des théories scientifiques à des croyances collectives. Le paradigme est un « *choix d'usagers* ». La science est privée de toute autonomie : les théories ne deviennent scientifiques que de manière résiduelle, à partir de la mise en convergence de croyances individuelles. Pour Canguilhem, ce qui pose problème dans la sociologie de Thomas Kuhn, ce n'est pas tant le problème du relativisme, mais plutôt le caractère purement passif du processus d'élaboration de la science. Il n'y a pas à proprement parler de découverte ni d'avancée, puisque la seule pratique de la délibération collective conduit au fond à approuver et acquiescer aux croyances partagées par le plus grand nombre.

C'est ici que se joue la différence cruciale, car, pour Canguilhem, la science engage bien un acte de création, une « *intention* » initiale, établissant à la fois de nouveaux contenus de connaissance et de nouveaux critères (des « *actes de régulation* ») pour la production du savoir scientifique. C'est en ce sens que la science implique un « *décollage* » par rapport à ses objets. Mais le même problème revient: comment défendre l'idée d'une « *intention* » et d'un processus créateur dans les sciences tout en maintenant la stabilité de ces dernières? En outre, cette création théorique renvoie nécessairement à un sujet créateur, auquel est associé un contexte historique spécifique, des besoins pratiques, *etc.* Comment concilier donc la

singularité d'un ensemble d'intentions subjectives et l'objectivité du discours scientifique ?

Deux réponses possibles sont envisageables. La première solution maintient une compatibilité logique avec la sociologie des sciences. Pour Canguilhem, la science progresse en effet par « erreurs et tâtonnements ». Les idéologies scientifiques peuvent dès lors se comprendre comme des approximations initiales, voir carrément comme des erreurs, qui se révèlent néanmoins nécessaires à long terme pour la science. En d'autres termes, la construction de la science est soumise à des impératifs socio-économiques, mais le résultat de ce processus -le savoir scientifique- est en lui-même objectif et indépendant. Les besoins pratiques sous-jacent à un questionnement scientifique peuvent donc stimuler et orienter les recherches mais ne déterminent pas directement le contenu des théories. L'évolutionnisme, par exemple, dans sa version moderne, est confirmé par la génétique ; pour autant, il ne valide pas les doctrines de Herbert Spencer, qui font figure aujourd'hui de croyances un peu datées.

Le darwinisme de Spencer, en tant qu'illustration du concept d'idéologie scientifique fournie par Canguilhem lui-même, favorise donc une telle interprétation. Cette solution apparaît donc comme cohérente avec les exigences initiales de Canguilhem. Néanmoins, il apparaît nécessaire, au regard notamment de la critique de la sociologie de Kuhn, et, plus généralement, à l'appui de la pratique effective de l'histoire des sciences par Canguilhem, de maintenir un statut purement théorique à l'idéologie scientifique, et de désolidariser cette notion de toute forme de besoin ou de contraintes pratiques. Le geste créateur du scientifique est sans doute, au départ, guidé par un impératif de la pratique. Il n'empêche que le savoir qui en résulte s'émancipe de ses conditions matérielles de production; mieux: le succès technique de la science dépend précisément de cette émancipation par rapport aux exigences immédiates de l'action.

L'idéologie scientifique peut donc être tournée vers la résolution de problèmes pratiques, mais l'intention réelle qui la dirige est une intention théorique. L'exemple du darwinisme social peut donc faire l'objet de deux lectures différentes. La première consiste à présenter le discours de Spencer comme une doctrine politique, qui a permis de « préparer le terrain » au darwinisme proprement dit, en accordant l'idée d'évolution aux circonstances sociales. Le darwinisme social est ainsi la déformation opportune d'une théorie scientifique. Rien n'empêche toutefois d'envisager la pensée de Spencer comme un projet purement théorique, visant à faire communiquer deux domaines de savoir nettement distincts, les sciences sociales alors naissantes d'un côté, et la biologie évolutionniste de l'autre côté. Certes, ce projet a bien sûr été conçu, au départ, comme une doctrine politique, dans l'intention de constituer une croyance collective adaptée aux exigences de l'époque. En

revanche, la pure intuition d'une synthèse entre logique biologique de l'évolution et réflexion socio-politique est une proposition théorique qui a indubitablement largement dépassé ses condition locales d'énonciation, et a rendu possible la production d'une multitude d'énoncés scientifiques, dépassant largement le cadre étroit des thèses de Spencer.

L'idéologie scientifique désigne au départ une intention, qui ne se définit pas par sa fin pratique, mais qui doit se comprendre plutôt comme une ébauche de pensée, un préalable à la production d'énoncés scientifiques. L'enjeu consiste alors à identifier le point d'origine d'une problématique scientifique qui, certes, se comprend elle-même (à tort) comme projet purement technique, mais qui précisément en tant que tel, est toujours et déjà théorique. L'application, à la neuroéconomie, de la notion d'idéologie scientifique ainsi comprise dans ce sens théorique permet ainsi de poser une contrainte méthodologique forte sur l'interprétation. L'hypothèse proposée ici, à partir de la notion de « psychiatrie économique », consiste en effet à expliquer l'émergence de la neuroéconomie par le projet d'ouverture de l'asile sur le monde social, et, plus précisément, par la volonté d'apporter un traitement « économique » de la maladie mentale. Il faudra cependant être vigilant à propos du rôle joué par d'éventuels besoins pratiques dans l'émergence de la neuroéconomie :si ces besoins, liés à la pratique du soin mental, ont pu ouvrir des problématiques théoriques qui ont ultérieurement donné naissance à la neuroéconomie, l'examen de ces éléments contingents ne doit pas permettre de présumer de la nature ou du contenu des connaissances produites pour résoudre ces problèmes.

La mise en œuvre de cette démarche soulève une néanmoins une dernière difficulté. La mise en évidence d'un intérêt ou d'un besoin pratique sous-jacent au questionnement théorique ne doit pas être appuyé par une étude du contexte. Cet intérêt informe de manière latente les textes scientifiques. Pourtant, il serait possible d'objecter que l'idéologie scientifique, bien que théorique, demeure, en tant qu'intuition pré-scientifique, étrangère à la science elle-même. Comment mettre en rapport le texte de la science avec un élément « hors-texte » sans en passer par le contexte ? Tout l'enjeu consiste à comprendre au fond en quoi consiste la nature de la spéculation pour l'historien. Canguilhem ne donne pas de réponse explicité à cette question, mais l'ensemble de ses travaux laisse à penser que la spéculation, en histoire des sciences, revient non pas à introduire de l'imaginaire dans le récit historique mais à établir un réseau de nécessités internes aux textes eux-mêmes. Canguilhem rejoint (ou anticipe) l'historiographie contemporaine en assignant à l'historien la tâche de « dramatiser », c'est à dire de mettre en récit, son corpus (*cf.* introduction au premier chapitre).

La spéculation ne réside donc pas dans un plus grand degré de liberté accordé à

l'interprétation : les lectures de Canguilhem sont toujours extrêmement précises et fidèles aux textes. C'est d'abord dans l'exhaustivité de sa connaissance théorique que se mesure la qualité de l'historien. Pourtant, le récit historique se veut critique, et il faut bien admettre, chez Canguihem, un recul de l'historien que le scientifique ne possède pas. Si l'histoire spécule, c'est peut-être dans l'organisation des textes à analyser, à travers la mise en rapport de textes qui n'entretiennent pas de liens évidents *a priori*, ou dans la mise en avant d'auteurs méconnus ou oubliés. C'est essentiellement dans le choix, et non dans son traitement du corpus, que l'historien dispose d'un degré de liberté.[51]

A propos de la neuroéconomie, l'enquête sur l'origine de la discipline a donc porté sur les sources intellectuelles des pionniers de la neuroéconomie, et non sur leurs intentions personnelles, leur « vision du monde », *etc.* Ce travail a fait apparaître la forte influence, en amont de ces recherches, de chercheurs comme George Ainslie par exemple, initialement issus de la psychiatrie, qui tentent d'élaborer une compréhension des maladies mentales dans des termes inspirés de l'analyse économique. Cet effort pour codifier économiquement, c'est-à-dire pour quantifier et formaliser les troubles mentaux[52], a donné naissance à un courant de recherches que nous avons appelé « néo-comportementalisme ». Le caractère spéculatif de cette interprétation réside donc dans le choix d'un corpus élargi, puisqu'il a fallu remonter dans les années 1960, et aborder un ensemble de travaux expérimentaux dépassant le cadre restreint de la neuroéconomie des années 2000. Ce choix n'est pas pourtant arbitraire, car ces recherches antérieures font figure d'une part de références théoriques des auteurs principaux de la neuroéconomie, et intègrent d'autre part l'économie générale de notre récit historique.

L'approche historique proposée ici se distingue ainsi d'autres modalités d'analyses en

51 S'il existe une tradition française propre à la philosophie des sciences contemporaines, celle-ci se caractérise ainsi probablement par cette association d'une connaissance extrêmement poussée de l'histoire des sciences, d'un souci du détail porté à l'extrême; et, en même temps, d'une volonté de contester les lectures établies, en diminuant le rôle des auteurs le plus souvent étudiés, et en insistant sur des penseurs considérés comme moins influent. Par exemple, dans son *Histoire du concept de réflexe*, Canguilhem critique la thèse historique traditionnelle (qui sera défendue par Glimcher, voir Glimcher, 2003) selon laquelle Descartes serait l'inventeur du réflexe, pour mieux souligner le rôle joué par le biologiste Hartley. Michel Foucault héritera de cette tendance à remettre en question les influences généralement admises par une « fréquentation assidue » des auteurs. La meilleure défense de ce genre d'étude réside toujours dans la (très) bonne connaissance de la pensée scientifique par l'historien. De ce point de vue, Canguilhem et Foucault ne proposent pas, à proprement parler, de méthodes pour l'histoire et la philosophie des sciences: la connaissance des sources est la seule exigence que doit remplir l'historien, son inventivité ou son « originalité » n'en étant qu'une conséquence contingente. Pourtant, ces mêmes auteurs sont souvent lus comme des représentant d'un « postmodernisme », introduisant à la fois un relativisme et un style plus «lyrique» en histoire de la pensée: Foucault et Canguihem sont généralement envisagés, notamment aux États Unis, comme les ambassadeurs de la « french theory », aux côtés de Deleuze, Derrida, Lacan ou Baudrillard. Sur la contestable réception de Foucault dans les pays anglo-saxons, voir Cusset, 2003.

52 Encore une fois (*cf.* introduction et *infra*), il convient de souligner qu'il s'agit là seulement d'une inspiration, et non d'un emprunt direct à la modélisation économique: en l'occurrence, les processus d'optimisation à l'œuvre dans l'apprentissage de la récompense ont été formalisés par des psychologues et des neurobiologistes, et constituent un apport théorique propre à la psychologie néo-comportementaliste.

histoire de la pensée, qui, elles aussi, veulent trouver un moyen terme entre internalisme et externalisme. C'est le cas par exemple des études portant sur les métaphores de la science économique, qui visent aussi à identifier un élément hors-texte qui informerait les théories économiques. Sur un sujet proche de la neuroéconomie, Philip Mirowski considère en particulier, dans une étude célèbre, que l'économie de l'après-guerre, notamment la théorie des jeux et ses prolongements en économie expérimentale, consacrerait la figure du *cyborg* (Mirowski, 2001). Mirowski considère cependant que cette influence s'exerce de manière *métaphorique*: l'agent économique est modélisé *à la manière* d'un cyborg, c'est à dire d'un individu mi-homme, mi-robot. Or cette image du cyborg n'est jamais définie clairement, et renvoie de manière confuse à tout ce qui est associé dans l'imaginaire collectif à la figure du robot: une forme d'automatisme, la science fiction, la guerre froide, etc. Ce type d'analyse renvoie donc *in fine* à un « air du temps », à une « vision du monde » cybernétique[53].

En revendiquant un tel degré de liberté dans l'interprétation, Mirowski soulève, pour l'historien « internaliste » le problème de savoir ce qui, initialement, motive le choix de la métaphore elle-même: pourquoi parler d'une économie du cyborg, et non d'une économie de la guerre froide, ou, plus généralement, d'une économie caractérisée par un élément quelconque de l'imaginaire collectif des années 1950 et 1960? De notre point de vue, l'agent économique n'est pas modélisé par les neuroéconomistes « à la manière d'un malade mental »: la neuroéconomie est l'héritière d'un programme de recherche qui se comprend initialement comme une solution théorique à l'échec pratique des techniques de régulation asilaire de la maladie mentale dans les années 1960 et 1970. L'élément hors-texte -lié ici à la gestion de la maladie mentale- n'est pas une simple source d'influence ou d'inspiration pour les neuroéconomistes, mais rend directement possible la collaboration entre neuroscience et économie autour de la question des comportements déviants. Quoique façonnée indirectement par des facteurs contextuels, la neuroéconomie n'est donc pas, de notre point de vue, influencée par l'imaginaire individuel ou collectif des neuroéconomistes.

L'approche économique, c'est-à-dire quantitative et fonctionnelle, des comportements déviants[54] désigne donc d'emblée un programme de recherches théorique, soulevant immédiatement des problèmes analytiques à résoudre: dans quelle mesure peut-on considérer que ces comportements sont maximisateurs? Qu'est-ce qui est alors maximisé? Les références à la médecine mentale en neuroéconomie ne doivent pas par conséquent être comprises de

53 Ceci est d'ailleurs parfaitement assumé par Mirowski puisque ce dernier multiplie les références à la culture populaire, au cinéma, à la littérature, *etc.*

54 Encore une fois, cette expression de « contrôle économique des comportements déviants » doit être comprise dans un sens purement théorique, c'est-à-dire comme une théorie économique des comportements déviants.

manière métaphorique: la question de la santé mentale est en elle-même, et directement, productrice d'énoncés théoriques qui ont été par la suite intégrés à l'économie.

C. Idéologie scientifique et neuroéconomie: esquisse d'un récit historique

La notion d'idéologie scientifique, telle qu'elle a été conçue par Canguilhem, désignera donc, à propos de la neuroéconomie, la tentative initiale, en psychiatrie, visant à codifier en des termes économiques des comportements relevant de la pathologie mentale. Nous désignons ce projet par l'expression de « psychiatrie économique ». Il convient de bien rappeler cependant que cette approche économique en psychiatrie n'a pas consisté à emprunter et appliquer des modèles fournis par l'économie: si les termes utilisés sont influencés par l'économie, à travers notamment la référence à des processus d'optimisation dans la formalisation des mécanismes du *reward learning*, la modélisation a bien été réalisée par des psychologues ou neurobiologistes appartenant au courant néo-comportementaliste. L'approche économique des troubles du comportement a donc été élaborée dans des disciplines extérieures à l'économie, et n'a été intégrée à celle-ci qu'*a posteriori*, au moment où apparaît un débat au sein des *behavioral economics* à propos de la régulation des comportements.

Mais, en devenant un sous-domaine de l'analyse économique, le néo-comportementalisme a introduit au sein de l'économie le problème de la maladie mentale et le concept de pathologie. C'est la raison pour laquelle nous avons également utilisé, plus haut, l'expression d'« approche pathologique du comportement économique » (*cf.* introduction). Celle-ci désignera donc plutôt le résultat scientifique, pour l'économie, du programme de recherche néo-comportementaliste, dont le projet de psychiatrie économique est à l'origine. Dans les deux cas, pour la neuroéconomie aussi bien que pour sa préhistoire théorique, la notion de pathologie constitue une élément structurant, dont l'importance apparaît notamment dans les discussions normatives, lorsqu'il s'agit d'évaluer le degré de rationalité des décideurs ou des comportements. L'approche économique de la maladie mentale ou psychiatrie économique a donc bien au une influence directe sur ce nouveau domaine de savoir, qui, du point de vue des économistes, apparaît comme une approche pathologique du comportement économique. Ce projet théorique peut donc être appréhendé comme une idéologie

scientifique, en reprenant ses trois caractéristiques principales fournies par Canguilhem (Canguilhem, 1977, p.44).

i. Les idéologies scientifiques « *sont des systèmes explicatifs dont l'objet est hyperbolique* » (Canguilhem, 1977, p.44)

En tant qu'elle engage un « dépassement présomptueux » de la science par le scientifique lui-même, l'idéologie représente une tentative pour élargir et étendre certains résultats théoriques, certains concepts, au delà de leur domaine d'application reconnu. C'est parce qu'elle anticipe ainsi sur la possibilité d'une telle extension que l'idéologie est de nature spéculative.

A propos de la psychiatrie économique, c'est-à-dire de l'idéologie scientifique qui à l'origine de la neuroéconomie,plusieurs gestes spéculatifs seront mis en évidence. On observe tout d'abord un brouillage constant de la frontière entre l'homme et l'animal. Plutôt que de réduire le premier au second, le « dépassement présomptueux » consiste ici à projeter dans l'animal des expériences spécifiquement humaines, et notamment celles liées à la maladie. Cela a pour conséquence théorique d'élargir la notion de récompense et/ou de motivation, qui servent à désigner, chez l'homme comme chez l'animal, toute forme de gain subjectif pouvant être attendu d'une action. Cela explique par ailleurs que la notion de comportement addictif et/ou impulsif, qui peut se comprendre comme un trouble dans l'apprentissage des récompenses, fasse elle aussi l'objet d'une très forte extension au sein de ce programme de recherche.

Les notions de récompense et d'addiction peuvent ainsi se comprendre comme des « *objets hyperboliques* », faisant écho à l'intérêt porté par Canguilhem au concept de régulation au XVIIIème et XIXème siècles. Canguilhem observe que le terme apparaît dans plusieurs champs: en physiologie, mais aussi en astronomie, en théologie, en économie (avec Malthus), en philosophie avec Leibniz. Or, pour Canguilhem, de telles régularités énonciatives suggèrent l'idée d'une rupture importante, d'une cassure dans la pensée scientifique, remettant en cause l'organisation générale du savoir et le partage établi des disciplines. Le concept de régulation est un « *objet polyscientifique ou interscientifique. N'entendons pas par là un objet traité en commun par plusieurs disciplines, mais un objet construit expressément comme effet de leur collaboration* ». En ce qui concerne la neuroéconomie, il faudra donc étudier la dimension « transgressive » des concepts de

motivation, de récompense et d'addiction, tout en soulignant la fragilité de ces objets « *hyperboliques* ». Le brouillage des frontières entre l'animal et l'homme sera de ce point de vue un élément important de l'idéologie scientifique dont il sera question ici.

ii. « *Il y a toujours une idéologie scientifique avant une science dans le champ où la science viendra s'instituer; il y a toujours une science avant une idéologie, dans un champ latéral que cette idéologie vise obliquement* » (Canguilhem, 1977, p.44)

Cette remarque est double. D'une part, l'idéologie scientifique est antérieure à la science proprement dite, qui viendra prendre sa place dans le champ du savoir. De ce point de vue, l'idéologie scientifique remplit une fonction positive dans le progrès des connaissances, en stimulant l'intérêt des chercheurs, en orientant vers des problèmes à résoudre, *etc.* Par exemple, à propos de Claude Bernard, Canguilhem observe: « *on peut même, à la rigueur, admettre que l'obstination de Claude Bernard à identifier maladie et empoisonnement, à chercher la maladie dans l'altération toxique, sous l'action du système nerveux, des éléments du milieu intérieur où baignent les cellules, a pu préparer les esprits à comprendre que l'infection consistait dans la libération par les microorganismes de chaque espèce d'une toxine propre* ». De la même façon, au sein de la préhistoire théorique de la neuroéconomie, la référence à un vocabulaire économique dans les travaux de Glimcher des années 1990, ou dans ceux de la science quantitative de la motivation dans les années 1960 et 1970, ont d'une certaine manière préparé les esprits des économistes à envisager une forme de collaboration avec les neurobiologistes, même si la forme exacte de cette collaboration reste à définir.

D'un autre côté, la science -ici, l'économie- est antérieure à l'idéologie, car cette dernière s'inspire toujours d'un programme de recherche mieux établi, dont elle essaye d'étendre la portée théorique au delà de son champ d'application initial. Les neurobiologistes, les psychiatres et les psychologues néo-comportementalistes ont effectivement commencé à considérer dans les années 1990 que l'analyse économique constituait un nouveau paradigme théorique adéquat pour leur discipline. Ils ont alors essayé d'appréhender des troubles du comportement à partir de processus d'optimisation. Glimcher fournit par exemple l'illustration d'un neurobiologiste qui, pour reprendre les termes de Canguilhem, « *louche du côté d'une science déjà instituée, dont elle reconnaît le prestige et dont elle cherche à imiter le style* ». La revendication d'une proximité avec l'économie apparaît également chez Antonio Damasio, qui affirme la possibilité de diagnostiquer des troubles neuro-comportementaux à partir de

protocoles inspirés de la théorie des jeux (Damasio, 2008).

iii. « *L'idéologie scientifique ne doit pas être confondue avec les fausses sciences, ni avec la magie, ni avec la religion* »(Canguilhem, 1977, p.44)

L'inspiration économique revendiquée par les chercheurs néo-comportementaliste est en fait assez trompeuse. En dépits de l'utilisation de termes empruntés au vocabulaire économique (maximisation sous-contrainte, utilité espérée notamment), la formalisation des processus du *reward learning* n'a pas fait l'objet auparavant de recherches approfondies par les économistes. Ce problème est nouveau pour la théorie économique. Les expressions d'« *utilité espérée physiologique* » chez Glimcher (Glimcher, Dorris et Bayer, 2005) ou de « *cerveau bayésien* » (Wacongne, Changeux et Dehaene, 2012; Moreno-Bote *et al.*, 2011) ont donc été à l'origine d'incompréhensions chez les économistes, car les questions abordées échappent au domaine couvert par la microéconomie. Une deuxième source de confusion a été l'apparition, au début des années 2000, de l'économie comportementale dans le scanner (Ross, 2008), qui peut se comprendre comme une tentative des économistes comportementalistes pour importer dans leur programme de recherche certains résultats expérimentaux des neurosciences. Or le néo-comportementalisme est largement incompatible avec l'approche kahnemanienne, et l'assimilation des neurosciences au sein des *behavioral economics* s'est souvent effectué au prix d'importantes déformations et approximations théoriques.

Pour autant, l'économie comportementale dans le scanner, et la référence à l'économie chez les neurobiologistes et psychologues, quoique hasardeuses, n'étaient pas non-scientifiques, en tant précisément qu'elles ont permis d'amorcer un dialogue théorique entre neuroscientifiques et économistes. Le propre de la non-science, souligne Canguilhem, est de ne jamais rencontrer le faux: c'est la raison pour laquelle la non-science n'a pas d'histoire. L'idéologie scientifique occupe quant à elle une place dans le champ de la connaissance. Cette place est toutefois déplacée, et la science qui y succédera la déplacera sensiblement. Ainsi, il faudra considérer, à propos de la neuroéconomie, que les approximations théoriques initiales sont des conditions de possibilité du discours scientifique: c'est en respectant cette exigence qu'il sera possible d'échapper à la visée prescriptive et évaluatrice selon laquelle une discipline en construction ne saurait être scientifique.

Conclusion du chapitre 1

La neuroéconomie est le plus souvent abordée dans la littérature secondaire selon un point de vue interne à la discipline. Il s'agit d'évaluer le contenu et l'apport de la neuroéconomie à la pensée scientifique (économique ou neurobiologique). Un premier groupe d'auteurs a d'abord cherché valoriser et promouvoir la neuroéconomie comme nouvelle branche de l'analyse économique (Camerer, Loewenstein et Prelec, 2005;Glimcher *et al*., 2008 ; Bourgeois-Gironde, 2008 ; Schmidt, 2010; Glimcher, 2010). Ces lectures, souffrant de plusieurs lacunes, ont nourri un certain scepticisme. Quasi-simultanément, la neuroéconomie a fait l'objet de vives critiques de la part d'économistes expérimentalistes et/ou comportementalistes, qui ont pointé du doigt les faiblesses théoriques de ce nouveau programme de recherche (Gul et Pesendorfer, 2005; Harrison, 2008-a; Rubinstein, 2008-a). Quoique parfois virulentes, ces remises en question méthodologiques ont néanmoins joué un rôle important dans la structuration ultérieure de la discipline. La critique méthodologique interne est ainsi un moteur théorique essentiel de la neuroéconomie. La perspective internaliste peut donc produire, à partir d'un matériel limité aux seuls éléments analytiques, une description cohérente et structurée de la neuroéconomie. Les récits internes ne visent cependant nullement à répondre à un questionnement historique. Il s'agit de structurer les résultats théoriques, sans expliquer l'origine même de cette structuration.

La démarche proposée ici vise à rendre compte de l'émergence de la neuroéconomie d'un point de vue historique. Elle repose sur l'application de la notion d'idéologie scientifique, proposée par Canguilhem (1977), au projet théorique visant à élaborer une modélisation économique des troubles mentaux. Cette idéologie a donné naissance dans les années 2000 à une nouvelle branche de l'analyse économique, qui se comprend du point de vue des économistes comme une approche pathologique du comportement économique (*cf.* introduction).

L'expression d'idéologie entretient une équivoque et doit être soigneusement distinguée de la rhétorique, mise en avant par McCloskey (1983), ou plus récemment par Mäki (2010) à propos de la neuroéconomie. La rhétorique ne désigne en effet qu'un habillage d'un discours scientifique qui peut expliquer certaines contingences de la vie académique, certains « effets de mode », mais ces auteurs considèrent néanmoins qu'il est possible d'identifier, en deçà des artifices de persuasion, un noyau de « bonne » théorie, objective et respectueuse des normes de la conversation académique. Or l'attention portée au problème de

la pathologie n'est pas du tout une déformation de la neuroéconomie. Elle désigne bien au contraire sa condition de possibilité théorique. Il n'y aurait pas eu de neuroéconomie sans, au départ, cette intention d'appréhender dans des termes économiques le problème de la maladie mentale.

Une autre solution consiste alors à rapprocher la neuroéconomie de son contexte d'apparition et de diffusion. Effectivement, le projet théorique de la psychiatrie économique est en lien avec une nouvelle conception de la maladie mentale, qui se développe à la suite de l'échec d'un mode de gestion asilaire en psychiatrie. Pourtant, la neuroéconomie n'apporte *in fine* aucune solution directe à ce problème, pas plus qu'elle n'apporte de débouchés intéressants pour l'industrie pharmaceutique. La modélisation économique des troubles mentaux a été élaborée dans une intention purement théorique. Elle n'a donc pas été dictée par des contingences sociales. Au final, l'idéologie scientifique qui est à l'origine de la neuroéconomie désigne la pure intuition théorique d'une synthèse entre deux domaines de savoir hétérogènes, l'économie et la médecine mentale. Ce rapprochement a été au départ assez hasardeux, car l'économie n'a joué aucun rôle théorique dans ce mouvement; mais l'inspiration économique revendiquée des chercheurs appartenant au mouvement néo-comportementaliste a bien conduit, dans les années 2000, à l'inscription de la neuroéconomie comme nouvelle branche de l'analyse économique.

PREMIERE PARTIE – PSYCHIATRIE ECONOMIQUE ET NEO-COMPORTEMENTALISME: LA PREHISTOIRE THEORIQUE DE LA NEUROECONOMIE (1955-1999)

La neuroéconomie trouve son origine théorique dans un ensemble de travaux expérimentaux en psychologie et en économie portant, depuis les années 1960, sur l'analyse des décisions séquentielles. Les études dans ce domaine sont essentiellement réalisées sur des animaux de laboratoire. Ce courant de recherche, que nous avons choisi d'appeler « science quantitative de la motivation » ou « néo-comportementalisme », constitue la préhistoire de la neuroéconomie (*cf.* introduction).

La filiation théorique entre neuroéconomie et ce programme de recherche antérieur est assez claire. Initialement, les neuroéconomistes ont en effet essayé de reproduire, chez l'homme, certains résultats obtenus sur le pigeon, en particulier la « *loi d'égalisation* » des rendements (*matching law*) formulée par Richard Herrnstein en 1961 (Herrnstein, 1961). Les neurobiologistes se sont inspirés à la fois des protocoles expérimentaux, impliquant des choix répétés entre deux options générant des récompenses variables, et de formalisations « d'inspiration économique »[55] (maximisation sous contrainte, utilisation de fonctions d'actualisation hyperboliques, algorithmes d'apprentissage). La plupart des ingrédients formant le noyau dur de la neuroéconomie sont donc déjà repérables au sein de cette tradition d'analyse psychologique des choix dynamiques. La dette théorique à l'égard des pionniers de cette approche est évidente et pleinement assumée par les neuroéconomistes contemporains (voir notamment Glimcher, 2003).

L'ensemble des travaux expérimentaux abordés dans cette première partie suggère donc que, préalablement à l'introduction de la neuroimagerie, un certain nombre de réflexions analytiques propres à la discipline sont déjà présentes. Celles-ci ont rendu possible l'application ultérieure de ces nouveaux outils dans le domaine. C'est en particulier au sein de ce programme recherche que sont élaborés les concepts de motivation et d'apprentissage de la récompense. Sans anticiper sur les développements analytiques, plusieurs points communs à tous ces travaux expérimentaux et à la neuroéconomie permettent d'identifier la structure

55 Comme cela a été signalé en introduction, cette inspiration économique est trompeuse, puisque les formalisations en question n'ont pas été développées par des économistes, mais bien par des psychologues ou des neurobiologistes.

générale de ce domaine de recherche.

(i) un nouvel agent économique: l'animal de laboratoire

Un premier élément important concerne la forte teneur biologique et « animale » de ce programme de recherche. Les expériences portent en effet très majoritairement sur des animaux de laboratoire, et les phénomènes biologiques observés (comportements, choix, activations neuronales) sont considérés comme directement pertinents pour la théorie économique de la décision (voir par exemple Glimcher, 2003). Il y a là une référence à la biologie en économie qui est tout à fait originale et novatrice, en tant qu'elle n'est pas, comme c'est le cas par exemple chez Schumpeter ou chez Alchian, simplement métaphorique. Il ne s'agit pas de penser le processus concurrentiel *à la manière de* l'évolution biologique, ou par analogie avec celle-ci: l'observation expérimentale du comportement doit servir de théorie universelle du choix économique[56]. Le biologiste a vocation à statuer directement sur des questions d'ordre économique.

L'importance du biologique et de l'animal constitue ainsi une spécificité forte de ce programme de recherche, mais elle est également une importante source de confusion. En effet, cette approche paraît au premier abord extrêmement naïve: comment l'observation d'un pigeon qui donne des coups de bec sur un levier pour obtenir de la nourriture pourrait-elle nous dire quoi que ce soit du comportement humain? Les économistes (et la plupart des individus) sont ainsi généralement rétifs à admettre l'idée selon laquelle ces expériences pourraient révéler un trait, même le plus grossier, de la rationalité individuelle. Or en fait les études ne visent pas à réduire le comportement humain à des processus biologiques partagés avec l'animal. La démarche est inverse: l'objectif pour les psychologues et biologistes est de montrer que les animaux, placés dans les conditions idéales du laboratoire, exhibent une forme de rationalité, ou, en d'autres termes, que l'instinct animal se constitue comme intelligence, en tant qu'il est adapté à son environnement.

56 Pour éviter toute ambiguïté relative à la notion de « décision économique » (*cf.* introduction), nous utiliserons les expressions de choix ou de comportement économique pour faire référence à la décision économique au sens matériel, c'est-à-dire au sens d'actes économiques réels (achat, vente, négociations, coopération, *etc.*) pouvant être simulés en laboratoire. Cette acception se distingue de celle qui est impliquée dans l'expression de « théorie de la décision »: ici, le sens est formel et le terme de décision renvoie à un problème abstrait de la théorie économique ou mathématique, pouvant ou non avoir une applicabilité dans la sphère des comportements économiques. L'analyse de la motivation et de l'apprentissage n'appartenait pas, au départ, à la théorie de la décision et n'y a été intégrée qu'*a posteriori*. Lorsque les psychologues néo-comportementalistes considèrent que leurs travaux relève de l'économie (voir par exemple Ainslie, 2001, p.113), c'est donc au sens où les comportements étudiés en laboratoire peuvent être pris comme une simulation d'actes économiques réels, sans que la théorie ou le modèle proposé ne fasse référence directement à un problème d'analyse économique.

Cette approche vise donc moins à simplifier le comportement humain qu'à souligner la richesse de l'intelligence animale. Cependant, le rapprochement entre cette intelligence animale et la rationalité économique, c'est-à-dire la rationalité telle qu'elle est comprise dans les termes de l'analyse économique, est une nouvelle fois source de confusion. En effet, la première est de nature évolutive : un comportement intelligent exige, de manière minimale pour le biologiste, que les réponses de l'organisme aux stimuli externes soit adaptées à l'environnement *sur le long-terme*. L'idée d'une rationalité évolutive, qu'elle soit qualifiée d'« *adaptative* » dans les termes du psychologue Gerd Gigerenzer (2000) ou d'« *écologique* » chez Vernon Smith (2007), constitue ainsi un deuxième élément d'identité théorique important.

(ii) une approche évolutionniste et séquentielle

La référence à l'évolution entraine une compréhension complètement différente de la rationalité, puisqu'elle oblige à un traitement non seulement dynamique, mais séquentiel de la décision. Kahneman et Tversky ont quant à eux consacré notamment d'importants travaux au problème de l'incohérence dynamique (Kahneman, Tversky, et Slovic, 1990), qui porte principalement sur l'écart pouvant exister entre la disposition à payer (*willingness to pay*) et la disposition à vendre (*willingness to sell*): par exemple, je suis disposé à acheter la loterie *A* à un prix plus élevé que la loterie *B*, mais, si je possède ces deux loteries, je souhaite vendre la loterie *B* à un prix plus élevé que la loterie *A*. La théorie kahnemanienne de la décision n'est donc nullement limitée à une simple analyse des choix statiques.

La dynamique des décisions, chez Kahneman, est ainsi envisagée à travers le problème de la compatibilité entre plusieurs choix discrets, mais la dimension séquentielle n'est pas prise en compte. Dans les problèmes séquentiels, les décisions doivent être prises par étapes successives, dans un environnement pouvant ou non être aléatoire. A chaque étape, une alternative doit être choisie parmi un nombre fini d'alternatives. Dans le cas aléatoire, les « machines à sous multi-jeux » (*multi-armed bandit problems*) constituent un type de problème séquentiel qui a beaucoup été étudié dans la littérature . Comme leur nom l'indique, ces jeux consistent en un choix successif entre plusieurs machine à sous. A chacune d'entre elles correspond un gain moyen qui lui est propre. Lorsque l'une d'entre elles est sélectionnée, celle-ci produit un gain aléatoire. L'individu doit maximiser son gain total, en supposant qu'il dispose soit d'un temps limité, soit d'un nombre de jetons limité. Les *multi-armed bandit problems* peuvent être étudiés, chez l'animal, en utilisant des récompenses alimentaires. Ils

forment le principal type de protocole pour les expériences qui sont étudiées dans cette première partie.

La dimension séquentielle de ce type de problèmes a deux conséquences théoriques importantes, qui distinguent nettement cette approche de la psychologie d'inspiration kahnemanienne. Tout d'abord, la décision séquentielle en environnement aléatoire correspond non pas à un choix risqué, mais un choix incertain, car les probabilités sont au départ inconnues du décideur (Cohen et Tallon, 2000, p.36). Ne connaissant pas au départ les distributions de gain associés à chaque machine à sous, le joueur doit en effet progressivement collecter de l'information pour améliorer ses décisions et détecter éventuellement la machine la plus avantageuse. Cependant, l'incertitude se réduit au fur et à mesure du déroulement du jeu, jusqu'à se réduire à du quasi-certain après un grand nombre répétitions. Par conséquent, la modélisation d'un problème séquentiel implique la modélisation des dynamiques acquisition et de traitement d'information. En particulier, un compromis doit être trouvé entre l'exploitation -choisir la machine qui semble pur l'instant offrir le gain optimal- et l'exploration -tester les machines peu connues, qui pourraient se révéler être plus avantageuses. Il y a donc, en amont, une réflexion sur la procédure de prise de décision (explorer ou exploiter) qui semble la plus adaptée.

A l'inverse, chez Kahneman, la procédure est toujours la même, et consiste à appliquer une fonction d'espérance d'utilité aux alternatives considérées. La dynamique, comprise comme la compatibilité entre plusieurs choix discrets, est donc traitée comme un problème purement calculatoire: par exemple, l'application de ma fonction d'espérance d'utilité me conduit à préférer A à B en t_0 et B à A en t_1, mais ce phénomène s'explique sans que la fonction d'espérance d'utilité soit modifiée entre t_0 et t_1. Dans un problème séquentiel, le joueur peut préférer la machine A à la machine B en t_0, et B à A en t_1, sans qu'il s'agisse nécessairement d'incohérence dynamique au sens où l'entend Kahneman, puisque ce choix en apparence incohérent peut s'expliquer par la nécessité de collecter de l'information relative aux gains de chaque machine. L'étude de la rationalité procédurale[57], c'est-à-dire de la dimension

57 Cette expression fait bien sûr référence à la notion de « rationalité procédurale » proposée par Herbert Simon, que celui-ci distingue de la rationalité substantive (*cf.* Simon, 1976). Pour Simon, la rationalité « substantive », telle qu'elle est analysée par les économistes néo-classiques, suppose que l'individu dispose de toute l'information pour prendre une décision optimale. A l'inverse, la rationalité procédurale définit des mécanismes, appelés heuristiques, par lesquels les individus acquièrent et traitent l'information pour prendre des décisions non pas optimales mais satisfaisantes (cf. Simon, 1976, p.132). Herbert Simon est une source d'inspiration importante pour la psychologie évolutionniste dont il est question ici. Vernon Smith et Gerd Gigerenzer en particulier le citent comme référence importante dans leurs ouvrages respectifs et lui reprennent cette distinction entre rationalité procédurale et substantive (Smith, 2007, p.39; Gigerenzer, 2000, p.730; Berg et Gigerenzer, 2010, p.159). En nous inspirant de la proposition de Philippe Mongin (1984), nous substituerons cependant ici aux expressions de rationalité procédurale et substantive celles de rationalités

délibérative de la décision, constitue donc un premier élément caractéristique des problèmes séquentiels qui sont étudiés par la psychologie évolutionniste

Par ailleurs, l'approche évolutionniste de la décision se distingue également dans sa conception de probabilités, qui correspond à ce que Hacking appelle une conception « *fréquentialiste* », par opposition à la conception -partagée, entre autres, par Kahneman- « *subjectiviste* » (Hacking, 1975). Dans la théorie kahnemanienne dite des perspectives (Kahneman et Tversky, 1979), les choix portent, typiquement, entre des loteries financières et/ou des gains discrets: par exemple, choix entre une loterie qui offre un gain de *X* euros avec une probabilité *p*, et un gain certain de *Y* euros. Ici, le terme de probabilité renvoie au degré de réalisation possible d'un état du monde (la loterie est gagnante, je gagne *X* euros). Tout l'enjeu consiste ensuite, dans la théorie des perspectives, à expliquer comment ces probabilités dites objectives (probabilité *p* de gagner la loterie) sont transformées en probabilités subjectives par une fonction individuelle de pondération des probabilités. Cependant, qu'elles soient objectives ou subjectives, les probabilités reflètent toujours la réalisation possible d'un événement discret (l'issue de la loterie m'est favorable).

Les psychologues évolutionnistes, comme Gerg Gigerenzer notamment, défendent une conception fréquentialiste. L'opposition entre ces deux approches des probabilités est au cœur de la controverse entre Kahneman, Tversky et le psychologue évolutionniste Gigerenzer (*cf.* Kahneman et Tversky, 1996; Gigerenzer, 1996)[58]. Pour les partisans du fréquentialisme, l'attribution d'une probabilité au sens subjectif, c'est-à-dire comme degré de confiance accordée à la réalisation d'un événement discret n'a pas de sens. Un événement discret doit toujours se réaliser ou ne pas se réaliser. Par exemple, un individu qui considère ses chances de gagner une loterie ne devrait pas utiliser l'expression de « gain probable », car il va de

calculatoire et délibérative. Comme le souligne Mongin, ce que Simon appelle « approche substantive » peut très bien prendre en compte, ou plutôt prendre pour objet d'étude les procédures de décision. Le modèle que Stigler a proposé en réponse à Simon, par exemple, incorpore au sein du modèle néoclassique les méthodes de recherche d'information envisagées par Simon. Mongin y voit un « *modèle néoclassique du joueur d'échec* » (Mongin, 1984, p.575). Or, comme le souligne Mongin, les objections les plus sérieuses de la critique simonienne portent non pas tant sur la reconnaissance ou l'existence de procédures d'acquisition et de traitement d'information, mais plutôt sur la manière de les modéliser. Dans l'approche substantive -celle de Stigler- la recherche et le traitement d'information représentent uniquement un cout supplémentaire, qu'il convient d'incorporer au calcul d'utilité; ce qui revient en fait à ignorer la dimension délibérative de la prise de décision, c'est-à-dire l'existence de plusieurs stratégies possible pour traiter l'information. Comme l'écrit Mongin, « *le propre du raisonnement calculatoire est qu'il opère sur des symboles qui sont physiquement disponibles: il transforme du donné en donné […] Le raisonnement délibératif n'opère pas exclusivement sur des données, il consiste pour partie en une invention d'objets* » (Mongin, 1984, p.598). De la même façon, le courant néo-comportementaliste traite de problèmes d'apprentissage qui peuvent tout à fait être incorporés dans un modèle microéconomique élémentaire. L'originalité des travaux bordés ici ne consiste donc pas à proposer un modèle économique de l'apprentissage, mais à proposer des règles formelles -des algorithmes- pour rendre compte des différentes stratégies de *learning*.

58 Pour une analyse de cette controverse, voir Jullien et Vallois, 2012.

manière certaine, avec cette loterie particulière, gagner ou ne pas gagner. C'est au regard de plusieurs observations, ou plusieurs participations à une loterie identique, que le terme de probabilité doit s'appliquer. Comme l'écrit Gigerenzer, « *dans la conception fréquentialiste, on ne peut parler de probabilité à moins qu'une classe référence ait été définie* » (Gigerenzer, 1993, p.292-293). Pour un fréquentialiste comme Richard von Mises par exemple, « *l'expression « probabilité de décès », lorsqu'elle est référée à une unique personne, n'a aucun sens* » (von Mises, 1928, p.11), puisque chaque individu finit tôt ou tard par mourir. La probabilité de décès renvoie plutôt à la fréquence observée des décès parmi un groupe d'individus ou une population de référence.

De la même façon, les auteurs étudiant des problèmes séquentiels du type *multi-armed bandit* désignent par l'expression de « *probabilité de récompense* » (*reward probability*) non pas un degré de confiance accordé par l'individu dans un choix discret (probabilité *p* que l'alternative A soit gagnante à l'étape *n*) mais la fréquence avec laquelle l'alternative considérée a délivré un gain au cours de l'ensemble du jeu. Il n'est pas nécessairement absurde de considérer, à l'inverse de von Mises, que l'individu puisse attribuer une probabilité au sens subjectif, c'est-à-dire à une réalisation discrète : à l'étape *n*, le joueur considère que la machine A est gagnante avec une probabilité *p*. Pourtant, à proprement parler, il ne s'agit pas d'un aléa ou d'un choix risqué, car les choix discrets ne satisfont pas les lois de probabilité[59]. Un individu ne choisit pas toujours l'alternative qui lui apparaît comme

59 De ce point de vue, il est possible de distinguer entre le fréquentialisme « radical » de von Mises, pour qui les probabilités appliquées aux évènements discrets n'ont aucun sens, de positions plus modérées exprimées par des psychologues évolutionnistes comme Gigerenzer. Comme le souligne Vranas (2000), la véritable opposition entre Kahneman, Tversky et Gigerenzer ne porte pas sur le fait de savoir si les probabilités subjectives et/ou fréquentielles ont un sens ou non. Gigerenzer reconnaît que les individus puissent concevoir les probabilités comme des croyances subjectives (Gigerener, 1996, p.594), et, inversement, Kahneman et Tversky ne rejettent pas forcément l'idée selon laquelle les individus puissent dans certaines circonstances concevoir les probabilités en termes de fréquences (Kahneman et Tversky, 1996, p.583). Le cœur de la controverse porte en fait sur le statut des normes, c'est-à-dire sur les règles que les individus doivent adopter pour prendre des décisions probabilistes optimales. Pour Kahneman et Tversky, les lois gouvernant le choix probable, comme la règle de Bayes ou la règle de conjonction par exemple, ont une validité normative indiscutable, ce qui signifie que ces règles doivent toujours être respectées pour que le choix soit rationnel. Par exemple, si le bulletin météo m'informe qu'il y aura du beau temps demain avec une probabilité de 80% et si je sais par ailleurs que la météo est fiable à 50%, je dois m'attendre à ce qu'il fasse beau demain avec une probabilité de 40%. Pour Gigerenzer, ce calcul peut avoir un sens pour l'individu, mais la violation de ces règles de raisonnement n'est pas, pour les portabilités subjectives, toujours irrationnelles. Une règle de décision comme la règle de Bayes est pour Gigerenzer une « règle de manuel » (t*exbook norm*), ce qui signifie qu'elle est dépourvue de toute signification concrète (*cf.* Gigerenzer, 1996, p.593). En situation de choix réel, il n'est pas toujours pertinent de prendre en compte le taux de base (*base rate*) ou degré de fiabilité des informations à disposition. Celui-ci est ajusté selon les circonstances Si tous les bulletins météo convergent vers la même prédiction, alors qu'ils offrent habituellement des prédictions divergents, je peux raisonnablement m'attendre à ce qu'il fasse beau avec une probailité proche de 80%. Pour Gigerenzer par ailleurs, i toutes les chaines de météo convergent De la même façon, dans les machines à sous multi-jeux, un fréquentialiste radical comme von Mises considèrerait que la probabilité de gain appliqué à un choix discret n'a pas de sens, parce qu'à chaque étape, la loterie considérée doit nécessairement être gagnante ou perdante, il n'y a pas de degré de réalisation possible d'un événement unique. Pour un fréquentialiste plus modéré, l'a

étant la plus probable en l'état actuel de ses connaissances, puisque le choix en faveur d'une alternative moins probable peut améliorer ses connaissances et ainsi ses choix futurs (stratégie d'exploration).

(iii) la notion de pathologie

Enfin, et de manière décisive, la référence à la notion de pathologie constitue un élément fédérateur déterminant au sein de ce programme de recherche. L'évidence absolue selon laquelle certains comportements, chez l'homme comme chez l'animal, sont fondamentalement inadaptés joue le rôle à la fois de pierre fondatrice et d'horizon régulateur pour l'analyse. En partant de la considération de cas manifestes de déviances, les expérimentateurs sont ensuite en mesure de montrer comment, dans les cas jugés normaux, l'intelligence de l'instinct se manifeste. Sans rentrer dans les détails, on voit apparaître ici une complication des rapports entre rationalité et irrationalité, normalité et pathologie, économie et biologie.

Cette référence triple -à l'animalité, à l'évolution, à la pathologie- constitue donc le noyau dur de cette approche interdisciplinaire, qui se forme à la fin des années 1950. Il s'agit ici de vérifier notre hypothèse interprétative, en montrant comment ces trois références ont progressivement donné naissance à une approche pathologique du comportement en économie. Notre analyse débouche par ailleurs sur deux propositions corollaires signalées plus haut: d'une part, le programme néo-comportementaliste dans lequel s'inscrivent les neuroéconomistes se distingue dès le départ de la psychologie kahnemanienne; d'autre part, le rôle joué par la modélisation économique est assez mineur, et ce en dépits d'une inspiration économique revendiquée par les chercheurs concernés (*cf.* introduction).

Cette approche pathologique du comportement économique a été élaborée par des chercheurs ayant l'ambition de comprendre la maladie mentale dans des termes empruntés à l'économie. La conception économique de la maladie mentale sera considérée comme une idéologie scientifique au sens de Canguilhem et constitue donc la structure déterminante de notre enquête (*cf.* chapitre 1, section II). Cette structure théorique a été élaborée progressivement dans les années 1960 et 1970 autour des expériences pionnières de la « science quantitative de la motivation ». Un champ de recherche commun à l'économie, la

probabilité appliqué au choix discret peut avoir un sens pour l'individu, mais il est absurde d'exiger que celui-ci applique un véritable raisonnement probabiliste à ces portabilités, puisqu'il peut être rationnel, dasn une logique d'exploration, de ne pas choisir l'alternative la plus probable.

psychologie et la biologie émerge ainsi autour de protocoles expérimentaux et de concepts (impulsivité, *self control*, *etc.*) communs et, notamment, autour d'une vision originale de la rationalité. Cette ouverture à l'économie génère aussi les premières confusions théoriques (chapitre 2. Impulsivité et choix inter-temporel : la naissance d'une idéologie scientifique en sciences comportementales).

Dans les années 1990, une importante innovation technologique transforme ce programme de recherche. L'utilisation de micro-électrodes permet en effet, chez le singe et le rat, d'étudier l'activité neuronale dans des protocoles relativement proches. Cette transformation technologique implique ainsi un nouveau niveau d'analyse -neurobiologique- et un nouveau type de données expérimentales -l'activité neuronale. Néanmoins, cette approche qui débouche directement sur les premières expériences de neuroéconomie au début des années 2000, doit être comprise dans la continuité de la science quantitative de la motivation, en tant qu'elle participe d'une idéologie scientifique commune (chapitre 3. Les années 1990 : La vocation économique de la neurobiologie – Paul Glimcher et l'« utilité espérée physiologique »).

Chapitre 2. Impulsivité et choix inter-temporel : la naissance d'une idéologie scientifique en sciences comportementales

La neuroéconomie trouve son origine théorique dans un ensemble de travaux expérimentaux en psychologie des années 1960. Ces études, réalisées en laboratoire sur des animaux (sur le rat, et surtout sur le pigeon), visent à mettre en évidence une relation quantitative entre les choix et les récompenses. Nous avons choisi de désigner ce courant de recherche par l'expression de « science quantitative de la motivation » ou de « néo-comportementalisme », car les auteurs concernés comprennent leurs travaux, comme un approfondissement des recherches comportementalistes de Benjamin Skinner sur le conditionnement.

Néanmoins, si le néo-comportementalisme s'inscrit lui-même dans la continuité du comportementalisme, et si, donc, la préhistoire de la neuroéconomie a elle-même sa préhistoire, le programme de recherche étudié dans le cadre de ce chapitre se distingue néanmoins du béhaviorisme traditionnel de Skinner. A la différence de ce dernier, l'analyse quantitative de la motivation vise à établir un rapprochement avec l'économie. La relation choix/récompenses étudiée par les psychologues fournit en effet un éclairage original du choix inter-temporel et de l'impulsivité. Or ces problématiques constituent en parallèle un objet de recherche laissé à l'abandon par les économistes, depuis l'article fondateur de Robert Strotz, qui, dès 1955, fournit un cadre de réflexion analytique pour la dynamique des décisions. La modélisation de la relation choix/récompenses par les néo-comportementalistes a permis d'approfondir le problème des décisions dynamiques. Ce projet aboutit finalement, dans les années 1980, à l'émergence d'une nouvelle branche de l'analyse économique consacrée à l'étude expérimentale du choix inter-temporel et de l'impulsivité.

Ultérieurement, les électrophysiologistes dans les années 1990, (*cf.* chapitre 3), et les neuroéconomistes dans les années 2000, reprennent cette même démarche, qui consiste à utiliser une approche économique, c'est-à-dire des processus d'optimisation sous contrainte, pour interpréter des données biologiques et comportementales. La dette théorique à l'égard des pionniers de la science quantitative de la motivation est évidente et pleinement assumée par les neuroéconomistes contemporains (voir notamment Glimcher, 2003). Le projet théorique visant à construire une approche de l'impulsivité inspirée de l'économie est donc

porteur des germes théoriques de la neuroéconomie.

Toutefois, il ne s'agit pas ici simplement de mettre en évidence un héritage théorique entre néo-comportementalisme et neurosciences, réalisé sous le patronage de l'économie. La compréhension économique de l'impulsivité sera comprise comme une « idéologie scientifique », au sens qui en a été donné dans le chapitre précédent. S'il y a idéologie, ce n'est pas parce que ce programme de recherche sert des intérêts socio-économiques (pharmaceutiques par exemple), mais parce qu'il suppose au départ, pour reprendre les termes de Georges Canguilhem, un « *dépassement présomptueux* » du scientifique (Canguilhem, 1977, p.23). Le geste spéculatif qui précède la réflexion théorique consiste ici à envisager la possibilité d'une articulation entre psychopathologie, biologie évolutionnaire et économie, désignée par l'expression de « psychiatrie économique » (*cf.* chapitre 1). Les modèles néo-comportementalistes de l'impulsivité représentent une tentative pour comprendre dans des termes économiques un phénomène qui, en psychologie et en biologie, est envisagé comme un trouble mental ou une déviance. Un tel rapprochement ne va pas de soi car il suppose la transgression des frontières admises entre l'homme et l'animal, et entre rationalité biologique, évolutive d'une part, économique d'autre part.

L'intuition théorique d'une psychiatrie économique provient notamment de la projection et de l'identification d'expériences humaines sur des anomalies biologiques. C'est une intuition similaire qui est reprise et approfondie par la neuroéconomie. La démarche consiste à isoler une tendance biologique (neuronale ou comportementale), et à interpréter celle-ci comme mécanisme descriptif d'un trouble mental humain (*cf.* P.H. Castel, 2010, p.26-27), en l'occurrence un trouble de l'impulsivité. La psychiatrie économique avance ainsi l'idée d'un dérèglement animal (ou neuronal) de l'impulsivité chez l'homme. Elle repose, par conséquent, sur un saut de l'animal à l'homme (et plus tard du cerveau à l'esprit), dont la pertinence peut être bien sûr discutée, mais qui est autorisé par la théorie économique, instrument de généralisation à l'individu et au social des résultats biologiques. L'essentiel du débat théorique dans ce projet concerne donc la modélisation économique des données expérimentales.

Le caractère idéologique de ce champ d'études expérimentales, compris ainsi de manière théorique, implique alors une assez large marge d'interprétation des données obtenues en laboratoire. Pour autant, l'existence d'une marge de manœuvre interprétative n'implique nullement que ce programme de recherche soit dépourvu de cohérence interne. La modélisation économique par des processus de maximisation conduit bien à une nouvelle forme de rationalité de la maladie mentale. Il conviendra d'en explorer les implications

théoriques et pratiques, à la frontière entre santé mentale et politique économique. Ce travail est donc complémentaire de celui des sociologues, qui, dans les années 1970, pointaient l'émergence d'une gestion économique de la santé mentale, se substituant au modèle de l'asile (Castel, Castel et Lovell, 1979).

Une présentation générale de la science quantitative de la motivation et de ses ambitions théoriques permet tout d'abord de mettre en évidence, au sein de ce programme de recherche, une ambition théorique en économie. L'idéologie scientifique dont est porteur le néo- comportementalisme peut à ce titre être rapprochée avec les travaux du sociologue Robert Castel portant sur l'idée d'une « conception économique de la maladie mentale ». Si l'éclairage avec la sociologie se révèle ici pertinent, il reste néanmoins à préciser la manière avec laquelle ce projet de « maximisation de la normalité », pour reprendre les termes de R. Castel, se constitue en tant que théorie (I. La psychiatrie économique comme idéologie scientifique). Au départ, l'analyse expérimentale et quantitative de la motivation fournit une explication originale de l'incohérence séquentielle et du choix inter-temporel (II. Choix inter-temporel et impulsivité : la reprise d'un problème théorique en économie par la psychologie néo-comportementaliste). Malgré d'importantes critiques internes, le ralliement à la modélisation économique, souhaité par certains psychologues, aboutit dans les années 1980 à faire du choix inter-temporel et de l'impulsivité un objet d'étude central pour l'économie dite comportementale, en plein essor. Cet intégration pose cependant problème, en raison des écarts théoriques importants existants entre d'une part le programme de recherche de l'économie comportementale, largement influencé par Daniel Kahneman, et la psychologie néo-comportementale d'autre part, d'inspiration biologique et évolutionnaire (III. Néo-comportementalisme et économie : un projet inter-disciplinaire contesté). Le néo-comportementalisme est ainsi traversée par une tension originaire: en s'inspirant de la modélisation économique, son ambition théorique consiste à proposer une approche quantitative d'un trouble mental. Mais l'intuition fondatrice de l'analyse repose sur l'évidence psychologique selon laquelle l'impulsivité est néfaste pour l'individu. Si la science quantitative de la motivation peine à s'inscrire dans le champ théorique de l'économie, c'est parce que le projet de psychiatrie économique est porteur, fondamentalement, d'une remise en question radicale de toute théorie normative du choix rationnel: la pathologie fournit un substitut négatif à l'absence de norme (IV. La disparition de la norme économique : la pathologie comme évidence fondatrice de l'analyse).

I. La psychiatrie économique comme idéologie scientifique

La science quantitative de la motivation est née de la volonté d'imiter le style de modélisation de l'économie, en faisant référence notamment à des mécanismes de maximisation sous-contraintes et à la notion d'utilité espérée, dans des disciplines extérieures à l'économie. Pour bien comprendre la démarche théorique des expérimentateurs, il semble important de souligner que ce rapprochement avec l'économie est d'abord souhaité par des psychologues et des biologistes, qui souhaitent reproduire dans leur discipline un type de formalisation utilisé en économie. Ce champ de recherche se construit ainsi à partir de l'intuition originaire d'une « médecine mentale à vocation économique » (A). Il faudra en suite préciser la portée et la signification de cette ambition, que l'on qualifiera ici d'idéologie biomédicale en référence aux « idéologies scientifiques » de Georges Canguihem (*cf.* chapitre 1). La volonté d'analyser la maladie mentale dans des termes économiques correspond en effet à une évolution récente de la médecine mentale, qui a fait l'objet de plusieurs études sociologiques. On s'intéressera en particulier aux travaux de Robert Castel. La sociologie n'est pas ici convoquée pour expliquer l'origine et le contexte d'élaboration d'un courant de pensée économique. Les analyses de Castel fournissent le point de départ pour un travail complémentaire d'approfondissement, sur le versant purement théorique, de la notion de psychiatrie économique (B).

A. La naissance d'une médecine mentale à vocation économique: la science quantitative de la motivation

Le rapprochement qui s'opère à partir des années 1960 entre économie et psychologie du choix impulsif s'effectue sous l'impulsion déterminante des psychologues. En effet, c'est à la faveur d'une reprise en main par la psychologie expérimentale du problème de la « myopie temporelle » des décisions qu'un important champ de recherches en économie sur l'impulsivité s'est développé (*cf.* section II.). D'emblée, les psychologues néo-comportementalistes concernés ont nourri l'ambition de se rapprocher de l'économie (1). Pourtant, les rapports de

ce programme de recherche à l'économie sont ambivalents. L'importance des préoccupations médicales et biologiques distingue le néo-comportementalisme d'une autre branche de l'économie influencée par la psychologie, l'économie comportementale (2).

1. Quantification des récompenses et utilité espérée

La nouvelle école du comportementalisme, qui se développe d'abord à Harvard sous l'impulsion de Richard Herrnstein dans les années 1960, puis par l'intermédiaire de ses élèves et successeurs (Howard Rachlin, George Ainslie, Goerge Loewenstein, Drazen Prelec notamment) entend développer une « *science quantitative de la motivation* » (Commons, 2001), à partir de l'hypothèse selon laquelle la fréquence ou le montant des récompenses (*rewards*) modifie quantitativement l'apprentissage et les comportements. Cette idée a d'abord une portée en psychologie expérimentale, car elle approfondit de manière important la théorie classique du conditionnement de Benjamin Skinner, et constitue ainsi, selon Howard Rachlin, un principe déterminant de comportementalisme « *moderne* »[60].

Or, cette hypothèse quantitativiste dans l'étude comportementale de la motivation suggère d'emblée un possible rapprochement avec la notion d'utilité espérée. En effet, si l'amplitude attendue des récompenses détermine la nature du comportement, l'individu doit bien, *ex ante* prédire ou anticiper le montant de récompense pouvant être obtenu en effectuant la tâche demandée par l'expérimentateur. Cette anticipation des récompenses évoque naturellement l'utilité espérée de la théorie économique.

Les travaux quantitatifs sur le conditionnement instrumental dans les années 1960 invitent donc spontanément à un rapprochement avec l'économie, et expliquent comment des psychologues venant d'horizons les plus divers, ont pu ainsi se convertir en économistes. Cependant, il importe ici de bien rappeler que la science quantitative de la motivation se distingue nettement d'un autre programme de recherche -les *behavioral economics*-, qui s'inspire de psychologie expérimentale, mais plus précisément de la recherche comportementale de la décision de Daniel Kahneman.

60 *Cf* Rachlin, Howard. 1991. *Introduction to Modern Behaviorism*. 3e éd. W.H. Freeman & Company,

2. Les rapports ambivalents du néo-comportementalisme à l'économie comportementale

La référence spécifique à la psychologie évolutionniste et aux sciences du vivant distingue l'approche néo-comportementaliste de celle proposée par Kahneman à partir de trois éléments caractéristiques: recours à des animaux de laboratoire dans les expériences, étude de problèmes séquentiels et de dynamiques d'acquisition d'information, mise en avant de la notion de pathologie empruntée au domaine médical (voir l'introduction à la première partie). Plus spécifiquement, les chercheurs néo-comportementalistes s'intéressent au départ à l'impulsivité, qui constitue une pathologie ou un trouble du comportement, et qui s'observe aussi bien chez l'homme que l'animal (voir par exemple l'expérience d'Ainslie, 1974). Cette pathologie consiste, pour reprendre les termes d'Ainslie à « *prendre une décision* […] *qui est par la suite regrettée* » (Ainslie, 2001, p.9).

Or, et c'est là la nouveauté du point de vue des chercheurs travaillant sur le conditionnement animal, cette anomalie comportementale peut être réfléchie dans les termes de l'analyse économique. En économie, l'impulsivité peut se concevoir comme un problème d'incohérence dynamique ou de « renversement des préférences ». La cohérence dynamique ou temporelle est respectée si les plans élaborés au début de la séquence sont maintenus au fur et à mesure de l'écoulement du temps, ce qui signifie, en d'autres termes, que les agents « *réalisent leurs intentions* » (*cf.* Cohen et Tallon, 2000, p.36). A l'inverse, les décideurs sont incohérents s'ils regrettent et modifient leurs choix exprimés *ex ante*.

La cohérence temporelle est donc généralement considérée en économie comme une exigence minimale de rationalité, dont la violation indiquerait une irrationalité manifeste (*cf* Machina, 1989, p.1637-1638). Les économistes estiment généralement que la cohérence doit constituer la norme de la rationalité individuelle, car des décideurs incohérents s'exposent à être « *exploités* » par des décideurs cohérents[61]. L'irrationalité que représentent les conduites impulsives peut donc être analysée dans le langage de l'économie théorique.

Il importe cependant de bien distinguer entre deux approches bien différentes de

61 Cet argument fait référence à ce qui est connu en théorie de la décision comme des cas de *Dutch book*, c'est-à-dire des paris qui assurent un gain certain quelle que soit l'issue du pari. De la même manière, un décideur cohérent peut réaliser un gain certain en échangeant avec un décideur dont les préférences se « renversent » au cours du temps. Supposons un tel individu A, avec une dotation initiale Y, et qui préfère, en t_0, X à Y, mais qui préfère ensuite, en t_1, Y à X. Par définition, il existe un montant minimal ε tel que $(X - \varepsilon)$ soit préféré par ce même individu A à Y en t_0. Un décideur cohérent pourra donc facilement tirer parti des déficiences de A, en lui proposant d'échanger en t_0 $(X - \varepsilon)$ contra sa dotation initiale Y. Il suffit ensuite d'attendre jusqu'en t_1, et de proposer à A d'échanger X contre sa dotation initiale Y. A n'ayant reçu dans l'échange précédent que $(X - \varepsilon)$, il devra rendre plus que ce qu'il avait obtenu, et aura donc perdu au final ε. A l'inverse, le décideur cohérent aura « exploité » A et lui aura extorqué ε.

l'incohérence dynamique en économie. La première correspond au traitement qu'en proposent Kahneman, Tversky et Slovic (1991). Le renversement des préférences renvoie notamment à l'observation selon laquelle les individus accordent une valeur plus élevée aux objets ou aux loteries qu'ils possèdent qu'à ceux qu'ils désirent acquérir: par exemple, je suis disposé à payer une somme plus élevée pour acquérir la loterie *A* plutôt que la loterie *B*, mais cette relation s'inverse dans l'hypothèse d'une vente. Cette anomalie peut s'expliquer par une fonction d'utilité espérée non-cumulative (Kahneman, Tversky et Slovic, 1991). Pour plus de clarté, nous désignerons ici par l'expression d' « incohérence dynamique » ce type d'anomalie purement calculatoire, dans laquelle la séquentialité ou la dimension temporelle de la décision ne joue aucun rôle. Si je préfère *A* à *B* en t_0, et l'inverse en t_1, cela n'a rien à voir avec l'écoulement du temps mais avec la modification es conditions du choix (position d'acquéreur ou de vendeur).

A l'inverse, les chercheurs néo-comportementalistes s'intéressent à une forme de renversement des préférences qui est liée à l'écoulement du temps: en t_0, je préfère *A* à *B*, et parce qu'un certain délai s'est écoulé entre t_0 et t_1, je préfère désormais *B* à *A* en t_1. La science quantitative de la motivation propose une explication temporelle ou séquentielle de l'incohérence, dans laquelle celle-ci résulte non pas d'une forme particulière de la fonction d'utilité, mais d'un type de procédure de décision adoptée fréquemment dans les problèmes de *multi-armed-bandit*, appelée routine de « mélioration ». Nous utiliserons donc spécifiquement pour ce type de renversement de préférence, afin de le distinguer des problèmes étudiés pour Kahneman (et par d'autres) les termes d'« incohérence séquentielle » ou d'« impulsivité ».

Le néo-comportementalisme ne traite donc pas au départ des mêmes problématiques théoriques que l'économie comportementale, en ce qui concerne le renversement des préférences. Il s'agit de deux démarches théoriques différentes, qui ont chacune leur propre histoire. Les premiers études néo-comportementales sont réalisées dans les années 1960, alors que l'économie comportementale ne se développe qu'à partir des années 1980 (Heukelom, 2009). Cependant, les rapports de la science quantitative de la motivation aux *behavioral economics* sont assez confus, car les néo-comportementalistes eux mêmes se considèrent assez souvent comme des économistes comportementalistes, ou plutôt se sont plutôt eux mêmes assimilés, plus tard, aux *behavioral economics*. En effet, la modélisation des comportements incohérents dans des machines à sous multi-jeux débouche, on va le voir, sur une analyse de ce que les économistes appellent « actualisation inter-temporelle hyperbolique ». Ce concept a été repris par la suite par les *behavioral economics*, qui en ont fait alors une notion fondamentale de leur approche du choix inter-temporel. Pourtant, il

faudra montrer que ce problème théorique apparaît en fait comme un élément hétérogène au sein du programme de recherche de Kahneman et Tversky (*cf.* section II).

Attraction et confusion sont ainsi les règles du rapport théorique qui unit la science quantitative de la motivation à l'économie. Les chercheurs au sein de ce champ jouissent d'un statut relativement incertain, puisqu'ils sont à la fois psychologues, voir même psychiatres, biologistes, économistes comportementalistes d'un genre bien spécial. C'est le cas notamment de George Ainslie, qui est une figure (théorique) incontournable dans le champ des études psycho-économiques sur l'impulsivité et le s*elf control*. Après avoir été formé comme médecin psychiatre, George Ainslie s'intéresse à partir des années 1960, ce qui est assez rare à l'époque pour un médecin, aux sciences comportementales. Il participe alors aux recherches menées par Richard Herrnstein, et travaille ensuite avec Howard Rachlin. Ses travaux, à la fois théoriques et expérimentaux, ont toujours porté sur l'impulsivité et son contrôle, chez l'homme comme chez l'animal. Néanmoins, on peut mettre en évidence une évolution progressive vers l'économie, qui culmine en 1992 avec la publication de son livre intitulé *Picoéconomie : l'interaction stratégique des états motivationnels successifs intra-personnels*. Dans cet ouvrage, Ainslie défend une représentation des conflits d'intérêts intra-individuels inspirés de la théorie des jeux. L'effort pour transcrire dans des termes économiques des phénomènes concernant les troubles de l'impulsivité a donc été constant chez cet auteur, qui a parallèlement toujours conservé une préoccupation initialement « médicale », puisqu'il a dans le même temps occupé un poste de chef-psychiatre au centre médical des vétérans à Coatesville, Pennsylvannie.

Dans les travaux d'Ainslie, les références à l'économie sont de deux types: à la théorie de l'utilité espérée d'une part (représentation du choix impulsif comme maximisation d'une fonction d'actualisation temporelle hyperbolique) et à la théorie des jeux évolutionnaire d'autre part (représentation des conflits d'intérêts sous la forme d'un dilemme du prisonnier répété). Il convient cependant, encore une fois, de bien insister sur le fait que cette inspiration économique ne désigne pas l'emprunt direct de modèles conçus par des économistes, car le travail de modélisation a bien été réalisé par les psychologues. En particulier, la représentation de l'incohérence temporelle sous la forme de la maximisation d'une fonction d'actualisation hyperbolique a été élaborée par des psychologues (Ainslie, 1975; Herrnstein et Ho, 1967), bien avant que cette idée ne soit reprise en économie comportementale (Laibson, 1997). De la même façon, en dépits de la référence faite à la théorie des jeux, l'approche en termes de « sois multiples » (*multiple selves*) est trompeuse, car ces modèles ont bien été développés d'abord par Ainslie (1975), puis diffusés en économie, rencontrant alors un large succès

(Elster, 1979; Schelling, 1984; Thaler et Sheffrin, 1981; Winston, 1980)[62].

La vocation « économique » de George Ainslie, comme des autres chercheurs néo-comportementalistes, renvoie donc simplement à l'utilisation de modèles reposant sur la maximisation de fonctions d'utilité sous contrainte. Il s'agit de transcrire les analyses traditionnelles des pathologies mentales, en particulier celles qui concernent les troubles de la volonté, en termes fonctionnels et quantitatifs : l'apport principal de la recherche sur la motivation des comportements consiste selon Ainslie à faire de la volonté un phénomène scientifique, c'est-à-dire quantifiable (Ainslie, 2001, p. 3-4). Bien que non-métaphoriques, ces références techniques à l'économie sont donc motivées initialement, chez George Ainslie, par des problèmes relatifs à la maladie mentale. Cette idéologie scientifique, que nous désignons par l'expression de psychiatrie économique, affiche ainsi l'ambition d'apporter un traitement original à certains troubles du comportement.

B. Science du diagnostic et maximisation de la normalité: les ambitions théoriques de la psychiatrie économique

La conception économique de la maladie mentale dont il est question dans ce chapitre sera décrite dans sa dimension purement théorique. Pourtant, la psychologie expérimentale à « vocation économique » vise avant toute chose une transformation pratique du soin psychiatrique. Elle trouve donc un écho direct dans un certain nombre de changements ayant affecté les techniques de gestion de la maladie mentale. Ce projet théorique doit être mis en rapport en particulier avec l'abandon progressif d'un mode de gestion « *asilaire* » de la maladie mentale, tel qu'il a été étudié notamment par Goffman (1968), au profit d'approches naturalistes, plus ouvertes. Ces évolutions récentes ont fait l'objet de plusieurs études historiques ou sociologiques (Ehrenberg, 2006, 2008 et 2010 ; Ehrenberg et Lovell, 2001 ; P.H Castel, 2009).

Les analyses proposées par Robert Castel méritent une attention particulière (Castel, 1973, 1977, 1981) car elles portent spécifiquement sur les techniques néo-comportementales

62 Par ailleurs, l'approche en termes de sois multiples ne fait pas l'unanimité chez les neuroéconomistes, car elle suppose l'existence de zones du cerveau spécialisées dans des horizons temporels d'évaluation particulier, ce qui remet en casue l'hypothèse d'une monnaie neuronale unique (*cf.* chapitre 5). Pour une critique de la théorie des choix multiples, voir Bernheim et Rangel, 2009, p.69; Monterosso et Luo, 2010)

de la médecine mentale moderne, qui ont été étudiées dans ce chapitre, et que Castel envisage dans *La société psychiatrique avancée* comme des « *techniques de la programmation économique des sujets* » (Castel, Castel et Lovell, 1979). Sans chercher à identifier dans cette étude sociologique portant sur le contexte le facteur explicatif d'un courant théorique, l'enjeu consiste ici à approfondir la suggestion avancée par Robert Castel d'une « *conception économique de la maladie mentale* ». En effet, Castel envisage, à partir d'une étude socio-historique, la possibilité d'une nouvelle approche théorique du soin mental ; la science quantitative de la motivation qui est analysée dans ce chapitre rejoint certaines intuitions de Castel mais permettra surtout d'en préciser le contenu analytique.

Dans *La société psychiatrique avancée,* Robert Castel s'intéresse aux difficultés rencontrées progressivement par le modèle dit « *asilaire* », aux États Unis, dès les années 1920. Ces problèmes sont à l'origine d'un mouvement de « déshospitalisation », qui a parfois pu être comprise comme l'amorce d'un vaste « désenfermement », en référence à l'«*enfermement*» des fous à l'âge classique que Michel Foucault avait étudié dans son *Histoire de la Folie à l'Age Classique*[63]. Castel interprète donc le développement de la psychologie expérimentale, et, plus généralement, de l'ensemble des sciences et techniques dites néo-comportementales comme une tentative de réponse technique et théorique à l'échec du modèle asilaire.

Le néo-comportementalisme correspond ainsi pour Castel à une forme d'ouverture du monde clinique et hospitalier sur le monde social. Cette nouvelle médecine mentale élargit son périmètre d'intervention, car elle traite non seulement des fous, ou des aliénés, mais «*des problèmes de socialisation, c'est à dire, au sens large de la conformité de la conduite de l'individu à l'ordre social* » (Castel, Castel et Lovell, p.68). Or, la thèse défendue par Castel est que cette ouverture effective de l'asile implique une ouverture théorique : la psychiatrie ne doit plus être considérée comme un savoir clinique, élaboré dans l'intimité du rapport malade-médecin, mais doit être conçue sur le modèle des autres sciences humaines, et notamment l'économie. Ce basculement théorique implique dès lors un changement objectif : poursuivant le rapprochement avec l'économie, les psychiatres ne doivent plus prioritairement soigner des

63 Michel Foucault n'a cependant jamais envisagé l'asile comme un modèle historique indépassable, ou comme le point d'achèvement de l'histoire des rapports entre raison et folie : il a lu les travaux de Castel et était, semble-t-il, informé et pleinement concerné par les changements de la médecine mentale liés aux sciences comportementales: «*sur ces techniques comportementales, il y a un peu de littérature en France. Dans le dernier livre de Castel,* la Société Psychiatrique avancée*, vous avez un chapitre sur les techniques comportementales et vous verrez comment c'est, très exactement, la mise en œuvre, à l'intérieur d'une situation donnée -en l'occurrence un hôpital, une clinique psychiatrique- de méthodes qui sont à la fois des méthodes expérimentales et des méthodes impliquant une analyse proprement économique du comportement* » (Foucault, [1978-1979], 2004, p. 274).

malades mais « maximiser la normalité ». Nous désignerons donc désormais cette approche économique de la maladie mentale par l'expression de « psychiatrie économique »[64].

La visée de « *renforcement de la normalité* » correspond à ce que Castel appelle aussi la « *thérapie pour les normaux* » ; celle-ci « *est cette quête qui s'appuie sur une batterie de techniques spirituelles et surtout corporelles pour maximiser le rendement humain de chacun au lieu de tenter, comme dans les thérapies classiques, de restaurer la santé. Une plus-value de santé (de jouissance, de sentiments, de conscience de son corps, etc.) est l'objectif à atteindre. Le modèle de la croissance psychique remplace celui de la tutelle médicale […]. C'est en fait la normalité qui devient le symptôme à traiter* ».

A première vue, l'analyse de Castel semble s'appliquer aisément à la littérature expérimentale sur l'impulsivité qui constitue notre corpus théorique. Les analogies entre santé psychique et capital humain y sont en effet très courantes. On trouve par exemple chez Ainslie l'affirmation selon laquelle « *chaque personne dispose […] d'un stock de capital humain – en sobriété, santé, bonne volonté, réputation, et de nombreux autres biens en dehors de l'argent lui même. Cette personne doit décider chaque jour de dépenser ou non une partie de ce capital pour une consommation immédiate. Chaque jour elle a le même intérêt à préserver son capital pour le futur, mais elle est aussi attirée par des plans de consommation qui en useraient une part importante dans le présent. Ses états motivationnels successifs peuvent ou non coopérer pour leur intérêt commun à long terme ou peuvent abandonner cet intérêt au profit de leurs intérêts individuels à court terme*. (Ainslie, 1992-b, p. 155).

Par delà son caractère séduisant, l'analyse de Castel mérite cependant d'être précisée. Il convient d'expliquer plus en détails la portée d'un tel rapprochement, opéré à partir d'une analogie capital humain/santé mentale. On peut penser qu'il s'agit là, chez Ainslie ou chez d'autres auteurs, plus d'une topique récurrente que d'une structure analytique déterminante : les modèles comportementaux de l'impulsivité n'impliquent pas nécessairement un rapport avec la théorie de Gary Becker. La référence au capital humain apparaît donc comme une figure de style, c'est à dire comme un élément de rhétorique, au sens de McCloskey (1985, *cf.* chapitre 1)[65]. Il est nécessaire de fournir davantage de consistance à ce concept de « psychiatrie économique ». Que signifie « *maximiser la normalité* »?

64 Cette expression n'est pas utilisée par Robert Castel, qui parle plutôt de « *conception économique de la maladie mentale* » (*cf. supra*)

65 Par ailleurs, comme le soulignent Orléan et Grenier, la notion de capital humain en économie est assez équivoque et peut s'appliquer à des branches de l'analyse économique. Suite aux célèbres analyses de Gary Becker et de l'école de Chicago par Michel Foucault, les théories du capital humain ont fait l'objet d'un intérêt renouvelé, qui s'explique en fait par la facilité avec laquelle ce concept permet d'opérer des rapprochements tous azimuts avec le biopouvoir foucaldien. (Grenier et Orléan, 2007)

Sans anticiper sur l'analyse des expériences, on peut retenir deux traits essentiels de cette visée de « renforcement de la normalité » identifiés par Castel, qui pourront servir à la fois de fil directeur et de principe problématique pour décrire l'idéologie biomédicale étudiée dans ce chapitre. La première caractéristique de cette approche néo-comportementale de la santé mentale concerne les rapports entre diagnostic et traitement. Selon Castel, la médecine mentale tend à se spécialiser dans l'expertise et la science du diagnostic, au détriment du soin proprement dit. En effet, Castel montre que, d'un côté, la remise en question du modèle de l'asile renvoie à une diffusion extrêmement large de toute une série de thérapies et de techniques de traitement, plus ou moins cautionnées par les autorités publiques, conçues pour l'usage immédiat du grand public : thérapies de groupes, *self-help*, analyse transactionnelle, médiation transcendantale, cri primal, *etc.* De l'autre côté, à l'hôpital, cette évolution permet aux psychologues de se concentrer sur la « recherche pure », chargée uniquement de détecter des troubles fonctionnels, et donc de se détacher des enjeux de soins et de prise en charge directe des patients.

Au sein du néo-comportementalisme, on observe également un effort de formalisation, de sophistication et d'innovation théorique, via notamment des références à la théorie des jeux, à la théorie de l'utilité espérée, qui semblent aller à rebours de leurs implications pratiques. Dans ses deux ouvrages, Ainslie, par exemple, n'indique jamais explicitement des méthodes de soin possibles pour les « échecs de la volonté ». Il se contente le plus souvent de renvoyer aux traditionnelles thérapies cognitives, fondées sur le conditionnement instrumental classique. Il s'agit là d'une constante, au sein de ce programme de recherche associant psychologie, néo-comportementalisme et modélisation économique: la tendance à imiter le « style » des sciences dures va de pair, paradoxalement, avec une très faible applicabilité des recherches. Cette observation vaut également, on le verra, pour la neuroéconomie.

La faible applicabilité des recherches n'appelle pas nécessairement un éclairage analytique plus approfondi. La deuxième caractéristique est en revanche plus problématique. Pour Castel, l'approche « *économique* » implique avant toute chose un effort de quantification des troubles mentaux. Cette tentative a une conséquence paradoxale, puisqu'elle aboutit, en remettant en question les définitions qualitatives de la maladie par les catégories cliniques, à brouiller la frontière entre le pathologique et le normal : le premier est envisagé dans la continuité quantitative du second. Dans cette perspective, on peut dire qu'à la limite, personne n'est complètement sain d'esprit, et tout le monde est un peu fou ; c'est la raison pour laquelle, pour Castel, la psychiatrie économique gomme la référence première à des pathologies

-celles-ci étant envisagées secondairement comme des « *déficits fonctionnels* » ou des handicaps- pour se concentrer sur l'amélioration des performances des sujets normaux. Castel, Castel et Lovel y voient un « *décrochage de la médecine par rapport à l'idée de maladie et sa prétention d'avoir à intervenir dans les problèmes de santé.* [Ce projet] *dépasse même la prévention : il vise à renforcer la santé par des techniques médicales ou psychologiques* » (Castel, Castel et Lovell, 1979, p.53).

La quantification des troubles mentaux par les techniques néo-comportementales conduit donc à élargir le périmètre d'action du soin mental dans deux directions : d'une part, dans les individus concernés -des fous aux normaux- et, d'autre part, dans les comportements visés. En effet, les définitions des troubles élaborées en laboratoire s'avèrent en pratique extrêmement souples, et, comme le souligne Castel, le néo-comportementalisme vise plus à renforcer la santé qu'à traiter des maladies nettement identifiées.

C'est ici que le propos de Castel appelle une précision théorique, car l'équivalence introduite entre quantitativisme et « *thérapie pour les normaux* » soulève une difficulté. Castel, Castel et Lovel considèrent en effet que l'utilisation de techniques néo-comportementales, qui se veulent objectives, conduit à élaborer des catégories nosographiques de plus en plus indistinctes et englobantes. Cela est pourtant paradoxal : à proprement parler, la sous-détermination empirique d'une catégorie clinique (telle que celle d'impulsivité par exemple) devrait plutôt être comprise comme un échec de la quantification. L'enjeu consiste alors à comprendre comment l'application d"un formalisme économique en psychiatrie a pu aboutir, conformément à l'intuition de Castel, à simultanément quantifier et brouiller la notion d'impulsivité.

II. Choix inter-temporel et impulsivité: la reprise d'un problème théorique en économie par la psychologie néo-comportementale

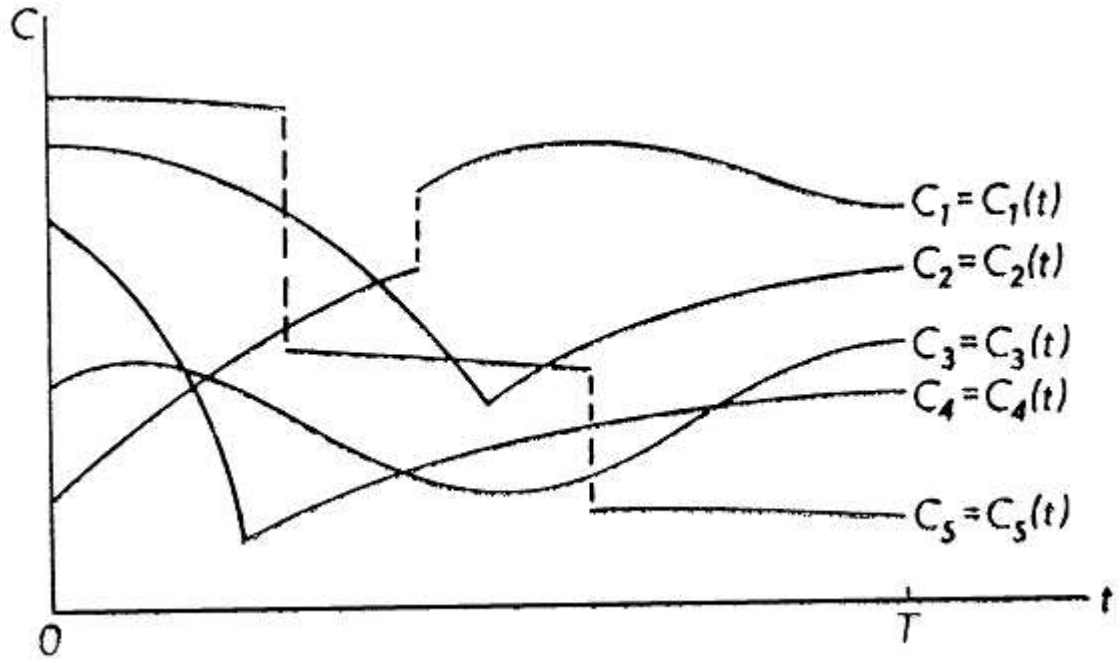
Cette section décrit la manière avec laquelle les psychologues néo-comportementalistes ont repris en main, à partir des années 1960, un problème théorique laissé à l'abandon par les économistes. L'impulsivité a fait l'objet d'une réflexion théorique en économie, dès les années 1950, avec la publication d'un article pionnier par l'économiste Robert Strotz (Strotz, 1955) (A). Pourtant, ce continent théorique demeure alors largement inexploré par les économistes. L'impulsivité ou incohérence séquentielle n'est pas étudiée comme telle. C'est plutôt le problème de l'incohérence dynamique qui attire l'attention des économistes (B). C'est donc à l'initiative de chercheurs issus de la psychologie néo-comportementale que se constitue, à la fin des années 1960, un champ d'étude des conduites impulsives commun à la psychologie et à l'économie. Les études en laboratoire montrent que l'incohérence séquentielle est une tendance biologique élémentaire, c'est-à-dire une constante comportementale du choix répété (C).

A. Robert Strotz et la myopie temporelle (1955) : une première approche économique de l'impulsivité

Dès 1955, dans un article pionnier intitulé « Myopie et inconsistance dans la maximisation dynamique de l'utilité », l'économiste Robert Strotz propose un premier modèle de l'incohérence séquentielle (Strotz, 1955). En suggérant que les comportements impulsifs résultent d'une actualisation temporelle des utilités futures non-proportionnelle au délai, contrairement à ce que suppose le modèle de Samuelson (1938), Strotz avance une proposition qui sera centrale pour les théories ultérieures du choix impulsif étudiées dans le cadre de ce chapitre.

L'article de Strotz repose sur l'idée simple selon laquelle la possibilité, pour un individu, de réévaluer de manière continue des plans de consommation inter-temporels qui lui

sont proposés, peut conduire celui-ci à modifier ses choix entre ces mêmes plans au cours du temps. Strotz considère donc dans son modèle un agent qui dispose d'un certain stock de consommation K_0. Ce stock qui doit être alloué au cours du temps peut se comprendre comme une contrainte budgétaire. L'agent doit à chaque période *t* choisir parmi un nombre potentiellement infini de plans de consommation inter-temporelle. A chacun de ces plans correspond une certaine distribution du stock initial de consommation K_0 au cours du temps, de $t = 0$ à $t = T$, *T* désignant l'horizon temporel de la séquence. Le choix auquel fait face l'individu peut être représenté de la manière suivante:



Strotz, 1955, p.166

Ici, chaque plan *C1, C2, C3, C4, C5* assure bien le même niveau de consommation global égal à K_0, mais tous ces plans n'offrent pas les mêmes niveaux de consommation à chaque instant *t*. Lorsque l'agent cherche, à l'instant τ, à déterminer le plan de consommation inter-temporel optimal, il cherche donc à maximiser la fonction d'utilité inter-temporelle Φ_τ suivante:

$$\Phi_t = \int_0^T \lambda(t-\tau).u[C(t),t]dt$$

sous la contrainte:

$$\int_0^T C(t)dt = K_0$$

Dans l'équation ci-dessus, l'expression *u [C(t), t]* est une fonction d'utilité instantanée, qui assigne à chaque instant *t* une valeur *u(t)* à *C(t)*. L'expression λ *(t)* est une fonction d'actualisation dont la valeur dépend du délai *t*.

Généralement, les économistes, à la suite de Samuelson (1938) supposent que le taux d'escompte des utilités futures est constant. Samuelson (1938, p.156) propose ainsi, dans son modèle, « *pour plus de simplicité* », de retenir une fonction d'actualisation exponentielle, qui vérifie cette condition, du type:

$$\lambda(t) = e^{-r \cdot t}$$

Le paramètre *r* peut se comprendre comme un taux d'actualisation ou un taux d'escompte, qui correspond, en mathématique financière, au rendement d'un placement continu à intérêts composés[66].

Samuelson considère donc une actualisation en temps continu avec une fonction d'utilité inter-temporelle prenant la forme d'une intégrale, mais la plupart des travaux postérieurs (et antérieurs) envisagent plutôt une actualisation en temps discret. Nous nous restreindrons donc ici au cas discret, avec une fonction d'utilité *U(t)* possédant la forme suivante:

---

66 En mathématiques financières, la valeur finale VF d'un placement VA (valeur actuelle) à taux d'intérêt composé i est égale à VF= VA $(1+i)^t$ où t désigne le nombre de périodes pendant lesquelles durent le placement. A chaque période, les intérêts sont incorporés au capital: c'est ce que l'on appelle la capitalisation composée. Cette formule renvoie donc à une actualisation du type VA = VF $(1+i)^{-t}$. Mais il s'agit ici cependant d'une capitalisation en temps discret. Lorsque l'actualisation se déroule en temps continu, c'est-à-dire lorsque la périodicité de paiement des intérêts notée s tend vers l'infini,

$$VF = \lim_{s \to +\infty} VA(1+\frac{i}{s})^{s \times t}$$

Si l'on pose m = s/i, on obtient:

$$VF = \lim_{s \to +\infty} VA(1+\frac{1}{m})^{m^{i \times t}}$$

Si $s \to +\infty$, $m \to +\infty$ donc

$$\lim_{m \to +\infty} (1+\frac{1}{m})^{m^{i \times t}} = e^{i \times t}$$

On a donc VF = VA.$e^{i.t}$ et VA = VF.$e^{-i.t}$

Dans la fonction d'actualisation proposée par Samuelson (1938) du type $\lambda(t) = e^{-r.t}$, le coefficient r peut donc se comprendre comme le taux de rendement d'un placement continu au taux d'intérêt composé r=i.

$$U_t(c_t,...,c_T) = \sum_{t=0}^{T} \lambda(t) \cdot u(c_t)$$

U_t peut donc se comprendre comme l'utilité d'une suite d'utilités futures instantanées $(c_\tau, ..., c_T)$ du point de vue de l'instant présent t. La fonction d'actualisation $\lambda(t)$ décroît en général au cours du temps ($\lambda'(t)<0$), ce qui signifie que la valeur actuelle d'une utilité future décroit avec le temps. Le taux d'actualisation $r(t)$ se définit comme l'opposé du taux de croissance de la fonction d'actualisation $\lambda(t)$:

$$r(t) = -\frac{\lambda(t) - \lambda(t-1)}{\lambda(t-1)}$$

Le facteur d'actualisation $\delta(t)$ est le rapport entre le coefficient d'actualisation attaché à l'instant t et celui attaché à l'instant précédent $(t-1)$:

$$\delta(t) = \frac{\lambda(t)}{\lambda(t-1)} = 1 - r(t)$$

$\delta(t)$ mesure donc la perte d'utilité liée à l'écoulement du temps. Dans ce cadre, la cohérence temporelle est vérifiée si le taux d'actualisation reste constant, ce qui signifie que la fonction d'actualisation $\lambda(t)$ décroît de façon linéaire avec le temps. En effet, pour reprendre l'exemple de Strotz, une actualisation proportionnelle au délai implique que l'importance relative de 1957 par rapport à 1956 soit la même en 1955 et 1956 :

$$\delta(1956) = \delta(1957) = \delta(1958) = ... = \delta$$

De cette manière, la liberté offerte à l'individu de réévaluer ses plans de consommation inter-temporel ne modifie pas ses préférences: l'agent choisit en $t = 1$ le même plan qu'il avait choisi en $t = 0$. Ce cas, que Strotz appelle le cas « harmonieux » (Strotz, 1955, p.167), est notamment rempli si la fonction d'actualisation est de la forme

$$\lambda(t) = \delta^t$$

où δ est une constante, ce qui peut se comprendre comme une approximation de la fonction d'actualisation exponentielle de Samuelson[67]. Dans ce modèle d'actualisation exponentielle, l'utilité inter-temporelle $U(t)$ est donc donnée par la relation suivante :

---

67 En effet, $\lambda(t) = \delta^t = (1-r)^t$

Or, pour r proche de zéro, $(1-r) \approx e^{-r}$

d'où $\lambda(t) \approx (e^{-r})^t = e^{-rt}$

$$U_t(c_t,...,c_T) = u(c_t) + \delta \cdot u(c_{t+1}) + \delta^2 \cdot u(c_{t+2}) + \delta^3 \cdot u(c_{t+3}) + ... + \delta^T \cdot u(c_T)$$

où $\delta = 1 - r$ est un paramètre compris entre zéro et un. Dans ce modèle d'actualisation exponentielle, on a :

$$\delta(t) = \frac{\lambda(t)}{\lambda(t-1)} = \frac{\lambda^t}{\lambda^{(t-1)}} = \lambda$$
$$r(t) = 1 - \delta(t) = 1 - \lambda$$

En d'autres termes, le taux et le facteur d'actualisation restent constant. Mais rien n'indique que ce cas harmonieux corresponde à la réalité. Selon Strotz, c'est plutôt l'inverse qui se produit, car la plupart des gens manifestent une « myopie temporelle »: l'actualisation est proportionnellement plus importante pour des utilités consommées à court-moyen terme que pour des utilités consommées à long-terme. Concrètement, cela implique que, très souvent, un agent peut *ex ante* se comporter de manière réfléchie et économe (*thrift*) et chosir un plan de consommation inter-temporel assurant une répartition équilibrée au cours du temps. Toutefois, lorsqu'une possibilité de consommation immédiate lui est offerte, l'agent peut être amené à dilapider et consommer l'intégralité de son budget, car la valeur actuelle des utilités à moyen-terme lui apparaît alors beaucoup moins importante qu'*ex ante*; l'utilité relative qui est alors liée à une consommation disponible immédiatement, au moment de la réévaluation, lui apparaît alors beaucoup plus attractive (Strotz, 1955, p.164).

Dans son article fondateur de 1955, Strotz fournit donc une explication originale des comportements impulsifs, à partir de deux hypothèses: la non-proportionnalité de l'actualisation temporelle au délai, et la possibilité pour l'individu de se « désengager », c'est-à-dire de réévaluer et de modifier ses plans de consommation inter-temporels au cours du temps. Strotz critique ici l'axiomatique de Samuelson (1938), qui présuppose une fonction d'actualisation exponentielle, proportionnelle au délai. Pour Strotz, le modèle de Samuelson (1938) peut éventuellement rendre compte d'une forme faible de myopie temporelle, liée à simple existence d'une fonction d'actualisation: les individus préfèrent en règle générale consommer immédiatement plutôt que plus tard. Ce phénomène est connu depuis longtemps en économie sous le nom d'impatience. La véritable nouveauté du modèle de Strotz consiste de ce point de vue à mettre en évidence une forme forte de myopie temporelle -l'impulsivité- entendue comme incohérence des choix: les agents manifestent non seulement une tendance à consommer une trop grande part de leur revenu immédiatement en raison de leur préférence pour le présent, mais ils regrettent par la suite leur manque de réserve.

B. Un problème théorique abandonné par les économistes? Incohérence séquentielle et incohérence dynamique

L'actualisation dite hyperbolique implique une fonction d'actualisation de la forme: $\lambda\ (t) = 1 / (1 + k.t)$. Strotz suggère que l'incohérence temporelle résulte d'une d'actualisation non-proportionnelle au délai, qui n'est donc pas de la forme $\lambda\ (t) = \delta^t$. L'idée d'utiliser des fonctions d'actualisation hyperbolique n'apparaît donc pas dans son article fondateur de 1955 Elle a été proposée plus tard par des chercheurs néo-comportementalistes (Ainslie, 1975; Chung et Herrnstein, 1967). Ce n'est qu'encore plus tardivement, dans les années 1990, que l'actualisation hyperbolique commence à être étudiée en économie comportementale. Dans les années 1950, le problème théorique posé par Strotz reste largement ignoré par les autres économistes.

Le relatif manque d'intérêt des économistes pour le problème des comportements incohérents peut d'abord s'expliquer, comme l'avance l'économiste comportementaliste Laibson, par le caractère « *problématique* » des préférences incohérentes. Celles-ci étant considérées comme contraire à une théorie « *classique* », de référence (celle de Samuelson, 1938), les économistes, soucieux de préserver la validité de leur théorie, auraient répugné à prendre en compte ces phénomènes contre-prédictifs (Laibson, 1997, p.450).

Cette interprétation est contestable. Le caractère contre-prédictif de ces comportements pour la théorie économique n'a en fait nullement empêché leur discussion par des économistes. En 1971, les psychologues Lichtenstein et Slovic réalisent une expérience dans un casino, avec des joueurs professionnels. Lichentenstein et Slovic observent que les préférences de ces individus sont régulièrement incohérentes, selon qu'il s'agisse ou de vendre des loteries financières (Lichtenstein et Slovic, 1971). Ce résultat a été beaucoup discuté en économie, notamment par Grether et Plott (1979). Depuis les années 1970, un important programme de recherche s'est développé autour de ce problème théorique, dans le sillage des travaux de Tversky, Slovic et Kahneman (1990).

Néanmoins, la notion de préférences incohérentes n'a pas le même sens dans l'article de Strotz (1955) et dans le programme de recherche kahnemanien. En effet, chez Strotz, l'incohérence dont il est question résulte de la simple possibilité de modifier séquentiellement, au fur et à mesure, les choix portant sur les plans de consommation future. Pour éviter toute confusion, nous avons choisi d'appeler ce type d'incohérence « incohérence temporelle », « séquentielle », ou impulsivité. Nous avons distingué ce problème théorique de celui de

l'incohérence dite « dynamique », lié à un effet de dotation (*endowment effect*) (*cf. supra*).

Il convient de souligner, par ailleurs, que le renversement de préférences qui est mis en évidence dans l'expérience Lichtenstein et Slovic, ne porte pas, au départ, sur des biens économiques sans risques, mais sur des loteries. A l'inverse, Strotz évacue, dans son article, le problème du risque, ou, plutôt, sépare ce problème de celui de l'incohérence séquentielle. La non-prise en compte du risque dans son article n'est en effet pas une simplification, qui pourrait conduire éventuellement à une complexification ultérieure de son modèle de base. Au contraire, le choix risqué représente pour Strotz un cas spécial, une limitation à son modèle qui se veut être applicable à la fois en présence et en l'absence d'aléa sur les utilités futures: « *nous faisons abstraction de toutes considérations concernant le risque et l'incertitude. Cela pourrait déranger ceux pour qui l'essence des problèmes dynamiques est ainsi ignoré; mais je pense qu'il apparaitra clair au fur et à mesure du développement que l'introduction du risque et de l'incertitude ne ferait que limiter l'analyse et empêcher d'avoir une vision claire des problèmes envisagés ici* » (Strotz, 1955, p.166)

L'incohérence séquentielle fait intervenir la notion d'actualisation temporelle; celle-ci peut ou non inclure une prise en compte du risque selon que le taux d'actualisation incorpore une représentation par l'agent de l'aléa sur les utilités futures. On peut donc faire l'hypothèse selon laquelle l'incohérence dynamique, telle qu'elle est étudiée au sein du programme kahnemanien dans les années 1970 et 1980, a étouffé et pris la place de cette vision large et séquentielle des préférences incohérentes[68]. Ce n'est que dans les années 1980 que l'économie comportementale redécouvre le problème théorique de l'actualisation hyperbolique et des conduites impulsives (*cf.* section III). Ce problème avait pourtant fait déjà l'objet d'une analyse par Strotz. Surtout, ce modèle théorique a stimulé un programme de recherche expérimental sur l'animal, validant bien avant les premières expériences sur l'homme, les intuitions théoriques de Robert Strotz.

68 Le problème de l'incohérence dynamique, c'est-à-dire d'un écart entre la propension à payer et à accepter, n'est cependant pas non plus limitée au choix risqué. Il est possible de mettre en évidence un effet de dotation avec des loteries, comme dans l'expérience de Slovic et Lichtenstein (1968), mais aussi avec des biens sans risques, comme des tasses à café et crayons, comme dans l'expérience de Kahneman, Knetsch et Thaler (1990).

C. Les apports de la psychologie expérimentale : l'impulsivité comme constante comportementale

Dans son article de 1955, Robert Strotz remet en question la pertinence le modèle de Samuelson (1938), en suggérant que la plupart des individus n'actualisent pas, le plus souvent, les utilités futures à un taux constant. Cette conjecture demeure cependant purement théorique. Elle demande à être confirmée empiriquement, pour valoir comme objection sérieuse au modèle de Samuelson. Dans les années 1960, des psychologues néo-comportementalistes travaillant sur le pigeon montrent que l'inhérence temporelle, au sens défini par Strotz, est une constante comportementale, qui s'explique par la séquentialité des décisions. Ces travaux ne valent cependant pas seulement comme une simple confirmation empirique puisqu'ils contribuent à approfondir le problème théorique posé par de Strotz, en modélisant les liens entre séquentialité et actualisation.

Il pourrait néanmoins paraître étrange que des expériences de laboratoire réalisées sur le pigeons puissent valoir comme validation empirique d'une théorie conçue pour représenter le comportement humain. En fait, les psychologues néo-comportementalistes qui étudient le choix répété chez l'animal ne prennent pas au départ le renversement des préférences comme une erreur de raisonnement, comme c'est le cas en économie comportementale. L'incohérence est vue comme l'effet non-voulu d'une forme de rationalité biologique adaptée aux décisions séquentielles. En d'autres termes, il s'agit de montrer d'une part que les animaux exhibent dans les choix répétés des comportements pouvant être qualifiés de rationnels, mais, que, d'autre part, cette rationalité biologique élémentaire implique aussi une actualisation non-proportionnelle au délai. La rationalité biologique, pour laquelle l'apprentissage et l'exploration jouent un rôle important, appelle ainsi un réaménagement de la rationalité économique.

Le point de départ de ces recherches consiste donc à montrer, et c'est important de le souligner, la richesse de l'intelligence animale, et non son caractère borné ou son possible conditionnement. L'expérience fondatrice dans ce domaine a été réalisée en 1961 par Richard Herrnstein (1961). Dans cette étude, Herrnstein utilise, comme Benjamin Skinner -dont il est alors l'élève à Harvard- des pigeons ayant appris une tâche simple lors d'un conditionnement instrumental: presser, à l'aide du bec, un levier délivrant une récompense alimentaire. Or Herrnstein ne cherche pas ici simplement à observer le comportement conditionné chez l'animal, mais plutôt à se servir de celui-ci comme d'un instrument d'étude des stratégies

d'apprentissage et d'exploration. En effet, l'originalité de l'expérience de Herrnstein est d'offrir la possibilité aux pigeons, conditionnés à presser sur un levier pour obtenir une récompense, de choisir (et non pas réagir de manière automatique par un réflexe conditionné) entre deux leviers. Lorsque l'un des deux leviers est choisi (par un coup de bec), l'autre levier est désactivé, jusqu'à l'obtention de la récompense. Chaque levier génère des montant fixes de récompenses, mais avec un délai plus ou moins important. Le délai d'obtention de la récompense associé à l'une des deux options est en moyenne plus court (donc plus avantageux), mais il varie à chaque décision.

Cette tâche correspond donc à un problème de « machines à sous multi-jeux » (*multi-armed bandit problems*), ici réduit à deux machines, dans lequel il s'agit de déterminer la machine (ici, le levier) la plus avantageuse à long-terme. Le résultat principal auquel Herrnstein aboutit est que les réponses (nombre de choix en faveur de l'un des deux leviers) reflètent les rendement relatifs de chaque levier. En d'autres termes, les pigeons allouent les fréquences de choix en faveur de chaque levier en fonction de la récompense moyenne relative de chaque levier, qui varie en sens inverse du délai d'obtention. La loi d'égalisation des récompenses (*matching law*) s'exprime ainsi sous la forme :

$$\frac{s_1}{s_1 + s_2} = \frac{r_1}{r_1 + r_2}$$

où s_1 et s_2 désignent respectivement le nombre de réponses en faveur de l'option (levier) *1* et *2* ; r_1 et r_2 les récompense moyennes obtenues en choisissant l'option *1* et *2*. C'est là le principal résultat de cette expérience, qui doit être comprise comme un approfondissement décisif de la théorie classique du conditionnement instrumental de Skinner : il existe une relation quantitative entre la réponse conditionnée et le montant de récompense obtenu. Plus un comportement est récompensé, plus il sera adopté par l'individu. Herrnstein démontre ainsi que les réponses varient proportionnellement avec le montant de la récompense. Mais, encore une fois, il importe de souligner que ce résultat ne signifie pas que le comportement des pigeons soit borné: au contraire, on aurait pu s'attendre à ce que les choix s'effectuent systématiquement en faveur de l'une des deux options. Ici, la régularité avec laquelle les pigeons parviennent à égaliser la fréquence de choix avec le montant relatif de récompense associée cette option démontre l'existence d'une forme de rationalité dans le choix répété.

Ce type de routine dite de « mélioration », dans laquelle la fréquence des choix est adaptée au coup par coup aux récompenses obtenues, est adaptée dans le cadre de l'expérience de Herrnstein, car elle permet effectivement aux pigeons d'obtenir un montant maximum de récompense. Mais dans d'autres situations, celle-ci peut aussi générer des incohérences

séquentielles. En effet, la formalisation de cette loi d'égalisation (des rendements, en anglais: *matching law*) débouche sur une fonction d'actualisation temporelle des récompenses hyperbolique (Chung et Herrnstein, 1967).

Skinner avait, comme le rappelle Herrnstein, avancé l'idée d'un « ratio d'extinction » (*extinction ratio*), défini, pour des protocoles de choix similaires à délai variable (*variable interval reinforcemnt schedule*) comme le rapport constant entre nombre de réponses non-récompensées et le nombre total de réponses (Skinner, 1938, p. 130). Toutefois, Herrnstein présente cette relation quantitative non comme la constante d'un rapport, mais comme une fonction linéaire entre comportement et récompense. Herrnstein considère que le nombre de réponses, noté *s*, est une fonction linéaire de la récompense obtenue, notée *r* :

$$s = k \cdot r$$

où *k* est une constante

Skinner quant à lui envisage une relation similaire, mais sous la forme :

$$k = \frac{s}{r}$$

Les deux formulations sont bien sûr strictement équivalentes d'un point de vue mathématique, mais celle de Herrnstein a une importance capitale car elle suggère plus nettement une fonction d'actualisation des récompenses de type hyperbolique. Dans un article ultérieur, Chung et Herrnstein (1967) montrent que cette loi d'égalisation des rendements implique une actualisation temporelle des récompenses de type hyperbolique. En effet, comme les choix sont limités à deux options, $(s_1 + s_2)$ est égal au nombre total de décisions prises lors de l'expérience, et $s_1 / (s_1 + s_2)$ peut se comprendre comme la fréquence ou la probabilité de choix[69] en faveur de l'option *1*, notée p_1 :

$$p_1 = \frac{r_1}{r_1 + r_2} = \frac{1}{1 + \frac{r_2}{r_1}}$$

Or, et c'est là l'une des caractéristiques importantes du protocole de Herrnstein, la durée totale de l'expérience et le montant de récompense obtenue à chaque décision étant fixes, r_1 et r_2 sont compris comme l'inverse du délai d'obtention moyen de la récompense attaché à chaque option. On a donc $r_1 = 1/t_1$ et $r_2 = 1/t_2$ avec t_1 et t_2 désignant respectivement les délais d'obtention moyens de la récompense attachés à l'option *1* et *2* respectivement. Cela implique:

69 Les portabilités sont donc comprises ici implicitement comme des fréquences. Sur le lien entre néo-comportementalisme et conception fréquentiste des probabilités, voir l'introduction à la première partie.

$$p_1 = \frac{1}{1+\frac{r_2}{r_1}} = \frac{1}{1+\frac{1/t_2}{1/t_1}} = \frac{1}{1+\frac{t_1}{t_2}} = \frac{1}{1+(\frac{1}{t_2})\cdot t_1}$$

Cette formulation renvoie à une fonction d'actualisation hyperbolique, du type:

$$\lambda(t) = \frac{1}{1+r\cdot t}$$

où *r* désigne le taux d'escompte. Ici, dans le cadre de l'expérience de Herrnstein, on a

$$r = \frac{1}{t_2}$$

Le taux d'escompte est égal à l'inverse du délai d'obtention associée à l'autre option: les récompenses futures associées à l'option 1 font l'objet d'une actualisation d'autant plus forte que le délai associé à l'autre option est court. L'actualisation hyperbolique correspond à la capitalisation par intérêt simple, et non à intérêts composés comme c'est le cas pour l'actualisation exponentielle (*cf. supra*) .

Une fonction d'actualisation temporelle hyperbolique possède la propriété de décroître à un taux décroissant, et non constant. Au fur et à mesure que la perspective d'une récompense s'éloigne dans le temps, son attractivité diminue, mais dans une proportion de plus en plus faible. Inversement, une option qui génère une récompense jugée faiblement attractive à moyen-long terme peut devenir très attractive si le délai d'obtention devient quasi-nul : l'actualisation hyperbolique explique donc pourquoi les individus renversent leurs préférences et préfèrent parfois choisir des options moins attractives lorsqu'elles sont associées à une récompense immédiatement accessible.

L'équivalence qui est introduite ici entre la loi d'égalisation des rendements et l'article antérieur de Robert Strotz mérite cependant d'être précisée. L'expérience de Herrnstein (1961), et sa formalisation en terme d'actualisation hyperbolique par Herrnstein et Ho (1967) représentent à la fois moins et plus qu'une confirmation empirique du modèle de Strotz. Tout d'abord, il convient de souligner que l'expérience de Herrnstein n'implique aucune incohérence séquentielle à proprement parler. Les pigeons se contentent de distribuer leurs choix en fonction des délais moyens observés dans le passé. Il n'y a pas à de plans de consommation inter-temporelle, c'est à dire de planification fixant à un certain moment les choix futurs. Pour qu'il y ait incohérence séquentielle dans les termes de Strotz, il faudrait qu'on propose *ex ante* au pigeon (et que celui-ci accepte) une répartition temporelle des gains différente de celle associée à un comportement de *matching*. C'est seulement dans ce cas que l'on pourrait conclure à une planification incohérente des choix.

Cependant, l'expérience d'Herrnstein laisse à penser qu'un tel phénomène pourrait être observable, puisque dans le cadre de ce protocole, la loi d'égalisation implique logiquement une actualisation non-proportionnelle au délai. Une étude réalisée ultérieurement par Georges Ainslie (1974) a par la suite apporté une confirmation à cette thèse. Ainslie montre que les pigeons, dans des protocoles de choix similaires, sont capables *ex ante* de s'engager dans un plan de consommation inter temporel plus avantageux que l'ensemble des récompenses reçues en adoptant un comportement de *matching*. C'est en ce sens que l'expérience d'Herrnstein dépasse la simple confirmation empirique, puisqu'elle montre que dans des choix répétés entre deux options délivrant une récompense fixe à intervalle variable, l'individu qui se contente d'égaliser les fréquences de choix aux récompenses obtenues actualise de manière non-proportionnelle au délai, et peut donc révéler des préférences incohérentes.

Cependant, le passage de la loi d'égalisation des rendements à la fonction d'actualisation temporelle hyperbolique repose, on l'a vu, sur la mesure de la récompense par l'inverse du délai d'obtention de cette même récompense (c'est-à-dire $r_1 = 1/d_1$ *et* $r_2 = 1/d_2$). Cela est justifié dans l'expérience de Herrnstein, car, les montants de récompense étant constants, et la durée totale de l'expérience étant fixe, le montant relatif de la récompense associée à l'une des options peut se comprendre comme l'inverse du délai d'obtention : un levier est d'autant plus intéressant qu'il débouche rapidement sur une récompense. En revanche, pour des protocoles plus complexes, dans lesquels, par exemple, les fréquences de récompenses varient avec les réponses, les choses se compliquent. Par conséquent, l'actualisation hyperbolique ne vaut comme conséquence théorique de la loi d'égalisation des rendements qu'à l'intérieur du protocole spécifique de l'expérience d'Herrnstein de 1961.

Les recherches menées par Richard Herrnstein dans les années 1960 ne permettent donc pas d'affirmer que le comportement animal est universellement caractérisé par une actualisation hyperbolique des récompenses futures. Néanmoins, ces travaux suggèrent que, dans les problèmes de machines à sous multi-jeux, l'attitude apparemment raisonnable qui consiste à égaliser les fréquences de choix aux délais relatifs conduit, implicitement, à des incohérences séquentielles au sens de Strotz. Le modèle économique de l'impulsivité proposé par Strotz n'est donc pas directement confirmé par ces expériences ; néanmoins celles-ci ouvrent un programme de recherches destiné à approfondir les implications de la loi d'égalisation des rendements, en étudiant son application dans d'autres protocoles. La signification économique de la loi d'égalisation des rendements reste donc alors assez largement indéterminée. L'idée, proposée par Strotz, selon laquelle la maximisation d'une fonction d'utilité inter-temporelle à taux non-constant permet de rendre compte du

renversement des préférences a suscité par la suite de vives controverses chez les psychologues néo-comportementalistes.

III. Néo-comportementalisme et économie : un projet interdisciplinaire contesté

Le néo-comportementalisme vise à s'affranchir du cadre étroit de la psychologie, pour, à terme, se rapprocher de l'économie. A la fin des années 1960, la représentation de la loi d'égalisation des rendements sous forme d'un problème d'actualisation temporelle hyperbolique jette une passerelle théorique vers l'économie (Herrnstein et Ho, 1967). Pourtant, le ralliement de ce programme de recherche à l'économie pose d'importants problèmes, à la fois en interne (chez les psychologues) et pour les économistes. La modélisation économique des comportements soulève d'abord un important débat au sein du courant néo-comportementaliste. Pour certains psychologues, ces résultats expérimentaux invalident la théorie de l'utilité espérée. D'autres auteurs analysent au contraire l'incohérence temporelle des préférences comme la conséquence de la maximisation d'une fonction d'actualisation hyperbolique. Quoique contesté, le formalisme économique finit par s'imposer, non pas pour des raisons théoriques, mais plutôt par opportunisme : une vision purement fonctionnelle de l'hypothèse de maximisation, acceptable par les deux camps, lance l'idée d'une réconciliation et collaboration théoriques entre économistes, psychologues et biologistes. (A) A partir des années 1980, le choix inter-temporel et l'actualisation hyperbolique des récompenses deviennent donc des objets d'analyse de la science économique. Toutefois, l'intégration de la science quantitative de la motivation au sein de l'économie comportementale, placée sous l'influence déterminante de Kahneman et Tversky, implique aussi une dissolution de son identité théorique. Les concepts d' « anomalies » ou de « *framing effects* » mobilisées au sein du programme kahnemanien peinent à rendre compte de la signification originaire de la notion d'incohérence séquentielle (B)

A. Melioration *versus* Maximisation : le débat interne sur la signification économique des expériences néo-comportementales

L'équivalence introduite par Herrnstein et Ho (1967) entre loi d'égalisation des rendements (*matching law*) et fonction d'actualisation temporelle hyperbolique a soulevé un important débat parmi les psychologues néo-comportementalistes. Ceux-ci ont alors suivi deux stratégies de recherche opposées. La première, défendue par Herrnstein lui-même, consiste à rejeter la généralisation de ce formalisme économique à la théorie expérimentale du choix répété. Herrnstein soutient en effet que la maximisation d'une fonction d'utilité – fût-elle de forme hyperbolique- est en fait impossible; les sujets se contentent, au coup par coup, d'améliorer comme ils le peuvent leurs décisions. La théorie de Herrnstein dite de la « mélioration » postule donc que la séquentialité est un obstacle à la rationalité économique, entendue comme maximisation d'une fonction d'utilité (1). D'autres psychologues ont au contraire visé une réconciliation avec l'économie via le modèle de Strotz. L'étalement dans le temps des décisions n'est plus une limite fondamentale posée à la maximisation, mais engage une rationalité au second degré, portant sur l'intelligence des séquences (*patterns*) (2).

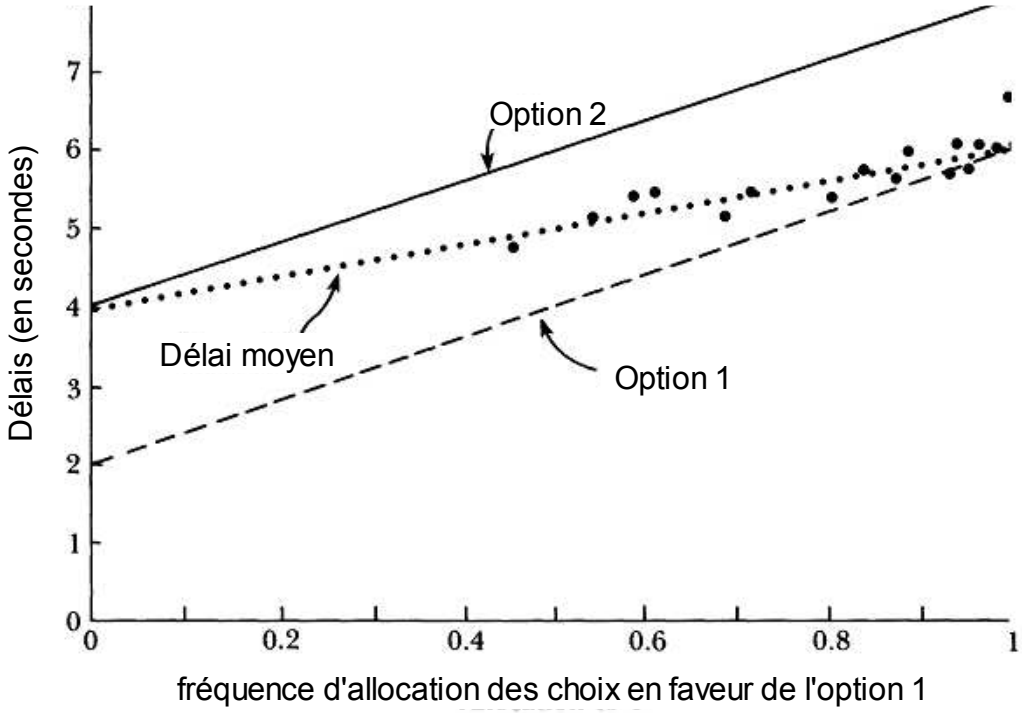
1. La séquentialité comme obstacle à la rationalité : Richard Herrnstein et la théorie de la mélioration

Dans les années 1970, Herrnstein développe sur la base de ses travaux de laboratoire une théorie de ce qu'il appelle « choix distribué », qu'il conçoit comme une critique directe de la théorie (économique) de l'utilité espérée. Pour Herrnstein, la plupart des décisions que les hommes ou les animaux prennent ne portent pas sur des actes isolés, mais sur des séries d'actes ou choix distribués. Or, la séquentialité du choix est un obstacle à la rationalité. En effet, dans les choix répétés, la plupart des individus ne peuvent, *ex ante*, maximiser le montant agrégé des récompenses, car, ne connaissant pas la distribution des récompenses à chacune des options, ils se contentent d'enregistrer une « *comptabilité mentale* » des rendements obtenus dans le passé, et d'égaliser, au coup par coup, la fréquence relative de choix en faveur d'une option à son rendement relatif (Herrnstein et Prelec, 1991).

Ce type de stratégie, que Herrnstein qualifie de mélioration, n'est nullement spécifique à l'animal et s'observe également chez des sujets humains. A partir des années 1980, les

résultats obtenus sur le pigeon sont étendus à l'homme. Herrnstein et Prelec (1991) montrent notamment que la plupart des individus se contentent de « méliorer » au coup par coup, dans des tâches de machines à sous multi-jeux, et que cette règle suivie par les sujets peut conduire à une allocation des choix globalement sous-optimale. Dans cette étude, les sujets choisissent entre deux machines à sous, générant à chaque fois une récompense monétaire fixe (un cent), avec un délai variable. Ils savent en outre que l'expérience se déroule en temps limité et qu'ils doivent gagner un maximum d'argent pendant la durée de la tâche. Les délais associés à l'option (machine à sous) *1* sont en moyenne de deux secondes plus courts que ceux attachés à la seconde option ; toutefois, ces délais varient selon les derniers choix effectués par le sujet. Pour l'option *1*, le délai augmente à mesure que le choix se répète en faveur de cette même option ; et c'est la logique inverse pour la seconde option. Les rendements relatifs de chaque option sont représentés sur le graphique ci-dessous:

« Choix individuels dans une expérience de mélioration »



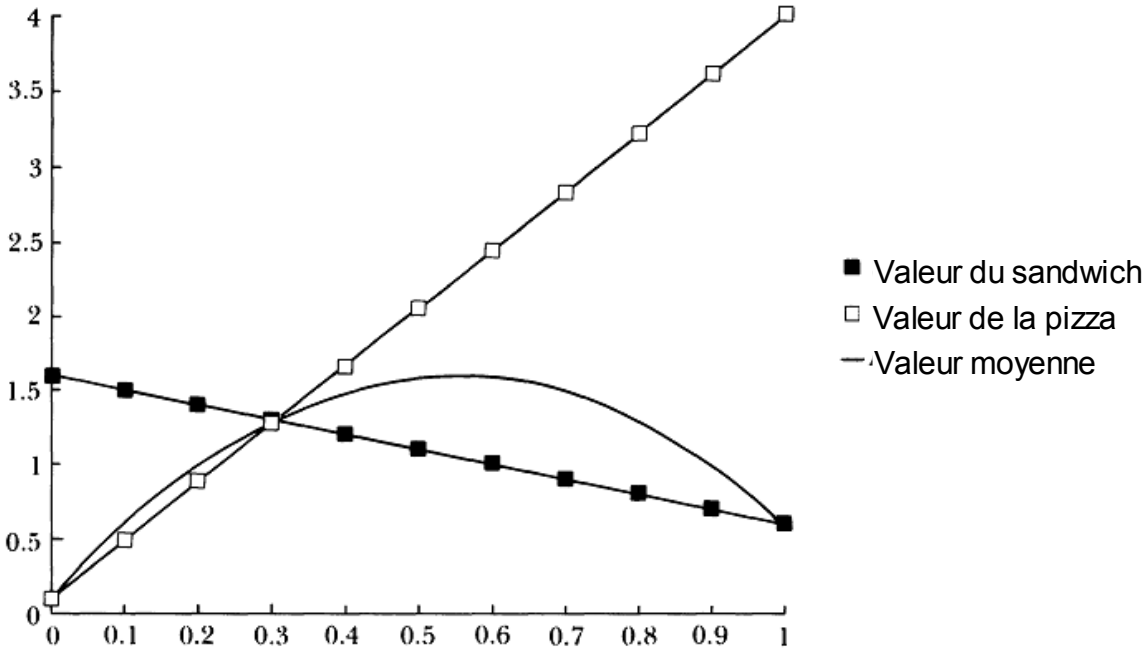
Herrnstein et Prelec, 1991, p. 143

La stratégie optimale consiste à toujours choisir l'option *2*, ce qui offre un délai moyen de 4 secondes[70]. Pourtant, aucun sujet n'a été capable de déterminer le niveau optimal d'allocation des choix. Les individus suivent une stratégie de mélioration qui n'est ici pas compatible avec la maximisation du gain total. En effet, pour n'importe quelle allocation des décisions, l'option *1* apparaît toujours plus rentable pour le sujet que l'option *2* : il est ainsi incité à allouer l'intégralité de ses choix à l'option *1*, pourtant moins favorable globalement que l'option *2*.

C'est donc la structure de la distribution des récompenses entre les deux options qui, en favorisant un comportement de « mélioration », est à l'origine d'une allocation sous-optimale. Dans l'expérience précédente, on observe que l'inefficience résulte en fait de l'interdépendance temporelle des satisfactions attendues : si je choisis *1* en *t*, la récompense attendue en choisissant à nouveau *1* en *t+1* est plus faible que si j'avais choisi *2* en *t*. Or, Herrnstein souligne que cette structure s'observe pour la plupart des choix : en général, les individus manifestent une préférence pour des séquences variées, c'est à dire qu'une trop grande répétition des choix en faveur d'une option tend à diminuer la satisfaction attendue de cette option. Par exemple, dans un choix quotidien et répété entre une pizza et un sandwich, on peut imaginer que la valeur attachée à chacun de ces biens varie en fonction inverse de la consommation passée de ce même bien :

70 Il faudrait éventuellement tenir compte d'un « effet de fin » (*end effect*), qui peut inciter, à la fin de l'expérience, à effectuer quelques choix en faveur de l'option 1 : l'équilibre optimal exact ne consisterait donc pas exactement à accorder 100% de choix en faveur de l'option 2. Cet effet est jugé cependant négligeable par les auteurs (Herrnstein et Prelec, 1991, p.142) et ne modifie pas la conclusion de l'expérience selon laquelle la stratégie de mélioration ne conduit pas à une allocation optimale des décisions.

« Valeur de la pizza et du sandwich, en fonction des repas des 10 derniers jours »



Herrnstein et Prelec, 1991, p. 140

Ici encore, la stratégie de mélioration conduit l'individu à une allocation sous-optimale, puisqu'il choisit l'équilibre situé à l'intersection des deux droites : toute allocation située en dehors de cet équilibre conduit à faire apparaître au moment du choix l'une des deux options comme plus avantageuse, et donc à faire varier la proportion des choix, jusqu'à retourner au point d'intersection des deux droites, dans lequel l'individu est indifférent entre les deux options. Pourtant l'allocation optimale se trouve au sommet de la courbe représentant la valeur agrégée des deux options, à 60% de choix en faveur du sandwich.

Les travaux de Richard Herrnstein reposent donc sur une idée simple, qui vaut comme objection immédiate contre les théories de la maximisation d'utilité : l'incohérence temporelle est une constante comportementale, liée à la nature dynamique et fréquentielle du choix répété. La plupart des individus en effet, observe Herrnstein, se montrent capables d'estimer correctement *ex ante* la satisfaction pouvant être attendue de l'accomplissement différentes activités, et de classer leurs préférences de manière absolue : par exemple, l'individu A préfère en règle générale la lecture à la télévision. Toutefois, ces préférences qui sont exprimées de

manière abstraite ne correspondent pas aux choix réels effectués par l'individu dans les circonstances concrètes d'un choix répété. Le même individu peut donc très bien décider d'allouer 80% de son temps de loisir à la lecture et 20% à la télévision : les fréquences observées diffèrent des préférences absolues, car elles ne correspondent pas du tout au même type de problème décisionnel.

La prise en compte d'une dimension temporelle et délibérative dans la prise de décision complique considérablement et modifie qualitativement la nature des choix. Il s'agit alors de distribuer des fréquences d'activité à chaque option. Il faut donc d'apprécier des satisfactions relatives, qui dépendent de leur ordre d'apparition dans la séquence du choix. En faisant ainsi appel à l'expérience commune, les théories du psychologue Herrnstein connurent un important succès dans la littérature, en psychologie expérimentale mais aussi en économie. Ses modèles du choix distribué et de la mélioration, ont fait l'objet de publications dans les plus grandes revues d'économie au début des années 1990 (Herrnstein, 1991; Herrnstein et Prelec, 1991). Cependant, cette stratégie de recherche qu'Herrnstein qualifie lui-même d'« *hétérodoxe* » (Herrnstein, 1990) à l'égard des modèles à espérance d'utilité a été progressivement abandonnée au profit d'approches psychologiques moins isolationnistes, visant au contraire une conciliation théorique avec la modélisation économique.

2. L'adoption d'un formalisme économique en psychologie comportementale et ses critiques

La stratégie de recherche hétérodoxe, résolument opposée au formalisme économique, poursuivie par Richard Herrnstein rencontre une opposition interne dans les années 1970 et 1980. Des psychologues issus du même courant de recherche néo-comportementaliste contestent en effet, à propos des choix distribués dits asymétriques, la distinction entre mélioration et maximisation (a). Ils défendent par conséquent une vision fonctionnelle de l'hypothèse de maximisation, afin de développer un formalisme d'inspiration économique en psychologie comportementale (b). La succession des décisions dans le temps n'est plus considérée comme un obstacle à la maximisation instantanée des récompenses. Cette attitude plus conciliante vis-à-vis de la modélisation économique suggère que la rationalité engage une intelligence au « second degré », portant non directement sur les choix eux-mêmes, mais sur les séquences (*patterns*) de choix (c).

a. La controverse Rachlin / Herrnstein

Richard Herrnstein lui même avait introduit une équivalence entre loi d'égalisation des rendements et actualisation temporelle hyperbolique, pour des protocoles de choix répétés à intervalle variable (Herrnstein et Ho, 1967 ; *cf. supra*). Toutefois, plutôt que de poursuivre ce rapprochement avec la modélisation économique, et, notamment avec le modèle d'actualisation temporelle non-linéaire de Phelps et Pollack (1968), Herrnstein a par la suite cherché à restreindre la portée de ce résultat. Il a, dans plusieurs publications, défendu l'idée ée selon laquelle l'égalisation observée des rendements par les sujets dans les expériences était incompatible avec l'hypothèse de maximisation dans des protocoles de choix répétés dits « asymétriques » ou à ratio et intervalle variables (*variable ratio - variable interval*).

L'expérience initiale de 1961 portait en effet sur un choix symétrique : le pigeon choisit d'appuyer sur l'un ou l'autre levier, ce qui lui permet d'obtenir une quantité fixe de nourriture. En revanche, dans un choix asymétrique, les deux options ne sont pas équivalentes :le pigeon doit choisir, par exemple, entre appuyer ou non sur un levier. Dans les protocoles typiques de ratio et intervalle variables (*variable ratio - variable interval*) étudiés par les néo-comportementalistes dans les années 1970 et 1980, le choix porte en fait sur deux options. La première option est à « ratio variable » (*variable ratio*, VR), c'est à dire que la récompense dépend d'une réponse du pigeon. En d'autres termes, le pigeon doit appuyer sur ce levier, afin de déclencher une minuterie, qui, après un délai fixe, peut délivrer un montant fixe de récompense si le pigeon appuie à nouveau à l'aide de son bec sur le levier. La seconde option, à intervalle variable (VI), ne fait pas dépendre la récompense d'une réponse du pigeon. La minuterie est enclenchée automatiquement et indépendamment des actions du pigeon. Celui-ci n'a donc qu'à patienter jusqu'au terme du délai avant de collecter la récompense.

La loi d'égalisation des rendements prend ici une forme plus complexe, du fait de la spécificité des protocoles. Baum (1974) a proposé une formulation logarithmique de cette loi pour les protocoles de choix asymétriques :

$$log\ (s_{VI}/s_{VR}) = a.\ log\ (r_{VI}/r_{VR}) + log\ b$$

où s_{VI} et s_{VR} désignent respectivement le nombre de réponses ou le temps passé sur les options à intervalle variable et à ratio variable ; r_{VI} et r_{VR} sont les récompenses obtenues en choisissant respectivement *VI* et *VR* ; *a* et *b* sont deux paramètres.

Dans ces expériences, la maximisation du gain total suppose un fort biais en faveur de l'option *VR*. Rachlin, Green et Tormey (1988) en fournissent une explication intuitive. Le

choix entre les options *VR* et *VI* peut se comprendre à la manière d'un choix entre deux tâches, l'une étant payée « à la pièce » (*VR*), c'est-à-dire que la récompense dépend de l'accomplissement de l'action demandée, l'autre étant payée au temps passé (*VI*). Or, le temps de travail « à la pièce » (choix de *VR*) compte également comme temps de travail pour *VI*. Un individu maximisateur doit donc, selon Rachlin, Green et Tormey, passer l'essentiel de son temps à travailler à la pièce (choisir *VR*), et, de temps à autre, passer à l'autre option pour simplement « collecter » les récompenses qui se sont accumulées au fil du temps.

Le problème est qu'un tel biais en faveur de *VR* n'est pas vérifié expérimentalement. Les pigeons se concentrent en général quasi-exclusivement sur l'option *VI* (Herrstein et Heyman, 1979). Dans l'expérience d'Herrnstein et Heyman (1986), les résultats expérimentaux permettent d'estimer les valeurs des paramètres *a* et *b* à respectivement 1,04 et 0,11, confirmant ainsi la préférence accordée par les pigeons à *VI*. Il n'y a donc pas de maximisation du gain total, puisque celle-ci suppose ici pour les mêmes paramètres des valeurs de 0,72 et -0,38.

Herrnstein a donc ultérieurement utilisé ces résultats comme preuves supplémentaires de l'inadéquation de l'hypothèse de maximisation (Herrnstein, 1990). Pourtant, d'autres auteurs ont essayé de concilier ces tendances comportementales observées en laboratoire avec la théorie de la maximisation d'utilité (Battaglio, Green, et Kagel, 1981). Ces « *partisans de la maximisation* » comme les appelle Herrnstein et Heyman (1986, p.210), qui, eux aussi n'observent pas de biais en faveur de *VR* dans leurs expériences, ont proposé deux arguments théoriques en faveur de la maximisation.

Rachlin, Kagel et Battalio (1982) ont d'abord souligné que, dans le cadre des protocoles à ratio et intervalle variables, la fréquence d'obtention de la récompense n'est pas la seule variable à prendre en compte pour évaluer l'attractivité de chaque option. Il y a en effet, dans la définition de l'expérience, un coût spécifique lié à *VR*, qui peut se comprendre comme la non-obtention d'un loisir associé à l'option *VI* : lorsque je choisis *VR*, je me prive du plaisir de ne rien faire en choisissant *VI*. Rachlin, Kagel et Battaglio (1982) proposent donc d'intégrer une variable supplémentaire dans la formule, qualifiée de « coût du loisir ». Néanmoins, cette solution est un échec, car les comportements observés sont maximisateurs dans cette expérience si et seulement si la valeur du loisir est trois fois supérieure à la valeur de la récompense, ce qui peut être légitimement être considéré comme fantaisiste (Herrnstein et Heyman, 1986).

Une solution plus radicale est avancée par Rachlin, Green et Tormey (1988) dans un

article important intitulé « y-a-t-il un test décisif entre matching et maximisation ? ». Selon les auteurs, l'option *VR*, qui théoriquement devrait être préférée par les sujets maximisateurs, a pour particularité de dissocier la période passée à produire, à « travailler » (*earning period*) et l'obtention du gain à proprement parler (*obtaining period*). Par conséquent, le choix en faveur de *VR* est associé à la représentation *subjective* d'une récompense *décalée dans le temps*. A l'inverse, l'obtention des récompenses dans l'option *VI* ne suppose qu'une action à la fin de la période, ce qui explique pourquoi un tel décalage entre l'accomplissement de la tâche et l'obtention de la récompense n'est pas perçu *subjectivement* par le pigeon.

Pour Rachlin, Green et Tormey (1988), le choix entre *VI* et *VR* porte donc sur deux options qui, respectivement, délivrent des récompenses immédiates ou décalées dans le temps [71]. Les auteurs proposent une modification du protocole afin de valider leur hypothèse. Ils montrent en effet que l'ajout d'une période d'attente supplémentaire lorsque le pigeon passe de l'une à l'autre option (*switching delay)* modifie considérablement les résultats : 25% des réponses sont désormais en faveur de *VR* (contre presque 100% dans le protocole initial) (Rachlin, Green et Tormey, 1988). La variable cachée qui explique la faible attractivité objective de VR est donc le facteur d'actualisation temporelle (*temporal discounting factor*) qui, au moment du choix de VR, réduit le montant perçu derécompense pouvant être attendue. De cette manière, l'hypothèse de maximisation devient compatible avec toutes les valeurs observées des paramètres *a* et *b* de l'équation de Baum (1974). Un sujet maximisateur est susceptible d'adopter toutes les stratégies possibles, dans la mesure où son comportement effectif dépend d'un facteur individuel de *discount* temporel.

71 Pour être précis, cette représentation subjective de la structure du choix n'est valable que lorsque le pigeon, ayant choisi VR, peut collecter une récompense immédiatement disponible en se tournant vers VI. L'argument de Rachlin, Green et Tormey (1986) porte spécifiquement sur ces situations dans lesquelles le sujet peut obtenir une récompense immédiate en changeant d'option. On pourrait en effet objecter que l'option *VI* est également associée à la représentation d'un décalage dans le temps, puisque les récompenses sont distribuées après l'écoulement d'un intervalle de temps régulier. Deux défenses sont possibles. Rachlin, Green et Tormey semblent suggérer que les récompenses associées à *VI* peuvent aussi, effectivement, faire l'objet d'une actualisation temporelle lorsque le pigeon « attend » simplement la fin de la période. Néanmoins, dans les situations évoquées ci-dessus (le pigeon choisit *VR* et la période d'attente de *VI* s'est écoulée), la récompense est perçue comme immédiate *au moment du choix*, ce qui explique un surcroît relatif d'attractivité de *VI* par rapport à *VR*. Cependant, cette solution est encore critiquable : si le pigeon choisi tout le temps *VI*, ce qui est souvent observé dans les faits, un tel supplément d'attractivité relative pour *VI* n'a pas lieu d'être. Si je ne fais qu'attendre passivement l'écoulement des périodes de *VI* ,les récompenses associées ne m'apparaîtront pas comme étant disponibles immédiatement. Une autre interprétation, plus convaincante, pourrait être envisagée. Il semble que la perception, ou non, d'une attente avant l'obtention d'une récompense dépende ici de l'implication, ou de son absence, du sujet dans une tâche. Lorsque je choisis *VR*, la récompense m'apparaît effectivement décalée dans le temps, car je m'implique dans la réalisation de la tâche requise. En revanche, *VI* n'est peut être pas perçue comme une tâche à proprement parler, récompensée par un gain personnel, mais plutôt comme un arrière-fond neutre : les récompenses associées à *VI* ne sont ainsi pas *attendues*.

b. Une vision fonctionnelle de la maximisation

La controverse entre Herrnstein et les partisans de la maximisation portait sur l'interprétation d'une fonction paramétrique réponses/récompenses, estimée à partir de résultats expérimentaux. Il n'y a pas à proprement parler de « vainqueur » à l'issue de cet affrontement théorique, puisque les interprétations proposées sont toutes deux compatibles avec les résultats obtenus en laboratoire. Comme le soulignent Rachlin, Green et Tormey, « *l'égalisation (maximisation) et la maximisation ne sont pas des théories concurrentes sur la nature fondamentale du choix, mais elles sont des points de vue compatibles qui peuvent révéler la fonction de l'environnement et la structure du comportement* » (1988, p. 113). En effet, pour ces auteurs, on l'a vu, ces deux hypothèses ne supposent pas de valeur fixe pour les paramètres de l'équation de Baum (1974).

La solution de Rachlin a néanmoins un avantage décisif sur celle d'Herrnstein, en tant qu'elle conserve une référence à la notion économique de maximisation. Cela permet en effet d'ouvrir la voie à un possible rapprochement avec la théorie économique, en reprenant et en approfondissant le problème posé par Strotz (1955). Ce ralliement à l'économie a cependant un coût, en imposant un assouplissement des conditions d'application de la théorie de l'utilité espérée. Rachlin adopte une vision purement fonctionnelle de l'hypothèse de maximisation : celle-ci est une forme vide. Elle ne prescrit ni de valeur fixe pour les paramètres, ni de stratégie déterminée, puisque ces dernières vont dépendent alors d'un facteur d'actualisation temporelle qui varie d'un individu à l'autre.

La stratégie défendue par le psychologue Howard Rachlin est donc guidée par la volonté *a priori* de donner une coloration économique à des résultats expérimentaux. La vision fonctionnelle de la maximisation n'est en fait pas à proprement parler une hypothèse théorique, car elle ne peut être réfutée par des résultats empiriques : comme l'écrit Rachlin, « *la maximisation, selon ce point de vue, est strictement relative et conditionnelle. La maximisation n'est pas un fait empirique qui doit être confirmé ou rejeté mais une technique d'analyse du comportement* » (Rachlin, 1995, p. 399).

Si la maximisation de l'utilité espérée n'est pas une description du comportement mais une simple technique de modélisation, c'est parce qu'elle est une pure forme logique dénuée de sens. Cette conviction est aujourd'hui largement répandue en théorie des jeux évolutionnaires (voir par exemple, Binmore, 2007 ; Ross, 2005) : la maximisation de l'utilité est une pure tautologie. Pour que cette proposition soit descriptive, il est nécessaire de préciser ce qui est maximisé : « *le problème n'est donc pas de savoir si les animaux*

maximisent la fréquence globale de récompense alimentaire. Ils ne le font manifestement pas. La question à poser est la suivante : qu'est-ce que les animaux maximisent ? Nous ne pouvons présumer ce à quoi un organisme (comme nous même, par exemple) attribue de la valeur. Les valeurs doivent être plutôt définies par les choix » (Rachlin, 1995, p. 399).

Le formalisme économique au sein du courant néo-comportementaliste a donc fait l'objet de critiques internes importantes. La vision fonctionnelle de la maximisation finit par s'imposer parce qu'elle représente un compromis dans ce débat, tout en conservant une orientation économique à ce programme de recherche, ce qui correspond à sa vocation originaire (*cf.* section I). Qu'elle soit dénoncée par Herrnstein comme abstraite, ou représentée par Rachlin comme formelle et non-descriptive, la représentation du comportement sous une forme maximisatrice fait l'objet d'un accord minimal. Surtout, si la question de savoir si les conduites impulsives sont, ou non, maximisatrices, peut faire débat, en revanche les psychologues convergent sur l'idée selon laquelle l'impulsivité est irrationnelle, car celle-ci est inefficace. Les psychologues néo-comportementalistes s'accordent sur le caractère irrationnel de l'impulsivité, au regard d'une rationalité plus élevée, au second degré, que manifestent les sujets normaux.

c. Vers une rationalité au deuxième degré : l'intelligence des séquences

Le programme de recherche emmené par Howard Rachlin affirme que l'impulsivité relève d'une forme de maximisation de la valeur. L'incohérence temporelle des décisions, ou l'incapacité à maximiser un gain financier total dans un choix répété, doit être considérée comme rationnelle, dans les termes de l'analyse économique. Pourtant, les psychologues, approfondissant cette idée d'actualisation temporelle hyperbolique, maintiennent que ces comportements sont inefficaces du point de vue d'une rationalité au second degré. C'est donc en mettant en évidence une telle intelligence temporelle, portant sur les séquences de choix, et non sur les choix isolément, que les psychologues peuvent affirmer qu'un comportement est simultanément maximisateur et inefficace.

Pour les psychologues néo-comportementalistes, une conduite impulsive est déviante dans la mesure où la décision qui en résulte ne correspond pas à celle qui aurait été prise *ex ante*, « à froid », dans des circonstances dans lesquelles aucune récompense n'est accessible immédiatement. En d'autres termes, le problème est que je préfère *A* à *B* en temps normal ; mais dans les situations où *B* est est disponible immédiatement, je peux céder à la tentation de

consommer *B* tout de suite, ce que je regretterai par la suite. Ce cas de figure est fréquent avec les fonctions d'actualisation hyperbolique, car celles-ci peuvent conduire à de brusques inversions de préférences entre deux options à délais différenciés. De ce point de vue, selon les psychologues, la meilleure preuve du caractère perturbant de ce type de comportement réside dans le fait que les individus, très souvent, choisissent *ex ante*, avant la possibilité d'être tenté, de se lier les mains en recourant à ce que l'on appelle des « mécanismes d'engagement » (*commitment devices*).

Les mécanismes d'engagement préalables sont extrêmement divers. Ils renvoient à toute forme de contrainte sur les choix, décidée par l'individu *ex ante*, permettant d'empêcher le choix en faveur de certaines options. Un plan d'épargne à prélèvement automatique empêche par exemple de consommer l'intégralité de son revenu. Ces mécanismes peuvent être plus ou moins formels. Ils peuvent également être adoptés par les animaux. Dans une étude intitulée « le contrôle de l'impulsivité chez le pigeon », Georges Ainslie a montré que les pigeons de laboratoire étaient également capables de suivre de telles stratégies d'engagement. Dans cette expérience, les pigeons devaient choisir soit de presser un bouton, qui délivrait une récompense d'un montant faible dans un délai court, soit de ne pas presser ce bouton, et d'obtenir ainsi une récompense plus importante après un délai d'attente plus long. La totalité des pigeons choisirent d'appuyer sur le bouton. Cependant, dans la deuxième partie de l'expérience, les pigeons ont réussi à adopter une stratégie dite d'engagement qui consistait, préalablement à presser sur un autre levier qui empêchait par la suite tout choix en faveur de l'option à court terme (Ainslie, 1974).

L'existence de mécanismes d'engagement chez l'homme comme chez l'animal, dans les domaines les plus variés, montre selon les psychologues que la plupart des individus préféreraient en règle générale ne pas agir de manière impulsive *s'ils le pouvaient*. L'enjeu consiste dès lors, dans un choix entre plusieurs options, à éviter toute possibilité de distorsion temporelle liée à la possible disponibilité immédiate de l'une des options. Pour cela, les individus doivent se placer dans un horizon de choix suffisamment long, de telle manière que les choix correspondent aux préférences de « long-terme », qui seraient choisies si l'actualisation était proportionnelle au délai. Le choix s'effectue alors en faveur de l'option qui serait préférée en l'absence de toute actualisation temporelle (Ainslie, 1992-a).

Comment un individu caractérisé par une fonction d'actualisation temporelle hyperbolique pourrait-il se comporter comme si cette même fonction était exponentielle ? Une première solution consiste à utiliser des mécanismes d'engagement. Ainslie (1992-a) suggère

également que la quantification stricte des récompenses - par le biais d'une évaluation monétaire notamment - permet souvent de rétablir une proportionnalité du coefficient d'actualisation. Les sujets deviennent alors plus attentifs aux montants absolus des rémunérations : je peux préférer une pomme aujourd'hui à une pomme demain, et préférer deux pommes dans un an et un jour à une pomme dans un an; ce cas classique d'incohérence temporelle peut parfois être corrigé chez certains individus lorsque les pommes sont remplacées par des euros[72]. Dans son article intitulé « Déduction du comportement économique « rationnel » à partir de courbes d'actualisation hyperboliques », Ainslie suggère ainsi que le sous-domaine des comportements concernés par l'estimation monétaire peut se comprendre comme une exception à la règle de l'actualisation hyperbolique, par analogie avec la physique newtonienne qui vaut comme « *cas spécial* » de la physique relativiste (Ainslie, 1992-a, p. 340).

La capacité à neutraliser les effets de distorsions de l'actualisation hyperbolique résulte d'abord et surtout, pour Ainslie comme pour Rachlin, d'une rationalité au second degré, qui porte sur la structure des choix répétés plutôt que sur ces choix pris isolément (Rachlin, 1995 ; Ainslie, 1992-a). Cette intelligence des séquences de décision peut se comprendre, pour reprendre les termes de Rachlin (1995), comme une faculté d'« *empaquetage* » (*bundling*). Supposons que je préfère *A* (manger sainement) à *B* (manger des hamburgers). Mais je succombe pourtant à la tentation de consommer *B*, lorsque l'on m'offre de consommer immédiatement cette option, alors que les bénéfices associés à *A* (santé, diététique) sont perçus à long terme. La solution, pour Rachlin et Ainslie, consiste à envisager chaque décision entre *A* et *B* non pas comme un choix discret, mais comme un choix portant sur la totalité de mes consommations futures. Si je choisis *B*, je dois m'imaginer que cette décision vaudra comme règle pour mes décisions futures, et donc que ce choix porte atteinte au principe selon lequel je préfère, « dans l'absolu », comme disent les sujets, *A* à *B*. Le sujet doit considérer les conséquences de ses décisions en tant qu'elles confirment ou infirment ses principes ou préférences.

En dépit d'importantes critiques internes, la psychologie néo-comportementale portant sur les conduites impulsives a donc progressivement adopté au cours des années 1970 un formalisme économique. Le concept d'impulsivité y est réduit à l'actualisation d'une fonction de *discount* temporel hyperbolique. La science quantitative de la motivation devient effectivement une branche de l'analyse économique à part entière dans les années 1980,

72 Ce n'est cependant pas la règle ; la plupart des sujets, comme on le verra dans la section suivante, maintiennent des choix incohérents avec des récompenses monétaires.

lorsque des expériences étendent à l'homme les premiers résultats obtenus sur les animaux. Néanmoins, l'intégration du néo-comportementalisme à l'économie comportementale pose problème. Celle-ci demeure en effet sous l'influence déterminante de la *behavioral decision research* de Kahneman et Tversky, aux orientations théorique largement opposées à celles de la psychologie évolutionniste (*cf.* introduction à la première partie). L'obstacle théorique le plus important au ralliement à l'économie ne réside finalement pas dans un écart supposé entre les données bio-comportementales et la théorie économique, mais plutôt dans les oppositions entre deux écoles de pensée en psychologie.

B. L'intégration de la science quantitative de la motivation à l'économie

Au cours des années 1980, un champ d'exploration commun à la psychologie et à l'économie se constitue, autour du problème de l'incohérence temporelle et de l'actualisation hyperbolique. L'issue du débat théorique entre Rachlin et Herrnstein a en effet montré une convergence possible entre les implications du modèle de Strotz et les expériences réalisées sur le pigeon. La portée de ces études demande donc, dans un premier temps, à être étendue à l'homme. C'est la raison pour laquelle les psychologues néo-comportementalistes tentent alors de reproduire certains résultats sur des sujets humains, comme chez Herrsntein par exemple (*cf. supra*). Mais, plus généralement, l'intégration du problème de l'impulsivité et de l'actualisation hyperbolique à la théorie économique se fait par l'intermédiaire d'un autre courant en économie influencée par la psychologie expérimentale, l'économie dite « comportementale », qui naît et se développe à la même période (1). Toutefois, l'influence dominante exercée par les travaux de Daniel Kahneman et Amos Tversky sur l'économie comportementale génère un certain nombre de malentendus sur la signification de l'incohérence séquentielle. L'identité originaire de la science quantitative de la motivation tend à se dissoudre lors de son intégration au sein de ce programme de recherche alternatif. La signification des expériences sur le pigeon fait l'objet de deux déformations. Tout d'abord, les économiste comportementalistes proposent des modèles d'actualisation dits quasi-hyperbolique pour rendre compte de l'actualisation hyperbolique observée sur le pigeon et chez l'homme. Cette simplification est trompeuse, car elle permet de représenter l'incohérence séquentielle à la manière d'un biais en faveur de l'utilité immédiate (2). L'impulsivité peut

alors se comprendre comme un effet de cadrage (*framing effects)* kahnemanien. Pourtant, la compréhension de l'impulsivité comme anomalie expérimentale, au sens de l'économie comportementale, soulève des difficultés théoriques, qui témoignent de l'hétérogénéité entre deux approches concurrentes en psychologie (3).

1. La reprise du problème de l'impulsivité par l'économie comportementale: mise en évidence et mesure d'une actualisation hyperbolique chez l'homme

C'est en 1981 que Richard Thaler, économiste de formation, publie dans une revue économique la première étude montrant empiriquement, chez l'homme, la non-proportionnalité de l'actualisation temporelle au délai, pour des récompenses monétaires (Thaler, 1981). Très vite, les études visant à mesurer, chez des sujets humains, les taux d'actualisation à différentes périodes, se multiplient. Ces travaux donnent naissance à un sous-domaine de la théorie économique consacré spécifiquement au choix et à l'arbitrage inter-temporel. Néanmoins, les chercheurs impliqués dans ces recherches n'appartiennent pas au courant néo-comportementaliste étudié dans les sections précédentes mais s'inscrivent, de près ou de loin, dans la filiation psychologique de Kahneman et Tversky. Il en résulte une importante divergence dans la manière de concevoir les protocoles chez l'homme et chez le pigeon.

Les résultats des études sur l'homme sont très variables d'une étude à l'autre, selon le type de récompense utilisée (monétaire, ou en nature) et l'horizon temporel adopté[73]. Néanmoins, l'ensemble de ces recherches a conduit les psychologues et les économistes à conclure que les préférences dynamiques des agents sont régulièrement incohérentes, à partir de trois types de preuves (Laibson, 1997, p. 445).

Tout d'abord, dans de nombreuses expériences en laboratoire, comme par exemple chez Thaler (1981), les sujets qui préfèrent « une pomme aujourd'hui » à « deux pommes demain » préfèrent également, le plus souvent, « deux pommes dans un an et un jour » à « une pomme dans un an ». Des résultats de ce type, qui invalident directement la thèse d'une actualisation exponentielle (Samuelson, 1938)[74], ont été reproduits avec d'autres types de

73 Pour une recension générale des mesures expérimentales du taux d'actualisation, on se référera à Frederick, Loewenstein, O'Donoghue, 2002.

74 Toutefois, s'ils invalident l'hypothèse de l'actualisation hyperbolique, ces résultats ne confirment pas pour autant, comme le souligne avec justesse Rubinstein (2003), l'hypothèse d'une actualisation hyperbolique : ces anomalies peuvent être expliquées par d'autres hypothèses alternatives, notamment par l'idée selon laquelle

biens (monnaie réelle, hypothétique, nourriture, accès à des jeux vidéos, *etc.*), et avec des profils variés de sujets.

Un autre type d'expérience consiste à proposer aux sujets une certaine somme d'argent maintenant contre une plus grande somme disponible après un délai déterminé. En faisant varier progressivement le délai et/ou les montants, il est possible de faire apparaître une incohérence de ces choix chez la plupart des sujets (Frederick, Loewenstein et O'Donghue, 2002). Enfin, le raisonnement sur des séquences de décisions montre que le taux d'actualisation n'est pas constant, du fait de deux traits typiques observables chez la plupart des individus : la préférence pour les séquences ascendantes, qui peut conduire à vouloir retarder le plus longtemps possible une récompense importante, et la non-additivité des préférences dans le temps, qui peut se comprendre comme un goût pour la variété (et/ou une aversion à la monotonie des choix répétés) (Loewenstein et Prelec, 1991). Ces études convergent avec le résultat similaire de l'expérience célèbre de Loewenstein (1987), qui montre comment des « *affects d'anticipation* » (excitation d'une récompense, crainte d'une punition) peuvent avoir un effet de distorsion important sur le coefficient d'actualisation temporelle. Certains individus préfèrent en effet retarder l'obtention d'un plaisir (dans l'expérience de Loewenstein, il s'agissait d'obtenir un baiser avec une célébrité choisie par le sujet) car l'attente elle-même dans ce cas peut être plaisante.

Ces études pourraient être comprises au premier abord comme des confirmation empiriques des résultats obtenus sur le pigeon, dans la mesure où elles montrent également, chez l'homme, la violation de l'hypothèse de constance du taux d'actualisation. Cependant, le protocole de ces expériences ne correspond absolument pas à celui des expériences réalisées sur le pigeon, qui consistaient à répéter un choix entre deux options à délais différenciés. L'intention théorique n'est pas la même. Dans les expériences sur le pigeon, l'actualisation hyperbolique est envisagée comme la conséquence logique d'un comportement d'égalisation (*matching*) des rendements dans une tâche de machines à sous multi-jeux. C'est ce lien logique qui est perdu de vue ici : chez l'homme, il s'agit de démontrer *directement* que le facteur d'actualisation n'est pas constant. A l'inverse, le psychologue néo-comportementaliste Herrnstein a quant à lui conçu une expérience pour des sujets humains qui adapte directement le protocole utilisé pour le pigeon, qui peut donc valoir comme confirmation de la loi d'égalisation des rendements (Herrsntein et Prelec, 1991, *cf. supra*).

Les divergences dans les intentions théoriques s'expliquent donc simplement par le fait

---

les sujets choisissent dans ces arbitrages en fonction de relations de « *similarité* » dans les options proposées.

que ce ne sont généralement pas les mêmes chercheurs qui ont successivement étudié le choix inter-temporel, en laboratoire, sur le pigeon puis chez l'homme. Les économistes et psychologues qui, dans les années 1980, s'emparent du problème de l'incohérence séquentielle appartiennent majoritairement au courant qui est désormais connu sous le nom d'économie comportementale. Lors de son transfert au sein de l'économie comportementale, la notion d'incohérence séquentielle fait ainsi l'objet d'une déformation, non pas tant par l'économie, mais par une école de pensée alternative en psychologie. Cette déformation consiste à présenter l'incohérence temporelle comme la conséquence d'une fonction non pas hyperbolique, mais quasi-hyperbolique. Cette simplification est trompeuse, car elle tend à faire voir le problème du choix inter-temporel comme un biais (*framing effect*) de raisonnement, et non comme la conséquence logique d'un comportement de *matching*.

2. Le modèle d'actualisation quasi-hyperbolique: l'incohérence temporelle comme anomalie

Dans les années 1980 et 1990, les études empiriques mettant en évidence, chez l'homme, l'absence de constance du taux d'actualisation inter-temporel se multiplient. Les modèles proposés par les économistes comportementalistes pour rendre compte de ces violations observées du modèle de Samuelson (1938) ne s'appuient pas sur une fonction d'actualisation hyperbolique, mais quasi-hyperbolique. Cette simplification apparemment innocente est cependant trompeuse, car elle permet de concevoir l'incohérence temporelle comme l'effet d 'un biais excessif pour les utilités immédiates, identifié par un paramètre *ad hoc* de la fonction. Cette approche caractéristique de l'économie comportementale (*cf.* Berg et Gigerenzer, 2010), qui consiste à intégrer un nouveau paramètre dans une fonction d'utilité pour rendre compte d'un écart ou d'une anomalie, va cependant à l'encontre des intentions théoriques des travaux initiaux réalisés sur le pigeon.

En 1997, l'économiste David Laibson propose un modèle « classique » du choix inter-temporel qui a eu par la suite une très grand influence sur la littérature (*cf.* Heukelom, 2009). Dans son modèle, Laibson considère que les individus sont caractérisés par des fonctions d'actualisation ni exponentielles, ni hyperboliques, mais quasi-hyperboliques. La fonction

d'utilité inter-temporelle y prend la forme suivante[75]:

$$U_t(c_t,...,c_T) = u(c_t) + \beta\,[\sum_{k=1}^{T-t} \delta^{k} \cdot u(c_{t+k})]$$

β, compris entre 0 et 1, désigne un biais en faveur du présent, ou, plus précisément, un biais en défaveur des utilités futures. En fait, dans l'actualisation quasi-hyperbolique, l'utilité disponible immédiatement ou à très court-terme $u(c_t)$ ne fait l'objet d'aucune actualisation. Les utilités futures font l'objet d'une actualisation exponentielle, et d'une dévalorisation, liée à un paramètre β, qui marque donc un biais supplémentaire en faveur du présent. L'actualisation quasi-hyperbolique permet effectivement de rendre compte de l'incohérence temporelle. Au fur et à mesure de l'écoulement du temps, une option qui n'est pas préférée dans le plan initial peut devenir proportionnellement plus attractive lorsqu'elle devient (quasi) immédiatement accessible au moment *t*.

Les économistes comportementalistes considèrent généralement, à l'appui des nombreuses études réalisées en laboratoire, aussi bien chez l'homme que chez l'animal, que l'actualisation inter-temporelle est plutôt de forme hyperbolique (voir par exemple Chabris, Laibson et Schuldt, 2008). Ils défendent pourtant l'utilisation de fonctions quasi-hyperboliques pour décrire les comportements, celles-ci étant plus simples à manipuler (Angeletos *et al.*, 2001). D'une manière similaire, Laibson considère que l'actualisation quasi-hyperbolique offre une approximation commode de l'actualisation hyperbolique (Laibson, 1997, p.450).

Le modèle de Laibson permet effectivement de régler des problèmes importants, en supprimant les difficultés liées à l'interprétation normative de l'actualisation hyperbolique. En effet, dans la perspective d'une fonction quasi-hyperbolique, la non-proportionnalité de l'actualisation au délai est liée d'une part à l'absence d'actualisation des utilités immédiates, et à l'existence d'un bais β en défaveur du futur. En d'autres termes, il « suffirait » de supprimer ce biais pour l'actualisation prenne la forme du modèle standard exponentiel. En économie comportementale, l'actualisation quasi-hyperbolique offre l'avantage de comprendre l'impulsivité à la manière d'un effet de cadrage kahnemanien, c'est à dire comme une déformation subjective, à l'origine d'un écart par rapport à un modèle de référence (Samuelson, 1938).

75 Laibson reprend en fait à son compte une fonction d'actualisation temporelle qui avait été proposée en 1968 par Phelps et Pollack (1968) dans un problème d'épargne et d'allocation inter-générationnelle. Le modèle de Laibson (1997) peut donc se concevoir comme une application de la fonction Phelps-Pollack au choix individuel (inter-temporel).

La séparation kahnemanienne entre le descriptif et le normatif (Tversky et Kahneman, 1986, p.252) est maintenue: d'un côté, les *behavioral economics* décrivent les choix réels à l'aide du modèle d'actualisation quasi-hyperbolique; de l'autre côté le modèle exponentiel fournit la norme du comportement rationnel. Tous les choix incohérents sont considérés comme irrationnels. Par exemple, pour Laibson, l'incohérence temporelle représente une erreur que les individus doivent chercher à corriger. L'investissement dans des actifs financiers illiquides permet selon ce même auteur, *ex ante*, de se « lier les mains » et d'empêcher tout renversement des plans initiaux (Laibson, 1997). Inversement, les innovations financières récentes ont eu, selon Laibson, des conséquences négatives sur le bien-être des individus, en facilitant l'accès à la liquidité et en renforçant ainsi les bais en faveur du présent (Laibson, 1997).

Le problème théorique de l'incohérence temporelle a donc fait l'objet d'une incorporation théorique réussie par l'économie comportementale. En promouvant un rapprochement avec l'économie, les psychologues appartenant au courant néo-comportementaliste se sont peu à peu, au cours des années 1980, transformés en économistes (comportementalistes). Georges Ainslie, Howard Rachlin et même Richard Herrnstein sont considérés et se considèrent désormais comme appartenant à l'économie comportementale (voir par exemple, Ainslie, 2001, p.33 ; Rachlin, 1995, p.397; Herrnstein et Prelec, 1991, p.137; Ross *et al.*, 2008, p.65). Pourtant, leur parcours théorique est sensiblement différent de celui des économistes et des psychologues kahnemaniens, qui, à la même époque, accueillent ces nouveaux arrivants et font du choix inter-temporel le pilier de l'économie comportementale, discipline alors en plein essor. Un élément fondamentalement hétérogène résiste à l'assimilation complète à l'intérieur du programme kahnemanien: la psychologie néo-comportementale ne reconnaît en effet pas le partage normatif/descriptif admis par Kahneman, ce qui laisse présager d'importantes oppositions théoriques au sein de l'économie comportementale.

IV. La disparition de la norme économique: la pathologie comme évidence fondatrice de l'analyse

Dans les années 1980, le choix inter-temporel et l'impulsivité deviennent des objets d'étude centraux pour l'économie comportementale. Cette incorporation des thèmes étudiés par la science quantitative de la motivation peut se comprendre comme un renoncement, mais elle laisse également la possibilité, au sein même de l'économie comportementale, d'une rupture avec le programme kahnemanien. En effet, l'incohérence temporelle constitue une anomalie au statut problématique pour les *behavioral economics*. L'impulsivité observée en laboratoire ne semble en effet pas nécessairement irrationnelle, ce qui remet en cause le partage normatif/descriptif établi par Kahneman (Tversky et Kahneman, 1986, p.252) (A). Ces difficultés s'expliquent en considérant, dans une perspective évolutionniste que l'incohérence représente une conduite adaptée à son environnement. Le modèle d'actualisation quasi-hyperbolique représente donc une simplification trompeuse de l'actualisation hyperbolique, car rien ne permet de justifier la définition économique de la rationalité normative à partir du modèle de Samuelson (1938) (B). Si cette disparition de la norme économique n'empêche nullement les psychologues de formuler des jugements appréciatifs, c'est parce que le rôle d'étalon normatif est joué, négativement, par le concept de pathologie. En d'autres termes, c'est en désignant certains comportements comme irrationnels qu'il est possible de définir et d'identifier les comportements rationnels (C).

A. L'impulsivité, une anomalie d'un genre problématique pour l'économie comportementale: la remise en cause du partage normatif/descriptif

Les paradoxes étudiés en économie comportementale (paradoxe d'Allais, d'Ellsberg par exemple) mettent en évidence des défauts de raisonnements, des raccourcis logiques utilisés par les individus pour simplifier la résolution de problèmes décisionnels. Typiquement, la plupart des biais cognitifs mis en évidence par Kahneman et Tversky dans la théorie des *prospects* (1979) appartiennent à cette catégorie d'erreurs logiques. Selon Kahneman, la maximisation de l'utilité espérée représente ainsi une norme ou un optimum,

qui, dans les faits, n'est généralement pas respectée. Largement influencée par cette perspective kahnemanienne, l'économie comportementale se donne pour vocation de construire une théorie descriptive de la décision et délègue à l'économie la tâche d'établir des modèles normatifs. Or l'impulsivité constitue une anomalie expérimentale d'un genre problématique pour l'économie comportementale, car elle tend à remettre en cause le partage établi, au sein du programme kahnemanien entre le normatif et le descriptif (Tversky et Kahneman, 1986, p.252).

Les biais de type kahnemanien sont conçus comme des anomalies, c'est-à-dire comme des erreurs: c'est la raison pour laquelle ils ont une validité descriptive mais non normative. Ils disparaissent en général lorsque, à la réflexion, les sujets comprennent le caractère illogique de leur choix. A l'inverse, le renversement dynamique des préférences n'est pas nécessairement perçu comme une erreur logique par les sujets. Même lorsque l'on indique aux individus l'incohérence de leurs décisions, ceux-ci ne considèrent en général pas avoir commis d'erreurs. Par conséquent, si l'incohérence séquentielle peut être considérée comme une anomalie pour le modèle de Samuelson (1938), elle ne semble pas nécessairement violer les principes normatifs adoptés par les individus: le modèle d'actualisation exponentielle est faux sur le plan descriptif, mais aussi sur le plan normatif.

L'incohérence temporelle représente un écart par rapport au modèle de Samuelson, mais elle peut souvent être justifiée intuitivement. Il n'apparaît pas illogique de préférer à la fois « une pomme aujourd'hui » à « deux pommes demain » et dans le même temps, « deux pommes dans un an et un jour » à « une pomme dans un an ». L'économie comportementale manifeste donc un certain embarras quant au statut normatif à accorder à l'actualisation hyperbolique. D'un côté, le caractère « intuitif » et légitime, du point de vue de l'agent, de l'incohérence temporelle des décisions est parfois admis et a fait l'objet de divers travaux expérimentaux. Par exemple, Loewenstein et Sicherman (1991) proposent à des sujets le choix entre deux séquences de versement de salaires, l'une descendante ($27,000, $26,000,…. $23,000), et l'autre ascendante ($23,000, $24,000,….$27,000). Bien que le total nominal des deux séquences soit égal, les individus devraient toujours choisir la seconde, car celle-ci permet de consommer plus dans chaque période. Pourtant, même après avoir expliqué pourquoi la séquence ascendante devrait toujours être préférée, les individus qui choisissent cette séquence répètent leur décision. Les auteurs proposent deux interprétations possibles. Il se peut tout d'abord qu'un privilège soit accordé *de facto* à toute séquence ascendante sur une séquence descendante similaire. Les individus pourraient également anticiper un futur effondrement de leur volonté conduisant à gaspiller des premiers versements initiaux élevés.

De manière similaire, à propos de l'effet de « magnitude », c'est-à-dire de la variation du taux d'actualisation, pour un même délai, selon l'ampleur du gain en question, les expériences réalisées par Shane Frederick suggèrent que les individus considèrent comme appropriés des taux d'actualisations différenciés selon l'importance du gain ou de la perte.[76]

Certains économistes comportementalistes admettent donc que la constance du taux d'actualisation soit une règle qui, en fait, aille à l'encontre du sens « intuitif » de rationalité des agents. Comme l'écrivent Frederick, Loewenstein et O'Donghue , « *la plupart des anomalies à l'actualisation de l'utilité sont considérées comme des anomalies en référence à un modèle qui a été construit sans aucune attention portée à sa validité descriptive, et qui n'a aucun fondement normatif ou prescriptif* » (Frederick, Loewenstein et O'Donghue, 2002, p.21-22). Cette reconnaissance signifie donc l'abandon d'un dogme important de l'économie comportementale, lequel accorde, à la suite de Kahneman, au moins à la théorie économique une validité normative. Dans le cas de Georges Loewenstein, cette attitude s'explique peut être par une influence moins grande de Kahneman, comme le suggère Heukelom (2009).

Cependant, d'un autre côté, l'incohérence temporelle peut aussi être considérée comme une erreur lorsqu'elle conduit à prendre des décisions manifestement irrationnelles. S'il peut apparaître raisonnable de préférer une pomme aujourd'hui à deux pommes demain, et préférer deux pommes dans un an et un jour à une pomme dans un an, la consommation de drogue dure représente pour les psychologues et les économistes comportementalistes un cas manifeste d'irrationalité. Une perspective évolutionniste sur ces comportements permet de comprendre pourquoi l'impulsivité est rationnelle ou irrationnelle selon les circonstances.

76 L'expérience de Frederick est décrite dans la recension de Frederick, Loewenstein et O'Donghue (2002). Cette expérience montre que l'effet de magnitude est plus prononcé lorsque les sujets évaluent à la fois de faibles et de larges montants, que lorsqu'ils évaluent uniquement de l'un ou l'autre type de gains. La différence dans les taux d'actualisation entre un faible montant (10$) et un montant important (1000$) est plus grande lorsque les deux jugements sont réalisés successivement que lorsqu'ils sont réalisés indépendamment. En ayant en tête le taux d'actualisation utilisé implicitement pour évaluer le choix précédent, les individus continuent de maintenir, et même accentuent la différence entre les taux d'actualisation. Cela suggère donc que les sujets considèrent cette variation comme légitime ou justifiée. Des résultats identiques sont obtenus à propos des gains et des pertes dans une expérience ultérieure (Frederick, Loewenstein et O'Donghue, 2002).

B. Justification biologique de l'actualisation hyperbolique: l'impulsivité comme dérèglement sans norme

Dans le modèle d'actualisation quasi-hyperbolique de Laibson (1997), le comportement rationnel donc celui qui serait adopté si l'individu avait une fonction d'actualisation exponentielle. La rationalité du choix inter-temporel est ainsi définie par la proportionnalité de l'actualisation au délai. Or, pour les psychologues néo-comportementalistes, l'absence de constance du facteur d'actualisation n'est pas toujours irrationnelle, dans la mesure où elle est la conséquence nécessaire, dans les tâches type machines à sous multi-jeux, de routines de mélioration (*cf. supra*). Dans une perspective évolutionniste, l'impulsivité résiste donc à un traitement simple en terme d'anomalie. Rien ne permet de justifier alors *a priori* que l'actualisation exponentielle représente la norme du choix rationnel. Le modèle d'actualisation quasi-hyperbolique est donc une simplification trompeuse de l'actualisation hyperbolique, en présupposant que toutes les conduites incohérentes représentent des anomalies comportementales.

L'adoption d'un modèle hyperbolique (et non pas quasi-hyperbolique) débouche donc sur des problèmes normatifs complexes, relatifs à la question de savoir si l'impulsivité est « bonne » ou non pour l'individu considéré. Ces problèmes seront approfondis dans le septième chapitre. Sans anticiper sur ces développements, il convient de souligner ici que l'approche évolutionniste du problème de l'incohérence temporelle aboutit à brouiller la frontière entre la rationnel et l'irrationnel.

Selon les biologistes et les psychologues qui travaillaient sur le pigeon, les conduites impulsives sont, dans une certaine mesure, adaptées à leur milieu : pour les animaux, un certain de gré d'impulsivité est par exemple nécessaire pour profiter des opportunités de l'environnement, en particulier lorsque celui-ci est très instable. Par ailleurs, le biais positif en faveur des possibilités de récompense à court terme permet aussi de simplifier la prise de décision, en réduisant le nombre d'options envisagées (Ainslie, 1992-b, p.71). L'actualisation hyperbolique constitue un avantage évolutif, ce qui explique sans doute pourquoi elle constitue une régularité comportementale, observable aussi bien chez l'homme que chez l'animal (*cf.* section 2). En d'autres termes, ce qui est regardé comme un biais dans le cadre d'une psychologie du choix statique, d'inspiration kahnemanienne, est regardé comme avantage évolutif dans le cadre de la psychologie évolutionniste.

Cet avantage évolutif peut néanmoins se transformer en inconvénient, dans certains

environnements, dont la structure tend à rendre plus saillantes les utilités et les récompenses pouvant être obtenues à très court-terme. Toutefois, à la différence de la notion d'effet de cadrage, cette conception environnementaliste du choix n'explique pas l'irrationalité par une déformation subjective, individuelle: le défaut ou le vice n'est pas dans l'individu, mais dans le milieu. C'est la raison pour laquelle Georges Ainslie considère que l'impulsivité, ou la « *défaite de la volonté* » pour reprendre ses termes, est liée à une pathologie non pas des êtres humains, mais de la civilisation dans son ensemble. La « *pathologie du loisir* » (*pathology of leisure*) désigne ainsi la tendance, au sein des sociétés et des environnements construits par l'homme, à offrir de manière excessive et disproportionnées des moyens de satisfactions immédiats de nos désirs (Ainslie, 1992-b, p.3).

L'actualisation hyperbolique sert donc, dans le néo-comportementalisme, à formaliser un phénomène très ambigu, qui peut être par certains côtés être considéré comme rationnel et par d'autres comme irrationnel. Une telle ambiguïté apparaît par exemple dans les travaux de Richard Herrnstein. Celui-ci considère en effet que les stratégies de mélioration dans les problèmes de choix dits distribués (*cf.* Herrnstein, 1990) sont compréhensibles intuitivement, si l'on se met à la place du sujet qui n'a pas de vision globale du jeu, et se contente d'améliorer comme il le peut, au coup par coup, ses décisions. Toutefois, Herrnstein déduit également de cette théorie de la mélioration et des choix distribués un modèle explicatif de l'addiction (Herrnstein et Prelec, 1992). Cela suggère qu'il n'y aurait pas au fond de différence de nature entre l'individu normal et l'*addict*, et que ce dernier ne serait qu'un joueur moyen parmi les autres, confronté à la possibilité de consommer une substance addictive.

Encore est-il possible de mettre en évidence, chez Herrnstein, une frontière assez claire entre le rationnel et l'irrationnel. En effet, la maximisation ou non du gain total permet de décider de la rationalité ou de l'irrationalité de la stratégie suivie par l'individu. Pour Herrnstein, la quasi-totalité des choix répétés sont irrationnels au sens économique du terme, c'est-à-dire que l'immense majorité des individus sont en pratique incapables de maximiser leur utilité subjective. De ce point de vue, il existe chez Herrnstein une ligne de démarcation nette entre les comportements rationnels et irrationnels, même si ces derniers peuvent avoir des conséquences plus ou moins dommageables, entre perdre une pomme ou devenir dépendant à l'héroïne.

Les choses se compliquent en revanche dès lors que l'on adopte un formalisme économique pour rendre de compte de l'incohérence séquentielle, en présentant celle-ci comme le résultat de la maximisation d'une fonction d'actualisation temporelle hyperbolique. Les psychologues néo-comportementaliste ne peuvent pas, en effet, soutenir de manière

cohérente que le comportement rationnel consiste à actualiser de manière exponentielle, puisque l'actualisation hyperbolique est au départ un avantage évolutif.

Les psychologues néo-comportementalistes manifestent donc des difficultés lorsqu'il s'agit de définir précisément le contenu d'une décision inter-temporelle rationnelle. Étant eux-mêmes devenus économistes comportementalistes, ils soutiennent, comme c'est l'usage dans cette discipline, qu'un choix inter-temporel « sophistiqué » requiert la prise en compte d'une trahison possible, dans le futur, des plans de consommation adoptés *ex ante*. Dans cette perspective, la rationalité dépend d'une intelligence au second degré, c'est-à-dire d'une aptitude à évaluer la séquence des choix et non les choix pris isolément (*cf. supra*).

Pour autant, l'enjeu ne consiste pas nécessairement à privilégier exclusivement le long-terme sur le court-terme, en élargissant autant que possible l'horizon temporel des séquences (*patterns*). En effet, les psychologues néo-comportementalistes qui ont travaillé au départ que le pigeon savent que les conduites qui résultent de l'actualisation hyperbolique servent au départ des intérêts biologiques. Par conséquent, une trop grande sophistication temporelle peut aussi être néfaste, dans la mesure où elle devient contraire à cet intérêt biologique élémentaire. Une trop grande attention portée à des *patterns* de moyen-long terme -ce que Strotz appelle « *parcimonie* » (Strotz, 1955)- peut aussi s'avérer pathologique. Comme l'écrit Rachlin, « *l'abstinent qui boit un verre ou le travailleur acharné (workaholic) qui prend un jour de congé doivent tous deux faire face à un affreux conflit interne* » (Rachlin, 1995, p. 403).

Si le bien-être réel ne peut être compris en référence aux seuls intérêts de long-terme de l'individu, il peut sembler délicat de comprendre ce qui constitue au fond les préférences rationnelles de individu. Pour les psychologues Ainslie et Rachlin, un décideur rationnel se caractérise donc non pas par un taux d'actualisation constant, mais plutôt par la capacité à maintenir un équilibre suffisamment stable entre les différents horizons temporels. L'individu doit réussir à résoudre le conflit interne vécu entre les multiples dimensions temporelles de la décision. Ceci passe notamment par l'établissement de règles personnelles (*personal rules*) : par exemple, un alcoolique définit un niveau acceptable pour sa consommation d'alcool.

Il serait alors tentant de définir la rationalité de la prise de décision par l'existence de ces règles[77]. Néanmoins, celles-ci ont également un inconvénient : lorsqu'elles deviennent trop

77 Comme nous le verrons dans le chapitre 7, les économistes comportementalistes considèrent précisément que les règles de planification inter-temporelle prises par les individus, et, plus généralement, tous les mécanismes d'engagement auxquels souscrivent les individus démontrent l'existence d'une tendance irrationnelle chez la plupart des agents, que ceux-ci tentent de contrôler (*cf.* chapitre 7). Implicitement, la rationalité est définie ainsi, pour un individu caractérisé par un taux d'actualisation non-constant, par

strictes, elles tendent à rigidifier les choix et peuvent paradoxalement donner lieu à des conduites pathologiques, que les psychologues qualifient de « compulsion ». Le cas des troubles du comportement alimentaire fournit une bonne illustration : il est bien connu que l'utilisation de règles trop strictement suivies pour la consommation alimentaire quotidienne (chiffrage calorique par exemple) est à la source de rituels alimentaires qui sont souvent liés à l'anorexie et la boulimie. La compulsion est un effet de ces règles personnelles (Ainslie, 1992-b, p. 217). Comme le souligne Ainslie, ces dernières peuvent éventuellement protéger les intérêts de moyen-long-terme contre de possibles tentations, mais elles peuvent également à leur tour devenir des « *prisons* » (Ainslie, 1992-b, p. 217). Distincte à la fois de l'abstinence et de l'impulsivité, la rationalité semble alors indéfinissable. Faut-il y voir le signe d'une aporie pour l'économie comportementale ?

C. La pathologie comme solution à l'absence de norme

Le néo-comportementalisme contourne les difficultés liées à la définition d'une norme de la rationalité pour le choix inter-temporel en adoptant un critère négatif de la norme: plutôt que de déterminer explicitement ce qu'est le choix rationnel, les chercheurs envisagent un certain nombre de comportements comme irrationnels ou impulsifs parce que pathologiques. A l'inverse du programme kahnemanien qui considère des écarts descriptifs à une norme déterminée par la théorie *économique* (Tversky et Kahneman, 1986, p.252), le programme de recherche néo-comportementaliste part à l'inverse de l'"intuition *psychologique* d'un possible caractère pathologique des conduites impulsives. L'enjeu ici consiste à analyser cette évidence paternaliste selon laquelle l'impulsivité constitue une menace pour l'intégrité de l'individu.

L'incohérence temporelle n'est pas nécessairement irrationnelle dans la mesure où elle constitue un avantage évolutif. D'un point de vue économique, elle peut aussi recevoir une justification descriptive par leur logique maximisatrice. Or, les psychologues et les économistes s'accordent pour considérer que certaines conduites incohérentes, disqualifiées comme impulsives, sont néanmoins incontestablement déviantes. Cette distinction s'appuie sur le sentiment d'évidence selon lequel les individus impulsifs agissent à l'encontre de leur intérêt réel : l'impulsivité est donc comprise comme une pathologie au sens médical du terme.

---

l'adoption de règles, c'est-à-dire par des mécanismes qui permettent précisément de rétablir un taux d'actualisation uniforme pour tous les horizons temporels

L'impulsivité est donc un concept ambigu. D'un côté, les études quantitatives en laboratoire et les modèles de l'économiste établissent des critères descriptifs stricts. De l'autre côté, l'idée paternaliste d'une menace contre individu est projetée sur ces essais de définitions formelles, ce qui explique que sa nocivité soit unanimement admise par le psychologues et les économistes. Le cas des comportements impulsifs montre ainsi comment l'objet scientifique fournit à la pensée théorique à la fois une matière et un imaginaire : l'incohérence temporelle est un concept objectif, mesurable, et il est pourtant, dès le départ, implicitement associé à des représentations de « conduites à problème ».

L'article de 1955 de l'économiste Robert Strotz sur la myopie temporelle en fournit une bonne illustration (Strotz, 1955). Strotz y propose, on l'a vu, une explication théorique de l'inconstance temporelle par le caractère séquentiel de la prise de décision et par la modification progressive des plans de consommation à chaque période. Or, si ces conduites sont « excusables » d'un point de vue de la théorie de la rationalité, elles mènent selon Strotz à conséquences tout à fait dommageables pour les individus, en particulier chez les classes populaires :

> « *C'est surtout parmi les classes à bas revenu, dans lesquelles les prédispositions et l'éducation sont généralement limitées, que l'on peut s'attendre à trouver des comportements imprudents de cette nature. En Amérique, les individus ayant de faibles revenus tendent à se gaver de nourriture après avoir obtenu leur paye; à chauffer trop fort leur maison lorsqu'ils ont de l'argent pour acheter un sac de charbon, à se comporter de manière extravagante, à faire la fête le jour de paye, à ne pas faire de budget, à s'engager dans de lourds paiements différés, à ne pas garder leurs enfants à l'école, et à laisser libre cours à leurs pulsions agressives et sexuelles. Leur fort taux de natalité est bien connu. Toutes ces caractéristiques comportementales peuvent être expliquées par l'échec à comprendre de manière intelligente le problème du choix inter-temporel* » (Strotz, 1955, p.177-178)

Dans une perspective relativement proche, Richard Herrnstein considère lui aussi que de nombreux cas de dérive résultent de la dynamique du choix. L'incohérence temporelle est là aussi associée à l'idée d'une faillite personnelle: « *ces phénomènes* [concernant l'impulsivité] *partagent un élément commun : ils sont tous des instances du choix distribué* […]. *Une personne ne prend pas la décision, une bonne fois pour toutes, de manière isolée, de devenir droguée, débauchée, avare, gloutonne, ou joueur compulsif ; elles se laisse plutôt aller à une série de choix innocents, ou presque innocents, ne portant isolément qu'à de faibles conséquences* » (Herrnstein, 1990, p.149).

Ce registre paternaliste est donc également utilisé par des psychologues qui, comme Richard Herrnstein, ont pourtant d'abord étudié le choix dynamique sur des pigeons de

laboratoire. La notion de pathologie est de ce point de vue équivoque, car elle permet de qualifier l'impulsivité de nocive *aussi bien chez l'homme que chez l'animal*. Georges Ainslie parle ainsi d'une « *pathologie du loisir* » pour évoquer la tendance à la surconsommation impulsive dans les sociétés occidentales marquées par une abondance de biens matériels (Ainslie, 1992-b, p.3). Le terme apparaît aussi chez les économistes. Akerlof, par exemple, se réfère dans son article sur la procrastination à des « *modes de comportement pathologiques* », à des individus qui prennent des décisions sérieusement néfastes (*seriously wrong decisions*). De telles pathologies doivent être étudiées par l'économiste car elles affectent la performance économique et sociale de certains institutions et certains individus (Akerlof, 1991).

La référence à la maladie laisse ainsi apparaître la possibilité d'une psychiatrie économique. Cette dernière ne repose pas tant sur une supposée réduction de l'homme à l'animal (ou, dans le cas de la neuroéconomie, sur une réduction de l'homme au cerveau), mais plutôt sur la projection d'expériences humaines liées à la pathologie sur des données biologiques et comportementales. Comme l'écrit Pierre-Henri Castel à propos des études faites sur les animaux, « *ce n'est donc jamais* […] *l'animal qui émerge « sous » l'homme, comme si on épluchait des couches successives jusqu'à trouver un noyau essentiel, mais l'homme qui est toujours déjà présent au cœur de l'animal testé* » (Pierre-Henri Castel, 2009, p. 50). Cette projection est essentielle, car elle permet d'identifier une simple tendance du comportement à des conduites jugées déviantes.

Si les économistes et les psychologues s'accordent donc pour rejeter l'impulsivité comme déviance, c'est parce qu'ils ne remettent pas en question le caractère déplaisant des troubles de l'impulsivité. L'utilisation du terme médical de « pathologie » sert d'argument d'autorité. Il faut donc parler d'évidence *psychologique* (ou médicale), car la simple description dans le langage neutre de l'économie, sous la forme de processus d'optimisation, ne saurait fournir la justification d'une telle disqualification. Pourtant, l'originalité de ce programme de recherche consiste aussi, en partant de cette intuition fondatrice de déviance, à utiliser cette modélisation d'inspiration économique. L'absence de définition précise et quantifiée de la rationalité peut se comprendre comme une limite de cette approche pour l'économie. Mais cette ambiguïté est aussi une caractéristique constitutive du néo-comportementalisme, lequel viserait à développer une approche économique d'un trouble mental défini dans les termes de l'analyse psychiatrique.

V. Impulsivité et maximisation de la normalité – Les objets hyperboliques du néo-comportementalisme

Les problèmes symétriques rencontrés respectivement par une trop grande rigidité et une trop forte incohérence du choix inter-temporel rendent difficile toute tentative pour définir la notion de rationalité. Comme l'écrit Ainslie, « *l'actualisation hyperbolique et les règles personnelles conçues pour contrôler celle-ci ont toutes deux des effets déformants* […] *Par conséquent, la rationalité devient un concept indéfinissable* […]. *Rationnel signifie être systématique, mais à la condition que le système n'aille pas trop loin et ne devienne pas compulsif* » (Ainslie, 1992-b, p. 154).

La quantification d'un trouble mental *via* l'emprunt d'un formalisme économique aboutit donc à un résultat contrasté. D'un côté, l'impulsivité fait l'objet d'une définition analytique mesurable en laboratoire. Mais de l'autre côté, la quantification de cette conduite *présumée* déviante a pour effet d'estomper la ligne de démarcation entre rationalité et irrationalité, car le comportement irrationnel n'est plus différencié qualitativement de l'irrationnel. Si le pathologique peut encore être saisi par l'évidence médicale de sa nocivité, la notion de rationalité économique est en revanche indéfinissable, parce qu'elle ne peut être comprise comme une quantité, mais plutôt comme la capacité à varier librement à l'intérieur d'une marge acceptable : comme l'écrit Ainslie, « *il n'y a pas de formule pour la rationalité* » (Ainslie, 1992-b, p. 154).

Cette limite fondamentale posée à la quantification peut se lire comme un constat d'échec. Or, dans les faits, cette difficulté n'est pas conçue comme telle par des psychologues comme Ainslie. Ce problème théorique se résous en pratique par un jugement d'évidence. L'obstacle à la quantification est tout à fait admis par Ainslie, par exemple, qui reconnaît qu'il est nécessaire de recourir à des données subjectives, relatives au vécu de l'individu, en particulier lorsqu'il s'agit de distinguer des règles de conduite saines et des rites compulsifs : « *le fait que la plupart des autres pathologies compulsives n'aient pas encore été décrites s'explique peut être par la difficulté à quantifier leur dommage. L'échec à « dépasser une règle qui est devenue trop rigide pour le cas présent » est au mieux un échec relatif, associé au nécessaire contrôle d'une pulsion qui s'est montrée antérieurement dommageable pour l'individu* » (Ainslie, 1992-b, p. 222). Si l'équilibre entre régulation saine, compulsion, et impulsivité est précaire, chacun sait néanmoins, selon Ainslie, distinguer un individu sain d'un

individu malade. En d'autres termes, l'impulsivité est une catégorie comportementale très souple, mais qui n'est pas pour autant dénuée de signification : son application peut être adossée à un jugement d'autorité de type médical.

Le projet de la psychiatrie économique conduit donc simultanément à quantifier et élargir la notion d'impulsivité, conformément à l'idée d'une « *maximisation de la normalité* » proposée par Robert Castel (*cf.* section I.) L'échec de la quantification n'est que relatif, car il est la condition d'un déploiement tous azimut d'un trouble comportemental. En effet, c'est parce que les techniques comportementales ne parviennent pas à chiffrer la maladie mentale qu'un brouillage apparaît entre le rationnel et l'irrationnel, et que, par suite, il est possible de considérer que tous les individus, en tant que normaux, sont susceptibles d'être malades. La catégorie de l'impulsivité permet de renvoyer à un spectre extrêmement large de troubles comportementaux, des plus bénins aux plus dangereux, de l'habitude de se ronger les ongles à l'addiction aux drogues dures. Il est dès lors aisé de comprendre que tout le monde est susceptible d'être menacé par un éventuel « *échec de la volonté* », pour rependre le titre de l'ouvrage de George Ainslie (2001)[78].

La proposition de Castel peut néanmoins faire l'objet de deux interprétations différentes. La première revient à considérer, comme semble le suggérer Castel lui-même, que, dans le cadre de la psychiatrie économique, la condition de cette application extrêmement large des catégories cliniques repose sur une absence de définition claire des troubles mentaux, qui sont ramenées à de vagues « handicaps fonctionnels ». Dans cette perspective, l'impulsivité constitue un objet hyperbolique, parce que sous-déterminé empiriquement. En jouant sur l'indétermination de ce concept, les psychologues seraient ainsi en mesure de cibler un plus grand nombre de patients potentiels.

A l'analyse, le développement théorique du néo-comportementalisme favorise cependant une deuxième interprétation, qui, semble-t-il, n'a pas été envisagée par Castel. Le problème peut en effet être déplacé, de la définition du trouble -l'impulsivité- vers celle de la récompense. Rien n'empêche de penser que le concept d'impulsivité est bien logiquement déterminé à l'intérieur du modèle du conflit intrapersonnel de Ainslie (1991). Il permet de maintenir au niveau empirique une frontière claire entre le rationnel et l'irrationnel, *à condition de définir explicitement ce que doit constituer une récompense « rationnelle »*. Dans les expériences de Herrnstein (1961), c'est le choix d'une mesure de la récompense par le temps qui a permis d'introduire une équivalence entre loi d'égalisation des rendements et

78 Sur l'extraordinaire diversité des comportements impulsifs, on se référera à Ainslie, 1992-b.

fonction d'actualisation temporelle hyperbolique. L'équivocité de la notion de récompense attendue a ainsi rendu possible le rapprochement avec l'économie, *via* le concept d'utilité espérée (*cf.* section 1).

Pour les biologistes, il paraît évident qu'une récompense soit une substance qui favorise le développement de l'organisme, qui peut donc être clairement identifiée par certaines propriétés physico-chimiques, comme par exemple les propriétés nutritives d'un aliment. Or, et c'est là l'un des éléments central de sa théorie, Ainslie ne reconnaît aucune récompense « objective » de ce type : une substance est reconnue comme récompense si et seulement si l'organisme la reconnaît comme telle. En d'autres termes, il n'y a de récompense que subjective, et ce même pour les formes les plus élémentaires de récompense (alimentaire par exemple) :« *l'on ne se récompense pas soi-même en ingérant une substance identifiée comme récompense, mais selon sa propre théorie de ce que doit être une récompense [...] même si la valeur d'une substance identifiée comme récompense doit d'une manière ou d'une autre être associée à un certain type de récompense objective [alimentaire, monétaire, etc.], lorsque l'on demande néanmoins aux individus ce qui est nécessaire pour constituer une récompense, ceux-ci ne peuvent de manière fiable identifier les éléments objectifs mis en évidence par la recherche en laboratoire. En effet, la récompense en tant que telle est indéfinissable* » (Ainslie, 1992-b, p.272).

Par delà l'apparente simplicité de sa définition biologique, la notion de récompense est en fait l'ultime atome d'évidence sur lequel s'appuie l'édifice de la psychiatrie économique. Cet obstacle peut, une nouvelle fois, être envisagé comme une limite fondamentale rencontrée par une analyse purement quantitative d'un trouble mental. Mais il est aussi possible de concevoir cette difficulté comme un moteur théorique : la circularité logique de la définition de la récompense comme « ce qu'un organisme considère comme récompense » rend nécessaire une théorie explicative de la manière avec laquelle un organisme en vient à accorder certaines valeurs aux objets de son environnement. L'enjeu, dès lors, pour ce programme de recherche, consiste non pas à mesurer la récompense, puisque cette mesure est par définition variable selon les individus mais à construire un nouveau concept permettant de rendre raison de sa formation. C'est ainsi que la notion d'*apprentissage de la récompense* a permis à la science quantitative de la motivation de franchir un nouveau cap important, à partir des années 1980.

Conclusion du chapitre 2

La psychologie néo-comportementale du choix inter-temporel se construit à partir d'éléments isolés : des comportements impulsifs et des tentatives de *self control* observées en laboratoire ; des techniques expérimentales ; une référence faite à l'hypothèse formelle de maximisation, inspirée de l'économie ; un problème pratique (l'échec d'un mode de gestion asilaire de la maladie mentale). L'idéologie scientifique fournit la pure intuition d'une synthèse entre ces éléments, à partir de la proposition suivante : l'impulsivité est un comportement pathologique qui résulte de la maximisation d'une fonction d'actualisation temporelle hyperbolique.

L'histoire de ce programme de recherche montre que cette idéologie s'est construite en psychologie expérimentale, par l'adoption progressive d'un formalisme d'inspiration économique. En introduisant une approche économique, c'est-à-dire quantitative et fonctionnelle, de la maladie mentale, la théorie néo-comportementale de la myopie temporelle jette un regard ambivalent sur les troubles du comportement. Étant subsumés sous le concept de maximisation, ces derniers représentent à la fois une déviance (psychologique) et un optimum (du point de vue de la maximisation de la valeur).

Il y a là incontestablement un *hiatus* théorique, qui explique l'incapacité des néo-comportementalistes à quantifier de manière absolue les déficits fonctionnels qu'ils étudient. En effet, s'il est admis que le pathologique n'est qu'une variation quantitative du normal, il est nécessaire, paradoxalement de s'en remettre à un jugement subjectif et qualitatif pour distinguer le rationnel de l'irrationnel. La psychiatrie économique s'est donc élaborée à partir de deux gestes spéculatifs décisifs : le premier, de l'animal à l'homme, a permis d'étendre la signification d'études biologiques en projetant sur l'animal des expériences humaines; le second, en séparant la déviance du normal, permet de maintenir une limite entre rationalité et irrationalité.

Ces « dépassements présomptueux » sont des présupposés théoriques, dont la validité n'est que rarement interrogée. Ils sont donc préalables à la recherche à proprement parler, parce que logiquement antérieurs. Ils pourraient valoir comme obstacle fondamental à la scientificité de ce programme de recherche. Effectivement, la sous-détermination empirique des concepts fait apparaître une importante marge d'appréciation subjectives des résultats. Néanmoins, si ce degré de liberté autorisé à l'interprète génère des controverses théoriques, il est aussi moteur du progrès scientifique. L'absence de définition objective de la *récompense*,

qui concentre au fond et explique toute l'imprécision de l'analyse, appelle en effet une théorie explicative de son *apprentissage*. La neuroéconomie hérite ainsi à la fois d'une idéologie scientifique, c'est-à-dire d'une programme de recherche structuré, et d'un problème théorique.

Chapitre 3. Les années 1990 : La vocation économique de la neurobiologie – Paul Glimcher et l'« utilité espérée physiologique »

Les psychologues appartenant au courant néo-comportementaliste étudié dans le cadre du chapitre précédent défendent un rapprochement de leur discipline avec l'économie, au nom d'un intérêt théorique commun portant sur le choix inter-temporel (*cf.* chapitre 2). La dynamique de recherche qui s'organise autour du projet de psychiatrie économique a également mobilisé, au cours de son développement, l'appui des neurosciences. Au cours des années 1980, le néo-comportementalisme annexe ainsi de nouvelles thématiques de recherches expérimentales, portant en particulier sur l'activité électrophysiologique du système nerveux. La neurobiologie est alors en profonde mutation, avec l'apparition des micro-électrodes qui permettent d'enregistrer *in vivo*, chez l'animal, l'activité de neurones individuels. En s'inspirant de la mutation des sciences comportementales, plusieurs équipes de recherche en électrophysiologie en viennent à défendre l'idée d'un rapprochement entre neurobiologie et l'économie. Ces travaux posent ainsi les jalons fondateurs ce de qui a été appelé ultérieurement, au cours de la décennie suivante, « neuroéconomie ».

L'approche néo-comportementaliste a donc été enrichie par l'introduction de ces nouvelles techniques expérimentales (les micro-électrodes). Or, il apparaît assez clairement que cette innovation technologique a joué en fait un rôle secondaire par rapport aux débats théoriques, en amont, en biologie, portant sur l'interprétation et la signification des données elle-même. En d'autres termes, la mesure des potentiels d'action neuronaux n'a pas été à l'origine directe de changements théoriques majeurs. En revanche, la possibilité d'une articulation entre ces données et la loi d'égalisation des rendements (*matching law*) s'est révélée être une heuristique de recherche féconde. L'enjeu de ce chapitre consiste donc à décrire la manière avec laquelle les micro-électrodes ont été « mises au service » d'une convergence entre neurophysiologie et le programme de recherche portant sur le *reward learning*.

Les études neurobiologiques abordées dans ce chapitre se donnent donc comme un prolongement de la science quantitative de la motivation, tout d'abord parce que ce sont des tâches du même type (machines à sous multi-jeux) qui sont étudiées dans ces expériences de

neurophysiologie des années 1990. Les pigeons conditionnés à donner des coups de bec sur des leviers sont remplacés par des singes conditionnés à effectuer des mouvements oculaires, mais l'objectif reste le même: l'observation des comportements de *matching*. Le néo-comportementalisme constitue ainsi une identité théorique commune à la fois aux psychologues étudiant le choix répété chez le pigeon ou chez l'homme, et aux neurophysiologistes, autour de deux principes fondateurs. Tout d'abord, sur le versant négatif, le néocomportementalisme se définit par opposition avec ce qui est considéré comme « comportementalisme traditionnel », associé à un schématisme (trop) étroit du type stimulus-réponse. A l'appui de nouvelles techniques expérimentales, psychologues et biologistes entendent eux aussi engager une rupture décisive par rapport à la théorie du réflexe conditionné des pères fondateurs de la science comportementale (Skinner, Pavlov). Sur le versant positif, le néo-comportementalisme se définit par l'adoption d'une modélisation d'inspiration économique, faisant référence à des processus de maximisation. Cette approche est conçue comme un paradigme alternatif au cadre béhavioriste traditionnel « stimulus-réponse ».

La neurophysiologie apporte néanmoins dans les années 1990 plusieurs changements décisifs par rapport aux études néo-comportementales antérieures. Celles-ci reposaient sur un appareillage expérimental assez simpliste: les psychologues étudiaient le réflexe conditionné chez l'animal en se servant des instruments classiques du béhaviorisme, avec en particulier les fameuses « boîtes de Skinner » (*Skinner boxes*). La neurobiologie fournit quant à elle de nouvelles techniques expérimentales à proprement parler, avec les micro-électrodes qui permettent d'enregistrer l'activité individuelle de neurones d'animaux conscients. Les interprétations qui s'appuyaient jusqu'alors sur la simple observation des comportements vont pouvoir s'effectuer à partir de données neuronales. Or, l'observation de l'activité neuronale précédant le choix permet de comprendre comment le sujet (ici, le singe) forme et attribue une valeur à chacune des alternatives (machines à sous) du jeu. Dans les études sur le pigeon, appuyées sur des données purement comportementales (nombre de choix en faveur de chaque option) les dynamiques de traitement et d'acquisition demeuraient inexpliquées. Richard Herrnstein supposait que les individus s'en remettent le plus souvent à des routines dites de mélioration dans les choix distribués. Or, les techniques neurophysiologiques ont mis en évidence que ces mécanismes qualifiés d'apprentissage de la récompense sont en fait bien plus complexes que cette heuristique postulée par les premiers psychologues néo-comportementalistes.

L'innovation dans les techniques expérimentales a ainsi débouché sur un nouveau

questionnement théorique, portant sur les stratégies d'apprentissage dans les tâches de machines à sous multi-jeux. Plus généralement, la neurophysiologie a offert la possibilité d'une rupture définitive avec le comportementalisme classique, en assurant un ancrage plus nettement biologique et évolutif à ce programme de recherche. Toutefois, la mise en avant d'une dimension biologique du choix apparaissait déjà au sein de la science quantitative de la motivation dans les années 1960. L'influence évolutionniste tendait ainsi (*cf* chapitre 2) à remettre en question la validité normative de la théorie économique de la rationalité, et portait en germes une rupture avec le programme de recherche kahnemanien. La neurophysiologie des années 1990 hérite manifestement de cette identité théorique ambigüe: poursuivant d'un côté un rapprochement avec l'analyse économique, son ancrage évolutionniste implique aussi, d'un autre côté, un écart croissant avec le cadre normatif de celle-ci. Les neurobiologistes comprennent la rationalité comme une intelligence du rapport au milieu. Cette rationalité peut recevoir une certaine explication économique par la théorie des jeux ; mais implique aussi que le choix ne s'effectue pas en fonction de préférences stables, mais plutôt à partir de processus d'apprentissage de la récompense. La convergence supposée entre les théories de l'utilité espérée et neurophysiologie a donc laissé apparaître d'importants hiatus théoriques, suggérant ainsi une possible reprise en main de la théorie économique de la décision elle-même par les économistes. Cela explique pourquoi certains neuroéconomistes ont pu, plus tard, nourrir des prétentions de réforme de l'analyse économique par les neurosciences (voir par exemple Camerer, Loewenstein et Prelec, 2005).

L'utilisation des micro-électrodes en neurophysiologie a été à l'origine, dans les années 1990, d'une importante remise en question le cadre de réflexion traditionnel stimulus-réponse associé à la réflexologie de Sherrington. En enregistrant l'activité des neurones pariétaux chez le signe, les biologistes ont en effet observé l'encodage d'un signal ni sonsoriel, ni moteur, que Paul Glimcher a qualifié d' « *utilité espérée physiologique* » (*cf.* Glimcher, Dorris et Bayer, 2005) (I. Nouveaux instruments, nouveaux modèles : les micro-électrodes au service d'une théorie de l'utilité espérée physiologique). La référence à la notion d'utilité espérée était pourtant trompeuse, car l'activité de ces neurones est en fait stochastique, et ne se comprend non pas comme l'encodage d'une utilité subjective, mais reflète un processus de choix et d'évaluations répétées plus complexes (II. Vers une théorie fréquentielle de la décision : la prédiction de la récompense comme processus stochastique). C'est donc autour de la modélisation d'algorithmes dits d' apprentissages de la récompense que s'effectue une convergence théorique. Ces processus impliquent aussi le circuit dopaminergique (III. De l'utilité espérée à l'apprentissage de la récompense). Il y a donc croisement et enchevêtrement

de deux approches : l'une centrée sur l'évolution et l'autre sur la maximisation. Le problème des comportements inadaptés, comme pour l'impulsivité dans le dernier chapitre, remet en question le partage normatif/descriptif établi par Daniel Kahneman (Tversky et Kahneman, 1986, p.252). La réflexion sur la pathologie permet ainsi de distinguer nettement la neurophysiologie de l'économie comportementale. Les travaux de Paul Glimcher servent ici d'exemple révélateur (IV. Glimcher et l'économie : des rapports ambivalents).

I. Nouveaux instruments, nouveaux modèles : les micro-électrodes au service d'une théorie de l'utilité espérée physiologique

Les avancées de l'électrophysiologie, liées à l'introduction de nouvelles techniques d'observation, provoquent dans les années 1970 de profonds changement en neurobiologie. Les micro-électrodes permettent en effet d'enregistrer l'activité de neurones individuels sur des animaux vivants (et conscients). L'analyse de ce signal, et sa corrélation avec des variables comportementales, ouvre et constitue ainsi un nouveau champ de recherches à part entière. Cette section décrit la manière avec laquelle une formalisation d'inspiration économique, faisant référence à des processus de maximisation, a permis aux neurobiologistes de penser les transformations techniques de leur discipline.

L'introduction des micro-électrodes n'a pas débouché immédiatement sur une innovation théorique. En effet, l'électrophysiologie peine au départ à s'émanciper du béhaviorisme traditionnel. Les activations neuronales sont interprétées comme des signaux encodant soit des commandes motrices, soit les propriétés sensorielles de stimuli externes (A). Il a donc fallu articuler un nouveau cadre de réflexion, inspiré de l'économie, à ces nouveaux instruments, afin de révéler leur importance théorique. Glimcher et Platt apportent de ce point de vue une rupture décisive, lorsqu'ils avancent l'idée d'un signal portant sur l'anticipation de la récompense, ni sensori ni moteur, qu'ils proposent d'appeler « *utilité espérée physiologique* » (Dorris, Bayer et Glimcher, 2005) (B).

A. Les débuts de l'électrophysiologie : l'influence déterminante du modèle stimulus-réponse

Si l'introduction des micro-électrodes a joué incontestablement un rôle important dans l'essor de l'électrophysiologie, celle-ci ne s'est pas pour autant pas développée uniquement à partir de la simple observation de potentiels d'action neuronaux. Comme le souligne avec justesse Glimcher (2010, p.151), les nouvelles techniques de mesure de l'activité cérébrale ont d'abord été utilisées dans la continuité des connaissances antérieures. Il y avait en particulier

une assez bonne compréhension anatomique de la manière avec laquelle le traitement de l'information s'organise dans le cerveau. C'est la raison pour laquelle les électrophysiologistes se sont beaucoup intéressés au cortex pariétal, car cette région du cerveau était alors considérée comme une zone d'association sensori-motrice. L'enregistrement de l'activité de neurones pariétaux visaient ainsi à mettre en évidence, dans le cadre théorique traditionnel du réflexe, le traitement d'informations sensorielles et /ou motrices. Cependant, les données recueillies par les électrophysiologistes ont progressivement mis en évidence l'inadéquation de ce modèle « stimulus-réponse ». Avant d'envisager ces limites théoriques, il convient ici de décrire en quoi consistait, sur le plan expérimental et pratique, cette théorie dite du réflexe.

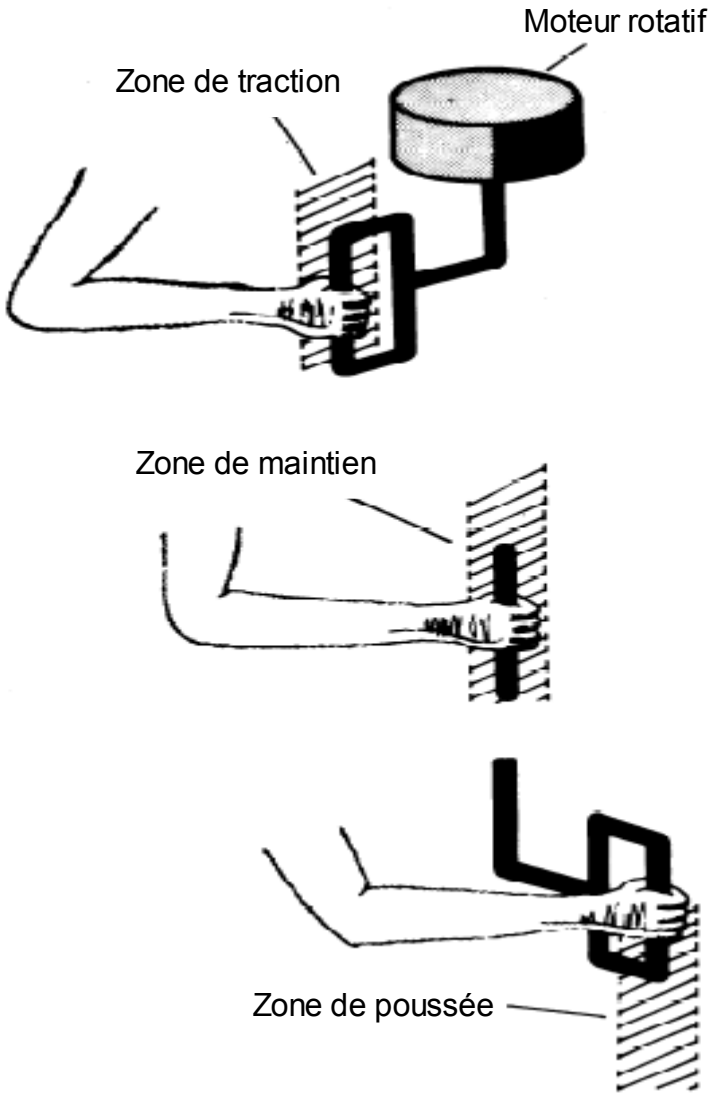
En neurophysiologie, la théorie dite du réflexe est associée principalement aux travaux de Charles Sherrington, au XIXe siècle. Sans rentrer dans le détail des recherches de Sherrington, il est possible de caractériser ce que Glimcher appelle réflexologie ou « *paradigme du réflexe* » à partir de l'idée simple selon laquelle la fonction du système nerveux consiste à connecter à un stimulus sensoriel une réponse musculaire (et non pas une intention d'action, comme l'ont supposé plus tard les neuroéconomistes). A chaque stimulus simple correspond une réponse approprié : un réflexe est ainsi une brique élémentaire dans le répertoire de comportements de l'organisme. La théorie du réflexe implique donc qu'un comportement complexe puisse être décomposé en comportements plus simples. Dans cette perspective, la neurophysiologie se donne pour tâche de décrire la manière avec laquelle le système nerveux décompose logiquement des signaux sensoriels et encode des commandes motrices[79].

Dans les années 1960, au moment où sont réalisés les premiers enregistrements d'activité neuronales par micro-électrodes sur des animaux conscients (Glimcher, 2003, p.95), la théorie du réflexe constitue le cadre théorique de référence[80]. Au début, les électrophysiologistes interprètent ainsi leurs données comme des signaux encodant les propriétés sensori-motrices de mouvements réflexes. L'étude de référence en la matière, réalisée par Tanji et Evarts en 1976, fournit une bonne illustration. Dans cette expérience, des singes apprennent par un conditionnement simple à réaliser deux mouvements opposés, qui

79 Une illustration de cette approche est le modèle dit du « neurone formel »de McCullochs et Pitt (1943). En s'appuyant sur la règle de Hebb d'apprentissage des réseaux de neurones formels, Franck Rosenblatt a quant à lui, avec son « perceptron », proposé une simulation informatique du réflexe classique (Rosenblatt, 1958). Pour une étude détaillée des modèles neuronaux inspirés par la théorie du réflexe, on se référera à Glimcher, 2003.

80 Encore aujourd'hui, selon Glimcher, en dépits des critiques dont elle a fait l'objet au cours des dernières années, la théorie du réflexe demeure encore aujourd'hui le « *cadre de référence des neurosciences* ». Un manuel communément utilisé (Kandel, Schwartz et Jessell, 1991) y accorde par exemple toujours une place importante (Glimcher, 2003, p.78).

consistent soit à tirer (*pull*) ou pousser (*push*) un levier (voir schéma ci-dessous). Au début de la tâche, une lumière informe le singe que l'expérience débute et qu'il peut donc obtenir une récompense en accomplissant le mouvement demandé. Le singe doit ensuite soit tirer soit pousser le levier en fonction de la couleur du second signal lumineux.



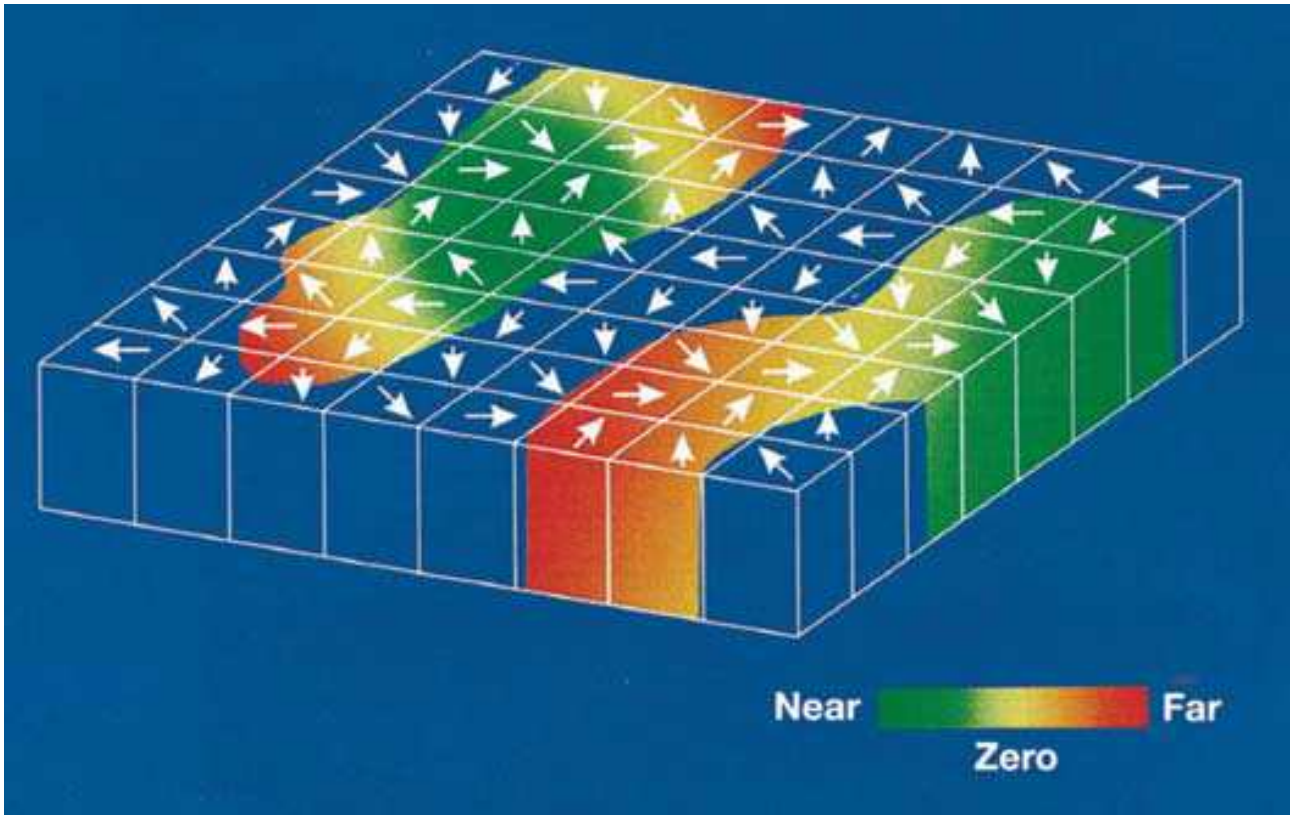
Tanji et Evarts, 1976, p.1063

Tanji et Evarts ont d'abord mis en évidence que l'activité de certains neurones pariétaux était responsable du déclenchement d'un des deux mouvements (tirer ou pousser). Pour ces neurones, les auteurs ont observé une activité faible pendant la période préparatoire, entre les deux signaux lumineux, puis une augmentation rapide après le second signal conduisant à une réponse motrice. Cette expérience semble donc confirmer le mécanisme traditionnel du réflexe. L'activité neuronale encode ici directement les propriétés visuelles

d'un stimulus qui débouche sur une réaction motrice automatique.

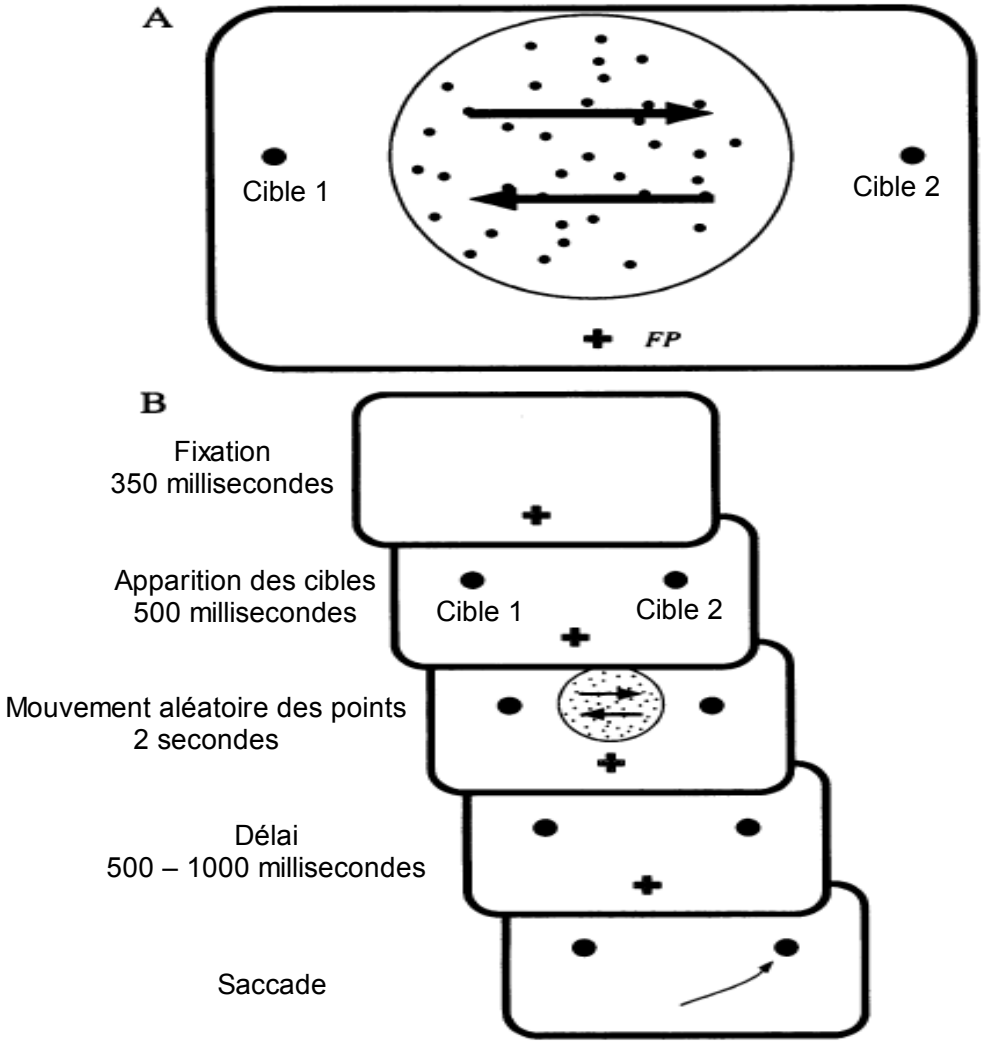
C'est surtout avec les travaux de Paul Newsome et de son équipe que le paradigme du réflexe en neuro-électrophysiologie trouve son apogée. Cet auteur a travaillé avec une méthode très proche de celle des neuroéconomistes, car, comme Glimcher par exemple, ses recherches visent à établir des corrélations entre les mesures de l'activité neuronales et des variables comportementales, par une technique qu'il qualifie de « *neurométrie-psychométrie* » (Parker et Newsome, 1998, p.249). Néanmoins, Newsome se distingue de la neuroéconomie en tant qu'il ne défend pas de rapprochement avec l'économie (Parker et Newsome, 1998, p.277), et qu'il reste fidèle au cadre de réflexion traditionnel stimulus-réponse.

Newsome et ses coauteurs ont approfondi de manière importante la portée de l'étude initiale de Tanji et Evarts, en proposant, à la suite de travaux antérieurs par Zerki (1974), l'hypothèse d'une organisation topographique du cortex pariétal pour l'encodage du mouvement : les neurones pariétaux (de l'aire moyenne temporale plus précisément) sont organisés par « *groupes* » (*patch*), qui chacun encodent spécifiquement une direction de mouvement. Pour les jugements perceptuels impliquant la détection d'un mouvement, le cerveau s'en remet à des réseaux spécialisés selon les directions possibles :



depuis Born et Bradley, 2005, p.163

Cette hypothèse a été confirmée par Newsome, Britten et Movshom, dans une étude réalisée en 1989 (Newsome, Britten et Movshom, 1989). Dans cette expérience, des singes regardent un signal lumineux. Celui-ci représente un cercle, à l'intérieur duquel plusieurs points lumineux bougent dans différentes directions, selon un degré de cohérence plus ou moins grand. La cohérence est maximale si l'ensemble des points se dirigent dans la même direction (à droite ou à gauche) ; elle est minimale lorsque le mouvement de l'ensemble est complètement désordonné. Le sujet dispose de deux secondes pour regarder le signal et doit ensuite effectuer une saccade oculaire en direction d'une des deux cibles (droite ou gauche) pour signifier le mouvement perçu de l'ensemble des points. Si le singe choisit la bonne cible, il obtient une récompense.



Shadlen et Newsome, 1996, p.629

Les données recueillies par les micro-électrodes ont permis aux auteurs d'établir une solide corrélation entre le degré de cohérence du stimulus, la probabilité des choix, et l'activité des neurones impliqués dans la détection de chaque type de mouvement. Tout semble donc se passer comme si les propriétés visuelles d'un signal, lié au mouvement, sont encodées par des groupes de neurones spécifiques. Plus le degré de cohérence du signal se renforce, plus l'activation du groupe de neurones concernés augmente, jusqu'à un seuil, mesurable, qui déclenche la saccade oculaire. Ce résultat expérimental robuste a été reproduit ultérieurement par Shadlen et Newsome (1996), qui ont montré que les neurones du cortex latéral intrapariétal, sur lesquels sont projetés les neurones moyens-temporals, effectuent une sommation des activations précédentes.

L'utilisation des micro-électrodes par Newsome ou par Tanji et Evarts favorise donc les interprétations classiques en termes de réflexes. L'activité des neurones du cortex sensori-moteur semble pouvoir être corrélée avec une grande précision avec des variables perceptuelles. Toutefois, les expériences de Newsome ne portent pas, à proprement parler, sur un mouvement-réflexe. En effet, la tâche expérimentale exige un traitement continu de l'information par le singe, qui doit progressivement réévaluer sa perception du degré de cohérence du mouvement, avant d'effectuer la saccade. L'activité observée pendant les deux secondes de présentation du signal ne représente donc qu'une activité préparatoire, antérieure à la commande motrice elle-même. On retrouve le même difficulté dans l'expérience de Tanji et Evarts, qui observent également une activité assez faible, mais non-négligeable, au début de la tâche, avant le déclenchement du mouvement. En interprétant ce signal comme une anticipation de la récompense, Platt et Glimcher ont essayé de s’affranchir du cadre théorique stimulus-réponse.

B. Platt et Glimcher, 1999 : De la réflexologie à la théorie de l'utilité espérée

L'étude neurophysiologique des réflexes trouve un débouché direct dans les expériences réalisées par Newsome et son équipe dans les années 1980. Pour autant, la notion de réflexe demeure assez controversée. Il soulève en pratique de nombreuses difficultés, bien connues des spécialistes du système nerveux, et dont certaines sont antérieures à l'apparition

des micro-électrodes. A l'appui de ces limités théoriques, Platt et Glimcher proposent d'interpréter ces données problématiques pour la théorie du réflexe comme des signaux d'anticipation de la récompense. Leur hypothèse ouvre la voie à un rapprochement avec l'économie, en suggérant l'idée d'une « *utilité espérée physiologique* » (Glimcher, Dorris et Bayer, 2005)

La théorie du réflexe, telle qu'elle est définie classiquement chez Sherrington, est certes dominante en neurophysiologie au moment de l'apparition des micro-électrodes, mais elle a toujours été assez controversée. Le problème ne porte pas vraiment sur la question de savoir si tous les comportements d'un organisme peuvent être considérés comme entièrement déterminés en tant que réflexes. Comme le souligne Glimcher, on peut tout à fait défendre, à la manière du dualisme cartésien, sur le plan neuronal, un schématisme stimulus-réponse, tout en maintenant au niveau de la conscience un domaine du libre-arbitre, entièrement indépendant du déterminisme biologique (Glimcher, 2003). En fait, en biologie, la théorie du réflexe se révèle trop réductrice même pour les mouvements les plus élémentaires, supposés être les illustrations directes d'une réaction de type stimulus-réponse. Au début du siècle, Graham Brown et ultérieurement Erik von Holtz, en s'appuyant sur le caractère cyclique de certains comportements observés chez les organisme vivants (dormir, se nourrir, etc.),, montrent que le système nerveux n'est pas qu'un simple véhicule de transmission d'information, mais peut aussi générer sa propre activité (Glimcher, 2003, p.87).

Le rythme biologique, reproduit à l'échelle neuronale remet ainsi en cause la pertinence de la théorie de Sherrington, car, selon cette dernière, le système nerveux a pour fonction unique de conduire l'information, d'un stimulus externe vers une commande motrice. Or, dans cette perspective, il est impossible d'expliquer toute forme d'activité nerveuse interne et spontanée, car le système nerveux ne saurait fonctionner de manière autonome.

Le principe dit de « réafférence » va encore plus directement à l'encontre de la théorie du réflexe. Ce principe postule qu'à chaque commande motrice, un signal moteur de *feedback* est émis pour informer le système nerveux de l'accomplissement du mouvement requis. Le signal de réafférence permet ainsi de comprendre comment, dans le cas de la vision par exemple, le cerveau parvient à distinguer les mouvements du champ visuel qui sont générés par un mouvement de l'œil, de ceux qui sont produits par un déplacement des objets dans le champ : lorsqu'une saccade oculaire est effectuée, le mouvement de l'œil déclenche un signal de réafférence qui informe le système visuel d'un déplacement de ses organes de perception.

Sherrington admet ce principe de réafférence, mais considère que le signal réafférent se propage selon un arc moteur réflexe : si je déplace mon regard, le mouvement de l'œil

déclenche en retour un réflexe qui donne lieu à un nouveau signal. Or, dès 1932, Kornmuller montre que ces signaux réafférents doivent être compris comme « *décharges corollaires* », c'est à dire qu'ils interviennent au moment même de la commande motrice : le signal efférent (moteur) est en fait d'emblée un signal afférent (sensoriel), ce qui signifie que le système nerveux peut fonctionner par boucles internes qui n'ont nullement besoin d'un stimulus externe pour être déclenchées (Glimcher, 2003, p.90).

La théorie du réflexe fait donc déjà l'objet d'importantes critiques, bien avant l'introduction des micro-électrodes. En 1941, Paul Weiss propose de considérer, dans une perspective assez proche de celle de Glimcher, que le système nerveux n'a pas pour fonction de relier un stimulus à une réponse musculaire mais plutôt d'associer un événement extérieur à une *intention* motrice. En refusant de considérer la possibilité d'une telle intention de mouvement préalable à l'agir, la théorie du réflexe soulève une série de difficultés qui sont liées, on l'a vu, à toute forme d'activité autonome du système nerveux par bouclage interne. Si les études de Newsome appliquent ainsi le paradigme du réflexe aux nouveaux instruments de mesure fournis par l'électrophysiologiques, ces expériences sont donc néanmoins d'emblée sujettes à discussion. A propos des études envisagées précédemment (Tanji et Evarts, 1974 ; Newsome, Britten et Movshom, 1989 ; Shadlen et Newsome, 1996), les controverses portent en particulier sur l'interprétation du signal préparatoire à l'action observé dans les études. Les auteurs observent une activité initialement faible, au début de la tâche, avant l'effectuation du mouvement. Chez Newsome, Britten et Movshom (1989), ce signal à l'intensité croissante est interprété comme encodant un degré de cohérence croissant du stimulus, conduisant, à partir d'un certain seuil, au mouvement. Or, ce signal préparatoire s'observe également lorsqu'aucun mouvement n'est effectué : dans l'expérience de Tanji et Evarts (1974), l'activité observée avant le signal indiquant la direction du mouvement oculaire à effectuer (*go cue*) ne saurait encoder un mouvement possible, pour la simple raison que le singe ne peut savoir, avant ce signal, de quel côté la saccade oculaire devra être réalisée. A l'analyse, il s'avère donc que ces signaux encodent quelque chose qui ressemble plus à une intention ou une anticipation de mouvement qu'à un mouvement à proprement parler.

Dans les années 1970 et 1980, une controverse se développe alors, autour de la question de savoir si les activités neuronales préalables à la réponse motrice encodent un signal d'attention ou d'intention (*cf.* Glimcher, 2003, p.247-250). En 1975, Mountcastle et ses coauteurs critiquent la théorie du réflexe en montrant que ces processus qualifiés de « commande » (*command process*) reflètent des déficits de volition et ne sont par conséquent ni sensoriels ni moteurs (Mountcastle *et al.*, 1975). Robinson, Goldberg et Stanton (1978)

plaident quant à eux en faveur du cadre de la réflexologie, en proposant de considérer ces activations comme des signaux sensoriels, modifiés par des facteurs attentionnels (*attentional factors*). Au delà du choix de vocabulaire, le débat attention-intention porte ainsi la possibilité (ou non) de comprendre l'activité des neurones pariétaux par leurs propriétés sensorielles et/ou motrices. Cette controverse perdure jusque dans les années 1990 (*cf.* Colby, Duhamel et Goldberg, 1996). L'étude réalisée par Gnadt et Andersen (1988) semble néanmoins apporter un solide démenti à l'interprétation de Goldberg, puisque cette expérience montre que le signal observé par Goldberg persiste en l'absence de stimulus sensoriel.

En faisant état des limites rencontrées par la théorie du réflexe, Michael Platt et Paul Glimcher avancent l'idée originale, à la fin des années 1990, que le signal préparatoire au mouvement reflète un signal d'« *utilité espérée physiologique* », c'est à dire un signal d'anticipation de la récompense (Platt et Glimcher, 1999). Cette proposition vise à mettre un terme au débat « attention-intention ». Elle a été inspirée par d'autres études relativement proches, notamment celle réalisée par Basso et Wurtz deux ans plus tôt (1997). Cette expérience reproduit le protocole de Newsome, dans lequel le singe doit fixer un nuage de points qui bougent, puis doit effectuer une saccade oculaire dans l'une des deux directions possibles direction (droite ou gauche) en fonction du mouvement perçu. En faisant varier le nombre de points-cibles (1, 2, 4, ou 8), Basso et Wurtz trouvent que l'activité pendant la période de fixation (avant le signal initiant la saccade (*go cue*)) est fonction inverse du nombre de cibles. Ce n'est pas tant ici la nouveauté du résultat qui importe mais l'originalité de son interprétation. Newsome, Britten et Movshom (1989) observent que l'intensité du même signal croît à mesure que le degré de cohérence du stimulus augmente. Cela semble logique puisqu'un plus grand (ou plus petit) nombre de points dans le nuage diminue (ou augmente) le signal, en affaiblissant (ou renforçant) le degré de cohérence du stimulus. Mais Basso et Wurtz proposent d'interpréter ce résultat non pas en termes de cohérence du stimulus, mais en terme d'incertitude d'une récompense probable : selon les auteurs, l'activité observée de ces neurones pariétaux change selon l'incertitude de l'environnement perceptuel (Basso et Wurtz, 1997).

A l'appui de cette notion d'incertitude, Platt et Glimcher ont avancé l'idée novatrice d'un signal d'anticipation de la récompense, incorporant à la fois sa probabilité et son amplitude (Platt et Glimcher, 1999). Leur expérience portait sur une tâche oculomotrice (*cue saccade task*) relativement simple. Le singe, dont l'activité du cortex pariétal est enregistrée au moyen de micro-électrodes, doit initialement fixer un point lumineux, situé entre deux cibles, l'une, à gauche, étant rouge, l'autre verte. Au bout d'un délai fixe, le signal lumineux

central devient soit rouge, soit vert. Le singe doit en conséquence tourner son regard vers la cible correspondante, à droite ou à gauche, afin d'obtenir une récompense (jus de fruit). L'étude est divisée en trois blocs au cours desquels cette même tâche est répétée cent fois. Dans les deux premiers blocs, la cible qui sera récompensée est indiquée au singe: après le délai d'attente, le signal central devient bleu ou vert, avec une fréquence constante, si bien qu'il est possible, au fur et à mesure de la répétition, d'estimer la probabilité de récompense associée à chacune des deux cibles avant que le signal central ne devienne rouge ou vert. Dans le premier bloc, le montant de récompense (volume de jus de fruit) est maintenu constant. Dans le deuxième bloc, la cible faisant l'objet d'une récompense est toujours la même (le signal lumineux devient toujours bleu ou vert) mais le montant de récompense varie à chaque essai.

Cette tâche peut se comprendre une version simplifiée des *multi-armed bandit problems*, dans laquelle les expérimentateurs indiquent, avant chaque choix, la machine à sous « gagnante ». Ici, les machines à sous correspondent à des mouvements oculaires. L'individu ne peut se tromper sur le choix de la bonne machine à sous, puisque celle-ci est désignée par le signal lumineux central. L'intérêt de cette expérience, sur le plan comportemental, est donc nul, puisque le choix est forcé. Ce qui intéresse les expérimentateurs est la manière avec laquelle l'activité neuronale, au cours de la période d'attente, et donc préalablement à la décision, anticipe sur l'indication de la saccade à effectuer. Platt et Glimcher proposent l'hypothèse selon laquelle il doit exister une activité neuronale qui serait susceptible d'encoder la probabilité de récompense associée à chaque cible dans le cadre du premier bloc (le montant étant fixe), et le montant de récompense associée à chaque cible dans le cadre du second bloc (la cible récompensée étant connue dès le départ).

Platt et Glimcher ont d'abord identifié des neurones du cortex latéral intra-priétal (LIP) associé à chaque mouvement oculaire (droite ou gauche). Ils ont observé que les activations de ces neurones étaient corrélées effectivement avec la probabilité de la récompense associée à chaque cible dans le premier bloc d'essais, et avec l'amplitude de cette même récompense pour le second bloc. Dans le troisième bloc, qualifié de « choix libre » (*free choice task*), le signal lumineux central ne devient ni rouge ni vert, et n'indique pas par conséquent la direction du mouvement oculaire à effectuer. La tâche correspond alors strictement à une machine à sous multi-jeux: le singe ignore la machine à sous gagnante et doit progressivement identifier la distribution des récompenses associées à chaque mouvement oculaire (machine à sous). Platt et Glimcher observent alors que l'activation tend à égaliser les utilités relatives associées à chaque option, d'une manière similaire à la loi d'égalisation des rendements de

Herrnstein (1967).

L'expérience de Platt et Glimcher montre donc que dans des *multi armed bandit problems*, l'activité des neurones pariétaux n'est pas à comprendre par ses propriétés sensorielles ou motrices, mais comme un signal permettant de prédire la probabilité et le montant d'une récompense anticipée. Cette étude suggère donc, en neurophysiologie, ce que Glimcher appelle une « *forme dure de théorie de l'utilité espérée* » (*hard form of expected utility theory*) :les activations neuronales sont le reflet direct de la théorie économique du choix incertain, une terminologie que Glimcher lui-même a rejetée par la suite (Glimcher, 2010, p.112). Toujours est-il qu'à la fin des années 1990, le signal dit de prédiction de la récompense est compris comme l'équivalent neuronal de la notion économique d'utilité espérée. Cette interprétation perdure dans la littérature, puisque Camerer par exemple, l'un des neuroéconomistes les plus en vue, analyse la neurophysiologie « *à la Glimcher* » comme une confirmation de la théorie économique de l'utilité espérée[81].

En 2005, Glimcher propose ainsi l'expression d'utilité espérée physiologique pour synthétiser l'apport de ses travaux (Glimcher, Dorris et Bayer, 2005). Cette formulation est en fait assez trompeuse, car, à proprement parler, l'activité neuronale observée dans l'expérience de Platt et Glimcher (1999) n'encode pas exactement une utilité espérée au sens économique du terme, c'est à dire la prédiction d'une satisfaction future probable. Il s'agit en effet plutôt d'un signal permettant de guider les choix, par une évaluation *a priori* de l'attractivité perçue de chaque option. Mais cette attractivité ne reflète pas nécessairement une satisfaction ou un plaisir effectif, car elle se produit de manière continue, y compris lorsqu'aucun choix n'est ultérieurement exprimé. Comme le souligne Glimcher lui-même, « *beaucoup d'économistes seront gênés par cette idée selon laquelle des signaux de prédiction de l'utilité espérée peuvent précéder l'action. Cela est gênant pour les économistes parce que cela suggère que l'attribution de valeurs précède les choix et que par conséquent ces signaux d'évaluation peuvent apparaître y compris lorsqu'aucun choix n'est exprimé ultérieurement -une idée que Paretto rejetait comme non-testable* » (Glimcher, 2003, p.173).

En outre, l'équivalence entre maximisation de l'utilité espérée et égalisation (*matching*) des rendements relatifs de chaque option dans les machines à sous multi-jeux pose problème, parce que, dans l'expérience de Platt et Glimcher (dans le troisième bloc, plus précisément), la stratégie de *matching* n'est pas une solution optimale. L'observation selon laquelle les choix et

81 « *l'ironie de la neuroéconomie est que les neuroscientifiques trouvent souvent les principes les plus élémentaires de la rationalité [économique] utiles pour expliquer le choix des êtres humains. Glimcher (2003) montre avec élégance comment le modèle simple de l'utilité espérée clarifie le type d'encodage réalisé par les neurones pariétaux* » (Camerer, 2005, p.449).

les activations neuronales tendent, au fur et à mesure des essais, à égaliser les rendements obtenus dans le passé vaut comme confirmation de la loi d'égalisation des rendements (*matching law*) de Richard Herrnstein (*cf.* chapitre 2). Toutefois, ce comportement n'est pas optimal au sens économique, puisque la maximisation de la récompense totale suppose en fait d'identifier et de choisir toujours l'option offrant en moyenne la plus grand récompense. Il y a donc ici, comme l'a reconnu ultérieurement Glimcher (2003, p.264-265) une confusion entre l'égalisation des rendements observée dans les *multi armed bandit problems* et la maximisation économique de la valeur qui est référée à des choix discrets. En mettant en évidence le caractère stochastique de l' « utilité espérée physiologique », les études ultérieures ont permis de comprendre ce signal neuronal comme un processus aléatoire conçu pour répondre à des problèmes d'apprentissage de la récompense

II. Vers une théorie fréquentielle de la décision : la prédiction de la récompense comme processus stochastique

La théorie de l'utilité espérée physiologique proposée par Glimcher (2005) réduit l'utilité, comprise comme valeur accordée par le sujet aux options de décision, à des variables neurobiologiques. Dans son ouvrage récent, Glimcher écrit ainsi « *les valeurs subjectives […] sont des objets neurobiologiques possédant une valeur cardinale allant de à 1500* », ce qui correspond en effet aux fréquences d'activation neuronales possibles (Glimcher, 2010, p.136). Cependant, Glimcher précise aussitôt que ces objets neurobiologiques se distinguent de l'utilité espérée cardinale des économistes car ces valeurs ont une variance importante. Le caractère aléatoire du signal de prédiction de la récompense témoigne donc d'une limite à l'analogie entre neurobiologie et théorie de l'utilité espérée: la maximisation ne porte en effet pas sur des choix discrets mais sur des séquences de choix (A). Il est possible néanmoins de maintenir une référence à l'économie, c'est-à-dire à un processus de maximisation, en présentant la stochasticité de ces processus neuronaux comme la stratégie optimale d'un jeu évolutionnaire (B).

A. Les aléas de la prédiction : une limite à la maximisation du choix discret

Glimcher, dans l'interprétation initiale son expérience fondatrice de 1999 réalisée avec Michael Platt, considère que le comportement d'égalisation des rendements (*matching behavior*) correspond à une forme de maximisation de la valeur (Platt et Glimcher, 1999). Or, cette observation s'appuie sur une erreur : comme le reconnaît ultérieurement Glimcher, si, dans certaines expériences réalisées sur le pigeon (*cf.* chapitre 2), « *l'égalisation des rendements est une stratégie optimale. Dans notre expérience* [Platt et Glimcher, 1999]*, cela n'était pas le cas. La stratégie optimale pour les singes consistait à identifier l'option générant la récompense la plus élevée et à continuer de choisir de cette option jusqu'à la fin du bloc de 100 essais* » (Glimcher, 2003, p.264). Les choix et les activations neuronales observés mettent donc en évidence un comportement d'égalisation des rendements (*matching*

behavior), qui ne peut donc pas être compris directement comme un comportement optimal dans cette tâche.

Pour éviter toute confusion, il convient de bien souligner que le signal de prédiction de la récompense sert à répondre à des problèmes type *multi armed bandit problems*. Par conséquent, cette activité neuronale porte non pas sur des décisions isolées mais sur des allocations de fréquences (de choix) entre deux options. Une lecture trop rapide de ces expériences peut ainsi facilement faire perdre de vue le cadre séquentiel de la prise de décision, qui n'est pas forcément évident pour un économiste[82].

Les réseaux de neurones impliqués dans la prédiction des récompenses sont donc destinés non pas seulement à estimer une satisfaction, mais à élaborer des stratégies d'apprentissage. Cela apparaît clairement à travers la nature aléatoire de ce signal neuronal. Dans l'expérience de Platt et Glimcher, les auteurs observaient que le signal de prédiction de la récompense était bien corrélé, *à chaque essai*, avec la probabilité et/ou l'amplitude de la récompense. Or, comme le souligne avec justesse Glimcher (2003, p.209), cette observation est en fait trompeuse : lorsqu'une décharge neuronale de 100Hz est enregistrée (et que cette valeur est prise comme une estimation de l'utilité espérée cardinale), cette valeur n'est en fait qu'une moyenne d'une variable aléatoire possédant une très forte variance. L'activité des neurones pariétaux suit en effet une distribution de Poisson, avec un coefficient de variance égal à un, ce qui signifie que la moyenne de cette variable aléatoire est égal à sa variance, et donc que celle-ci est très importante. En d'autres termes, le signal de prédiction de la récompense permet certes de prévoir l'amplitude et la probabilité d'une récompense incertaine ; mais ce signal est stochastique, et il peut donc prendre des valeurs différentes lors de la répétition d'un choix maintenu parfaitement identique (Glimcher, 2003, p.209). La stochasticité de ces processus neuronaux peut ainsi se comprendre comme une sorte de production biologique du comportement aléatoire (*cf.* Bayer, Lau et Glimcher, 2007).

L'« *utilité espérée physiologique* » n'est donc pas l'équivalent exact de l'utilité espérée « cardinale » des économistes. Cette conclusion pouvait aussi être déduite des expériences de Newsome (Newsome, Britten et Movshom, 1989 ; Shadlen et Newsome, 1996). Dans ces études, une activation continue et croissante des neurones associés à la détection d'un type de mouvement déterminé était observée, lorsque le degré de cohérence du stimulus augmentait, c'est à dire lorsque, implicitement, la probabilité de la récompense augmentait. Mais le choix final ne s'exerce pas sur le groupe de neurones ayant l'activation la plus forte, ce qui

82 Sur la spécificité de ce cadre séquentiel en théorie de la décision, voir l'introduction à la première partie.

équivaudrait à un choix en faveur de l'option offrant la plus grande probabilité de gain. Il y a plutôt un seuil d'activation neuronal à partir duquel le mouvement (saccade oculaire) est déclenché. Une fois ce seuil atteint, une seule option est retenue et toutes les autres sont rejetées. Ce principe est connu sous le nom du « vainqueur emporte tout » (*winner-take-all-principle*). Celui-ci n'est pas isomorphe à une fonction de maximisation, car une activation peut atteindre ce seuil sans être nécessairement la plus forte possible. Plutôt que de parler de maximisation de la valeur, il conviendrait en fait plutôt de considérer d'un niveau de récompense attendu, jugé satisfaisant par l'organisme.

En reprenant le protocole des expériences de Newsome, Roitman et Shadlen montrent que le seuil retenu dépend de la complexité de la décision. Lorsque le mouvement des points est rapide (et donc facile à percevoir), la décision s'effectue vite en faveur de la meilleure option. En revanche, lorsque le mouvement est lent, la décision perceptuelle est plus longue et perd en précision. Roitman et Shadlen suggèrent que le seuil à partir duquel la saccade oculaire est effectuée permet de réaliser un arbitrage entre rapidité et précision de la perception: s'il est faible, le choix est largement simplifié, au risque éventuellement de ne pas choisir la meilleure option ; s'il est élevé, la probabilité de l'erreur diminue, mais le choix est plus complexe et plus long (Roitman et Shadlen, 2002). Glimcher, d'une manière très caractéristique de sa démarche théorique, propose alors une stratégie de repli sur une autre théorie économique. Selon lui, cet effet de seuil peut se comprendre, dans le cadre de la théorie de la rationalité limitée d'Herbert Simon (1957), comme l'analogue d'un « *prix de réservation* » (Glimcher, 2003, p.203), c'est à dire comme un seuil d' « acceptabilité » à partir duquel les options jugées satisfaisantes sont retenues.

La référence à la théorie de Simon n'est cependant mobilisée par Glimcher qu'à titre illustratif. Il n'y a en effet nullement besoin de postuler une limitation de la rationalité pour interpréter ces résultats. Au contraire, comme l'avancent Roitman et Shadlen (2002), le seuil en question peut se comprendre comme l'optimisation d'un choix entre précision et rapidité de la perception. La prise en compte d'un coût de traitement de l'information permet ainsi de conserver une référence à la maximisation sous contrainte. La vocation économique de la neurophysiologie dans les années 1990 repose en fait plutôt sur un hiatus entre la maximisation utilité espérée au sens statique de l'économiste d'une part, et l'égalisation des rendements relatifs dans un choix répété entre des options générant des récompenses variables (en probabilité et en amplitude) d'autre part. Ces deux processus sont tout à fait distincts : le premier implique un calcul sur des valeurs discrètes, alors que le second repose sur une

allocation des fréquences de choix, et engage donc un raisonnement sur des variables aléatoires (*cf.* l'introduction à la première partie).

Tout l'enjeu consiste alors à maintenir une référence à l'économie, c'est-à-dire à la maximisation. Les psychologues néo-comportementalistes, qui, dans les années 1970, étudiaient le choix répété chez le pigeon, avaient introduit un rapport inverse entre délai d'obtention de la récompense et utilité, ce qui permettait de comprendre l'incohérence des choix, comme la conséquence de la maximisation d'une fonction d'actualisation hyperbolique (*cf.* chapitre 2). Les neurobiologistes avancent quant à eux l'idée selon laquelle les choix qui peuvent apparaître comme des erreurs d'un point de vue statique sont néanmoins justifiés, à long terme, pour les besoins de l'apprentissage. La recherche d'information est une variable stratégique que doit intégrer la fonction de maximisation. La mélioration, c'est à dire la tendance à l'égalisation des rendements obtenus, est ainsi comprise comme une forme d'optimisation dans le cadre d'une théorie des jeux dite « évolutionnaire », dans laquelle la stratégie optimale est une stratégie mixte, qui conduit à des comportements variés.

B. De la maximisation du choix discret à la théorie des jeux évolutionnaires: la stochasticité comme réponse optimale

Les neurophysiologistes envisagent le signal de prédiction de la récompense comme une variable quantitative aléatoire, dont la moyenne des réalisations dans un choix répété entre deux options tend à s'approcher de l'espérance de rendement. Par conséquent, la répétition d'un choix à l'identique entre deux options ne conduit pas nécessairement à chaque fois aux mêmes estimations. Le caractère stochastique de la prise de décision apporte ainsi un démenti à ce que Paul Glimcher une forme « dure » de la théorie de l'utilité espérée (*hard form of expected utility theory*), selon laquelle la maximisation de l'utilité d'un choix isolé pourrait être réduite directement au niveau neuronal (Glimcher, 2010). Ces écarts entre les prédictions successivespourraient être interprétés comme des erreurs d'évaluation dans le cadre étroit de la théorie économique du choix discret. Ils renvoient en fait à des valeurs différentes prises par une même variable aléatoire. La distribution de cette dernière peut alors (ou non) elle-même être adaptée au problème décisionnel envisagé, en fonction de deux exigences. La variabilité des décisions peut être comprise comme une stratégie optimale dans

la perspective d'une théorie des jeux évolutionnaires soit parce qu'elle est requise pour les besoins de l'apprentissage (1), soit parce que la meilleure stratégie est mixte, c'est à dire que la maximisation du gain nécessite d'adopter des comportements changeants et qu'aucune stratégie dite « pure » n'est gagnante (2). L'activité neuronale peut donc être dite stochastique en deux sens différents.

1. La variabilité des décisions comme moyen d'apprentissage: un premier sens de la stochasticité

La nature stochastique du signal d'utilité espérée physiologique s'explique d'abord par les besoins de l'apprentissage. En effet, la prédiction de la récompense vise toujours à anticiper sur un mouvement ou un choix à réaliser. Dans les expériences étudiées précédemment, les singes n'ont initialement aucune idée précise des récompenses associées à chaque option de choix, et des variables en fonction desquelles ces récompenses changent. Les singes ont été entraînés à effectuer une tâche précise -effectuer une saccade oculaire- pour laquelle ils ont obtenu une récompense (jus de fruit, eau). Au début de l'expérience, ils peuvent donc s'attendre à recevoir une gratification, mais sans connaître de manière certaine sa probabilité et son ampleur. Dans le troisième bloc d'essais dit de « *choix libre* » (*free choice task*) de l'étude de Platt et Glimcher (1999) par exemple, le singe doit effectuer par exemple une saccade oculaire vers la cible de droite ou de gauche sans savoir, au moment du choix, la probabilité de récompense associée à chaque cible. A partir d'un certain nombre de répétitions, l'observation des récompenses obtenues dans le passé permet d'établir une estimation relativement précise de ces probabilités.

Le caractère aléatoire du signal de prédiction de la récompense peut ainsi d'abord être requis, dans ces expériences, pour les besoins de l'apprentissage, pour acquérir une connaissance fiable des probabilités. L'absence de connaissance fiable sur les rendements espérés empêche de sélectionner l'option possédant le rendement espéré le plus élevé. Cette contrainte informationnelle oblige à allouer les fréquences relatives de choix en fonction des fréquences de récompense relatives observées dans le passé (comportement dit de *matching*). En d'autres termes, la correction progressive, par essais et erreurs, des prédictions peut être vue comme une forme de maximisation dès lors que ce processus d'amélioration des connaissances constitue une stratégie d'apprentissage optimale. En prenant en considération les problèmes liés à la recherche d'une information initialement inexistante, le tâtonnement et l'exploration deviennent en effet nécessaires. La formalisation de ces processus

d'apprentissage fera l'objet d'un approfondissement dans la section suivante.

Dans des tâches type machines à sous multi-jeux, la variabilité du signal de prédiction peut d'abord se justifier par l'impératif d'amélioration des connaissances :sachant que la dernière fois que j'ai choisi l'option *1*, j'ai obtenu un montant *X* de récompense alors que je m'attendais à *Y*, il est normal que je modifie, lors du choix suivant, ma prédiction en estimant cette fois-ci la récompense associée à l'option *1* entre *X* et *Y*[83]. L'absence de connaissance, au départ, quant à la distribution exacte des récompenses associées à chaque machine à sous appelle une modification continue de mes croyances. La variabilité du signal de prédiction de la récompense peut rendre compte de ce processus d'amélioration des connaissances relatives aux probabilités.

2. La variabilité des décisions comme stratégie mixte (optimale): un second sens de la stochasticité

Les processus d'optimisation que font apparaître ces expériences peuvent, en outre, être dits stochastiques dans un second sens, distinct de celui qui est impliqué par les besoins de l'apprentissage. Les choix effectués sont stochastiques en effet au sens où la maximisation de la valeur nécessite de varier les décisions : il peut en effet y avoir un intérêt stratégique dans le simple fait de ne pas jouer toujours la même stratégie. Une idée similaire apparait chez Herrnstein, pour qui les choix réels des individus qui sont guidés par des préférences exprimées sous forme de fréquences. Par conséquent, les préférences absolues n'ont en général aucune signification pratique, car les individus accordent en général une valeur supérieur à des choix variés. Un individu qui préfère *A* à *B* peut néanmoins, en pratique, choisir *A* dans 80% des cas et *B* dans 20% des cas. Si *A* et *B* désignent par exemple respectivement « regarder la télévision » et « lire un livre », il peut y avoir une lassitude ou une monotonie à toujours choisir *A* (*cf.* chapitre 2 ; Herrnstein, 1991).

Dans un ordre de considérations relativement proches, dans certains jeux dits « évolutionnaires », empruntés par les biologistes à l'économie, les stratégies optimales sont des stratégies mixtes, ce qui signifient qu'elles consistent à adopter des comportements systématiquement variés, car aucune stratégie pure ne permet de maximiser la valeur.

83 Comme on le verra dans la section suivante, l'ampleur de la correction dépend d'un paramètre appelé « coefficient d'apprentissage » : plus celui-ci est élevé, plus j'ai tendance à modifier mes prédictions en fonction des expériences passées immédiates. Dans notre exemple, la prédiction suivante sera d'autant plus proche de *X* que le coefficient 'apprentissage est important.

L'explication est légèrement différente de celle de Herrnstein : ici, je ne peux choisir tout le temps *A* plutôt que *B* car une telle stratégie s'avère *formellement* défavorable à long terme. Herrnstein suggère plutôt qu'une telle stratégie peut être optimale *in abstracto* mais ne convient pas aux choix tels qu'ils sont effectués dans la réalité : en pratique, la répétition des choix oblige les individus à varier leurs décisions. Néanmoins, le nerf de l'argumentation est identique. Il s'agit d'allouer des fréquences de choix en faveur de certaines options, et non pas d'établir un ordre de préférences entre ces options qui prévalent absolument.

La maximisation stochastique de la valeur tolérerait ainsi deux types d'erreur ou d'aléas, l'un étant lié à l'apprentissage, l'autre au caractère fréquentiel du choix. L'étude sur le jeu de l'inspection (*inspection game*) réalisée par Dorris et Glimcher (2004) en fournit une bonne illustration. Ce jeu, proposé initialement par un économiste (Kreps, 1990) est également connu sous le nom de « jeu du travail ou de l'esquive » (*work and shirk game*). Il oppose deux joueurs, un travailleur ou un inspecteur. Le premier choisit de travailler ou de ne pas travailler. S'il travaille, il obtient un gain certain ; s'il décide d'« esquiver », il peut obtenir un gain deux fois important (qui correspond au gain du travail plus le gain lié au loisir) à la condition toutefois que le second joueur ne réalise pas d'inspection. Celui-ci, en effet, doit choisir, au même moment que l'autre joueur, entre « inspecter » et ne « pas inspecter ». La modulation des coûts associés respectivement à l'inspection et à l'esquive possible du travailleur permet ainsi de modifier la matrice des gains possibles et les incitations de l'inspecteur (voir ci-dessous).

		ADVERSAIRE	
		Inspection	Pas d'inspection
SUJET	Travail	2 – C 0,5	2 0,5
	Esquive	1 – C 0	0 1

Matrice des gains du jeu de l'inspection, reproduit à partir de Dorris et Glimcher, 2004, p.366.

Les unités de gain représentent, pour le signe, 0,25 mL d'eau. Michael Dorris et Paul Glimcher proposent d'adapter ce jeu aux tâches oculo-motrices utilisées dans les études électrophysiologiques sur le système visuel du primate. L'expression des choix entre « travail » et « esquive » s'effectue par saccade oculaire : après une période de fixation, une saccade à gauche signifie « je choisis de travailler », une saccade à droite signifie « je choisis d'esquiver ». Bien sûr, le singe n'a pas conscience de ces expressions, mais il est supposé adopter le rôle du travailleur parce qu'il fait face à la matrice de gain exposée ci-dessus. Concrètement, après avoir répété la tâche pendant le conditionnement initial précédant l'expérience, le singe sait en effet qu'en regardant la cible de gauche, il obtient une récompense à tous les coups, de manière certaine, alors qu'un regard à droite permet d'obtenir de temps à autre une récompense deux fois plus importante.

Cette tâche correspond en fait, pour le singe, aux protocoles à intervalle et ratio variables (*variable intervalle (VI) variable ratio (VR)*) utilisés par les psychologues dans les années 1970 pour étudier le choix répété chez le pigeon (*cf.* chapitre 2). Du côté de l'adversaire, le comportement de l'inspecteur est simulé par un programme informatique (les singes jouent le rôle du travailleur). La matrice de gain de l'inspecteur dépend d'un paramètre spécifique, lié au coût de l'inspection, noté ici C. En effet, si j'inspecte le travailleurs alors que celui-ci travaille, je gagne $2 - C$, ce qui correspond au gain du travail réalisé (2), diminué du coût de l'inspection C. J'aurais donc pu gagner davantage (2) en n'inspectant pas. En revanche, si celui-ci ne travaillait pas, j'obtiens $1 - C$; j'ai donc ici bien fait d'inspecter puisque j'aurais obtenu un gain nul en n'inspectant pas. Bien que coûteuse, l'inspection peut donc dissuader le travailleur d'esquiver.

Les expérimentateurs ont ainsi conçu différents programmes informatiques correspondant au comportement optimal d'un inspecteur selon différentes valeurs possibles pour C. Les deux singes de l'expérience jouent 29 blocs d'une centaine d'essais contre ces différents opposants. Au cours de chaque bloc, l'opposant reste le même (la matrice des gains reste constante), si bien que le singe apprend, au fur et à mesure, à évaluer le profil de son opposant. En observant les conséquences de ses décisions passées en terme de récompense, il déduit des fréquences de récompense probables à chaque option, et en particulier pour l'option risquée. A partir d'un certain nombre de répétitions, on peut donc s'attendre à ce que le singe ait « compris » la matrice des gains associée à la tâche qu'il doit effectuer.

Lorsque ce jeu est répété, aucune stratégie pure n'est optimale, cela signifie que si je choisis toujours la même option, je ne peux maximiser mes gains. En effet, mon comportement devient trop prévisible et mon adversaire peut tirer parti de mon obstinations

sur l'une des deux options pour anticiper ms décisions. Comme dans les jeux de « pierre-feuille-ciseaux », la stratégie optimale est mixte, c'est à dire qu'elle consiste à varier les coups, pour empêcher que mon adversaire ne puisse prédire mes comportements à l'avance. Ici, on observe que pour le travailleur, les gains espérés de chaque option sont identiques (0,5). Selon la vigilance plus ou moins grande de l'inspecteur, le travailleur doit donc allouer des fréquences de choix en faveur d'une des deux options. Si par exemple l'observation des jeux passés me révèle qu'il y a inspection dans 50% des cas, de manière non-prévisible, je dois en conséquence travailler dans 50% des cas de manière non-prévisible.

Les résultats de l'expérience montrent que les singes, comme les sujets humains, sont tout à fait capables de ce genre de comportement stratégique qui consiste d'une part à estimer le profil de mon adversaire et à y adapter les fréquences de décisions en conséquence, et, d'autre part, à jouer de manière volontairement aléatoire afin d'empêcher que l'adversaire puisse lire dans mon jeu. Après quelques répétitions, les singes égalisent leurs fréquences de choix au programme informatique spécifique, et varient ensuite systématiquement leurs décisions tout en maintenant des fréquences stables de choix entre les options (Dorris et Glimcher, 2004, p.366-368).

Dans cette expérience, l'enregistrement de l'activité des neurones pariétaux par micro-électrodes génère un signal de prédiction de la récompense, observé dans les études neurophysiologiques évoquées plus haut. Avant d'effectuer la saccade, pendant la période de fixation, les neurones associés à un mouvement vers la droite ou vers la gauche s'activent en fonction de la récompense attendue, associée à chacun des deux mouvements possibles. La moyenne de ces observations montre que ce signal est effectivement prédicteur de la récompense, puisqu'il est corrélé avec les montants de gains déterminés par la matrice du jeu. Ces activations sont par ailleurs l'écho, au niveau neuronal, des choix ou comportements effectifs, puisque les singes allouent effectivement la fréquence de leurs choix en faveur de chaque option en fonction du gain espéré. Néanmoins, si la moyenne de ces observations neuronales et comportementales est stable à long terme, une forte variabilité au coup-par-coup est observée. Si par exemple la moyenne dex choix et des prédictions dans un bloc s'approche de la stratégie optimale qui consiste par exemple à travailler dans 50% des cas, le comportement observé ne consiste pas pour autant à alterner de manière répétitive esquive-travail-esquive-travail-*etc*. Il y a plutôt des séquences internes au cours desquelles les prédictions et les choix sont systématiquement biaisés en faveur d'une option, et l'amplitude de ces biais varie de séquences en séquences.

Comme le soulignent Dorris et Glimcher, la stochasticité des processus neuronaux (et

des comportements réels) est en fait double. La première est liée à l'apprentissage. Au fur et à mesure des répétitions dans un même bloc, j'améliore ma compréhension de mon adversaire, ce qui me conduit à modifier en conséquence mes prédictions de gains. La variabilité du signal de prédiction de la récompense s'explique alors par un un processus de tâtonnement empirique par lequel le travailleur découvre peu à peu la fréquence des inspections. A un second niveau, une fois que le travailleur a établi une estimation fiable de la stratégie (en termes de fréquences) de son adversaire, la stochasticité est requise afin de ne pas toujours faire les mêmes prédictions et les mêmes choix. La variabilité du signal neuronal permet alors une « *production biologique* » du comportement aléatoire, en tant que celui-ci est une stratégie maximisatrice (Dorris et Glimcher, 2004, p.368).

Ces résultats, on le voit bien, ne sont pas nécessairement incompatibles avec une approche économique, c'est-à-dire avec l'hypothèse de maximisation. Les problématiques liées à la stochasticité du choix ne sont pas neuves en économie, et plusieurs économistes ont essayé de développer depuis les années 1960 des modèles stochastiques de la décision (Lenfant, 2010). Glimcher (2003, p.252) propose de manière opportune un rapprochement avec deux de ces modèles. Il défend en particulier, en neurobiologie, l'utilisation de la théorie de « la main tremblante » de Selten (1975). Dans ce modèle, les utilités espérées sont fixes, mais le mécanisme de choix ou de prise de décision se comporte de manière stochastique. Ce type de processus peut avoir un intérêt stratégique dans le jeu de l'inspection ou dans d'autres jeux répétés dans lesquels aucune stratégie pure n'est optimale, et où il est nécessaire de varier les coups pour empêcher l'adversaire de prédire mon comportement. McFadden a quant à lui proposé un modèle dans lequel le mécanisme de choix est fixe (il consiste à choisir toujours l'option ayant la plus haute espérance de gain) mais les utilités attendues sont des variables aléatoires (McFadden, 1973). Ce type d'aléa correspond au premier type de stochasticité qui a été identifié dans le cadre du jeu de l'inspection (*cf.* supra), et qui est lié à l'apprentissage et à la nécessité de découvrir au coup par coup la matrice des gains.

Le signal neuronal de prédiction de la récompense peut donc être appréhendé comme un mécanisme assurant une maximisation de la valeur à la condition de bien préciser que cette valeur n'est pas une constante du choix répété, mais une variable aléatoire. Il est possible de prendre en compte la stochasticité du choix soit au niveau du processus de d'acquisition d'information et d'apprentissage, soit au niveau des utilités attendues. L'observation des mécanismes neuronaux sous-jacents aux choix répétés suggère l'idée d'une évolution optimisatrice à long-terme, reposant sur des processus stochastiques qui peuvent néanmoins déboucher à court terme sur des erreurs de prédiction et des comportements inadaptés.

Ce thème d'une rationalité « économique » de l'évolution biologique, selon lequel celle-ci serait un mécanisme d'optimisation à long-terme, a par la suite largement nourri un courant de recherches assez proches, appelé « *théorie des jeux évolutionnaire* » ou « *écologie comportementale* ». Dans les années 1980, des biologistes (John Maynard Smith, 1982 ; Stephen et Kreps, 1986 ; Kreps et Davies, 1993) analysent des problèmes concernant l'évolution des espèces à la manière de processus de maximisation sous contraintes. L'idée consiste à présenter l'évolution comme un ensemble de jeux stratégiques des espèces entre elles. Dans ces jeux, les stratégies optimales sont le plus souvent des stratégies mixtes, ce qui favorise une grande variabilité des comportements et des phénotypes.

En neurophysiologie comme en zoologie, ce genre d'approche conduit à accorder un avantage évolutif à l'erreur, ou, au moins, à la variabilité des choix et des comportements. L'instabilité des décisions lors de la répétition d'un choix à l'identique est ainsi « justifiée » comme la meilleure réponse d'un jeu stratégique. Cette théorie des jeux évolutionnaires permet donc de conserver une référence à la notion économique de maximisation de la valeur, en considérant celle-ci comme un processus stochastique, là où les psychologues néo-comportementalistes avaient pour cela modifié la forme de la fonction d'actualisation temporelle (*cf.* chapitre 2).

La convergence entre neurophysiologie et économie s'établit donc à la faveur d'une théorie de l'utilité espérée physiologique qui doit être comprise comme une théorie du choix stochastique. Mais le terme de stochastique, comme le montre l'expérience de Dorris et Glimcher (2004), peut s'entendre en deux sens différents, qui ne sont pas toujours bien distingués. La proposition selon laquelle la prédiction de la récompense est stochastique peut d'abord signifier soit que la prédiction, c'est-à-dire le processus de délibération, soit que les récompenses elles-mêmes, sont stochastiques. Dans un cas, la variabilité des comportements est une réponse à l'absence d'information fiable sur les gains pouvant être attendus, elle est un moyen d'acquérir des connaissances sur la distribution des récompenses : je choisis de manière variable pour améliorer mes prédictions. Dans l'autre cas, la variabilité est une stratégie souhaitable en elle-même : l'aléa devient un objectif en soi. Si l'étude neurobiologique d'inspiration néo-comportementaliste s'intéresse principalement aux processus d'apprentissage, elle conserve néanmoins un certain attachement à l'idée typiquement biologique selon laquelle la variabilité des comportements possède une valeur en soi.

III. De l'utilité espérée à l'apprentissage de la récompense

Au niveau biologique, la prédiction de la récompense est conçue pour s'appliquer à des choix dans lesquels les probabilités de gains associées à chaque option sont initialement inconnues et sont progressivement apprises. La notion d'utilité espérée physiologique est en cela trompeuse, car le signal neuronal correspondant ne porte pas sur des choix discrets, mais des séquences de choix répétés. Il ne génère donc pas des valeurs subjectives fixes, mais fournit des prédictions aléatoires dont la moyenne des réalisations autorise néanmoins, au fur et à mesure des répétitions, une convergence vers les valeurs optimales pour le problème considéré. Dire que le signal de prédiction de la récompense est stochastique implique donc, à terme, un mécanisme de correction *ex post* des erreurs d'appréciation. L'étude neurobiologique de l'utilité espérée physiologique débouche ainsi naturellement sur celle des processus d'apprentissage (*learning*) de la valeur ou « *apprentissage hédonique* » (*cf.* Sutton et Barto, 1998). Les travaux portant sur le système dopaminergique ont d'abord mis en évidence l'existence d'un signal d' « erreur de prédiction de la récompense » (*reward prediction error*) (A). Des algorithmes empruntés à l'intelligence artificielle (*machine learning*) permettent de modéliser ce signal comme un processus d'apprentissage hédonique, qui bouleverse les théories classiques du conditionnement (B).

A. les neurones dopaminergiques et l'erreur de prédiction de la récompense: un circuit neuronal spécialisé dans l'apprentissage hédonique

Les études neurophysiologiques portant sur l'apprentissage se sont nourries, comme, plus généralement l'ensemble de la neurobiologie contemporaine, à la fois de connaissances anatomiques assez anciennes, de nouvelles techniques de mesure fournissant des données expérimentales, mais aussi de nouvelles manières d'interpréter ces données. Les neurobiologistes avaient en effet depuis les années 1950 mis en évidence un certain nombre de fonctions associées à un neurotransmetteur qui joue un rôle important dans l'apprentissage, la dopamine. Cette dernière était notamment associée au plaisir, et, notamment, aux dérèglement du plaisir dans les addictions, en particulier dans la consommation de cocaïne.

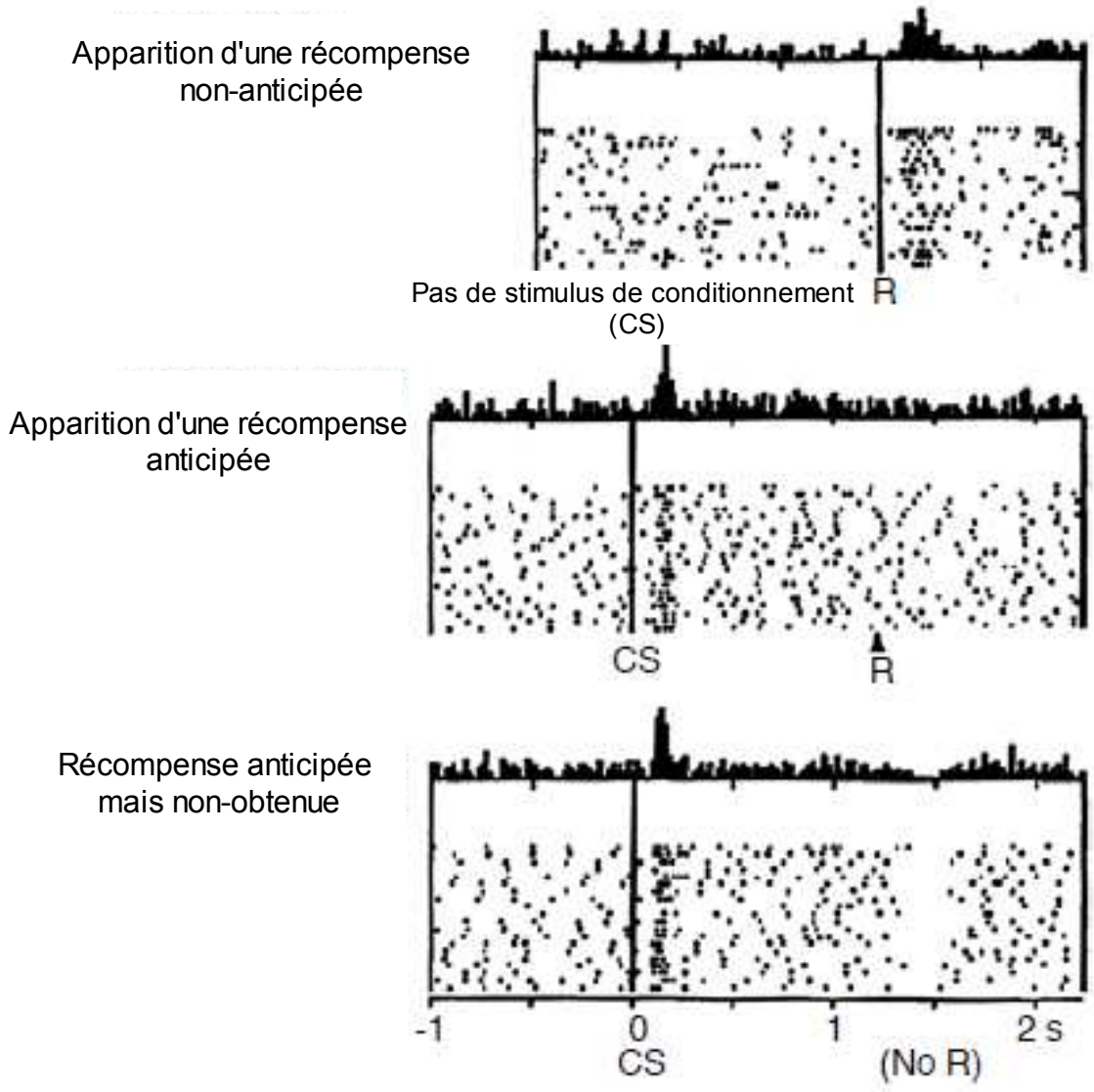
Des données anatomiques permettaient en outre de comprendre son mode de transmission dans le système nerveux. Les neurones émetteurs de dopamine, appelés neurones dopaminergiques, irriguent de très nombreuses régions du cerveau, tout en ayant des potentiels d'action de très longue durée (ce qui signifie que leur fréquence d'activation est faible). Le circuit dopaminergique était donc représenté comme un mécanisme d'inhibition et d'excitation non spécifique, très général, opérant à la manière d'une toile de fond pour les processus psychiques : comme l'écrit plus tardivement Glimcher, les neurones dopaminergiques « *ne peuvent pas dire grand chose au reste du cerveau, mais ce qu'ils disent doit être entendu par la plus large audience* » (Glimcher, 2003, p.307).

Le système dopaminergique était également associé à des fonctions motrices, puisque l'on savait notamment que les patients atteints de la maladie de Parkinson subissaient des déficits de dopamine. L'absorption de substances précurseurs de la dopamine (L-Dopa par exemple) permet ainsi de diminuer les symptômes moteurs chez les parkinsoniens. Néanmoins, au moment où les premières études électrophysiologiques sur des neurones dopaminergiques sont réalisées, aucune vision théorique d'ensemble ne permet de comprendre réellement le fonctionnement de ce circuit neuronal, dont les fonctions (plaisir, contrôle moteur) apparaissent hétérogènes. L'utilisation des micro-électrodes a peu à peu fait apparaître des interprétations originales sur le fonctionnement de ce module.

En 1986, Fibiger et Phillips montrent que, chez le rat, les neurones dopaminergiques sont actifs lorsque le sujet rencontre dans son environnement de la nourriture ou de l'eau. En répétant l'expérience toutefois, on observe que une progressive extinction de cette activation, au fur et à mesure des répétitions (Fibiger et Phillips, 1986). Certains ont alors proposé que cette activation représentait une sorte de signal attentionnel (voir par exemple Horvitz, 2000):la dopamine encoderait la salience des stimuli annonciateurs de récompense. Si j'identifie de la nourriture dans mon champ perceptif, cette perception, si elle suffisamment claire, me permet de prédire l'obtention d'une récompense à brève échéance. Cela entraîne une activation dopaminergique, qui débouche sur une action (se diriger vers la nourriture). Cependant, cette interprétation ne permet pas d'expliquer l'érosion du signal au fil du temps.

Les travaux de Wolfram Schultz et son équipe (Mirenowicz et Schultz 1994 ; Schultz, Dayan, Montague 1997) ont permis de mieux comprendre le fonctionnement du système dopaminergique. L'expérience la plus célèbre de Schultz (Schultz, Dayan, Montague 1997). Elle a joué un rôle très important dans la littérature, et a fait l'objet ultérieurement d'intenses discussions, porte sur une tâche de conditionnement simple. Une lumière signale à un singe le début de l'expérience : en appuyant sur un levier, il peut alors obtenir un jus de fruit. En

répétant cette tâche, le signal lumineux devient peu à peu stimulus de conditionnement, d'une manière similaire à la cloche qui, dans les expériences de Pavlov, annonce l'obtention de nourriture et fait saliver à l'avance le chien. Les expérimentateurs sélectionnent des neurones dopaminergiques et enregistrent leur activité au moyen de micro-électrodes.



Schultz, Dayan, Montague 1997, p.1594.

Les données recueillies suggèrent que l'activité des neurones dopaminergiques encode une « *erreur de prédiction de la récompense* ». En effet, au début de l'expérience, lorsque le singe ne s'attend pas à obtenir une récompense, l'obtention effective de celle-ci déclenche une forte activité des neurones dopaminergiques (*cf.* premier graphe ci-dessus): le singe est surpris et la récompense déclenche une erreur positive. Cette activation pourrait néanmoins correspondre au plaisir effectif, liée à la consommation du jus de fruit. Or, au fil du

conditionnement, le pic d'activité se déplace vers le stimulus de conditionnement (signal lumineux). Lorsque la lumière s'allume, le singe anticipe l'obtention de la récompense, ce qui produit donc à nouveau un signal positif d'erreur de prédiction de la récompense (*cf.* deuxième graphe ci-dessus). Ce dernier remplit donc une fonction purement prédictive et n'est pas nécessairement lié à l'obtention effective de la récompense. Si l'expérimentateur trompe le singe en ne délivrant pas de récompense après le signal lumineux, un pic d'activité est enregistré. Ce n'est qu'*ex post* que le singe corrige cette anticipation erronée par une erreur de prédiction qui est cette fois-ci négative. La non-obtention du jus de fruit attendu déclenche une forte diminution d'activité des neurones dopaminergiques. Le signal neuronal du circuit dopaminergique semble donc coder les modifications relatives de prédiction de récompense, qui peuvent être liées soit à l'obtention d'une récompense/punition (ou si l'on préfère, la non-occurence d'une satisfaction attendue) soit à des stimuli associés à des satisfactions/punitions. Cette idée simple fournit le point de départ à une théorie de l'apprentissage hédonique, qui emprunte à l'intelligence artificielle des algorithmes pour modéliser des processus du *learning*.

B. La formalisation des processus d'apprentissage: de Pavlov aux algorithmes du *machine learning*

Les expériences réalisées par Wolfram Schultz (Mirenowicz et Shultz 1994 ; Schultz, Dayan, Montague 1997) semblent au premier abord n'apporter qu'une simple confirmation de la théorie pavlovienne du réflexe conditionnée. Au niveau comportemental, la tâche similaire à celle des études de Pavlov sur le chien. La véritable nouveauté se situe au niveau neuronal. Les expérimentateurs observent en effet que c'est le même signal qui encode à la fois une satisfaction inattendue, et, par la suite, l'anticipation conditionnée de la récompense. Cette observation suggère que l'apprentissage, chez l'animal, ne procède pas passivement, par mémorisation des récompenses passées mais vise aussi à prédire le futur, et à anticiper, via des stimuli pertinents, les séquences d'actions à venir. Même si les données neuronales ne semblent ici simplement refléter des évènements extérieurs et/ou des comportements (1), elles ont néanmoins permis une transformation importante des théories traditionnelles du *learning*, d'inspiration pavlovienne (2).

1. Apports et limites des théories pavloviennes de l'apprentissage

Des années 1950 aux années 1970, la conception de l'apprentissage en psychologie est fortement liée à Pavlov et à la théorie du réflexe conditionné. Dans son expérience classique sur le chien, Pavlov montre qu'en exposant un chien à un stimulus de conditionnement (cloche) de manière répétée, avant de donner de la nourriture à ce même chien, un lien associatif se forme progressivement entre la perception du son de la cloche et la salivation, qui précède et anticipe l'obtention de nourriture (Pavlov, 1927).

En 1951, deux psychologues, Bush et Mosteller, proposent de formaliser cette observation sur le conditionnement sous la forme suivante:

$$A_{t+1} = A_t + \alpha\ (R_t - A_t)$$

où A_{t+1} désigne la probabilité de salivation au moment où la cloche sonne, ou, plus précisément, la force du lien associatif entre la perception de la cloche et la salivation; A_t désigne la même variable à l'essai antérieur ; R_t désigne le montant de récompense obtenu à l'essai antérieur, et α est un paramètre d'apprentissage. La variable *A* (probabilité de salivation) renvoie donc, plus généralement, à l'anticipation d'une récompense probable (ici, de la nourriture), au moment de la perception d'un stimulus de conditionnement (ici, la cloche). Cette probabilité est estimée à partir de la prédiction qui avait été faite lors de l'essai précédent, corrigée par une erreur de prédiction $(Rt - A_t)$ constatée *ex post*. L'ampleur de la correction dépend d'un paramètre d'apprentissage α compris entre 0 et 1. Plus α est grand, plus les prédictions dépendent des récompenses qui viennent d'être obtenues. Par exemple, dans le cas extrême où $\alpha = 1$, la prédiction à l'essai suivant est égale à la récompense qui vient d'être obtenue ($A_t = R_t$) (Bush et Mosteller, 1951).

La formalisation de l'expérience de Pavlov débouche donc naturellement sur une théorie de l'apprentissage, c'est-à-dire une théorie explicative de la manière avec laquelle les organismes parviennent à estimer les probabilités de récompenses futures. L'équation de Bush et Mosteller peut aussi servir de modèle formel de l'apprentissage. La forme mathématique est similaire, mais il ne s'agit plus ici de « probabilité de salivation », mais de valeur anticipée:

$$VA_{t+1} = VA_t + \alpha\ (R_t - Va_t)$$

d'où $VA_{t+1} = (1 - \alpha)\ VA_t + \alpha.R_t$

où *VA* désigne donc ici la valeur attendue de la récompense.

Ce type de règle d'apprentissage correspond à l'analyse ultérieure célèbre des deux psychologues Rescorla et Wagner (Rescorla et Wagner, 1972) et elle est ainsi connue dans la littérature sous le nom d'apprentissage « Rescorla-Wagner ». Cette formule semble adaptée à

l'expérience de Schultz étudiée plus haut. Les expérimentateurs observent en effet au début de l'expérience un pic d'activité des neurones dopaminergiques qui correspond à l'obtention d'une récompense inattendue, donc à l'erreur de prédiction $(R_t - VA_t)$. L'activité enregistrée après conditionnement, au moment du signal lumineux serait le signal de prédiction de la récompense VA_{t+1}, qui est indépendant de l'obtention effective d'une récompense R_{t+1}.

La formule d'apprentissage du type Rescorla-Wagner ne permet cependant pas de rendre compte d'un élément décisif de l'étude de Schultz. En effet, tout l'intérêt de cette expérience consiste à montrer que c'est le même signal (neuronal) qui encode l'erreur de prédiction, au début et à la fin de la tâche, et la prédiction elle-même : tout se passe comme si la prédiction VA_{t+1} correspondait en fait à une erreur de prédiction anticipée. En d'autres termes, VA_{t+1} ne se forme pas comme moyenne pondérée des prédictions et récompenses passées, telle qu'elle est exprimée sous la forme $((1 - \alpha)(VAt + \alpha.Rt))$ dans l'équation de Rescorla-Wagner, mais plutôt comme une prédiction de ce que va être l'erreur de prédiction future $(R_{t+1} - VA_{t+1})$. L'apprentissage a donc en fait pour fonction d'empêcher la génération d'erreurs de prédiction liées à des récompenses inattendues, en remontant dans le temps et en attribuant cette valeur non-prévue à un événement antérieur qui sert de signe annonciateur de la récompense.

2. L'apprentissage par différence temporelle (*Temporal Difference Learning* ou *TD learning*)

La limité principale de l'apprentissage du type Resorla-Wagner réside dans son caractère statique : l'individu apprend seulement au moment attendu de l'obtention de la récompense, en comparant le résultat à la prédiction. La mémorisation des prédictions et des récompenses passées est au fondement de l'apprentissage. Or, comme le suggère l'expérience de Schultz, l'apprentissage ne se déroule pas uniquement au moment attendu de l'obtention de la récompense, mais de manière continue: tout stimulus qui modifie mes anticipations produit un nouveau signal.

Ces difficultés théoriques ont justifié l'utilisation, dans les études neurophysiologiques sur l'apprentissage, de nouvelles formules d'apprentissage à différence temporelle (*temporal difference learning*). Ces modèles, empruntés à la robotique, ont fait l'objet d'une appropriation massive par les neurobiologistes. Le livre de référence en la matière est l'ouvrage de Sutton et Barto (1998), publié un an après l'expérience de Shultz.

L'apprentissage par différence temporelle se distingue de l'apprentissage Rescorla-Wagner en analysant le processus d'amélioration des prédictions non pas sur des choix discrets mais sur des séquences de choix. Par conséquent, les prédictions ne porte pas seulement sur la prochaine récompense anticipée, mais sur l'ensemble total de toutes les récompenses futures. Surtout, l'erreur de prédiction ne porte plus sur la constatation *ex post* d'un écart entre la récompense et la prédiction passée, mais, plus généralement, sur tout écart entre cette anticipation rationnelle des futures récompenses et toute information pertinente qui conduit à modifier cette anticipation.

A un niveau élémentaire, l'obtention d'une récompense inattendue est une « information » qui entraîne une réévaluation à la hausse de mes anticipations. La formalisation de l'apprentissage par différence temporelle ou par la règle Rescorla-Wagner ne se distinguent pas ici, en interprétant le pic d'activité observée dans l'expérience de Shultz avant conditionnement comme une erreur de prédiction de la récompense. Cependant, dans la perspective de Rescorla et Wagner, cette erreur est utilisée, de manière statique, au prochain essai, pour corriger la prédiction suivante. Elle est ainsi incorporée dans les prédictions suivantes, avec une ampleur variable selon le coefficient d'apprentissage α. Dans l'apprentissage par différence temporelle, l'erreur n'est pas prise en compte dans les futures prédictions, mais apparaît de manière de plus en plus précoce : elle est produite non plus au moment de l'obtention du jus de fruit mais à partir de l'occurrence d'un stimulus qui annonce de façon suffisamment sûre cette récompense. L'objectif consiste donc à identifier des prédicateurs stables de la récompense et remonter dans le temps. En cela l'erreur de prédiction « primitive » préfigure un procédé plus général d'apprentissage par lequel les anticipations sont réévaluées en permanence en fonction des événements extérieurs.

L'apprentissage par différence temporelle est ainsi tourné vers l'avenir, plutôt que vers la mémorisation des récompenses passées. Les signaux d'erreur de prédiction et de prédiction sont confondus dans une même et unique variable *V,* qui correspond à l'anticipation de la séquence totale de toutes les récompenses. Les études empiriques[84] montrent que le signal de prédiction de la récompense encode la prédiction V totale des récompenses, sous la forme suivante :

$$Vt = \gamma \ V_{t+1} + \gamma^2 V_{t+2} + \gamma^3 V_{t+3} + \gamma^4 V_{t+4} + \dots$$

$$Vt = \sum_{t=1}^{+\infty} \gamma^t V_{t+1}$$

γ est un coefficient d'actualisation temporelle, qui détermine la vitesse de

84 On se réfèrera à Schultz, 2007 pour une revue complète des études électrophysiologiques sur le système dopaminergique chez le singe.

l'apprentissage. Si γ est élevé (proche de 1), une faible variation dans les valeurs attendues des récompenses futures a un impact direct sur V_t. L'individu apprenant tend donc à émettre très souvent des signaux d'erreur de prédiction, et à anticiper très longtemps à l'avance l'occurrence des récompenses. L'inconvénient d'une telle stratégie réside dans sa moins bonne précision: plus je remonte dans le temps pour anticiper les récompenses futures, plus je m'expose à la possibilité d'une erreur. Un faible γ implique au contraire une forte diminution de la valeur actuelle des récompenses futures ; les signaux prédicteurs de récompenses ne conduisent qu'à une faible modification des anticipations et ne suffisent pas à produire un signal d'erreur de prédiction. L'apprentissage est donc beaucoup plus lent mais il est plus fiable.

Comment les récompenses futures V_{t+1}, V_{t+2}, V_{t+3} et suivants peuvent-elles être estimées en l'absence de connaissance fiable? Au début de l'expérience, après réception de la première récompense, rien n'indique au singe que cette récompense se répètera dans le futur. Ce n'est qu'au fur et à mesure des répétitions que le singe apprend progressivement à former des signaux prédictifs de V_{t+1}, V_{t+2}, V_{t+3} et suivant. En pratique, les récompenses futures sont donc bien sûr estimées à partir de l'observation des récompenses obtenues dans le passé (V_{t-1}, V_{t-2}, V_{t-3}, *etc.*). Si par exemple j'ai obtenu un même montant de jus de fruit, à échéance régulière, toutes les heures, à chaque fois que la lumière s'allume, je m'attends à l'instant *t* à obtenir à l'avenir des mêmes récompenses toutes les heures. Par conséquent, le signal lumineux déclenche une activation neuronale correspondant à

$$Vt = \gamma\ V_{t+1} + \gamma^2 V_{t+2} + \gamma^3 V_{t+3} + \gamma^4 V_{t+4} + \ldots$$

La différence essentielle de l'apprentissage par différence temporelle avec la règle de Rescorla-Wagner ne porte donc pas vraiment sur la nature des valeurs prédites (V_{t+1}, V_{t+2}, V_{t+3} et suivant au lieu de V_{t-1}, V_{t-2}, V_{t-3}, *etc.)*.C'est surtout la nature du paramètre d'apprentissage γ qui importe: γ n'est pas un coefficient de mémorisation, mais d'actualisation temporelle (des récompenses futures). Une convergence s'établit ainsi entre les problématiques du *learning* et celles concernant le choix inter-temporel, qui avaient fait l'objet de nombreuses études sur le pigeon dans les années 1960 (*cf.* chapitre 2).

Cependant, il convient de souligner que cette convergence entre apprentissage de la récompense et choix inter-temporel n'est pas encore exploitée, dans les années 1990, en électrophysiologie. Dans ces études, les expérimentateurs ne cherchent en effet pas à estimer une fonction d'actualisation et/ou à tester des prédictions sur la forme de cette fonction (hyperbolique, exponentielle), mais simplement à estimer un coefficient d'actualisation. Les

résultats neuronaux servent à mettre en évidence des écarts par rapport à un niveau d'activation standard (*neuronal base line*). Ces écarts sont interprétés comme des signaux d'erreur de la récompense; certaines études montrent que ce signal est affaibli en présence d'incertitude (Nakahara *et al.*, 2004) ou lorsque la périodicité d'obtention des récompenses s'allongent, confirmant ainsi la fonction d'estimation des récompenses futures associées à ce signal. Néanmoins, ces corrélations n'ont pas pour objectif d'estimer directement une fonction d'actualisation, ce qui supposerait des protocoles d'apprentissage plus complexes, avec des choix entre des options offrant des récompenses avec délai variable.

Les travaux en neurophysiologie portant sur l'apprentissage de la récompense rencontrent plusieurs problèmes théoriques dans les années 1990. Tout d'abord, l'encodage des erreurs négatives de prédiction fait l'objet de nombreux débats. Les données indiquent que les variations d'activité dopaminergique sont proportionnellement moins fortes pour les erreurs négatives que pour les erreurs positives (*cf.* Glimcher, 2010). En outre, l'expérience de Schultz suggère que le système dopaminergique encode un signal d'erreur de prédiction analogue à celui du TD *learning*, l'estimation complète des paramètres de cette fonction d'apprentissage reste à réaliser.

L'utilisation de protocoles plus complexes chez l'homme et de l'imagerie par résonance magnétique fonctionnelle a permis, au début des années 2000, de répondre à ces défis théoriques. Pourtant, la solution formelle de ces problèmes d'apprentissage était déjà connue des neurophysiologistes dans les années 1990: il manquait alors les techniques expérimentales nécessaires pour estimer les paramètres de ces fonctions. Sans anticiper sur le chapitre suivant, il est possible de suggérer ici, que, pour des problèmes d'apprentissage plus complexes, correspondant par exemple à l'expérience de Platt et Glimcher (199) précédemment évoquée, un algorithme d'apprentissage optimal peut être défini. Le critère d'optimalité de l'algorithme dépend de deux paramètres: le premier est le paramètre γ, qui est un coefficient d'actualisation temporelle (*cf. supra*). Le second, appelé « température », reflète un arbitrage entre ce que les informaticiens appellent l' « exploration » et l' « exploitation »: plus ce paramètre est élevé, plus l'individu aura tendance à varier ses choix pour améliorer ses connaissances concernant la distribution des récompenses, ce qui implique donc parfois de prendre volontairement des décision hasardeuses, en choisissant une option jugée moins favorable en l'état actuel des connaissances. A l'inverse, un coefficient de température faible indique que l'individu tend à se concentrer sur l'option jugée la plus rentable, au risque parfois de se tromper en établissant cette estimation sur la base de connaissances trop incertaines. Quelles qu'en soient ses limites empiriques dans les années 1990, cette manière originale de

formaliser les choix bouscule les cadres de pensée de la microéconomie.

C. Approfondissement et remise en question de la théorie économique

Les neurobiologistes comme Paul Glimcher comprennent leur approche au départ comme une tentative d'approfondissement, au niveau neuronal, de la théorie de l'utilité espérée (*cf. supra*). Leurs recherches visent à établir non pas une correspondance directe entre les variables neuronales et la maximisation de l'utilité espérée, mais à montrer que des mécanismes neuronaux d'apprentissage permettent d'approcher des solutions optimales sur le long terme. Cette approche fait de la maximisation de la valeur le cadre normatif du choix, comme idéal à atteindre sur le temps long, dans une perspective proche du programme kahnemanien.

Or, au fil des recherches, les références au cadre normatif de l'économie disparaissent. L'attention est portée en fait sur l'optimalité non pas des signaux de prédiction discrète, mais du processus d'apprentissage dans son ensemble. C'est donc finalement à partir de la comparaison entre les deux paramètres d'un algorithme optimal qu'un comportement donné peut, ou non, être décrit comme adapté à son environnement. La théorie biologique et computationnelle de l'apprentissage se substitue donc au cadre normatif de l'économie. Quelle place reste-t-il donc pour la théorie économique au sein de ce programme de recherche?

La neurobiologie à vocation économique qui se développe dans les années 1990 représente en fait un apport analytique pour la théorie microéconomique. En dépit de leurs ambitions initiales d'adopter le cadre normatif de l'économie, les études électrophysiologiques sur le système dopaminergique du singe aboutissent, comme les recherches sur le pigeon étudiées dans le chapitre précédent, à bousculer les référents normatifs de l'économie. Au lieu de se contenter d'emprunter à l'économie des modèles formels, la neurobiologie produit ainsi une théorie microéconomique du comportement apprenant. Cela confirme ainsi l'une de nos deux propositions d'interprétations de la neuroéconomie, selon laquelle l'économie n'a joué au fond aucun rôle moteur dans la consrtuction théorique de la discipline (*cf.* introduction).

L'adoption d'un cadre théorique complètement nouveau en économie, à partir de la formalisation des processus du *reward learning* est néanmoins source de difficultés à venir

pour cette discipline naissante. Les études neurobiologiques portant sur l'apprentissage de la récompense se trouvent en effet, dans les années 2000, intégrées au sein de l'économie comportementale, qui dans l'ensemble maintient une validité normative à la théorie économique, en adoptant le partage kahnemanien entre les théories normatives et descriptives (Tversky et Kahneman, 1986, p.252). Ces difficultés se donnent à voir notamment dans le problème des erreurs et des anomalies comportementales, qui fait apparaître la spécificité de ce programme de recherche en psychopathologie économique par rapport à l'analyse en termes de biais et d'heuristiques proposée par Kahneman et Tversky.

IV. Glimcher et l'économie: des rapports ambivalents

Les rapports qui unissent la neurobiologie à vocation économique des années 1990 à la théorie économique sont ambivalents. Le cas de Paul Glimcher peut servir d'exemple révélateur. Ses travaux sont d'abord caractérisés par un vocabulaire extrêmement confus pour les économistes (apprentissage de la récompense, utilité espérée physiologique) (A). La clarification de ces concepts permet de mettre en évidence, chez Paul Glimcher, une ambition théorique dépassant ses intentions initiales, puisqu'elle vise non pas à adopter le cadre normatif de l'économie mais à substituer à celui-ci des processus formels d'apprentissage de la récompense (B).

A. Utilité espérée physiologique, apprentissage de la récompense: un vocabulaire théorique confus pour l'économiste

Les travaux de Paul Glimcher ont eu une grand influence sur la neuroéconomie, et ont été abondamment commentés, aussi bien en économie qu'en neurosciences. Néanmoins, la démarche de Glimcher a fait l'objet d'une assez large incompréhension par les économistes. Ces confusions s'expliquent par le vocabulaire théorique assez confus utilisé par Glimcher dans les interprétations de ses expériences, ou dans les interprétations qu'il donne d'autres travaux expérimentaux.

Un premier malentendu important est lié à l'utilisation de l'expression d'utilité espérée physiologique pour designer le signal d'activation neuronal observé, dans l'expérience de Platt et Glimcher (1999), avant l'exécution de la saccade oculaire. Cette expression laisse sous-entendre en effet que ce signal représente la maximisation d'un choix discret. Or ce signal ne prend sens qu'à l'intérieur d'un problème séquentiel d'apprentissage de la récompense. En dépit du caractère équivoque de cette expression, Glimcher a pourtant toujours été assez clair sur ce point: l'utilité espérée physiologique ne correspond pas directement à son analogue en économie. La notion économique d'espérance d'utilité suppose un raisonnement sur des probabilités objectives, données à l'individu. Mais dès lors que ces probabilités ne sont pas connues mais doivent être progressivement découvertes, un critère de décision tel que le

théorème de Bayes devient inopérant:

> « *pour un économiste, des outils tels que le théorème de Bayes ou les fonctions d'utilité permettent d'identifier un idéal de comportement, ou ce que les économistes appellent le choix rationnel. De nombreux économistes supposent en outre que ces outils peuvent non seulement servir à identifier des solutions optimales mais aussi à prédire le comportement d'individus prenant des décisions économiques dans le monde réel. Malheureusement, cette approche en terme de rationalité globale* [le critère normatif servant à décrire à la fois les comportements optimaux et les comportements réels] *fonctionne assez mal pour expliquer la prise de décision dans le monde économique* » (Glimcher, 2003, p.199).

Cette citation laisse à penser que Glimcher s'inscrit là dans la continuité de la démarche de Daniel Kahneman: la théorie économique établit un critère normatif, les approches expérimentales fournissant une série d'écarts descriptifs par rapport à cette norme. Or c'est là où réside le plus grand malentendu à propos des travaux de Glimcher, car les anomalies n'ont pas du tout la même signification en neurobiologie et au sein du programme kahnemanien. Le malentendu renvoie à une confusion entre deux types d'anomalie chez Glimcher.

Les premières sont des fausses anomalies, c'est-à-dire qu'il s'agit de choix qui pourraient être considérées comme des erreurs, mais qui se justifient pour les besoins de l'apprentissage. Elles sont tolérées d'un point de vue normatif parce qu'elles sont liées à l'absence de connaissance initiale dans un problème dans lequel il s'agit d'apprendre au fur et à mesure la valeur des options. Dans ce type de situation, indépendamment du processus d'apprentissage dans son ensemble, des choix discrets ne renvoient à aucune maximisation: le système nerveux ne peut pas être optimal à chaque essai. Comme l'écrit Glimcher, « *nous ne pouvons supposer a priori que la sélection naturelle contraint le système visuel à déterminer une solution complète et optimale pour chaque problème computationnel qui relève de la survie de l'organisme* » (Glimcher, 2003, p.54).

Glimcher ne se place donc pas dans une perspective du choix similaire à Daniel Kahneman. Pour Glimcher, les individus réels et les animaux ne subissent pas nécessairement l'influence de déformations subjectives (*framing effects*) qui les empêcheraient de maximiser leur espérance d'utilité. Au contraire, pour Glimcher, « *une analyse de nombreux systèmes, cependant, semble suggérer que le système nerveux est très efficace pour réaliser certains buts computationnels. Ce que nous savons déjà montre que des buts bien spécifiés peuvent très bien servir à définir que que les animaux font effectivement* » (Glimcher, 2003, p.166). En

d'autres termes, le système nerveux est efficace dès lors qu'il s'agit d'atteindre des objectifs « bien spécifiés ». C'est au regard de ces « buts computationnels » qu'un système peut être dit rationnel ou irrationnel.

Pour Glimcher, l'évolution est un processus optimisateur sur le long terme. C'est la raison pour laquelle il considère que les définitions « économiques » de la rationalité fournies par le critère de maximisation d'utilité permettent d'approcher le comportement effectivement adoptés par les animaux et le organismes sur longue période. Or en fait, et c'est là l'une des sources de confusion les plus importantes, la théorie économique ne permet pas, dans les expériences de Glimcher, ou au mois pour les problèmes d'apprentissage, de spécifier de buts computationnels, c'est-à-dire ne permet pas de définir des objectifs quantitatifs pour un système apprenant rationnel. La citation évoquée plus haut, aux accents kahnemaniens, est donc trompeuse: la théorie économique, non seulement « *fonctionne assez mal pour expliquer la prise de décision dans le monde économique »* mais ne peut en outre servir de cadre normatif pour les problèmes d'apprentissage.

Cela ne signifie pas pour autant que, pour Glimcher, les animaux et les individus réels manifestent une rationalité parfaite. Glimcher considère bien un second type d'erreurs, qui constituent des déviances, des cas manifestes d'irrationalité. Il s'agit bien là d'anomalies au sens de l'économie comportementale, mais c'est par rapport à un processus d'apprentissage optimal, et non par rapport à une norme économique, que certains comportements peuvent être disqualifiés comme irrationnels. L'un des objectifs principaux de ce travail vise à montrer que, dans la neurobiologie d'inspiration économique, ces cas pathologiques servent de substitut à l'absence de norme. Il faudra analyser pour cela les insuffisances de la définition des algorithmes d'apprentissage (*cf.* chapitre 5). Glimcher, d'une certaine façon, pressent la portée théorique extrêmement importante de ces anomalies. Pour Glimcher, les échecs du vivant sont des « *outils critiques pour identifier les contraintes phylétiques et architecturales* » du système nerveux. Prenant l'exemple d'une bactérie qui adopte un comportement de forage (recherche de nourriture) aléatoire, qui peut ou non être adopté à son environnement, Glimcher écrit: « *d'une certaine manière, nous enrichissons plus nos connaissances à propos de la bactérie lorsque nous l'étudions dans des environnements dans lesquels son comportement n'est pas optimal. Ce sont les variables économiques que la bactérie ne peut prendre en compte qui peuvent en fait le mieux nous renseigner sur la physiologie de cet organisme. Cela suggère que les approches neuroéconomiques pourraient être, de manière paradoxale, le plus adaptées lorsque les animaux ne se comportent pas selon le schéma optimal prévu par la théorie* » (Glimcher, 2003, p.396).

Il pourrait encore une fois y avoir ici une confusion sur la nature de ce « *schéma optimal prévu par la théorie* ». Néanmoins, Glimcher identifie lui-même dans ce passage un élément spécifique de son approche, irréductible à l'économie: les anomalies « disent » quelque chose de plus que leurs phénomènes. Cette valeur ajoutée théorique apportée par l'observation de cas déviants signale donc les ambitions de réforme normative de ce programme de recherche.

B. Un rapport ambigu à l'économie

L'approche théorique de Paul Glimcher se distingue de celle de Daniel Kahneman, en tant qu'elle vise à remettre en question la théorie économique à la fois dans sa dimension normative et dans sa dimension descriptive (*cf. supra*). Les déclarations d'intentions et autres manifestes programmatiques sont, de ce point de vue, sources de confusion chez Glimcher car celui-ci y affirme souvent vouloir préserver le cadre normatif de la théorie économique. Dans son livre de 203, Glimcher écrit ainsi: « *les données comportementales fournies par les animaux dans ces expériences […] suggèrent que les sujets n'adoptent jamais de conduite optimale. Selon certaines personnes, cela prouve que les modèles économiques du comportement optimal ne sont pas utiles pour décrire le fonctionnement du cerveau. Je veux répondre ici à cette objection. Les modèles économiques décrivent les tâches que les animaux et les hommes accomplissent dans chaque situation décisionnelle. Ils définissent la manière avec laquelle un problème doit être résolu. Les animaux réels et les individus réels dévient de ces solutions: ils se comportent de manière sous-optimale* » (Glimcher, 2003, p.334).

En fait, les modèles économiques ne définissent pas la manière avec laquelle un problème d'apprentissage doit être résolu, pour la simple raison que les algorithmes du *reward learning* n'ont pas été élaborés par des économistes. Glimcher nourrit une ambition plus large que celle qui viserait simplement à énumérer une liste d'écarts par rapport à une norme économique de référence. Déjà, dans son ouvrage de 2003, Glimcher suggère que « *peut être de manière surprenante, dans une perspective neuroéconomique, des déviations se trouvent être plutôt des bonnes choses* » (Glimcher, 2003, p.396). Il souligne ainsi, dans son approche, le rôle théorique tout à fait spécifique des cas de déviance (*cf. supra*).

Les ambitions de réforme normative sont mieux assumée dans son livre plus récent,

publié en 2010. Glimcher y analyse le problème des anomalies en psychologie et économie comportementale. Il considère les différents traitements théoriques qu'elles sont susceptibles de recevoir, et que Glimcher rejette successivement (Glimcher, 2010, p.149). Une première solution consiste à montrer que ces anomalies peuvent être compatibles avec le cadre axiomatique de l'utilité espérée. Ces tentatives ont cependant été progressivement vouées à l'échec, les expériences mettant en évidence des anomalies toujours plus nombreuses et plus robustes. La solution inverse, en psychologie, revient à abandonner purement et simplement cette théorie et à adopter une approche en termes d'heuristiques. Le problème est que ces heuristiques manquent de généralité. Chacune d'entre elles vaut que pour un problème décisionnel donné. La troisième solution est celle qui est proposée par Kahneman et Tversky. La *prospect theory* représente une sophistication du cadre axiomatique de référence, en ajoutant notamment une fonction de pondération subjective des probabilités. La principale difficulté soulevée par cette approche est qu'elle multiplie l'ajout de paramètres *ad hoc*: dans sa forme complète, le modèle comprend cinq paramètres libres! Paradoxalement, la voie kahnemanienne est une solution qui sur le plan empirique n'est pas soutenable, du fait du trop grand nombre de paramètre à calibrer dans la fonction d'utilité.

Glimcher annonce quant à lui vouloir « *axiomatiser des concepts psychologiques inexistants dans la théorie standard* » (Glimcher, 2010, p.149), à la manière par exemple du modèle de Caplin et Leahy (2001) qui incorpore la notion de plaisir d'attente, de Loomes et Sugden avec la notion de regret (1987),de Gul et Pesendorfer et la notion de *self control* (Gul et Pesendorfer, 2001), ou encore à la manière de Rabin et Köszegi (2006) avec la notion de point de référence. On peut faire l'hypothèse, contrairement à ce qu'avance Glimcher, qu'il n'y a au fond nullement besoin d'axiomatiser ces concepts. La formalisation des processus d'apprentissage constitue de fait un nouveau sous-domaine de la microéconomie; comme l'écrit Glimcher, la neurobiologie « *pourrait permettre d'enrichir (ou si vous préférez, de complexifier) le modèle néoclassique original avec des contraintes algorithmiques* » (Glimcher, 2010, p.371). La vocation économique des neurobiologistes, dans les années 1990, est donc trompeuse: le problème du *reward learning* peut se comprendre comme un problème d'analyse économique, puisqu'il intervenir un processus d'optimisation sous contraintes, mais la modélisation de ce processus n'a pas été réalisée par des économistes.

Conclusion du chapitre 3

La science quantitative de la motivation, étudiée dans le cadre du chapitre précédent, débouche dans les années 1990 sur un ensemble d'études expérimentales portant sur le système nerveux du singe. Cet approfondissent neurobiologique du néo-comportementalisme est à l'origine de deux avancées théoriques majeures. Tout d'abord, l'influence du néo-comportementalisme permet à la neuro-électrophysiologie alors naissante de s'affranchir du cadre de référence de la réflexologie et du schéma stimulus-réponse. Ni sensoriel ni moteur, le signal d'utilité espérée physiologique témoigne de cette nouvelle approche en neurophysiologie, dans laquelle le système nerveux poursuit une fonction prédictive autonome.

Ce programme de recherche affiche au départ l'intention d'adopter le critère de maximisation de l'utilité espérée comme cadre normatif. Mais les neurobiologistes étudient en effet des protocoles qui permettent d'étudier des dynamiques d'apprentissage de la récompense, pour lesquels l'analyse économique standard est inopérante. Leur ambition économique initiale se transforme assez rapidement en un nouveau sous-domaine de la théorie économique.

Il y a là à la fois continuité et rupture avec la science quantitative de la motivation des années 1960. En effet, les expériences sur le pigeon ne révélaient que les choix des sujets, et non les mécanismes proprement délibératifs par lesquels ceux-ci acquièrent, développent et traitent l'information. L'observation de l'activité neuronale préalable au choix lui-même met en lumière les procédés délibératifs à l'œuvre dans les tâches de machines à sous multi-jeux. La délibération dans les *multi-armed bandit porblems* s'avère bien plus complexe que les routines de mélioration postulées par Herrnstein. L'étude et la formalisation des processus par lesquels la distribution des récompense sont progressivement découverts représente donc un apport théorique de la neurobiologie.

D'un autre côté, la neuro-électrophysiologie du *reward learning* s'inscrit dans le prolongement du néo-comportementalisme. Elle nourrit en effet des rapports ambivalents avec l'économie, en visant à la fois une intégration au sein de celle-ci, tout en élaborant un cadre normatif alternatif. Les psychologues néo-comportementalistes et les neurophysiologistes partagent une identité théorique commune, bien distincte de celle du programme de recherche kahnemanien.

Conclusion de la Première Partie

La neuroéconomie a pour antécédents théoriques un ensemble de recherches en électrophysiologie portant, dans les années 1990, sur l'apprentissage de la récompense chez le singe. Ces travaux ont eux-mêmes pour origine un courant de recherche plus ancien, connu sous le nom de science quantitative de la motivation. Mais celle-ci, à son tour, peut se comprendre comme un approfondissement du behaviorisme de Benjamin Skinner. Jusqu'où cette remontée dans le temps théorique peut-elle s'arrêter? L'identification des précurseurs d'une démarche scientifique semble conduire à une régression à l'infini.

L'approche de l'histoire des sciences défendue par Georges Canguilhem nous est apparue pertinente, ici, pour répondre à ces difficultés. La notion d'idéologie scientifique a permis de déterminer une identité commune à la neuroéconomie et à ses antécédents scientifiques. Contrairement à ce que le sens commun du terme d'idéologie peut laisser sous-entendre, il ne s'agit pas là de définir un ensemble de déterminations contingentes, liée au contexte d'apparition ou de diffusion, pouvant expliquer l'émergence de la neuroéconomie. L'identité commune à la neuroéconomie et à sa préhistoire est une identité purement théorique: de notre point de vue, ces recherches sont solidaires d'un projet de psychiatrie économique. Celle-ci désigne la synthèse *a priori* entre différents domaine de savoir: l'économie, la psychopathologie, et les techniques néo-comportementales du conditionnement et de l'apprentissage. Le programme de recherche ainsi constitué se distingue de chacune de ces composantes. L'utilisation d'une approche inspirée de l'économie distingue ce néo-comportementalisme économique du comportementalisme classique de Skinner, mais aussi de la description traditionnelle des troubles mentaux dans le langage de la psychopathologie traditionnelle. L'impulsivité, par exemple, est un trouble mental qui peut faire l'objet d'études comportementales, dont les résultats sont formalisés à l'aide de modèles empruntés à l'économie.

Mais cela signifie également que ce programme de recherche se distingue de l'économie. En dépit de son intention initiale de poursuivre un rapprochement avec l'analyse économique, le néo-comportementalisme se déploie progressivement comme un nouveau sous-domaine autonome de la théorie économique, spécialisé dans l'étude de l'apprentissage de la récompense. Ce rapport ambivalent à l'économie est source de difficultés. D'un côté, la discipline vise à s'intégrer au sein de la branche comportementale de l'économie sous influence kahnemanienne. De l'autre côté, elle sape implicitement ses fondements normatifs

en évaluant la rationalité non pas à partir d'une théorie économique de référence, mais à partir d'un algorithme d'apprentissage optimal. Les déclarations d'intention des précurseurs de la neuroéconomie sont ambigües. Le mariage entre néo-comportementalisme et le courant kahnemanien est ainsi porteur d'un divorce en puissance.

La prétendue vocation économique des néo-comportementalistes doit donc être nuancée. Il y a certes une volonté, au départ, d'imiter un style de modélisation inspiré de l'économie, parce que faisant référence à la notion de maximisation. Mais le problème du *reward learning* n'avait jusqu'alors pas été étudié par les économistes. Les chercheurs néo-comportementalistes, n'étant pas initialement économistes, peuvent revendiquer plusieurs apports propres à la théorie économique.

Les études sur le pigeon, dans les années 1960, ont d'abord permis de montrer que le modèle d'actualisation temporelle non-proportionnel au délai, proposé par Strotz, n'est pas une nécessairement une anomalie. Il peut naître des contraintes naturelles d'un problème de décision séquentielle. Dans un protocole impliquant un choix répété entre deux options offrant des récompenses avec des délais différenciés, un comportement visant à égaliser les fréquences relatives de choix en faveur de chaque option aux montants relatifs de récompense obtenus conduit implicitement à actualiser les récompenses de manière hyperbolique. Ces travaux expérimentaux fournissent donc un lien théorique entre choix séquentiel, inter-temporel et actualisation hyperbolique.

Les études sur le singe réalisés dans les année 1990 au moyen de micro-électrodes s'appuient sur des protocoles similaires. L'utilisation de techniques neurophysiologiques montre qu'à chaque étape d'un problème de décision séquentielle du type machine à sous multi-jeux, les montants attendus de récompense doivent faire l'objet d'un apprentissage, c'est-à-dire que la satisfaction obtenue dans le passé est mémorisée, et, surtout, que le système assurant cet apprentissage fonctionne par projection anticipative. Les algorithmes d'apprentissage par différence temporelle, empruntées aux sciences computationnelles pour modéliser ces dynamiques, convergent ainsi vers la problématique de l'actualisation temporelle.

Apprentissage de la récompense, choix inter-temporel, actualisation hyperbolique: la préhistoire théorique de la neuroéconomie constitue, dès les années 1990, un socle analytique solide pour le déploiement de cette dernière. Ces recherches ne portent cependant que sur l'animal,. Elles demandent encore à être confirmées sur l'homme. Effectivement, c'est en utilisant les ressources d'un nouvel instrument d'observation -l'imagerie par résonance magnétique fonctionnelle-que ce paradigme expérimental a pu, dans les années 2000, être

étendu à l'homme. Les premières expériences de neuroéconomie montrent ainsi, en s'appuyant sur la neuroimagerie, que les processus d'apprentissage de la récompense observés chez l'animal pour des récompenses alimentaires sont similaires à ceux observés sur des sujets humains, pour des types de récompense les plus divers (gustative, monétaire, symbolique, *etc.*).

L'innovation technologique a donc joué un rôle important dans le passage de l'animal à l'homme. Néanmoins, la généralisation des résultats aux sujets humains était déjà acquise dès les années 1990, avant l'apparition de l'imagerie fonctionnelle. En effet, les recherches sur l'animal montrent la richesse de l'intelligence animale. Elles établissent que celle-ci constitue une bonne approximation de conduites qualifiées de rationnelles chez l'homme. Il ne s'agissait donc pas du tout de pointer le caractère borné du comportement animal, et d'en évaluer ses possibles conséquences sur l'homme. Au contraire, la démarche s'appuyait en fait plutôt sur la projection, sur l'animal de laboratoire, de caractéristiques généralement considérées comme relevant spécifiquement de la rationalité humaine. Les expérimentateurs ont mis en évidence, par exemple, que les pigeons font preuve de sophistication inter-temporelle en étant capables de s'auto-contraindre et en anticipant le renversement de leurs choix de consommation; ou que le système nerveux chez le singe fonctionne de manière autonome, peut s'affranchir des contraintes externes immédiates, et ne se limite ainsi pas à la simple transmission de commandes motrice sou de signaux sensoriels, *etc.*

Cette approche ascendante, qui cherche dans l'animal des indices de la rationalité humaine, s'oppose ainsi nettement dans son esprit à la méthode de Daniel Kahneman, puisqu'elle n'aboutit pas du tout à montrer l'incapacité des individus, dans la vie réelle, à atteindre certains buts normatifs établis par la théorie économique (Tversky et Kahneman, 1986, p.252).

Si, finalement, le franchissement de la frontière entre l'homme et l'animal par la neuroéconomie s'appuie, pour reprendre les termes de Georges Canguilhem, sur un « *dépassement présomptueux* », l'origine de ce geste spéculatif se situe dans l'humanisation de l'animal plutôt que dans l'animalisation de l'homme. La préhistoire théorique de la neuroéconomie a donc permis de comprendre la genèse de cette idéologie scientifique de psychiatrie économique, fondée sur des déterminations biologiques. Les premières expériences de neuroéconomie, au tournant des années 2000, ne font que reproduire un paradigme plus ancien. Elles apportent un enrichissement essentiellement analytique, on va le voir, en approfondissant la nature du lien entre actualisation temporelle et apprentissage de la récompense, pour des protocoles plus complexes. Il reste donc à comprendre la signification

théorique, pour l'homme, de ce projet de psychiatrie économique, qui vise à décrire des troubles mentaux dans des termes économiques.

DEUXIEME PARTIE – D'UNE INNOVATION TECHNOLOGIQUE A LA CONSTRUCTION D'UNE NOUVELLE DISCIPLINE: LA NEUROECONOMIE DES ANNEES 2000 ET L'IMAGERIE FONCTIONNELLE PAR RESONANCE MAGNETIQUE

C'est en 2001 que sont réalisées les premières expériences de neuroéconomie à proprement parler (Berns *et al.*, 2001; Delgado *et al.*, 2001, Knutson *et al.*, 2001). Ces expériences mobilisent une nouvelle technique -l'imagerie fonctionnelle par résonance magnétique (IRM_f)- pour observer, chez l'homme, des processus cérébraux qualifiés d'apprentissage de la récompense (*reward learning*). L'IRM_f représente une innovation technologique importante. A la différence d'autres méthodes d'imagerie (tomographie par émission de positrons (TEP), magnéto-encéphalographie), l'IRM_f est en effet une technique non-invasive, qui a permis d'étendre à l'homme un paradigme expérimental utilisé dans les années 1990 sur le singe, à partir de micro-électrodes (*cf.* chapitre 3)[85].

L'apparition de l'IRM_f est antérieure à la neuroéconomie. La première image du cerveau obtenue par l'IRM_f est réalisée en 1992. Ce n'est que progressivement, au fil des années 1990, que ce nouvel outil conçu d'abord à des fins médicales commence à être utilisé en neurosciences cognitives. Assez rapidement, les études portant sur l'identification des fonctions du cerveau par l'IRM_f se multiplient. Mais ce n'est cependant qu'au cours de la décennie suivante que nait l'idée d'appliquer ce nouvel instrument à des protocoles d'apprentissage de la récompense, similaires à ceux utilisés chez le singe dans les années 1990.

Les expériences de neuroimagerie réalisées en 2001 (Berns *et al.*, 2001; Delgado *et al.*, 2001) s'inscrivent donc au départ dans le prolongement du programme de recherche néo-

85 L'IRM_f se distingue en outre par une assez bonne résolution temporelle et spatiale. Cette technique a néanmoins d'assez nombreux inconvénients. Elle impose en particulier d'assez lourdes contraintes sur le sujet lors de l'expérience: celui-ci ne doit pas bouger, le bruit de la machine peut être perturbant. Sa résolution spatiale reste en outre assez grossière, en comparaison des enregistrements obtenus par micro-électrodes à l'échelle du neurone. Les différentes limites liées à l'utilisation de l'IRM_f ont été soulignées par des économistes comportementalistes sceptiques quant à l'utilisation de la neuroimagerie en économie. Ces critiques seront abordées dans le chapitre 5 plus en détails. Pour une analyse complète des différentes techniques d'imagerie et de leurs avantages et inconvénients respectives, on se réfèrera à Filler, 2009.

comportementaliste décrit dans le chapitre précédent. Les travaux de Wolfram Schultz constituent en particulier une référence commune pour les premiers neuroéconomistes. Or, très vite, cette identité théorique se dissous au profit d'autres approches du comportement. Au début des années 2000, les études de neuroimagerie se revendiquant explicitement ou non de la neuroéconomie se multiplient, mais les protocoles utilisés n'ont plus grand chose à voir avec le cadre théorique originaire du *reward learning*.

Pourtant, l'IRM$_f$ a bien permis initialement de reproduire sur l'homme des résultats théoriques acquis antérieurement sur le singe. L'utilisation de ce nouvel instrument n'a pas abouti à rejeter le paradigme du *reward learning* : au contraire, les études sur l'homme ont permis d'en élargir sa portée descriptive mais aussi théorique, en approfondissant notamment les liens entre deux problèmes jusqu'alors disjoints, entre la prédiction de la récompense et l'actualisation temporelle. Si le cadre théorique lié à l'apprentissage de la récompense perd progressivement en influence dans les expériences de neuroimagerie, c'est d'abord parce que cette nouvelle technique expérimentale n'est pas utilisée uniquement par des neurophysiologistes, qui, comme Glimcher ou Schultz, s'inscrivent dans le courant néo-comportementaliste, mais aussi plus largement par des biologistes, puis par des économistes comportementalistes, qui appartiennent à des programmes de recherches distincts. La science quantitative de la motivation peine à affirmer son hégémonie sur l'IRM$_f$. Cet instrument fait l'objet de tentatives d'appropriation par des chercheurs issus de traditions théoriques différentes.

Les premiers travaux de neuroéconomie sont donc influencés par les pères fondateurs du néo-comportementalisme abordés dans la première partie, (Herrnstein, Schultz, Glimcher, *etc.*), mais aussi par d'autres traditions théoriques. En particulier, les travaux d'Antonio Damasio en neurosciences, ou ceux de Kahneman en économie comportementale, ont joué un rôle important dans l'orientation des recherches.

L'apparition de l'IRM$_f$ a donc été pour la science quantitative de la motivation un facteur à la fois de consolidation et de déstabilisation. L'irruption de cette nouvelle technologie a permis de confirmer sur l'homme les résultats acquis sur l'anima. Elle a aussi été l'origine d'une dilution de l'identité théorique propre au néo-comportementalisme. Par ailleurs, l'IRM$_f$ est censé élargir les capacités d'observation des expérimentateurs en apportant des données d'un nouveau type (signal BOLD représentant l'oxygénation du sang dans les différentes zones du cerveau) ; toutefois, l'interprétation de ces données génère également, chez les économistes comme chez les neurobiologistes,une forte suspicion quant aux pouvoirs

d'observation de ce nouvel appareillage. Entre enthousiasme débordant et scepticisme radical, l'IRM_f suscite les réactions les plus diverses. Les avantages liés à la nouvelle technique ne sont que les envers de ses inconvénients : si la multiplication des études de neuroimagerie signale un intérêt théorique accru pour la neurobiologie de la décision, l'éparpillement des recherches qui en résulte est aussi synonyme de fragilisation pour la discipline.

La fragilité des résultats théoriques de la neuroéconomie est ainsi inhérente au nouvel instrument -l'IRM_f- utilisé en laboratoire. Cela explique aussi, en retour, la grande diversité des approches théoriques mobilisées et du manque de cohérence caractéristique des débuts de la neuroéconomie. C'est en se re-saisissant de l'héritage théorique propre au néo-comportementalisme que les neuroéconomistes ont progressivement réussi à distinguer leur discipline de programmes de recherche concurrents et à faire de celle-ci un nouveau sous-domaine autonome de l'analyse économique. La seconde partie de notre étude a ainsi pour but de décrire le processus de construction de la neuroéconomie, à partir de ce qui a été conçu ici comme relevant d'un projet (et d'une idéologie) de psychiatrie économique. Nous limiterons ici au problème du choix inter-temporel et de son approche en termes de *reward learning*, qui constituent le cœur théorique à la fois de la neuroéconomie et de sa préhistoire. Le premier chapitre décrit la manière avec laquelle les premiers résultats de le neuroimagerie portant, chez l'homme, sur l'apprentissage de la récompense, ont été appropriés, assimilés et influencés par l'économie comportementale d'inspiration kahnemanienne et par les travaux du neurobiologiste Antonio Damasio sur les émotions (chapitre 5. Une discipline sous la tutelle de l'économie comportementale et des neurosciences, 2000-2005). Puis, dans la seconde partie des années 2000, la neuroéconomie s'est peu à peu affranchie de la tutelle des neurosciences et de l'économie comportementale, pour émerger comme une discipline autonome, caractérisée par une approche pathologique du comportement économique. L'idéologie scientifique du néo-comportementalisme a ainsi donné naissance à un nouveau sous-domaine de l'analyse économique, orienté spécifiquement par des considérations relatives à la maladie mentale (chapitre 5.De l'économie comportementale dans le scanner à la neuro-psychiatrie computationnelle : la constitution d'une discipline autonome, 2005-2010).

Chapitre 4. Une discipline sous la tutelle de l'économie comportementale et des neurosciences (2000-2005)

Au début des années 2000, une série d'études en neuroimagerie reproduisent sur l'homme des résultats expérimentaux obtenus dans les années 1990 chez le singe (*cf.* chapitre 3). Ces expériences mobilisent une nouvelle technique d'observation -l'imagerie par résonance magnétique fonctionnelle (IRM_f). Cette innovation technologique permet ainsi d'appliquer à l'homme les acquis principaux de la théorie du *reward learning* chez le singe, associée notamment aux recherches de Wolfram Schultz (*cf.* Schultz, Dayan et Montague, 1997). Assez rapidement, ces recherches en neuroimagerie donnent naissance à ce que Berns et Montague appellent alors, en 2002, « *l'économie neuronale* » (*neural economics*) (Berns et Montague, 2002). Les travaux en neurosciences dans ce domaine se multiplient.

Néanmoins, le paradigme du *reward learning* peine à s'imposer comme cadre théorique de référence. En effet, au cours de la première partie de la décennie 2000, les études par IRM_f des zones du cerveau impliquées dans l'apprentissage de la récompense de la récompense mobilisent peu à peu d'autres approches. En particulier, certains chercheurs issus de l'économie comportementale, et « convertis aux neurosciences », essayent d'appliquer aux neurosciences des cadres interprétatifs de leur programme de recherche. Ces économistes voient dans l'imagerie un moyen de confirmation directe de leurs propres explications du comportement. La neuroéconomie, à ses débuts, subit également l'influence d'autres courants de recherche en neurobiologie. Les travaux d'Antonio Damasio sur la notion d'émotion (Damasio, 1994, Bechara *et al.*, 1994) font notamment l'objet d'une réappropriation massive par les premiers neuroéconomistes.

La diversité des approches théoriques mobilisées constitue un frein pour le déploiement de la discipline, car aucun projet fédérateur ne s'impose. Les recherches ont un caractère très disparate. La neuroéconomie ne se constitue donc pas, dans un premier temps, comme une discipline autonome, mais se place sous la tutelle de discipline plus anciennes, et mieux reconnues sur le plan académique. La neuroéconomie fait notamment l'objet, on va le voir, d'un investissement massif par l'économie comportementale. Les premiers manifestes de la neuroéconomie sont ainsi conçus par des économistes comportementalistes. Cette mise sous tutelle d'un domaine de recherches initialement indépendant ne résulte donc pas

nécessairement d'une stratégie disciplinaire explicitement définie : le courant néo-comportementaliste tend à dissoudre son identité théorique au début des années 2000 simplement à cause de la circulation et de l'utilisation par d'autres programmes de recherches de l'IRM_f. L'innovation technologique a donc été un facteur de déstabilisation pour la science quantitative de la motivation.

Ce chapitre a ainsi donc pour objectif de décrire le magma théorique caractéristique des premières années de la neuroéconomie, et de mettre en évidence la progressive dissolution du cadre théorique néo-comportementaliste au sein de l'économie comportementale. Au départ, l'IRM_f permet pourtant d'étendre, du singe à l'homme, le domaine d'application de la théorie de l'apprentissage de la récompense (I. Les apports de l'imagerie fonctionnelle par résonance magnétique : extension et approfondissement d'un paradigme expérimental conçu initialement sur l'animal). Des économistes comportementalistes se ont ensuite intéressés à ces résultats. Néanmoins, cette assimilation s'est faite au prix d'une déformation de la perspective interprétative initiale, puisque l'idée d'une opposition entre raisons et émotions a été substituée à l'hypothèse d'une monnaie neuronale commune (II. La tutelle de l'économie comportementale ou « l'économie dans le scanner »). Cette « *économie comportementale dans le scanner* » (Ross, 2008) s'est par ailleurs largement appuyée sur les travaux du neurobiologiste Antonio Damasio, qui apparaît ainsi comme la deuxième figure tutélaire de la neuroéconomie au début des années 2000 (III. La tutelle des neurosciences : Antonio Damasio et la théorie des marqueurs somatiques).

I. Les apports de l'imagerie fonctionnelle par résonance magnétique : extension et approfondissement d'un paradigme expérimental conçu initialement sur l'animal

Il s'agit ici d'analyser les apports de l'IRM_f du point de vue du programme de recherche néo-comportementaliste qui a été étudié dans les chapitres précédents. L'imagerie par résonance magnétique fonctionnelle permet d'étendre à l'homme le paradigme expérimental du *reward learning*, qui avait été conçu initialement sur le singe dans les années 1990 (Schultz, Dayan et Montague, 1997). L'apprentissage de la récompense constitue un cadre théorique de référence pour de nombreux travaux en neuroimagerie au début des années 2000. En 2001, plusieurs expériences importantes mettent en évidence un signal d'erreur de prédiction de la récompense, similaire à celui qui avait été observé sur le primate à partir de micro-électrodes (A). Assez rapidement, les résultats de ces travaux pionniers sont reproduits et confirmés. Cette phase de convergence théorique donne ainsi naissance à l'économie neuronale (*neural economics, cf.* Berns et Montague, 2002) (B). L'innovation technologique liée à l'IRM_f a donc été à l'origine d'une double extension théorique: descriptive d'une part, en élargissant à l'homme des résultats acquis sur l'animal; et analytique d'autre part, car l'étude par neuroimagerie du cortex préfrontal a permis de mieux comprendre les liens entre l'apprentissage de la récompense et l'actualisation temporelle (C).

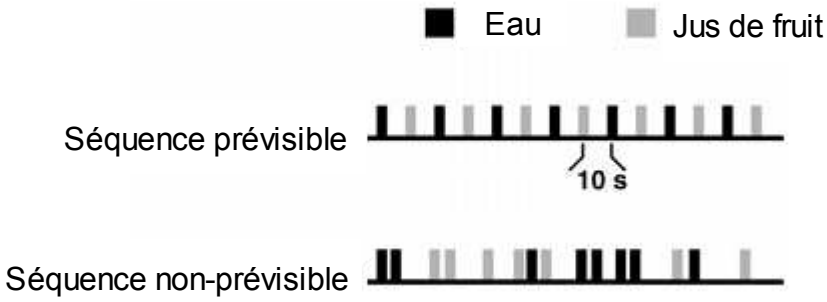
A. L'erreur de prédiction de la récompense: un paradigme expérimental prometteur pour les neurosciences en 2001

2001 est une année importante pour la neuroéconomie. Plusieurs études publiées au cours de cette année mettent en effet en évidence l'existence, chez l'homme, d'un signal d'erreur de prédiction de la récompense, analogue à celui qui est observé sur le singe à l'aide de micro-électrodes (*cf.* Schultz, Dayan et Montague, 1997). Il existe aujourd'hui une littérature extrêmement abondante sur le sujet. L'objectif n'est pas ici de rendre compte de l'ensemble des connaissances accumulées sur l'apprentissage de la récompense chez l'homme, mais de présenter trois expériences fondatrices ayant autorisé un rapprochement avec le

paradigme du *reward-learning*. Ces expériences dégagent plusieurs caractéristiques importantes du signal d'erreur de prédiction de la récompense: cette activité cérébrale remplit une fonction de nature projective et anticipatrice, qui se distingue à la fois du traitement des données sensorielles et des commandes motrices; en outre, l'activité observée est indifférente à qualité du stimulus, ce qui signifie qu'elle permet d'encoder la valeur de tout type de récompense.

L'étude publiée par Berns, McClure, Pagnoni et Montague (2001) peut être considérée comme la première expérience de neuroimagerie portant, chez l'homme, sur l'apprentissage de la récompense. Son apport spécifique consiste à montrer que certaines zones du cerveau remplissent une fonction de prédiction (des récompenses et des punitions), c'est-à-dire que leur activités est liée à la présence d'un aléa quant à l'obtention ou non d'une gratification.

L'expérience de Berns, McClure, Pagnoni et Montague a été explicitement conçue par ses auteurs comme une tentative de reproduction d'un travail de Schultz sur le singe, qui montrait que l'activité des neurones dopaminergiques varie selon la prédictibilité des récompenses (Schultz *et al.*, 1992). Berns et ses coauteurs affirment ainsi que leur protocole s'inspire très directement de celui utilisé par Schultz, d'abord parce qu'ils utilisent un même type de récompense, le jus de fruit (Berns, McClure, Pagnoni et Montague, 2001, p. 2797). Les sujets, placés dans le scanner, reçoivent ainsi de faibles montants (0,8 mL) de liquide (eau ou jus de fruit) par l'intermédiaire d'une paille. Deux conditions différentes permettent de mettre en évidence le rôle joué par l'incertitude. Dans la condition de contrôle, on alterne chaque type de récompense de manière régulière (eau-jus de fruit-eau-jus de fruit- *etc.*), et les intervalles de temps entre chaque récompense sont identiques (10 secondes). Dans la condition « imprévisible » (*unpredictable condition*), le délai d'attente entre deux récompenses varie à chaque fois, et l'alternance régulière entre les deux types de récompense n'est pas respectée. Chaque sujet est exposé à plusieurs séquences de récompense prévisibles et non-prévisibles, dans un ordre aléatoire.



Berns, McClure, Pagnoni et Montague, 2001, p.2794

Les expérimentateurs cherchent ensuite à corréler les activités observées dans les différentes zones du cerveau avec plusieurs variables comportementales. Les résultats montrent tout d'abord que les préférences entre les deux types de récompense, exprimées par un questionnaire, n'ont pas d'effet sur l'activité cérébrale. En revanche, la plus ou moins grande prédictibilité de la récompense modifie le signal observé dans la région du striatum ventral et dorsal (Berns, McClure, Pagnoni et Montague, 2001, p.2793). Cette région est largement irriguée par les neurones dopaminergiques, et elle est inscrite dans ce qui est appelé le circuit de la récompense. Cette zone du cerveau réagit donc à une incertitude face à une récompense attendue: ici, une forte augmentation de son activité est observée au moment de l'obtention effective de la récompense, mais seulement dans les séquences non-prédictibles. Lorsque l'obtention de la récompense devient trop certaine, c'est-à-dire trop régulière, dans les séquences prévisibles, l'activité observée initialement décroit au fur et à mesure du conditionnement, un résultat observé déjà par Schultz sur les neurones dopaminergiques du singe (Schultz, 1992).

Cette étude est importante car elle permet de mettre en évidence, à la manière des travaux antérieurs de Glimcher et de Schultz, un signal cérébral permettant de constater *ex post* une récompense non-anticipée. Cette activité implique donc la prise en compte d'une forme de plaisir, mais se distingue néanmoins de la simple expression d'une préférence subjective. L'activité observée ici dans le striatum remplit une fonction strictement prédictive: il s'agit d'une réponse à la prédictabilité, et non pas à un plaisir effectivement ressenti.

Berns, McClure, Pagnoni et Montague (2001) considèrent donc que leur expérience fournit un « *appui solide* » à l'application de modèles computationnels du *reward learning* sur l'homme (Berns, McClure, Pagnoni et Montague, 2001, p.2797). Une étude publiée deux ans plus tard par McClure, Berns et Montague (2003) renforce cette conclusion. Cette expérience s'appuie sur un protocole légèrement plus complexe, qui permet de montrer que le signal observé dans le striatum ne se limite pas à constater passivement l'obtention d'une gratification, mais permet aussi d'anticiper sur les récompenses futures. Dans cette version modifiée de leur première expérience, Berns, McClure, Pagnoni et Montague commencent d'abord par conditionner chaque sujet en exposant ceux-ci à une longue séquence prévisible, composée de 49 récompenses obtenues de manière régulière, 6 secondes après l'apparition d'un signal lumineux. Dans un second temps, les délais d'obtention de la récompense après l'apparition du signal lumineux sont imprévisibles et varient entre 1 et 10 secondes.

Deux types d'activité dans le striatum sont observées dans la deuxième partie de l'expérience, après conditionnement. Le premier intervient au moment où la récompense est

attendue, 6 secondes après l'apparition du signal lumineux. Cette activité correspond à la constatation d'un écart (négatif) entre la prévision d'une récompense habituellement obtenue à ce moment précis, et son absence à cet essai. Une deuxième activation est observée au moment de l'obtention effective de la récompense, de manière similaire à l'expérience précédente. Cette étude suggère donc l'idée d'un système neuronal impliqué dans l'apprentissage de la récompense, analogue à celui du singe, fonctionnant par la succession de signaux de prédiction et d'erreurs de prédictions.

Comme chez le primate, le *reward learning* chez l'homme s'appuie sur des circuits neuronaux spécifiques, qui fournissent une échelle commune d'évaluation à toutes les récompenses: « *les circuits de la récompense, chez l'homme, sont impliqués dans l'anticipation des récompenses indépendamment de leur modalité et fournissent ainsi un moyen pour comparer la valeur de stimuli hétérogènes* » (Berns, McClure et Montague, 2003, p.344). L'étude réalisée par Delgado et ses co-auteurs en 2001 -Delgado *et al.*, 2001) fournit un point d'appui important à cette idée, en montrant que les circuits de la récompense chez l'homme permettent d'estimer la valeur attendue de récompenses non seulement gustatives, mais aussi monétaires, ouvrant ainsi la voie à une très large extension des comportements humains susceptibles d'être concernés par le *reward learning.*

L'expérience de Delagado *et al.* vise explicitement, comme celle de Berns, McClure, Pagnoni et Montague (2001), à reproduire chez le sujet humain les résultats acquis précédemment sur le singe: « *cette étude a pour objectif d'identifier des régions du cerveau humain associés à la présentation d'une récompense et de corréler ces observations avec la perspective récente sur l'apprentissage de la récompense inspirée de la recherche sur les animaux* » (Delgado *et al.*, 2001, p.3072). Le protocole s'appuie ici sur un jeu de carte, impliquant des gains ou des pertes monétaires. Les sujets, placés dans le scanner, visualisent un carte sur laquelle figure un nombre masqué, compris entre 1 et 9. Ils doivent ensuite parier sur la valeur de la carte, en indiquant si celle-ci est selon eux inférieure ou supérieure à 5. Dans un second temps, la valeur de la carte est affichée, puis les sujets voient le gain ou la perte éventuelle associée à leur pari. Trois conditions du jeu sont utilisées par les expérimentateurs. Dans la condition de contrôle, il est indiqué aux sujets qu'ils n'obtiendront ni gain ni récompense. Dans la condition avec récompense (*reward condition*), les sujet débutent la partie avec un gain nul et obtiennent 1 dollar à chaque bonne réponse. Dans la condition avec punition, les sujets débutent avec 20 dollars de gains, et perdent 0,5 dollar à chaque mauvaise réponse.

D'une manière similaire à l'expérience précédente, cette étude met en évidence dans

les mêmes régions du cerveau deux signaux distincts, au moment de l'affichage des gains et des pertes, et au moment de la prédiction. Ces résultats suggèrent donc que ces zones du système nerveux sont impliqués dans l'évaluation des récompenses, indépendamment de leur nature ou de leur qualité, puisqu'elles traient chez l'homme à la fois des récompenses alimentaires et, ici, monétaires (Delgado *et al.*, 2001, p.3077).

L'étude de Knutson *et al.* (2001) améliore la compréhension des processus neuronaux d'apprentissage de la récompense chez l'homme et chez le singe. Cette expérience s'inscrit à nouveau explicitement dans le prolongement des travaux de Wolfram Schultz sur le singe : « *nos hypothèses sont inspirées par la recherche sur les primates (...] (Schultz et al., 1997)* (Knutson *et al.*, 2000, p.20). Knutson *et al.* soulignent par ailleurs que l'ensemble des régions du cerveau impliquées (noyau caudé, putamen, nucelus acumbens, cortex cingulaire antérieur, cortex préfrontal), communément regroupés au sein du circuit de la récompense, sont connus en anatomie pour être associées à l'anticipation et à la réponse aux stimuli extérieurs. L'IRM_f, comme les micro-électrodes sur le singe, vient donc confirmer et préciser des connaissances accumulées par d'autres moyens, antérieurement, ne neurophysiologie : l'innovation technologique ne fait pas table rase du passé des neurosciences, mais s'appuie au contraire sur la consolidation et la convergence progressive des observations. De ce point de vue, cette expérience améliore la compréhension des données recueillies, en montrant que le signal d'erreur de prédiction ne reflète ni le traitement de données sensorielles, ni la transmission de commandes motrices. A la frontière entre le sensoriel et le moteur, les processus du *reward learning* engagent une forme de cognition tout à fait spécifique, qui remet donc en cause le paradigme traditionnel du réflexe.

Le protocole vise ainsi à dissocier le signal d'erreur de prédiction de toute caractéristique sensorielle ou motrice. Après la visualisation d'un stimulus lumineux, les sujets doivent répondre le plus rapidement possible en pressant sur un bouton. Comme dans l'expérience précédente, trois conditions différentes sont utilisées : dans la condition de contrôle, aucune récompense n'est obtenue. Pour la condition avec récompense et pour celle avec punition, un gain ou une perte monétaire est calculé en fonction du temps de réaction de l'individu.

Les régions du circuit de la récompense s'activent au moment de l'affichage des gains et des pertes, seulement dans les conditions avec récompense ou punition. Le signal observé, qui encode un signal d'erreur de prédiction, est donc conditionnel à la présence d'une incitation (ici, monétaire) : « *plusieurs facteurs suggèrent que nous avons mis en évidence ici, avec ce paradigme expérimental, une activité cérébrale liée à une incitation. Tout d'abord, les*

activations observées ne peuvent être expliquées par les propriétés sensorielles du stimulus, puisque les séquences de stimuli dans la condition de contrôle ne provoquent aucune activation dans ces régions, alors que ces stimuli sont qualitativement identiques à ceux utilisés dans les séquences avec récompense et punition. […] Ce n'est que lorsqu'ils sont associés à des valeurs incitatives que ces stimuli déclenchent une activité. En outre, la préparation motrice ne peut pas non plus expliquer ces résultats. Les participants effectuent en effet une réponse motrice à tous les essais de chaque séquence, et pourtant les activations n'apparaissent que dans les séquences avec incitations. » (Knutson *et al.*, 2000, p.26).

Pour Knutson et ses co-auteurs, le signal de prédiction se situe à la frontière entre le sensoriel et le moteur, tout en se distinguant également de la pure cognition ou réflexion: le signal est impliqué dans l'évaluation d'une récompense, mais cette estimation est quasi-instinctive, et participe de près aux processus sensorimoteurs. Il y a ici quelque chose de très frappant pour les neurobiologistes: ce type d'activité neuronale va en effet à l'encontre de modèle traditionnel du réflexe, emprunté à Sherrington (*cf.* chapitre 3). Plus généralement, les trois travaux fondateurs étudiés ici suggèrent une forme tout à fait particulière d'activité neuronale, de nature essentiellement projective, distincte de l'encodage de plaisir hédonique à proprement parler, mais néanmoins susceptible de s'appliquer à des types de gratifications extrêmement variées. L'ensemble des expériences de neuroimagerie sur l'apprentissage de la récompense convergent alors, au début des années 2000, vers la notion fédératrice de « *monnaie neuronale commune* » (Montague et Berns, 2002).

B. La notion de monnaie neuronale commune :un résultat fondateur pour l'« économie neuronale »

Les expériences de neuroimagerie sur l'apprentissage de la récompense chez l'homme débouchent, en 2002, sur la publication, par Montague et Berns, d'un article de synthèse, intitulé « l'économie neuronale et les substrats biologiques de l'évaluation » . Cette étude fait assez rapidement figure de manifeste pour ce domaine de recherches alors en plein essor. Montague et Berns y avancent deux concepts importants : la notion de « *monnaie neuronale commune* » tout d'abord, pour rendre compte des schémas d'activité associés au circuit de la récompense chez l'homme ; et l'expression d'« *économie neuronale* », qui suggère une

ouverture possible des recherches à l'économie.

Comme pour Paul Glimcher, le succès de l'article de Berns et Montague dans la littérature neurobiologique est lié à l'utilisation d'un vocabulaire à caractère économique. Cette tonalité économique s'exprime tout d'abord à travers l'expression de « *monnaie neuronale commune* » (*common neural currency*). Selon ses auteurs, cette notion renvoie au caractère polymorphe de l'apprentissage de la récompense chez l'homme: les expériences montrent en effet que « *les réponses neuronales dans le circuit composé du striatum et du cortex orbito-frontal permettent de convertir des unités hétérogènes de récompenses futures en une sorte de monnaie interne, c'est-à-dire dans une échelle d'évaluation commune à l'évaluation de tous les stimuli ou décisions comportementales* »(Berns et Montague, 2002, p.265).

Si le terme de « *monnaie* » (*currency*) évoque déjà un rapprochement avec l'économie, cette ambition est plus explicitement formulée à travers l'expression d' « *économie neuronale* », qui est utilisée ici pour désigner l'ensemble des recherches expérimentales sur le *reward learning*. Pour Berns et Montague, les expériences sont liées aux travaux de Kahneman et Tversky, dans la mesure où, notamment dans l'expérience de Berns, McClure, Pagnoni et Montague précédemment citée, les résultats indiquent que le contexte dans lequel les récompenses sont obtenues influencent l'évaluation subjective de cette récompense (en particulier, l'évaluation est sensible à la prédictibilité de l'environnement, *cf. supra*). Berns et Montague ajoutent que ces effets peuvent se comprendre comme des effets de cadrage (*framing*) au sens kahnemanien (Berns et Montague, 2002, p. 270).

Il faudra revenir plus en détails sur cette proximité supposée entre *reward learning* et l'approche kahnemanienne du choix en incertitude. La référence aux travaux de Kahneman n'est en tout cas pas déterminante dans l'article de Berns et Montague. Les deux auteurs visent plus généralement à établir un lien avec la théorie économiqueau sens large, sans nécessairement inscrire leur démarche au sein d'un programme de recherche en particulier. Si l'économie neuronale a quelque chose à voir avec l'analyse économique, c'est tout simplement parce que les neurobiologistes s'occupent de décisions impliquant l'évaluation d'une récompense, que Berns et Montague qualifient d'« *évaluation économique* » : « *en utilisant le terme d'évaluation économique, nous nous référons aux problèmes auxquels fait face le système nerveux de l'individu lorsque celui-ci doit effectuer des décisions successives rapides, en prenant en compte les coûts réels immédiats et les retombées futures de ces actions, bonnes ou mauvaises* » (Berns et Montague, 2002, p.265).

L'article-manifeste de Berns et Montague de 2002 constitue un socle théorique

commun à l'ensemble des travaux de neuroimagerie portant sur l'apprentissage de la récompense, à partir d'une référence très large à l'économie, qui va plus loin qu'une inscription stricte au sein du programme kahnemanien. La notion de monnaie neuronale commune de Berns et Montague, tout comme l'utilité espérée physiologique de Paul Glimcher (*cf.* chapitre 3) se situe en fait dans le prolongement de l'approche théorique de Richard Herrnstein. Sur le plan théorique, l'économie neuronale trouve un socle analytique dans l'utilisation d'algorithmes d'apprentissage empruntés au *machine learning*. Si le contenu économique de ces recherches reste donc encore à définir, celles-ci trouvent donc dans leur regroupement au sein de l'économie neuronale l'occasion d'affirmer leur unité et leur cohérence.

Plus tard, d'autres travaux de synthèses ont poursuivi cette tâche d'unification des résultats expérimentaux. Dans un article intitulé « les substrats neuronaux de l'évaluation des récompenses chez l'homme : le rôle de l'IRM$_f$ », McClure, York et Montague dressent ainsi un large panorama des expériences réalisées, et identifient deux éléments théoriques communs à tous ces travaux : l'utilisation de l'IRM$_f$ d'une part, et la formalisation des données à l'aide d'algorithmes d'apprentissage hérités des sciences computationnelles[86] (McClure, York et Montague, 2004). On peut noter enfin la publication en 2009 d'un manuel consacré à la récompense et à la prise de décision, sous la direction de Jean-Claude Dreher et Leon Tremblay, avec notamment un chapitre spécifique sur l'apprentissage de la récompense dans le cerveau par McClure et D'Ardenne (McClure et D'Ardenne, 2009).

Parallèlement, les études sur le singe font aussi l'objet de nombreuses tentatives de synthèse (voir notamment Schultz, 2002). Le début des années 2000 voit donc la consolidation du paradigme expérimental du *reward learning*, désormais étendu à l'homme grâce à l'IRM$_f$. L'apport de cette nouvelle technique ne se borne pourtant pas à généraliser les observations obtenues sur le singe aux sujets humains. L'IRM$_f$ permet en outre un approfondissement analytique de ce paradigme, puisque l'activité de certaines régions frontales du circuit de la récompense éclaire sous un jour nouveau le lien entre apprentissage et actualisation temporelle.

86 Les auteurs font notamment référence à l'ouvrage de référence de Sutton et Barto (1998), qui a fait l'objet d'une étude approfondie dans le chapitre précédent consacré à Paul Glimcher.

C. Apprentissage de la récompense et choix intertemporel : le rôle spécifique du cortex pré-frontal

A partir des années 1990, les neurobiologistes commencent à emprunter aux sciences computationnelles des algorithmes d'apprentissage par différence temporelle (*TD learning*) pour formaliser les données neuronales fournies par micro-électrodes (*cf.* chapitre 3, section III). Au cours de la décennie suivante, l'utilisation de l'imagerie par résonance magnétique fonctionnelle a permis non seulement d'appliquer ce paradigme expérimentale au cerveau humain, mais aussi d'approfondir ses implications théoriques. En effet, chez l'homme, une zone spécifique appartenant au circuit de la récompense, localisée dans le cortex préfrontal, est connue pour avoir un rôle dans la mémorisation et la planification temporelle des décisions. L'étude par imagerie de cette région rend ainsi d'autant plus pertinente l'utilisation, en neurosciences, d'algorithmes du *reward learning*, en montrant comment cette monnaie neuronale commune incorpore un facteur d'actualisation temporel.

L'apprentissage par différence temporelle (*TD learning*) suppose (*cf.* chapitre 3), qu'un système apprenant évalue les récompenses susceptibles d'être obtenues en émettant un signal de prédiction V_t de la forme suivante :

$$Vt = \gamma \ V_{t+1} + \gamma^2 V_{t+2} + \gamma^3 V_{t+3} + \gamma^4 V_{t+4} + \dots$$

$$Vt = \sum_{t=1}^{+\infty} \gamma^t V_{t+1}$$

γ est un coefficient d'actualisation temporelle, ce qui signifie que chaque signal correspond à la valeur anticipée de toutes les récompenses futures. A la différence de la règle d'apprentissage classique du type Rescorla-Wagner, l'apprentissage par différence temporelle procède par anticipation projective, et non pas par la simple constatation passive d'un écart entre les prédictions et les récompenses effectivement obtenues (*cf.* chapitre 3, section III).

Les algorithmes du *TD learning* impliquent donc une forme d'actualisation temporelle, à travers le coefficient γ. Dans les études sur le singe, la question du *discount* temporel n'était cependant pas approfondie pour elle-même. Certaines études montraient que le signal de prédiction est affaibli lorsque la périodicité d'obtention des récompenses s'allongent (Nakahara *et al.*, 2004), mais ces résultats étaient doublement limités. D'une part, aucune région spécifiquement responsable de l'actualisation n'avait été identifiée. D'autre part, les corrélations mises en évidence entre le délai et l'intensité du signal (Nakahara *et al.*, 2004) n'avaient pas pour objectif d'estimer directement une fonction d'actualisation, ce qui aurait supposé des protocoles d'apprentissage plus complexes, avec des choix entre des options

offrant des récompenses avec délai variable.

L'IRM$_f$ permet de contourner ces difficultés, puisque, chez l'homme, les neurobiologistes sont en mesure de recueillir des données neuronales dans des choix entre deux récompenses avec des délais différenciés. Des corrélations quantitatives peuvent ainsi être étbalies entre les fonctions d'actualisation estimées à partir des choix et les signaux d'activité cérébrale. Surtout, le problème de l'actualisation temporelle renvoie chez l'homme à une zone spécifique du cerveau -le cortex pré-frontal- qui a fait l'objet de nombreuses études en neuroimagerie mais qui bénéficie aussi de l'appui de recherches plus anciennes.

Les travaux fondateurs d'Antonio Damasio en particulier ont dès le début des années 1990 établi que les régions frontales étaient impliquées dans la planification temporelle des décisions. A la suite de Damasio, l'activité du cortex préfrontal est souvent interprétée, notamment par les économistes comportementalistes, en opposition avec l'activité dans les régions limbiques, impliquées dans l'évaluation des récompenses immédiates ou à très court terme: il y aurait d'un côté des zones « *impulsives* », et de l'autre côté un territoire cérébral responsable du « *self control* ». Cette perspective dualiste sur la cognition tend ainsi à remettre en question l'hypothèse d'une monnaie neuronale commune. Sans anticiper sur la présentation des recherches réalisées par Damasio et son équipe (*cf.* troisième section), il est possible au moins de suggérer ici que cette représentation dualiste du fonctionnement du cerveau n'est pas la seule possible, et que l'identification d'une fonction de mémorisation et de planification temporelle peut tout aussi être compatible avec l'hypothèse d''un système neuronal unitaire d'évaluation des récompenses.

Bogacz, McClure, Li, Cohen et Montague (2007) proposent ainsi un modèle d'apprentissage par différence temporelle incorporant cette fonction de planification temporelle du cortex frontal. Les auteurs reprennent à leur compte un protocole utilisé par Richard Herrnstein, connu sous le nom de « *tâche à optimum croissant* » (*rising optimum task, cf.* Herrnstein, 1991, et son analyse dans le chapitre 2). Il s'agit en fait d'un problème type *multi-armed bandit*, mais dans lequel la distribution des récompenses associée à chaque machine à sous est conçue de telle manière qu'un comportement de *matching* ne permet pas de maximiser les gains sur le long-terme. En effet, le montant des gains obtenus à chaque fois en choisissant l'une des deux options dépend des choix passés: plus je choisis la même option, plus les récompenses obtenues diminuent, jusqu'à atteindre un niveau plancher. Cependant, le niveau plancher de la machine *A* reste toujours supérieur au gain moyen obtenu en suivant une stratégie de *matching*. Par conséquent, la stratégie optimale dans le *rising optimum task* consiste à découvrir que l'option *A* se révèle toujours être l'option la plus favorable à chaque

essai, et ce malgré la baisse progressive de son rendement.

L'étude de cette tâche avait abouti à des résultats similaires chez l'homme et chez le pigeon (*cf.* Herrnstein, 1991): la plupart des sujets se montrent incapables de maximiser et se contentent d'égaliser au coup par coup les rendements relatifs de chaque option (*cf.* chapitre 2). Bogacz, McClure, Li, Cohen et Montague retrouvent ici des résultats identiques dans la première partie de leur étude avec des sujets humains. Mais les auteurs observent qu'en diminuant le délai d'attente entre le choix et l'affichage de la récompense (et donc en augmentant la fréquence des choix effectués), un plus grand nombre de sujets parvient à identifier la stratégie optimale, consistant à toujours choisir l'option A (Bogacz *et al.*, 2007, p.3)[87].

Ce résultat peut apparaître contre-intuitif: plus la fréquence des décisions augmente, plus les individus semblent capables d'effectuer des décisions réfléchies. Tout se passe donc comme si la réduction du délai entre chaque choix améliore la capacité à raisonner sur des séquences globales de choix, plutôt que sur des successions de décisions envisagées indépendamment les unes des autres. McClure, Li, Cohen et Montague proposent une explication et un modèle orignal pour décrire ce phénomène. Selon ces auteurs, le comportement d'égalisation ou de *matching* témoigne d'une myopie temporelle du décideur: l'individu, en effet, se contente de continuer à choisir la même option lorsque la récompense qui vient d'être obtenue est supérieure à la récompense obtenue précédemment lors du dernier choix en faveur de l'autre option. Lorsqu'une option débouche sur des récompenses trop faibles, l'individu change d'option. Ce type de stratégie permet effectivement d'égaliser les rendements relatifs de chaque option sur le long-terme, mais il ne permet pas de raisonner sur ds séquences globales éventuellement plus avantageuses, comme dans le cas du *rising optimum task*. Le problème, dans la mélioration, est que le décideur n'envisage pas la moyenne des gains passés mais uniquement le résultat obtenu lors de la dernière décision. Les décisions prises en t_{-2}, t_{-3}, ...t_0 ont donc un poids décroissant et quasi-nul dans la décision.

Pour Bogacz *et al.*, le sujet peut être incité à adopter une perspective cumulée sur ses gains ou pertes dès lors que celui-ci « *garde une trace* » de l'ensemble de ses choix passés, et pas seulement du seul choix précédent la décision. Or les récompenses obtenues antérieurement sont susceptibles d'influencer d'autant plus la décision présente que l'éloignement dans le temps de ces récompenses est faible. C'est la raison pour laquelle le raccourcissement des délais d'attente entre chaque décision, en rendant le décideur moins

87 On pourrait penser, en effet, qu'une décision complexe, reposant sur une prise de perspective globale des gains et des pertes, suppose au contraire un délai plus long entre chaque choix, permettant au décideur de mieux « réfléchir » ses décision (Bogacz *et al.*, 2007, p.3)

oublieux de l'ensemble des gains obtenus dans le passé, incite à adopter une vision cumulée et prospective sur la tâche, et ainsi à évaluer une optimum pour toute la séquence, et non plus pour la seule décision présente (Bogacz *et al.*, 2007, p.2).

Bogacz *et al.* construisent, à l'appui de cette idée selon laquelle la plus ou moins grande mémoire des récompenses passées détermine l'horizon temporel du choix, un modèle descriptif du comportement, appelé modèle à traces d'éligibilité. Le modèle permet de rendre compte des choix observés dans le *rising optimum task* (Bogacz, McClure, Li, Cohen et Montague, 2007, p.7). Il s'appuie sur un algorithme standard d'apprentissage par différence temporel, corrigé d'un paramètre d'oubli, représentant la diminution constante, au cours du temps, du poids des récompenses passées dans l'évaluation des actions présentes.

Le modèle de Bogacz, McClure, Li, Cohen et Montague suggère que les capacités de *self control* et de planification temporelle résultent d'abord d'une capacité à garder une trace du résultat des actions passées. Il est en lien direct avec la théorie des marqueurs somatiques de Damasio, qui sera abordée dans la troisième section, et selon laquelle les expériences émotionnelles passées fournissent des marqueurs somatiques, c'est-à-dire des points d'ancrage pour l'action et l'évaluation des choix. Il importe de souligner ici que ces réflexions sur la planification inter-temporelle ou l' « intelligence au second degré » (*cf.* chapitre 2, section III) peut très bien s'intégrer dans le cadre analytique du *reward learning*, comme l'illustre le modèle de Bogacz *et al.*. Si, par conséquent, l'hypothèse d'une monnaie neuronale commune n'empêche nullement la prise en compte d'une hiérarchie entre les processus d'évaluation dans le cerveau, les économistes comportementalistes ont au contraire cherché à rompre avec ce cadre d'analyse, en insistant plutôt sur la dualité des processus cognitifs dans le cerveau.

II. La tutelle de l'économie comportementale ou « l'économie dans le scanner »

Les premiers développements théoriques de la neuroéconomie entre 2000 et 2002 s'inscrivent dans le prolongement direct du programme de recherches néo-comportementaliste portant sur le *reward learning*. Assez rapidement, des économistes comportementaux s'approprient les résultats expérimentaux décrits précédemment, pour en fournir leur propre interprétation. Certains d'entre eux, convertis aux neurosciences, participent directement à l'élaboration d'expériences de neuroimagerie, en collaboration avec des neurobiologistes[88]. Le contenu théorique de la neuroéconomie subit ainsi des effets de distorsion lors de son assimilation au sein des *behavioral economics*,. Ce que Ross appelle ainsi « *l'économie comportementale dans le scanner* » (Ross, 2008) a soulevé par la suite de nombreuses critiques. Les économistes comportementaux tendent en effet à se servir de l'IRM_f comme un instrument de confirmation, au niveau neuronal, de leurs modèles. Cet usage particulier de la neuroimagerie aboutit à l'élaboration de sous-domaines de recherche ayant vocation à mettre en évidence des théories comportementales pré-existantes « dans le cerveau » (A). A partir de 2005, ces travaux donnent naissance aux premiers manifestes de la neuroéconomie, qui sont donc caractérisés par une large influence de l'économie comportementale (B)

A. L'imagerie fonctionnelle comme instrument de vérification de l'économie comportementale

La neuroimagerie suscite un enthousiasme certain au cours des années 2000. Plusieurs économistes comportementalistes décident alors de collaborer avec des neurobiologistes. Bien que se voulant fidèles aux données des expériences, les économistes convertis aux neurosciences ne se bornent pas à importer au sein de leur discipline les acquis de la théorie du *reward learning*, mais traduisent également ses principaux résultats dans les termes de

88 En France, Giorgio Coricelli est sans doute le meilleur exemple d'économiste converti, puisqu'il est le premier économiste de formation ayant été recruté comme chercheur par un laboratoire de recherches en neurosciences, à l'Institut des Sciences Cognitives de Lyon.

leur propre approche. Cette interprétation tout à fait particulière vise à faire de L'IRM$_f$ un instrument de vérification, au niveau neuronal, des principales théories proposées par les *behavioral economics*. Il s'agit ici d'étudier plusieurs de ces tentatives, qui concernent respectivement l'actualisation quasi-hyperbolique (1), la théorie de la punition altruiste (2), la théorie des regrets (3) et la théorie des perspectives (4).

1. L'actualisation quasi-hyperbolique dans le cerveau

Les algorithmes d'apprentissage par différence temporelle (*TD learning*) incorporent un coefficient d'actualisation temporelle. La mise en évidence à l'IRM$_f$, de signaux d'erreur prédiction conformes à ces algorithmes, débouche donc naturellement sur l'étude du choix inter-temporel. Or, l'actualisation des récompenses en fonction du délai peut se comprendre de deux façons différentes. Le *discount* temporel peut d'abord être envisagé comme la conséquence d'un processus d'apprentissage lui-même étalé dans le temps. Cette conception néo-comportementaliste, qui correspond à celle du modèle à traces d'éligibilité de Bogacz *et al.*, 2007, suppose donc l'existence d'un signal de prédiction unique qui évalue à la fois les récompenses et leur délai d'obtention. A l'inverse, dans la perspective des modèles quasi-hyperboliques de l'économie comportementale, l'actualisation temporelle est envisagée de manière indépendante du problème de l'apprentissage. L'arbitrage inter-temporel porte sur des choix statiques, et met en scène une opposition entre deux systèmes d'évaluation, l'un portant sur les récompenses immédiates, et l'autre portant sur les récompenses à long-terme.

Les économistes comportementalistes intéressés par les neurosciences ont donc essayé de mettre en évidence, à l'aide de l'IRM$_f$ ce que l'on pourrait appeler une « théorie de l'actualisation quasi-hyperbolique dans le cerveau ». Les modèles d'actualisation quasi-hyperboliques, comme celui de Laibson (1997) (*cf.* chapitre 2) utilisent une fonction d'utilité ayant la forme suivante :

$$U(c_t,...c_T) = u(c_t) + \beta \sum_{k=t+1}^{T-t} \delta(t) + u(c_{t+k})$$

Ce type de modèle est également connu sous le nom d'actualisation temporelle « δ - β ». Les deux paramètres δ et β correspondent au poids relatif de chaque système d'évaluation. Le « système β » est le système orienté vers les récompenses immédiates. En effet, β, compris entre 0 et 1, désigne un biais en faveur du présent ou, plus précisément, un

biais en défaveur des utilités futures. Plus β est grand plus l'individu aura tendance à surestimer l'importance des récompenses disponibles immédiatement. Le « système δ » est le système orienté vers le futur, $\delta(t)$ correspondant à la fonction d'actualisation exponentielle, proportionnelle au délai, du modèle de Samuelson.

Les expériences décrites dans cette section visent précisément à mettre en évidence des régions cérébrales spécialisées dans chacun des deux systèmes d'évaluation. Une première étude, publiée en 2004, est réalisée par un neuroscientifique qui a joué un rôle important dans les recherches portant sur le *reward learning* chez l'homme, Samuel McClure. Celui-ci s'associe ici à deux économistes comportementalistes, dont notamment Donald Laibson, auteur du modèle d'actualisation quasi-hyperbolique précédemment cité.

Le titre de l'étude est explicite: « des systèmes neuronaux distincts encodent les valeurs respectives des récompenses immédiates et avec délai » (McClure, Laibson, Loewenstein et Cohen, 2004). Le protocole est inspiré directement des études classiques du choix inter-temporel en économie comportementale: on demande aux sujets d'effectuer des arbitrages entre différents montants d'argent (ici, 5 à 40 dollars), disponibles à différents délais (ici, compris entre 0 et 6 semaines).

Les résultats comportementaux révèlent, sans surprise, que les sujets n'actualisent pas les récompenses de manière proportionnelle au délai. Les résultats de l'imagerie sont doubles. D'une part, une dissociation apparaît au sein des régions appartenant au circuit de la récompense. Certaines zones -le *striatum* ventral en particulier, et plus généralement le système limbique- s'activent de manière beaucoup plus importante lorsque le choix inclut une option de récompense immédiate. Les autres régions sont actives uniformément pour tous les choix. Surtout, lorsque les sujets choisissent la récompense avec le délai le plus long, une activité beaucoup plus forte est observée dans le cortex préfrontal (latéral et postérieur).

L'association respective des systèmes δ et β au cortex préfrontal et au système limbique constitue un résultat solide, puisque l'étude de 2004 a été reproduite en 2007 par McClure *et al*. avec des récompenses alimentaires. D'une manière similaire aux expériences décrites dans le cadre de la première section, les sujets reçoivent ici des récompenses sous forme de goutes d'eau ou de jus de fruit. Le nombre de gouttes détermine le montant de la récompense. Les délais, beaucoup plus courts, varient de 0 à 25 minutes. Les participants doivent effectuer des arbitrages temporels et reçoivent ensuite les récompenses choisies en conséquence.

Les résultats observés sont similaires à ceux de l'étude précédente: les sujets actualisent les récompenses de manière non-proportionnelle au délai, et une augmentation significative de l'activité du cortex pré-frontal est associée aux choix en faveur des récompenses à long-terme. Selon les auteurs de ces études, les travaux de neuroimagerie montrent donc l'existence de deux systèmes neuronaux d'évaluation distincts. L'arbitrage inter-temporel révèle ainsi une « *compétition entre la cigale limbique impulsive et la fourmi préfrontale plus prévoyante* »(McCLure, Laibson, Loewenstein et Cohen, 2004). Or, c'est chez l'homme que les régions préfrontales, associées au système δ, sont les plus développées. Par conséquent, la capacité à différer l'obtention d'une récompense serait spécifiquement humaine:

> « *il y a un large écart entre l'actualisation temporelle chez l'homme et chez les animaux. Les êtres humains effectuent de manière routinière des arbitrages entre des coûts et des bénéfices immédiats, et des coûts et bénéfices futurs, avec des délais de plusieurs dizaines d'année. A l'inverse, même les primates les plus avancés, dont la taille du cortex préfrontal est néanmoins largement inférieure, ne semblent pas supporter des périodes d'attente non-programmées dans la gratification supérieures à quelques minutes. Même si certains animaux sont capables d'effectuer des arbitrages temporels sur des longues périodes (par exemple, en stockant de la nourriture pour l'hiver), un tel comportement semble invariablement être instinctif, à l'inverse de la nature plus générale de la planification chez l'homme. Par ailleurs, l'ensemble des études portant sur les lésions du cerveau […] convergent pour montrer que les lésions du cortex préfrontal ont pour effet d'accentuer l'influence des récompenses immédiates, ainsi que de dégrader la capacité à planifier les décisions* » (McClure, Laibson, Loewenstein et Cohen, 2004, 504).

McClure et ses coauteurs mettent donc ici en avant la spécificité des capacités humaines de planification. Le développement beaucoup plus important des régions frontales chez l'homme serait à l'origine d'une différence de nature entre les comportements humains et animaux. Cette perspective accorde à l'homme un privilège sur l'animal. Elle s'oppose nettement à celle du néo-comportementalisme qui vise au contraire à souligner la proximité des modes de cognition humain et animal[89]. Ceci explique pourquoi l'idée d'une opposition

89 A l'inverse, le modèle de Bogacz *et al.* (2007) peut s'appliquer à la fois à l'homme et à l'animal, dans les tâches à optimum croissant. Dans la perspective retenue par Bogacz *et al.*, la plus ou moins grande capacité de planification temporelle dépend des capacités de mémorisation. Bien évidement, on peut supposer que celles-ci sont plus élevées chez l'homme que chez le pigeons; mais cette conception maintient une continuité quantitative entre l'espèce humaine et les animaux, à l'inverse de McClure *et al.* qui considèrent ici une différence de nature et une spécificité radicale de la cognition humaine. On peut observer que, paradoxalement, Samuel McClure a participé également à l'étude de Bogacz *et al.* (2007). La position de McClure est de notre point de vue tout à fait significative. En effet, Samuel McClure est un neurobiologiste, qui a participé aux premières expériences de la neuroéconomie, dans le droit prolongement du programme néo-comportementaliste; puis, ses collaborations avec des économistes l'ont amené à soutenir des interprétations plus proches de l'économie comportementale. Cette trajectoire théorique n'est pas nécessairement incohérente et illustre notre interprétation du développement de la neuroéconomie.

entre deux modes d'évaluation, l'un affectif et instinctif, l'autre rationnel et réfléchi, a eu un grand succès dans l'étude de comportements traditionnellement considérés comme spécifiquement humain, parce que relevant d'un raisonnement interpersonnel.

b. La théorie de la punition altruiste dans le cerveau

Les deux expériences sur le choix inter-temporel de McClure *et al.* en 2004 et 2007 ont joué un rôle important dans la littérature, en popularisant l'idée d'une opposition entre deux systèmes cérébraux d'évaluation. L'hypothèse avancée par McClure et ses coauteurs selon laquelle le cortex préfrontal exercerait une fonction d'inhibition sur les inclinations dirigées vers des satisfactions immédiates trouve également à s'appliquer à l'étude du choix inter-personnel. La neuroimagerie a ici été utilisée pour corroborer une théorie préexistante de la coopération et de l'altruisme, connue sous le nom de « *théorie de la punition altruiste* » (*altruistic punishment theory*). Ce sous-domaine de l'« économie comportementale dans le scanner » (Ross, 2008) est aussi le plus controversé. Il fera l'objet d'un développement spécifique dans le sixième chapitre. Nous nous contenterons donc ici simplement d'en indiquer les grandes lignes théoriques.

La théorie de la punition altruiste a été proposée par des économistes comportementalistes (Fehr and Gächter, 2002). Elle postule que le comportement coopératif résulte de l'anticipation de sanctions qui sont généralement associées, dans les sociétés humaines, à la violation des normes sociales. La coopération s'explique ainsi par l'internalisation des normes (Gintis, 2003). Une telle perspective déplace le problème de la coopération vers celui de l'origine de la sanction: pourquoi certains individus décident-ils, en premier lieu, de punir le non-respect des normes, sachent que cette punition a un coût? Dans la théorie de la punition altruiste, la sanction devient ainsi une sorte de bien public de deuxième ordre (Fehr and Gächter, 2002, p.137).

Le succès de cette théorie en neurosciences provient de la découverte selon laquelle les régions du cerveau impliquées dans l'évaluation des récompenses (le striatum, en particulier) s'activent lorsque les sujets, dans des jeux de coopération, adoptent des « punitions altruistes », c'est-à-dire des punitions des individus non-coopératifs à un coût personnel (de Quervain *et al.*, 2004; Fehr and Camerer, 2007; Fehr, 2008).

A un second niveau, des travaux de neuroimagerie ont montré que l'activité du cortex

préfrontal latéral permettait d'inhiber cette tendance naturelle à punir les individus égoïstes (Spitzer *et al.*, 2007, Knoch *et al.*,, 2006). La dissociation, au sein du circuit de la récompense, entre deux systèmes d'évaluation distincts permet une nouvelle fois de proposer une explication dualiste: ici, le système limbique serait responsable d'une inclination instinctive à punir les individus non-coopératifs (et à respecter les normes); les régions préfrontales étant à l'origine de stratégies sociales plus complexes, impliquant ce que les chercheurs appellent une « *intelligence machiavélique* »(*cf.* Fehr, 2008). L'IRM_f semble donc confirmer une théorie comportementale préexistante.

Il est pourtant clair que l'utilisation de la neuroimagerie sur des sujets humains ne se borne pas dans ce domaine à valider un paradigme issu de l'économie comportementale, mais en bouleverse profondément les cadres interprétatifs. La neurobiologie propose notamment une explication aux comportements punitifs, dont l'origine demeurait inexpliquée sur le plan comportemental: la punition altruiste serait source de satisfaction.

Les neurosciences soulèvent aussi de nouvelles questions. En particulier, l'utilisation, dans ces jeux coopératifs, de sujets atteints de lésions cérébrales ou d'individus autistes, psychopathes *etc.* pose le problème de la norme de rationalité à l'œuvre dans ces comportements. En effet, d'un côté, il semble que la rationalité économique dans ces jeux est une rationalité « machiavélique », égoïste et calculatrice, qui suppose l'exercice de facultés spécifiquement humaines, associées au cortex pré-frontal. Mais, de l'autre côté, les sujets manifestant diverses pathologies neuronales se distinguent précisément par des stratégies excessivement machiavéliques, par une incapacité à éprouver les émotions « normales » liées à l'interaction avec d'autres sujets. Or ces individus sont généralement considérés par les neurobiologistes comme ayant ce qu'ils appellent des troubles de la cognition sociale (*cf.* chapitre 6): faut-il donc considérer que ces sujets sont irrationnels, et que le comportement rationnel consiste à suivre son instinct, et donc à punir et coopérer? Comment concilier les perspectives médicales et économiques sur ces questions? Les économistes comportementalistes hésitent, sur ces problèmes, entre une interprétation prudente qui voit dans les données neuronales un simple soutien à leurs théories, et un approfondissement plus audacieux des perspectives ouvertes par les nouveaux instruments. Cette tension se manifeste dans les études par IRM_f sur le regret.

3. La théorie des regrets dans le cerveau

L'hypothèse d'un contrôle du système limbique par le cortex préfrontal a permis aussi d'éclairer les mécanismes cérébraux du regret. Depuis le modèle fondateur de Loomes et Sugden (1982), la notion de regret a fait l'objet de nombreux travaux en économie comportementale. Les modèles incorporant le regret permettent en effet d'expliquer des anomalies du type du paradoxe d'Allais. Des chercheurs ont tenté, à l'aide de l'imagerie, d'éclairer les soubassements neuronaux du regret.

Plusieurs études, en particulier celles réalisées par Giorgio Coricelli et son équipe (Camille *et al.*, 2004; Coricelli, 2005) indiquent que l'expérience du regret implique l'activité du cortex orbito-frontal[90]. Le protocole de ces expériences repose sur une succession de choix entre deux loteries plus ou moins risquées. Il est plus rentable, sur le long terme, de choisir toujours l'une des deux loteries, mais les sujets ignorent au départ la distribution des gains associés à chaque option. Après chaque choix, les résultats de chaque loterie sont affichés à l'écran. Cela permet ainsi de générer un sentiment de regret. Si l'individu constate que le gain associé à la loterie non-choisie est supérieur au gain obtenu, alors celui-ci peut émettre ce que les psychologues appellent un « jugement contra-factuel », c'est-à-dire qu'il peut juger de ce qui se serait passé si son choix avait été différent.

La plupart des participants se montrent généralement capables, à partir d'un certain nombre de répétitions, d'identifier la loterie la plus rentable. Ils maintiennent alors leurs choix en faveur de celle-ci. L'expérience du regret joue ici un rôle important, puisque c'est en constatant à chaque coup les écarts entre la stratégie suivie et la meilleure stratégie que les sujets peuvent progressivement se rapprocher tendanciellement du meilleur rendement global.

Ce type de jeu ressemble beaucoup aux protocoles utilisés pour étudier l'apprentissage de la récompense chez l'homme ou l'animal: il s'agit de choisir de manière répétée entre deux options générant des récompenses d'un montant variable, l'objectif étant de collecter de l'information au fur et à mesure afin d'identifier la meilleure stratégie. L'affichage simultanée des résultats de chaque option apporte toutefois une modification importante, en offrant l'opportunité aux participants de regretter leurs choix. En effet, le regret, au sens strict, se distingue de l'erreur de prédiction: celle-ci porte sur un écart entre une prédiction sur le gain futur au moment du choix et le gain effectivement obtenu. Dans le cas du regret, le gain obtenu n'est pas comparé avec sa prédiction, mais avec le gain « contrefactuel » qui aurait été obtenu en choisissant l'autre option Le regret diffère donc de la déception, qui naît de la constatation d'un écart entre un gain et des attentes de gains.

90 Pour une synthèse complète des études dans ce domaine, on se réfèrera à Coricelli, Dolan et Sirigu, 2007.

Cependant, rien n'empêche de considérer que le regret constitue une forme élargie ou sophistiquée d'erreur de la prédiction. Les études précédemment citées (Camille *et al.*, 2004; Coricelli, 2005) montrent que l'activité de cortex orbito-frontal est corrélée avec l'amplitude du regret possible mesurée dans les situations de choix défavorables, par l'écart entre gain obtenue et celui qui aurait pu être obtenu en choisissant l'autre machine à sous (ici, une loterie). Or cette activité s'observe au moment du choix suivant: le cortex orbitofrontal exerce ainsi une fonction d'enregistrement étendue des pertes, mais toujours dans l'objectif d'améliorer les prédictions suivantes. Cette fonction associée au cortex orbito-frontal dans le cadre des expériences sur le regret est donc assez proche de celle qui lui est attribuée dans le modèle à « traces d'éligibilité » étudié dans la section précédente (Bogascz *et al.*, 2007): le cortex frontal fournit, dans les processus d'apprentissage de la récompense, une mémoire étendue des pertes, qui permet aux sujets d'adopter une perspective plus globale sur le jeu, et ainsi d'échapper aux stratégies de mélioration qui relèvent de choix au coup par coup.

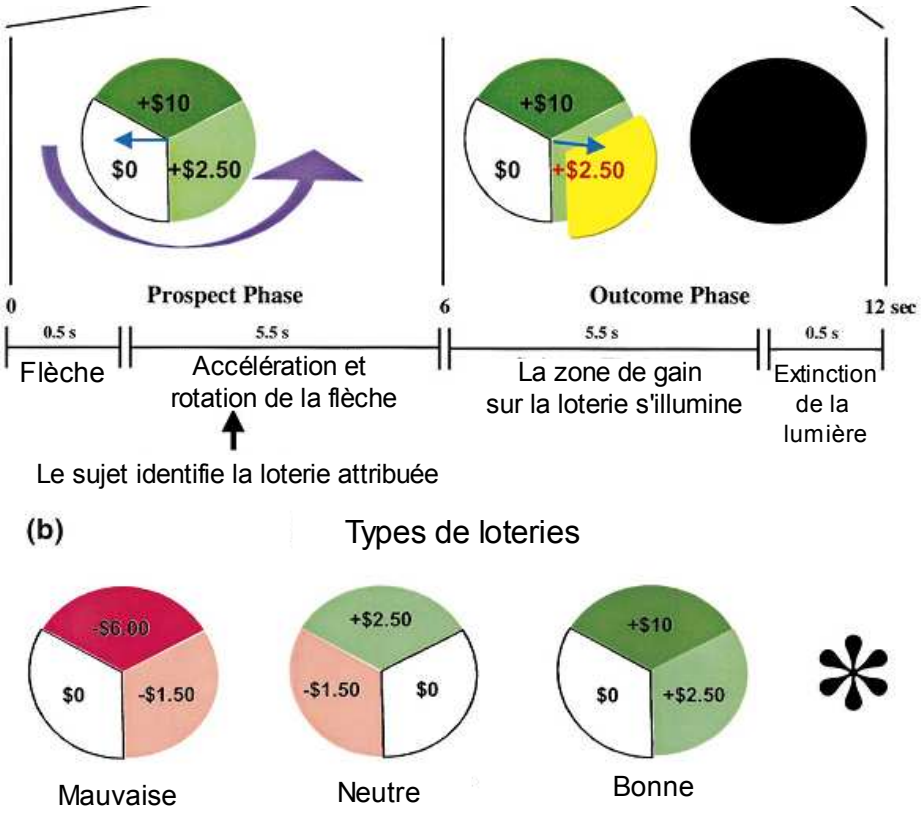
Les travaux sur le regret pourraient donc ainsi s'insérer dans le paradigme du *reward learning*, en considérant que le regret constitue une forme étendue d'erreur de prédiction. Les protocoles utilisés sont par ailleurs toujours du type *multi-armed bandit*. Néanmoins, plutôt que d'incorporer l'activité du cortex orbito-frontal au sein d'un signal unitaire d'évaluation des récompenses, les auteurs préfèrent souligner l'opposition entre deux types de raisonnement: l'un à court-terme, appuyé sur les régions du striatum, et l'autre contre-factuel, lié au cortex orbito-frontal (Camille *et al.*, 2004; Coricelli, 2005). Implicitement, ce type d'interprétation revient à abandonner la notion de monnaie neuronale commune pour accentuer une compétition supposée entre deux types distincts de processus neuronaux d'évaluation. L'idée d'un « contrôle *top-down* des émotions » (Camille *et al.*, 2004; Coricelli, 2005) par le cortex orbito-frontal modifie ainsi la signification d'expériences qui auraient pu être comprises comme un approfondissement du paradigme du *reward learning*.

4. La théorie des *prospects* dans le cerveau

Le dernier sous-domaine de l'« économie comportementale dans le scanner » (Ross, 2008) nait de tentatives diverses visant à confirmer, à l'échelle neuronale, les résultats principaux de la théorie des perspectives de Kahneman et Tversky (1979). Cette « théorie des perspectives dans le cerveau » témoigne donc d'un effort, de la part du programme de

recherche kahnemanien, pour investir le champ de la neurobiologie. Le ralliement de Daniel Kahneman aux neurosciences est effectif depuis sa participation, en 2001, à une expérience de neuroimagerie sur des sujets humains. Par la suite, d'autres études ont été réalisées pour vérifier les propositions principales de la *prospect theory* : l'aversion aux pertes, l'existence d'un point de référence dans les jugements probabilistes, la non-linéarité des jugements probabilistes. L'ensemble de ces travaux soulève *in fine* le même problème méthodologique: pourquoi utiliser la neuroimagerie, s'il s'agit simplement de confirmer une théorie préexistantes pouvant faire l'objet de tests comportementaux beaucoup plus simples? Cette difficulté se manifeste à travers l'ambiguïté avec laquelle les données neuronales sont utilisées ici, puisque les chercheurs paraissent hésiter à utiliser celles-ci comme d'un simple moyen de calibrage des paramètres du modèle de Kahneman, et leur attribuer au contraire une fonction théorique plus importante, en leur conférant un rôle explicatif à proprement parler.

En 2001, Daniel Kahneman participe et collabore à une expérience de neuroimagerie (Breiter *et al.*, 2001). Cette étude repose sur un protocole relativement proche de celui des premières expériences sur le *reward learning* chez l'homme (*cf.* section I). Placés dans le scanner, les sujets visualisent un écran sur lequel figurent trois loteries, qui sont représentées sous la forme d'un cercle, partagé en trois parties, qui correspondent chacune à des gains ou des pertes possibles. Les trois loteries sont respectivement « *bonne* » (gains de 0, 2,5 ou 10 dollars), « *neutre* » (gain nul, gain de 2,5 dollar, ou perte de 1,5 dollar) ou « *mauvaise* » (perte de 0 dollar, 1,5 dollar ou 6 dollars). Les sujets n'ont aucun choix à réaliser. Pendant la première phase, qualifiée de « *phase de perspective* » (*prospect phase*), on indique au sujet la loterie qui lui a été attribuée. La flèche indiquant le gain, située au centre du cercle, commence à tourner. L'arrêt de la flèche sur l'une des parties du cercle marque le début de la phase de résultat (*outcome phase*): le participant visualise la gain obtenu. L'expérience est répétée. On varie à chaque fois de manière aléatoire les loteries attribuées à chaque sujet. Le protocole comprend également des essais de contrôle pour bien différencier les activations au scanner. Pendant les essais de contrôle, les sujet ne visualisent qu'un point de fixation lumineux au centre de l'écran (Breiter *et al.*, 2001, p.620).



Breiter *et al.*, 2001, p.622

Les résultats convergent avec ceux des autres études portant sur l'apprentissage de la récompense. Les données fournies par l'IRM$_f$ montrent en effet l'activation spécifique, pendant les phases de perspective et de résultat, de régions du cerveau appartenant au circuit de la récompense,associées à l'évaluation de stimuli plaisants. En outre, cette activité est d'autant plus importante que le gain ou la perte possible (pendant la phase de perspective, selon la loterie attribuée) ou que le gain ou la perte réalisé (pendant la phase de résultat, selon l'issue de la loterie) est élevé. Cela suggère donc que ces régions encodent, pendant la phase de perspective, un signal de prédiction de la récompense, correspondant à l'espérance du gain ou de la perte pouvant être obtenu, et, pendant la phase de résultat, un signal d'erreur de prédiction, correspondant à l'écart entre la prédiction et le gain ou la perte effectif. Les expérimentateurs concluent ainsi à l'existence d'« *un circuit commun, généralisé à l'évaluation de toutes les catégories de récompense* » (Breiter *et al.*, 2001, p.627), qui se rapproche de la notion de monnaie neuronale commune proposée par Berns et Montague (2002, *cf. supra*).

Cependant, l'interprétation fournie ici par les expérimentateurs se caractérise aussi par une certaine coloration kahnemanienne, qui se distingue du cadre théorique du *reward*

learning. Deux résultats spécifiques sont compris en lien avec la *prospect theory*. Tout d'abord, l'activation, dans les deux phases, est symétrique pour les gains et les pertes, alors que les montant absolus de pertes sont pourtant plus faibles que les montants de gains (*cf.* figure ci-dessus). Cela indique donc que les pertes ont un impact plus fort que les gains. En outre, les résultats comportementaux montrent que les sujets ne parviennent pas à adopter un point de vue global sur la tâche et ne conservent pas en mémoire la somme cumulée de leurs gains et pertes. Cette tendance est d'une certaine manière confirmée sur la plan neuronal, puisque le signal d'erreur de prédiction pendant la phase de résultat est corrélé avec l'écart entre la prédiction et le gain ou la perte immédiatement obtenue : la correction s'effectue au coup par coup, sans évaluation de la position nette globale. Ce mode d'estimation, par prédictions et corrections successives, implique donc qu'une perte ou un gain donné soit apprécié *ex post* par rapport à un point de référence, qui est fourni par la prédiction ex ante. Une perte de 1,5 dollar ne produit pas toujours un signal de prédiction de la même ampleur : dans la loterie intermédiaire, cette perte sera considérée comme malchanceuse et sera source de déception, alors qu'elle sera envisagée comme normale dans la loterie « mauvaise » et ne fera pas apparaître d'écart par rapport aux attentes.

Les auteurs considèrent donc que deux aspects importants de cette expérience sont « *déduits de la théorie des prospects […]. Selon le premier principe, l'évaluation d'un actif risqué, comme une loterie, ne dépend que très faiblement de la somme cumulée des gains et des pertes (la position "patrimoniale"* [asset position]. *Cette évaluation est plutôt envisagée comme un gain ou une perte par rapport à un point neutre […]. Le second principe postule que l'impact d'une perte excède celui d'un gain de même ampleur* » (Breiter *et al.*, 2001, p.620).

Sans anticiper sur les critiques de l'« économie comportementale dans le scanner » qui seront envisagées dans le prochain chapitre, il convient de souligner que cette tentative de rapprochement entre la *prospect theory* et le paradigme du *reward learning* pose ici plusieurs problèmes. Tout d'abord, la correspondance établie par les expérimentateurs entre l'erreur de prédiction et la notion de point de référence est assez contestable. Chez Kahneman, l'existence d'un point de référence est comprise comme une déformation ou un biais dans les jugements probabilistes : les individus évaluent des loteries en fonction d'une position qu'il juge neutre, en rapport avec leurs attentes. Mais la *prospect theory* ne fournit pas d'explication de l'origine de ce point de référence (Kahneman et Tversky, 1979), qui varie selon la présentation des alternatives aux sujets. Or, les expériences sur l'apprentissage de la récompense ont

précisément pour objectif d'étudier les processus de formation des prédictions, et donc d'expliquer comment les attentes de gains ou de pertes naissent de dynamiques d'acquisition d'information. En d'autres termes, le point de référence est dans le cadre de la *prospect theory* un paramètre comportemental exogène qu'il est nécessaire d'estimer pour « calibrer » le modèle. La modélisation du *reward learning vise* au contraire à faire du point de référence une variable endogène, en montrant comment les attentes de gains et de pertes sont progressivement corrigées par les individus.

Le principe, également hérité des travaux de Kahneman, selon lequel les individus ne parviennent pas à adopter de position globale sur leurs gains et pertes cumulés soulève aussi des difficultés. En effet, dans les études portant sur le regret (*cf. supra*) ou dans les tâches à optimum croissant (*cf.* Bogasz *et al.*, 2007), les sujets parviennent à adopter une perspective temporelle élargie, en conservant en mémoire les pertes passées, et en évaluant les conséquences à long terme de leurs actions. Cette capacité à raisonner sur des séquences de choix plutôt que sur des récompenses prises isolément engage l'activité de zones situées dans le cortex pré-frontal. Dans l'expérience à laquelle participe Kahneman, il est raisonnable de supposer que les individus n'adoptent pas de perspective globale sur leurs gains et pertes parce qu'ils reçoivent les récompenses de manière passive, sans pouvoir choisir entre les loteries qui leur sont attribuées. Par conséquent, ce protocole n'est pas susceptible de générer des stratégies d'apprentissage plus sophistiquées. Si, au lieu d'attribuer aléatoirement une des trois loteries, un véritable choix est proposé entre des machines à sous dont le risque et l'espérance de gain sont initialement inconnues, les sujets vont vraisemblablement exercer des efforts de mémorisation des choix passés plus importants. Le premier principe d'inspiration kahnemanienne, auquel les expérimentateur font référence ici, n'est donc pas valable pour tous les protocoles et tous les individus.

Il ne s'agit pas de dire ici que les expériences de neuroimagerie réfutent la théorie des *prospects*. Force est de constater néanmoins que l'application d'une grille d'interprétation kahnemanienne aux problèmes type *multi-armed bandit* fonctionne assez mal (*cf.* introduction à la première partie). Bien que se chevauchant sur certaines idées -notamment celle selon laquelle les évaluations individuelles sont construites à partir d'attentes subjectives ou de point d'ancrage-, les approches inspirées de Kahneman et de la *prospect theory* d'un côté, et celle qui est issue du néo-comportementalisme et de l'étude du *reward learning* de l'autre côté ne se superposent pas exactement.

Kahneman et ses coauteurs n'ont pas, toutefois, exactement cherché dans cette étude à interpréter les résultats à l'aide de la théorie des *prospects*, puisqu'ils affirment plus

modestement s'être servis de cette théorie pour la conception du protocole: « *deux principes psychologiques dérivés de la prospect theory ont inspiré la conception de ce protocole* ». Cependant, en neurobiologie, et plus généralement dans toutes les disciplines expérimentales, la conception du protocole a un rôle théorique considérable, puisqu'elle oriente vers des problèmes à résoudre, et limite l'éventail des interprétations possibles. C'est la raison pour laquelle, en neurosciences, le terme de « paradigme » inclut des manières de formaliser les données expérimentales mais aussi, et peut être avant tout, des manières de construire les tâches expérimentales. C'est ainsi que ce qui est appelé ici « paradigme du *reward learning* » regroupe à la fois des algorithmes permettant de formaliser les données, et des protocoles expérimentaux permettant d'observer l'apprentissage chez le pigeon, le singe ou l'animal, du type *multi armed bandit problems*. Si les auteurs considèrent que la théorie de Kahneman a joué un rôle important dans la conception du protocole, cela signifie donc que cette théorie joue le rôle dans cette étude d'un « paradigme » (expérimental), ce qui apparaît d'ailleurs assez clairement dans la conclusion: « *le but de cette étude par neuroimagerie était de répertorier les réponses hémodynamiques à l'anticipation et l'expérience de gains et de pertes monétaires à l'intérieur d'un paradigme appuyé sur des principes psychologiques bien établis concernant l'anticipation et le choix en incertitude (Kahneman and Tversky, 1979)*» (Breiter *et al.*, 2001, p.620).

A la suite de la collaboration entre Kahneman et des neurobiologistes, plusieurs équipes de recherches ont tenté d'utiliser la théorie des *prospects* comme un paradigme expérimental en neurosciences. Cet ensemble de travaux a fait l'objet d'une synthèse par Fox et Poldrack (Fox et Poldrack, 2008). Tout en faisant état des découvertes importantes des neurosciences dans le domaine, cette étude rend compte aussi des difficultés liées au croisement des perspectives issues respectivement du *reward learning* et de la *prospect theory*. Ces difficultés sont liées au rôle ambigu attribué aux données neuronales dans ces expériences, puisqu'elles sont tour à tour envisagées comme simples paramètres du modèle de Kahneman, mais aussi comme variables explicatives.

Fox et Poldrack affirment au départ que la neuroimagerie remplit une fonction de calibrage du modèle de décision de Kahneman. En effet, la théorie des *prospects* suppose,, doit être « calibrée » pour être descriptive, c'est-à-dire que les divers paramètres du modèle doivent être estimées pour chaque individu. Pour cela, les économistes comportementalistes observent les choix de leurs sujets entre différentes loteries pour en dériver les paramètres du modèle. Ils utilisent le plus souvent la fonction de pondération des probabilités proposée par Prelec (1998) . Plusieurs méthodes de calibrage comportemental peuvent être utilisées mais

Fox et Poldrack constatent une forte hétérogénéité des résultats produits par chacune des méthodes d'estimation. Pour Fox et Poldrack, les données neuronales permettraient de fournir un nouveau type de calibrage du modèle, plus robuste (Fox et Poldrack, 2008, p.147).

Pourtant, ce programme n'est pas respecté dans la suite de l'article. En effet, il est difficile de comprendre en quoi les données de la neuroimagerie fourniraient des estimations paramètriques plus robustes que les données comportementales, puisque toutes deux dépendent du choix d'une méthode d'inférence similaire sur le plan comportemental. Au mieux est-il possible d'établir une corrélation entre des données comportementales et des données neuronales, ces dernières étant également valides uniquement pour un certain type de protocole expérimental. Les exemples de calibrage de la fonction d'utilité par la neuroimagerie avancés par Fox et Poldrack reposent ainsi sur de simples corrélations entre les choix et les activations de régions associées au circuit de la récompense. L'expérience de Hsu *et al.* (2009) s'appuie ainsi sur un protocole dans lequel les sujets choisissent entre des loteries de montants et de risque variables. Les auteurs en dérivent les paramètres de pondération des probabilités en utilisant la fonction de Prelec (1998). L'analyse des données produites par l'IRM_f montre une corrélation entre l'activité du striatum et le paramètre de non-linéarité estimé à partir des choix. Ici, les données neuronales sont donc utilisées pour confirmer une tendance comportementale, plutôt que de servir de nouvelle méthode de calibrage à proprement parler.

Les illustrations expérimentales founries par Fox et Poldrack laissent donc à penser que la neuroimagerie chez l'homme n'a pas vraiment pour vocation de constituer un nouvel instrument d'estimation des paramètres de la *prospect theory.* Elle établirait plutôt des convergences entre neurosciences et *prospect theory*, à partir de corrélation comportementales et neuronales. Fox et Poldrack suggèrent ainsi par la suite que la neurobiologie « *confirme* » à l'échelle neuronale l'existence d'un point de référence à l'appui notamment des expériences de de Martino *et al.*, 2006 et Windmann *et al.*, 2006 (Fox et Poldrack, 2008, p.155). Toutefois, même cette fonction plus modeste de confirmation ne va pas sans poser problème. L'application d'un paradigme expérimental directement inspiré de l'approche kahnemanienne produit des résultats contrastés, en particulier dans l'étude de l'aversion aux pertes. Weber *et al.* (2007) montrent, dans le cadre d'un protocole classique en économie comportementale, que l'asymétrie observée entre volonté d'acheter et volonté de vendre est corrélée avec l'activation de l'amygdale. Cette région est associée à l'expérience d'émotions négatives. Toutefois, ce résultat est problématique, car, comme le soulignent Fox et Poldrack (2008, p.155), d'autres études de neuroimagerie montrent également que l'amygdale encode aussi des

valeurs positives (Xue, Ghahremani and Poldrack, 2008). Par ailleurs, dans l'expérience de Weber *et al.*, il y a une incertitude quant au fait de savoir si l'activation de l'amygdale reflète une diminution d'utilité liée à la vente de l'objet, ou un gain lié à l'obtention d'une récompense monétaire en retour.

Ce genre d'étude met donc en évidence que les protocoles issus de l'économie comportementale ne fonctionnent pas toujours très bien en neurosciences, et peuvent soulever des difficultés interprétatives qui n'apparaissaient pas sur le plan comportemental. Fox et Poldrack assignent donc finalement à la neuroimagerie un rôle théorique bien différent de celui qui apparaissait au départ: selon eux, au lieu de simplement refléter des variables comportementales, les activations observées du cerveau peuvent fournir des variables explicatives de ces tendances comportementales: « *il sera important de déterminer à l'avenir les régions qui sont impliquées causalement dans ces distorsions (plutôt que de simplement refléter ces distorsions) en étudiant notamment les sujets avec des lésions ou des troubles mentaux. Si l'absence de linéarité dans la pondération des probabilités est le produit d'un système cérébral spécifique, alors il doit être possible de trouver des sujets dont les pondérations ne sont plus linéaires à la suite de lésions spécifiques dans ces régions* » (Fox et Poldrack, 2008, p.168).

Comme le soulignent Fox et Poldrack, les expériences de neuroimagerie sur sujets humains, appuyées sur le paradigme de l'apprentissage de la récompense, dépassent le cadre strict de la théorie des *prospects*. Les schémas d'interprétation de type kahnemanien posent des difficultés en neurobiologie, car ceux-ci prennent certaines observations comportementales (aversion aux pertes, existence d'un point de référence, etc.) comme des preuves d'un écart existant entre la rationalité économique « idéale » -associée à la maximisation d'une fonction d'utilité objective- et le comportement réel des individus. Or, dans l'approche néo-comportementaliste, la maximisation de l'utilité espérée n'est pas fausse ou inadéquate, mais elle est inopérante:c'est en référence à la dynamique de l'apprentissage que le comportement peut ou non être qualifié de rationnel. Toutefois, si Fox et Poldrack reconnaissent dans leur article certaines de ces difficultés et se limitent à signaler quelque points de rencontres locaux entre les deux approches, l'idée d'une « théorie des prospects dans le cerveau » a joui d'un succès certain dans la littérature. L'influence kahnemanienne sur la neuroéconomie à ses débuts se donne à voir en particulier dans les premiers manifestes de cette jeune discipline.

2. Les modèles duaux et les premiers manifestes de la neuroéconomie, entre simplification descriptive et rhétorique anti-économique

Dans la première partie des années 2000, le paradigme expérimental du *reward learning* perd progressivement son hégémonie au sein de la neurobiologie à « vocation économique » (*cf.* chapitre 3), au profit d'autres cadres théoriques hérités de l'économie comportementale. La mise sous tutelle de la neuroéconomie par les *behavioral economics* aboutit notamment à la promotion de modèles duaux, qui remettent en question la notion de monnaie neuronale commune (Berns et Montague, 2002). L'assimilation des études de neuroimagerie au sein de l'économie comportementale fait donc apparaître ce que John Davis appelle des « *biais de sélection* » (Davis, 2008, p.363)[91], dans la mesure où les économistes réduisent l'apport des neurosciences à l'idée d'un conflit entre raison et émotions alors que ce programme de recherche s'est développé à partir d'un paradigme d'apprentissage plus large, applicable à la fois à l'homme et à l'animal.

Il ne s'agit pas, bien sûr, d'affirmer ici que les économistes comportementaux ont réellement déformé les résultats théoriques des neurosciences, en truquant par exemple les données expérimentales. Les modèles duaux reçoivent bien un appui dans certaines expériences qui montrent par exemple que des régions duc cortex préfrontal remplissent des fonctions de *self control* et d'inhibition du système limbique (*cf. supra*). La déformation porte plutôt sur la conception des protocoles, mais aussi sur la manière de présenter les résultats, de les discuter et de les interpréter. Ce biais de sélection apparaît en particulier dans les articles de synthèse qui sont publiés au milieu des années 2000. Un panorama rapide de ces premiers manifestes de l'économie comportementale dans le scanner met en évidence deux éléments caractéristiques: une rhétorique hétérodoxe d'un part, dirigée contre ce qui est considéré comme relevant de l'« économie standard », et la mise en avant d'une dualité des systèmes cognitifs humains d'autre part.

Ce type de positionnement est notamment celui de Georges Loewenstein, qui publie en 2000 un article intitulé « les émotions dans la théorie économique et dans les comportements économiques » (Loewenstein, 2000). Loewenstein est un économiste comportementaliste

91 Pour Davis, cette idée de bais de sélection s'applique aux relations entre économie comportementale et psychologie: « *L'économie comportementale - un programme de recherche en économie, non en psychologie – emploie des importations en provenance de la psychologie, mais les recadre selon des préoccupations économiques* » (Davis, 2008, p. 363). On peut faire valoir que des biais similaires apparaissent dans les relations entre économie comportementale et neurosciences.

important, qui, au sein des *behavioral economics*, est l'un des chercheurs les moins influencés par l'approche kahnemanienne de la décision (*cf.* Heukelom, 2009). Il s'est intéressé très tôt à la psychologie néo-comportementale (*cf.* chapitre 2), puis aux neurosciences. Dans cet article, il défend l'idée selon laquelle l'économie dite standard aurait négligé l'importance des émotions ou de ce qu'il appelle « *facteurs viscéraux* » (*visceral factors*). Pour Loewenstein, les données empiriques indiquent que la rationalité délibérative et calculatrice des « *économistes* » ne joue qu'un faible rôle face aux affects et aux émotions. Loewenstein retient ainsi comme apport significatif des expériences de psychologie la notion d'émotion, dont il juge qu'elle a un caractère potentiellement révolutionnaire pour la théorie économique (Loewenstein, 2000, p.425-428).

Cette relecture de l'histoire de la pensée économique par Loewenstein pourrait bien sûr être discutée, mais force est de constater que cette lecture des neurosciences à partir de la notion d'émotion, a eu un grand succès dans les premières heures de la neuroéconomie. En 2004, Glimcher publie en collaboration un économiste, Aldo Rustichini, un autre article-manifeste intitulé « la neuroéconomie: la convergence du cerveau et de la décision » (Glimcher et Rustichini, 2004). Les deux auteurs estiment que la théorie économique doit être reconstruite à partir des neurosciences. Il faut néanmoins souligner que le positionnement de Glimcher y est ici beaucoup plus fidèle au paradigme du *reward learning* que celui de Loewenstein. Pour Glimcher et Rustichini, l'économie conserve un rôle en tant que cadre théorique: les individus sont considérés comme choisissant entre différentes options sur la base de la désirabilité relative de chaque option. Néanmoins, le formalisme économique jusqu'alors utilisé pour représenter ces processus mentaux doit être abandonné au profit d'algorithmes d'apprentissage, et une explication neuronale de la formation de cette désirabilité doit être fournie (*cf.* chapitre 3, Glimcher et Rustichini, 2004, p.452).

Glimcher et Rustichini admettent donc un partage des tâches entre l'économie et les neurosciences, plutôt que de promouvoir une « révolution » de la première par les secondes. Le positionnement plus prudent de ces deux auteurs, et plus fidèle au paradigme du *reward learning*, s'explique sans doute par la présence de Paul Glimcher, dont les rapports théoriques à l'économie sont assez ambivalents (*cf.* chapitre 3). Néanmoins, si l'économie fournit le cadre conceptuel général, Glimcher et Rustichini entendent bien modifier en profondeur les modèles utilisés par les économistes.

Une telle ambition de réforme de la théorie économique est tout à fait caractéristique du milieu des années 2000. L'article publié par Camerer, Loewenstein et Prelec en 2005 est

sans aucun doute la meilleure illustration de cette rhétorique résolument hétérodoxe, qui attribue aux neurosciences un potentiel révolutionnaire pour l'économie. Conformément à l'article de Loewenstein précédemment cité, les trois auteurs partent du constat d'une absence de prise en compte des émotions par la théorie économique (Camerer, Loewenstein et Prelec, 2005, p.10). Passant alors en revue un large nombre d'études de neuroimagerie, Camerer, Loewenstein et Prelec insistent alors sur la multiplicité des processus cognitifs mis en évidence par les neurosciences. Ils plaident en conclusion en faveur d'un changement « *radical* », plutôt qu'incrémental, de la théorie économique (Camerer, Loewenstein et Prelec, 2005, p.64) .

La rhétorique hétérodoxe chez Camerer, Loewenstein et Prelec s'appuie sur un enthousiasme résolu en faveur de la neuroimagerie (*cf.* introduction). Or, assez rapidement, l'utilisation de l'IRM_f en économie comportementale a soulevé de vigoureuses critiques méthodologiques qui en restreignirent sa portée théorique (*cf.* chapitre 5). En faisant état à la fois des avancées des neurosciences mais aussi de leurs limites, l'article de synthèse publié un an plus tard par Sanfey, McClure, Loewenstein et Cohen témoigne ainsi d'une ambition limitée, puisque les auteurs, économistes ou neurobiologistes n'entendent plus désormais révolutionner l'économie mais plus modestement améliorer la compréhension de la prise de décision (Sanfey *et al*., 2006). Plus généralement, comme le souligne avec justesse Mäki (2011, p.107), les économistes comportementalistes qui ont exprimé au départ un enthousiasme débordant en faveur des neurosciences, comme Colin Camerer notamment, ont progressivement modéré leurs positions.

Toutefois, même s'il ne s'agit plus de contester directement l'autorité de l'économie, mais de promouvoir « *un dialogue entre les deux disciplines* », Sanfey *et al.* (2006, p.108) continuent à voir dans la notion d'émotion l'apport décisif des neurosciences. En outre, même si la rhétorique s'est adoucie, et si les promesses de l'interdisciplinarité se substituent à celles de la révolution, il s'agit toujours de rhétorique: l'illusion consiste en effet à croire que la théorie économique peut statuer sur des protocoles décisionnels en neurobiologie qui relèvent pourtant de l'apprentissage de la récompense[92]. Dans leurs versions radicales ou plus

92 Cette erreur, volontaire ou non, est courante aussi bien en neuroéconomie, que dans sa préhistoire théorique comme on pu le voir notamment chez Glimcher (*cf.* chapitre 2). Elle consiste à supposer un dialogue possible entre la théorie économique de la décision, généralement associée de manière vague à la « théorie de l'utilité espérée », et les travaux en neurobiologie portant sur l'évaluation des récompense. La maximisation de l'utilité espérée serait censée fournir une norme pour les comportements biologiques, soit comme objectif effectivement atteint sur le long terme par des organismes évolutifs, soit comme objectif à atteindre, ou comme étalon de performance permettant de mesurer le degré d'efficacité des comportements réels: « *même s'il est admis que le cerveau (et par conséquent le comportement) ne fonctionne pas de manière parfaitement optimale, il y a néanmoins plusieurs raisons pour lesquelles une telle hypothèse est intéressante. Tout*

consensuelles, les manifestes de l'économie comportementale dans le scanner tendent donc à diminuer l'importance théorique du *reward learning*, en mettant en avant la dualité des processus cognitifs chez l'homme, supposément ignorée par les économistes. Cette lecture tout à fait particulière des travaux de neuroimagerie par l'économie comportementale est en fait largement influencée par le concept d'émotion, et en particulier par les travaux du neurobiologiste Antonio Damasio.

d'abord, si les formes complexes de comportement ne sont pas optimales, des mécanismes évolutifs plus simples peuvent plus facilement s'approcher de l'optimalité, ou au moins s'en être approchés au sein de l'environnement dans lequel ils ont évolué. En outre, une telle hypothèse d'optimalité permet de développer une théorie formelle, comme il est plus facile de définir et de caractériser précisément le comportement optimal d'un système. Cette théorie formelle, en retour, permet de générer des prédictions sur le comportement de ce système. Enfin, même lorsque le comportement (ou une fonction neuronale) se révèle sous-optimale, la définition de la performance optimale permet de fournir un étalon utile pour évaluer les comportements réels. L'identification des manières avec lesquelles le comportement dévie systématiquement de l'optimalité peut ainsi éclairer leurs mécanismes sous-jacents. L'utilisation du modèle d'utilité espérée est un exemple de cette approche, et a été appliqué de manière productive à la recherche portant sur les substrats neuronaux de la récompense et de la prise de décision » (Sanfey *et al.*, 2006, p.112). Le problème, ici, concerne l'absence de définition de ce que recouvre le terme d' « optimalité », qui semble associé à l'analyse économique, mais sans référence théorique, à l'exception de l'utilité espérée. Or on a montré, dans le chapitre 3, que la théorie économique ne permet pas de fournir de norme dans les problèmes d'apprentissage étudiés par les neurobiologistes; par conséquent, les deux disciplines n'ont pas vocation à dialoguer en la matière. Les travaux sur le *reward learning* doivent être conçus comme une nouvelle branche de la microéconomie à part entière (*cf.* chapitre 3).

III. La tutelle des neurosciences : Antonio Damasio et la théorie des marqueurs somatiques

Les recherches menées par Antonio Damasio et son équipe ont joué un rôle important dans le développement de la neuroéconomie. Cette jeune discipline a mobilisé à ses débuts l'appui des économistes comportementalistes, mais ceux-ci à leur tour se sont largement inspirés, en neurosciences, de la théorie des marqueurs somatiques de Damasio. Les modèles duaux élaborés par les différentes branches de l'économie comportementale dans le scanner renvoient, explicitement ou implicitement, à des conceptions damasiennes, en mettant en avant une dimension affective dans la prise de décision. Le ralliement de Damasio à la neuroéconomie est pourtant assez tardif. Celui-ci fait plutôt figure de référence externe au champ, plutôt que de neuroéconomiste à proprement parler. Damasio a incontestablement largement influencé la neuroéconomie, à travers sa théorie des marqueurs somatiques, et sa volonté de construire des protocoles de diagnostic en neuropsychiatrie inspirés de l'économie (A). Son opposition à la notion d'« *évaluation économique* » ou de monnaie générale commune (Berns et Montague, 2002, p.265) marque cependant une divergence avec les intentions théoriques des neuroéconomistes (B).

A. De la neuropsychiatrie à la théorie des jeux : les lésions du cortex préfrontal et l'*Iowa Gambling Task* (IGT)

Damasio a travaillé pendant 18 ans au département de neurologie de l'université de l'Iowa. Neurologue de formation, ses recherches sont d'abord orientées par des enjeux relatifs la médecine clinique et la thérapeutique. D'une manière similaire à Georges Ainslie (*cf.* chapitre 2), la question de la maladie mentale et de son traitement traverse l'œuvre de Damasio. Ce dernier a en particulier largement étudié les patient atteints de lésions du cortex préfrontal. Ces travaux cliniques ont permis de mieux comprendre la fonction cognitive de cette région du cerveau. Ils ont par la suite été une source d'inspiration importante pour les neuroéconomistes. En effet, Damasio a essayé de jeter un pont entre la clinique et l'économie

comportementale, en élaborant un protocole de diagnostic pour les patient lésés, inspiré[93] de la théorie des jeux.

Le succès des travaux de Damasio, dans le grand public aussi bien que dans le milieu académique, repose d'abord sur un talent certain dans la description de cas cliniques. Dans ses deux principaux ouvrages (Damasio, 1994, 2003), Damasio a su rendre compte avec finesse des traits de caractère des patients atteints de lésions du cortex préfrontal. Ces individus manifestent deux types de déficit, dans le raisonnement interpersonnel d'une part, et dans ce que Damasio appelle le « *comportement financier* » (Damasio, 2008, p.210). Ces deux déficits résultent de l'incapacité à prendre en compte la dimension temporelle de leurs actions. Les patients étudiés par Damasio ont en effet de grandes difficultés à suivre un programme ou un emploi du temps, mais aussi à tirer les leçons de leurs erreurs, à s'amender et se repentir. Sur le plan social, cette incapacité se traduit par l'irrespect des conventions et des règles de politesse, voire par de l'agressivité (Damasio, 1994, p.60).

L'explication des troubles du comportement observés chez les patients atteints de lésions dans les régions préfrontales est cependant délicate, car ceux-ci ne montrent aucun déficit dans les test classiques d'intelligence utilisés en neuropsychologie. Leurs capacités de mémorisation sont même parfois supérieures aux capacités moyennes des sujets sains[94]. L'absence de preuves médicales de leurs déficits pose d'importants problèmes pratiques pour ces individus, car ceux-ci ne peuvent ainsi obtenir des pensions ou des allocations (Damasio, 1994, p.64)

Pour Damasio, les troubles comportementaux des patients atteints de lésions préfrontales s'expliquent par un déficit non pas cognitif, mais affectif. Comme le souligne Damasio, ces individus manifestent un détachement émotionnel déroutant : ils semblent « *envisager la vie sur un mode neutre* » (Damasio, 1994, p.70). Étant en « *mesure de*

93 De la même manière que pour les travaux étudiés dans la première partie, la référence qui est faite à la théorie économique doit être comprise comme une simple inspiration, et non comme un emprunt de modèles formels élaborés par des économistes. L'*Iowa Gambling Task* peut en effet faire penser à des problèmes décisionnels étudiés en théorie des jeux, mais il a été conçu par Damasio, tout comme les modèles d'actualisation hyperbolique en psychologie qui, bien que faisant référence à un processus de maximisation, ont bien été développés par les psychologues.

94 Observant par exemple les performances d'un patient atteint de lésions préfrontales surnommé « EVR » dans divers tests neurocognitifs, Damasio écrit : « *en contraste saisissant avec ses handicaps dans le prise de décision de la vie réelle, l'intelligence générale et les capacités de résolution des problèmes dans un environnement de laboratoire de EVR restent intactes. Par exemple, il obtient un très bon score au jeu de carte de Wisconsin (Milner, 1963) ou dans des paradigmes d'évaluation de la cohérence personnelle (Petrides & Milner. 1982), d'estimations cognitives (Shallice & Evans, 1978), et ses jugements de fréquence et de récurrence (Milner, Petrides, & Smith, 1985) ne manifestent aucun défaut; il n'est ni obstiné ni impulsif; sa connaissance élémentaire est intacte, ainsi que sa mémoire de court-terme et sa mémoire de travail; ses réponses aux problèmes sociaux ou aux dilemmes éthiques sont comparables à celles des sujets normaux (Saver & Damasio, 1991)* » (Bechara *et al.*, 1994, p.8) .

connaître mais non de ressentir » (Damasio, 1994, p.73), ils sont incapables de faire appel dans les processus de choix réels à leurs facultés de raisonnement pourtant intactes. Dans le domaine financier par exemple, ils parviennent très bien à résoudre des calculs d'actualisation et de capitalisation, tout en poursuivant dans leur vie privée des stratégies extrêmement hasardeuse. Sur le plan social, la connaissance des règles et des conventions morales ne les empêche pas de violer celles-ci.

Sur la base de cette connaissance accumulée à partir de cas clinique, l'apport de Damasio à la neuroéconomie a été double. Damasio a d'abord développé un protocole de diagnostic pour les patients atteints de lésions préfrontales, appelé *Iowa Gambling Task* (IGT), qui a eu une large influence en avançant l'idée d'un test de détection « économique »[95] de troubles mentaux.

L'IGT fait l'objet d'une première publication par Damasio et ses coauteurs dans la revue *Cognition* en 1994 (Bechara, A. Damasio, H. Damasio et Anderson, 1994). L'IGT ressemble en fait aux protocoles utilisés pour étudier l'apprentissage de la récompense, du type machines à sous multi-jeux. Il repose sur un choix répété entre plusieurs options générant des récompenses, gains ou pertes, de montants variables, l'objectif étant d'acquérir de l'information pour découvrir la stratégie optimale. Les sujets font face à quatre tas de cartes et reçoivent au début de l'expérience 2000$[96]. Ils doivent ensuite tirer successivement des cartes parmi les quatre tas, sans savoir à l'avance lorsque le jeu s'arrête. A chaque fois qu'ils tirent une carte, les sujets obtiennent de l'argent. Parfois, après avoir tiré certaines cartes, ils doivent payer une punition.

Le but du jeu est de maximiser le gain total. Les cartes dans les deux premiers tas -*A* et *B*- produisent des gains plus élevés, 100 dollars à chaque carte, contre 50 dollars dans les deux derniers tas *C* et *D*. Néanmoins, les punitions associées aux tas *A* et *B* sont en moyenne plus importantes : en choisissant successivement 10 fois le tas *A*, les gains sont de 1000 dollars, mais le sujet subit également 5 punitions de 250 dollars, soit une perte globale de 250 dollars. En choisissant 10 fois dans les tas *C* et *D*, le sujet obtient 500 dollars de gains, mais les pertes sont inférieures (250 dollars), et par conséquent le gain global est de 250 dollars. Pour résumer, les tas *A* et *B* d'un côté, et *C* et *D* de l'autre, sont équivalents sur le long-terme en termes de gains ou pertes totales, les punitions étant soit plus fréquentes et de magnitude

95 Sur la nature de cette inspiration économique, voir la note précédente.

96 Il s'agit de monnaie factice. Dans cette étude, les gains et les punitions monétaires ne sont pas réels (c'est-à-dire qu'ils ne donnent pas lieu à des paiements à la sortie du laboratoire), ce qui a été beaucoup critiqué par la suite. Les différents problèmes méthodologiques pouvant être soulevés à l'encontre de l'*Iowa Gambling Task* feront l'objet d'un développement spécifique dans le chapitre suivant.

moins importantes ou moins fréquentes et de magnitude plus importantes.

Les tas *A* et *B* sont donc désavantageux parce qu'ils produisent des pertes nettes à long-terme. L'étude de Bechara *et al.* (1994) vise à comparer les performances de sujets sains et de patients lésés à l'*Iowa Gambling Task*. Les expérimentateurs observent que la plupart des sujets sains, après 30 ou 40 répétition, identifient la stratégie optimale et choisissent uniquement des cartes parmi les deux tas avantageux C et D[97]. Si l'on contraste les stratégies suivies respectivement par les 44 sujets « normaux » et les 7 sujets lésés, les résultats sont assez nets : les seconds prennent en moyenne plus de cartes désavantageuses que de cartes avantageuses, alors que les premiers parviennent à obtenir un gain global positif (Bechara *et al.*, 1994, p.9).

La performance individuelle à l'IGT (gain ou perte) peut donc servir de facteur discriminant. Les déficits à l'IGT peuvent faire l'objet de plusieurs interprétations, comme le suggèrent Bechara *et al.* : ils peuvent résulter soit d'une insensibilité aux pertes (les patients lésés ne se préoccuperaient uniquement des gains, et ne prennent pas en compte les punitions), soit d'une trop grande sensibilité aux récompenses (les patients lésés prendraient en compte les punitions mais sur-évalueraient les récompenses), soit de l'absence de prise en compte des conséquences futures (les patients lésés seraient sensibles aux pertes mais ne parviendraient pas à projeter dans le temps les émotions nées du sentiment de perte).

L'étude de Bechara, Tranel and Damasio publiée en 2000 permet de trancher plutôt en faveur de la dernière interprétation. Cette expérience repose sur une version inversée de l'*Iowa Gambling Task*, dans laquelle les punitions sont infligées à chaque choix de carte, et certaines cartes seulement produisent des gains (la structure des gains et des récompenses est inversée, mais il y a toujours deux tas plus avantageux que les deux autres). Les résultats montrent que les patients lésés, même dans cette version inversée, poursuivent des stratégies désavantageuses. Cela montre donc que le problème ne vient pas d'une trop grande sensibilité aux récompenses[98]. En outre, la mesure des réactions électrodermales[99] est comparable lors de

97 Encore une fois, ce résultat a été contesté par la suite. Dans l'étude de Glicksohn, Naor-Ziv et Leshem (2007), environ la moitié des sujets sains (46%) se montrent incapables d'identifier la meilleure stratégie après 100 décisions à l'IGT. Plus généralement, l'ensemble des critiques et des problèmes relatifs à l'IGT seront envisagés dans le chapitre suivant.

98 En effet, dans cette version modifiée de l'IGT, il s'agit d'apprendre à éviter de choisir les tas associés à des pertes de grande ampleur. Si donc le déficit dans la version originale de l'IGT (dans laquelle il s'agit d'éviter de choisir les tas associés à des gains ponctuels de grande ampleur) résultait d'une trop grande sensibilité aux gains (de grande ampleur), on devrait s'attendre à ce que ce déficit ne se manifeste pas dans la version inversée.

99 Le mesure de l'activité électrodermale est souvent utilisée en psychologie expérimentale comme un indice des émotions ressenties par le sujet. Une activité élevée indique une forte réactivité émotionnelle.

l'obtention des punitions chez les sujets lésés et sains, ce qui suggère que les premiers manifestent bien une sensibilité aux pertes (Bechara, Tranel et Damasio, 2000, p.2189) .

Les lésions du cortex préfrontal n'empêchent pas l'expérience d'émotions positives et négatives relatives aux gains et à la perte. Le déficit à l'IGT ne peut s'expliquer que par une incapacité à prendre en compte les conséquences futures des décisions, c'est-à-dire à anticiper précisément l'occurrence probable de ces émotions (Bechara, Tranel and Damasio, 2000, p.2189). Au moment de choisir, et avant de constater les résultats de ses choix, un sujet « normal » imagine son état émotionnel futur à partir des expériences passées.

Cette idée selon laquelle la délibération s'appuie sur l'histoire affective de l'individu est au cœur de la théorie des marqueurs somatiques de Damasio (*cf.* Damasio, 1994). Celle-ci ne doit pas être comprise comme une théorie des émotions, qui mettrait en avant le rôle des instincts primaires et des pulsions animales dans la pensée humaine. La théorie des marqueurs somatiques porte de manière plus complexe, non pas sur la pure sensibilité, mais plutôt sur la capacité des sujets à projeter leurs émotions dans le temps.

Si, pour Damasio, « *les émotions participent à la raison* », c'est parce que celles-ci jouent le rôle de valeur de référence ou d'objets préférés (Damasio, 1994, p.154). Au moment de choisir entre plusieurs actions, les émotions jouent le rôle de « *marqueurs somatiques* » au sens où elles encodent la valence (positive ou négative) des résultats possibles de chaque option. Ces marqueurs somatiques résultent de l'histoire de l'individu et s'expliquent par son expérience passée : par exemple, la peur intense suscitée par la menace d'un prédateur explique pourquoi, chez l'animal, la vision d'un prédateur provoque systématiquement une émotion (la peur), qui déclenche un programme d'action automatique (la fuite). En associant certaines expériences passées à des données du monde présent (présence d'un prédateur dans le champ de vision), les émotions font ressortir certains aspects d'une situation. Elles peuvent ainsi, dans les cas extrêmes, jouer le rôle de « signal d 'alarme », en conduisant à rejeter une action donnée ou au contraire à adopter un comportement instinctif (la fuite). Plus généralement, les marqueurs somatiques fonctionnent pour Damasio comme une intuition rapide. Ils permettent de choisir entre un nombre plus restreint d'alternatives, en évaluant les résultats des actions possibles (Damasio, 1994, p.204). Par exemple, dans l'IGT, au moment de choisir entre les quatre tas de cartes, les marqueurs somatiques sont constitués de l'ensemble des émotions éprouvées dans le passé lors de l'obtention de gain ou de perte dans le jeu. Après un nombre suffisant de répétitions, les marqueurs somatiques associés aux tas défavorables doivent être suffisamment aversifs pour pouvoir rejeter automatiquement de

choisir une carte dans ces deux tas.

Dans la perspective de Damasio, les émotions assistent donc le processus de raisonnement. Sans émotions, les sujets seraient incapables de traduire en actes leurs connaissances : en connaissant les deux cartes m'assurant un gain financier certain à long terme, je ne choisirais pas de tirer ces deux cartes si la perspective d'un gain financier m'est totalement indifférente et ne suscite chez moi aucun affect. Pour autant, il convient de souligner encore une fois que la rationalité n'est pas pour Damasio dépendante d'une pure émotivité. Ce qui compte, c'est à la fois la capacité à conserver une trace d'une expérience affective originaire, mais d'abord et surtout de « réveiller » et de projeter cette marque de manière pertinente au moment présent et futur. Certes, cette expérience affective originaire renvoie à des pulsions et des instincts primaires, qui constituent ce que Damasio appelle le « *répertoire de représentations potentielles innées* » (Damasio, 1994, p.154). Mais la rationalité dépend précisément des facultés de mémorisation et d'association de ces représentations à de nouvelles relations (Damasio, 1994, p.100). Du point de vue de la théorie des marqueurs somatiques, les émotions ou marqueurs somatiques sont donc des instincts primaires détournés de leur objet. Plutôt que de déboucher sur l'idée d'une opposition entre les émotions et les capacités réflexives et de planification temporelle, Damasio met donc en avant une continuité évolutive entre les processus cognitifs humains et animaux.

La référence aux travaux de Damasio dans les manifestes de l'économie comportementale dans le scanner est donc en règle générale assez éloignée de l'esprit général de la théorie des marqueurs somatiques. En particulier, Camerer, Loewenstein et Prelec (2005) prennent les recherches de Damasio et de son équipe comme un appui à l'idée selon laquelle il y aurait une spécialisation des zones cérébrales impliquées dans l'évaluation des récompenses en fonction de l'horizon temporel de la décision. Pourtant, pour Damasio, le cortex pré-frontal n'a pas seulement pour rôle d'inhiber les inclinations en faveur des satisfactions immédiates, mais aussi de conserver une trace de l'expérience affective passée et de projeter celle-ci dans le futur. Damasio rejette par ailleurs explicitement les interprétations du fonctionnement du cortex préfrontal en termes d'inhibition (Bechara, Tranel and Damasio, 2000, p.2189).

Le « détournement » de la théorie des marqueurs somatiques par les économistes n'est cependant pas complètement illégitime, et n'a jamais été critiqué par Damasio lui-même. Cette théorie est en fait assez indéterminée. En laboratoire ou dans la vie réelle, les interprétations comportementales susceptibles d'être données à partir des travaux de Damasio

sont assez ambiguës, y compris dans l'IGT, et sont généralement compatibles avec plusieurs hypothèses concurrentes (*cf.* chapitre 5). Quoiqu'il en soit, l'innovation théorique incontestable de Damsio repose sur l'idée d'utiliser une technique néo-comportementale -un protocole type machine à sous multi-jeux- comme outil de diagnostic clinique. Cette avancée théorique est aussi une avancée thérapeutique, car, comme le souligne Damasio, l'IGT « *fournit le premier diagnostic de laboratoire pour les patients ayant des lésions au cortex préfrontal ventromédian -une avancée plutôt utile, étant donné que ces patients passent en général avec succès tous les autres tests neuropsychologiques et manifestent leurs défaillances uniquement dans la vie réelle* » (Damasio, 2008, p.211). Cette idée d'une neuropsychiatrie d'inspiration néo-comportementale a eu une large influence sur la neuroéconomie. Damasio a cependant toujours gardé ses distances à l'égard de ce nouveau champ disciplinaire, en opposant à la notion d' évaluation économique celle de marqueur somatique.

B. Émotions versus évaluation économique : les rapports ambivalents de Damasio à la neuroéconomie

Damasio a joué un rôle de première importance dans le développement de la neuroéconomie. L'IGT constitue une avancée majeure, non seulement parce qu'il offre un moyen de détection de certaines pathologies associées à des troubles de l'impulsivité, mais aussi parce que, plus généralement, l'idée d'utiliser des protocoles type machines à sous multi-jeux comme outils de diagnostic a constitué par la suite, selon nous, un principe théorique déterminant de la neuroéconomie. Par ailleurs, l'influence de Damasio sur la neuroéconomie est également visible à travers les nombreuses références qui y sont faites à sa théorie des marqueurs somatiques. Les économistes comportementalistes qui, comme Georges Loewenstein, sont convaincus de l'importance des émotions se sont beaucoup inspirés des travaux de Damasio, qu'il souvent à l'appui de leurs modèles duaux (voir par exemple Loewenstein, 2000, p.432).

Pourtant, Damasio ne participe pas directement au développement de la discipline. Ce n'est qu'en 2008 que Damsio rédige un chapitre dans un ouvrage collectif rassemblant les

contributions importantes de la neuroéconomie (Glimcher *et al.*, 2008): son ralliement est donc relativement tardif. Cet article fait apparaître assez clairement que Damasio n'est pas vraiment neuroéconomiste, ou ne se considère pas exactement comme tel. Damasio y discute en effet la notion de valeur (économique), mobilisée abondamment dans les interprétations des travaux de neuroéconomie, en particulier dans le cadre théorique de l'économie neuronale de Berns et Montague (2002).

Damasio reconnaît que le concept de valeur ou d'estimation économique peut être pertinent d'un point de vue biologique, comme le postulent Berns et Montague (2002), dans la mesure où « *les évaluations que nous produisons quotidiennement dans nos activités sociales et culturelles ont toujours un lien, direct ou indirect avec la biologie et […] et les processus de régulation vitale* » (Damsio, 2008, p.212). Toutefois, Damasio regrette que dans la littérature de neurobiologie expérimentale, ces valeurs soient systématiquement réduites à « *des molécules ou à la machinerie de récompenses et de punitions* » (Damsio, 2008, p.212). En d'autres termes, ce qui pose problème pour Damasio est que ces valeurs économiques soient comprises, dans le cadre de protocoles de jeux financiers, comme des prédictions de gains ou de pertes (financières ou alimentaires), pouvant être corrélées à des variables neuronales voire neurochimiques.

Pour Damasio, cette notion de valeur économique ainsi comprise est trop étroite. Elle ne correspond pas à son propre concept d'émotion, qui va plus loin qu'un simple mécanisme de calcul des gains et des pertes. Même s'il admet les difficultés de cette entreprise -« *distinguer l'émotion de ses composantes en termes de récompenses et punitions pose un problème conceptuel majeur* » (Damsio, 2008, p.211)- Damasio manifeste donc une réticence vis à vis d'une approche considérée selon lui comme « *économique* », c'est-à-dire comme purement quantitative. Comme chez Georges Ainslie, l'importance des préoccupations thérapeutiques est ainsi à l'origine d'une reconnaissance des limites posées à la quantification des troubles mentaux et des comportements humains. De son point de vue, la neurobiologie expérimentale purement quantitative, appuyée sur l'IRM_f, tend à négliger le rôle des émotions, qui constituent une « *intelligence précognitive mais complexe* » (Damasio, 2008, p.213).

Conclusion du chapitre 4

Au début des années 2000, l'utilisation des techniques de neuroimagerie sur des sujets humains permet d'étendre à l'homme les résultats théoriques des recherches néo-comportementalistes. Les premiers travaux par IRM_f portant sur le *reward learning* confirment l'hypothèse d'une rationalité évolutive élémentaire, commune à l'homme et à l'animal, par lesquels les organismes apprennent à former des anticipations sur la conséquence probable de leurs actions.

Ces expériences suscitent l'intérêt des économistes comportementalistes. Or, ces derniers y voient au départ essentiellement une opportunité pour donner un appui empirique supplémentaire à leurs propres modèles. La neuroéconomie, ainsi soutenue et défendue par les *behavioral economics*, fait figureà ses débuts d'« *économie comportementale dans le scanner* » (Ross, 2008). Dans les cas du regret, du choix inter-temporel ou inter-personnel, les économistes comportementalistes convertis aux neurosciences appliquent des schémas d'interprétation dualistes aux données fournies par la neuroimagerie. La neuroéconomie alors naissante s'affirme alors comme une approche mettant en avant le rôle des émotions , en s'inspirant largement des travaux de Damasio.

L'économie comportementale dans le scanner soulève cependant des problèmes d'interprétation, liés à l'application d'une grille interprétative kahnemanienne sur des tâches type machines à sous multi-jeux. L'enthousiasme des premières années a assez rapidement été refroidi par des critiques virulentes adressées par des économistes comportementalistes sceptiques quant à la pertinence des techniques de neuroimagerie pour leur discipline. Cela ne signifie pas pour autant que l'apport théorique de ces travaux soit nul. L'étude du fonctionnement des zones préfrontales chez l'homme a notamment permis d'améliorer la compréhension des fonctions de mémorisation et d'apprentissage. La portée médicale du néo-comportementalisme commence à être reconnue, avec l'utilisation de problèmes type *multi-armed bandit* comme outils de diagnostic de l'impulsivité. C'est en assumant plus explicitement son héritage néo-comportementaliste que la neuroéconomie a pu par la suite résoudre ses difficultés interprétatives et s'affirmer comme une discipline autonome.

Chapitre 5. De l'économie comportementale dans le scanner à la neuro-psychiatrie computationnelle: la constitution d'une discipline autonome (2005-2010)

Au cours de la deuxième partie des années 2000, la neuroéconomie s'émancipe de ses disciplines de tutelle et s'affirme progressivement comme un nouveau sous-domaine autonome de l'analyse économique. Un basculement important s'opère en 2005. C'est en effet au cours de cette année qu'est publié dans le *Journal of Economic Literature* l'article de Colin Camerer, Georges Loewenstein et Drazen Prelec, intitulé « la Neuroéconomie: comment les neurosciences peuvent guider l'analyse économique » (Camerer, Loewenstein, et Prelec, 2005). Cette publication marque l'apogée de l'économie comportementale dans le scanner. Les auteurs, tous trois économistes comportementaux de renom, veulent promouvoir l'utilisation des neurosciences dans leur discipline (*cf.* chapitre 4).

S'il a incontestablement rempli son objectif de promotion et de diffusion des thèmes de recherche de la neuroéconomie, l'article-manifeste de Camerer, Loewenstein et Prelec a également été à l'origine d'un grand nombre de critiques méthodologiques. Le rapprochement entre économie et neurosciences a alors fait l'objet d'une réévaluation par des économistes non-convertis aux neurosciences. Ces derniers, bien qu'étant sceptiques à l'égard de la neuroéconomie, ont apporté d'importantes contributions au débat théorique, en invalidant certaines hypothèses, mais aussi et surtout en précisant les concepts théoriques utilisés par les neuroéconomistes. *A posteriori*, les critiques méthodologiques externes ont joué un rôle positif dans le développement de la neuroéconomie (*cf.* introduction).

L'article de Camerer, Loewenstein et Prelec (2005) initie ainsi indirectement une rupture dans le développement de la neuroéconomie. En exposant plus clairement à leurs détracteurs les principaux défauts de l'approche qu'ils se proposaient de défendre, Camerer, Loewenstein et Prelec ont sapé les fondements de l'économie comportementale dans le scanner. Les diverses controverses méthodologiques qui apparaissent dans la littérature ont alors incité les neuroéconomistes à se démarquer de l'économie comportementale. La thèse principale avancée dans ce chapitre est que la défense proposée par les neuroéconomistes n'a pas consisté à concevoir un nouveau paradigme *ad hoc*, plus robuste, mais plutôt à se réinscrire leur approche au sein du courant néo-comportementaliste.

Il s'agit donc ici d'étudier comment la neuroéconomie, à partir de 2005, s'est

construite comme une discipline indépendante à partir de ce qui a été conçu dans la première partie comme étant le noyau dur de sa préhistoire, c'est-à-dire un paradigme de psychiatrie économique. L'économie comportementale dans le scanner a d'abord été soumise à une importante remise en question méthodologique (I. Les critiques externes comme moteurs théoriques). La mise en avant du rôle de la pathologie (II. La construction d'un paradigme autonome (1) : la référence à la pathologie) et de la notion de monnaie neuronale commune (III. La construction d'un paradigme autonome (2) : la mise en avant d'une monnaie neuronale commune à l'homme et à l'animal) permet de répondre à ces critiques. L'accusation de réductionnisme a conduit les neuroéconomistes à se différencier des *behavioral economics* et à construire un paradigme autonome. La neuroéconomie peut alors être comprise comme un projet de psychiatrie économique, qui se donne à voir en particulier dans l'analyse des comportements addictifs (IV. La neuroéconomie comme projet de psychiatrie économique le cas des comportements addictifs).

I. Les critiques externes comme moteurs théoriques

Même si elle apparaît à ses débuts comme une discipline en vogue, la neuroéconomie fait débat. En économie comme en neurosciences, son potentiel théorique est régulièrement remis en question. Chez les économistes, le scepticisme constitue presque la règle. L'usage oblige l'économiste publiant une étude sur la neuroéconomie à émettre des réserves sur la méthode employée par les neuroéconomistes. En France, les rares avocats de cette approche (Schmidt, 2010; Bourgeois-Gironde, 2008) peinent à convaincre la communauté des économistes, qui expriment souvent de la curiosité pour les neurosciences, mais toujours teintée de circonspection. En neurosciences, les critiques sont moins acerbes, et portent plus spécifiquement sur des problèmes plus techniques concernant les protocoles, les manipulations statistiques liées à la production de données en neuroimagerie, ou sur l'interprétation de ces données (voir par exemple Kable, 2011; Platt et Huettel, 2008; Huettel, 2010). Néanmoins, l'idée de fournir une théorie générale du comportement, à la manière d'Antonio Damasio (1994) fait aussi l'objet de critiques sévères.

L'enjeu ici consiste à montrer que ces critiques ont joué un rôle de moteur théorique, en mettant en évidence plusieurs travers méthodologiques caractéristiques des premières expériences de neuroéconomie. Les critiques visent à la fois l'économie comportementale dans le scanner (A) et la notion d'émotion dans les travaux de Damasio (B).

A. Les critiques de l'économie comportementale dans le scanner

L'économie comportementale dans le scanner vise à expliquer directement les choix économiques et leurs mécanismes psychologiques sous-jacents à partir de données neuronales. Cette approche soulève plusieurs difficultés. Il est d'abord possible de faire valoir que l'économie ne traite que des choix et des comportements observés et non de processus psychologiques: c'est le traditionnel argument des préférences révélées, avancé notamment par Wolfgang Gul et Gustav Pesendorfer (1). Cette critique n'est cependant pas spécifique à la neuroéconomie, mais vaut plus généralement pour toute tentative d'explication psychologique

du comportement économique. Les reproches formulés par Ariel Rubinstein (2) et Glenn Harrison (3) remettent en cause quant à eux plus directement le potentiel théorique, pour les *behavioral economics*, des instruments d'observation neuroscientifiques, et de l'IRM_f en particulier.

1. Le traditionnel argument des préférences révélées et de l'« économie sans pensée » *(mindless economics*) (Gul et Pesendorfer. 2005)

Faruk Gul et Wolfgang Pesendorfer sont deux économistes de Princeton, connus notamment pour leurs modèles dit des « tentations » et du « *self control* » (Gul et Pesendorfer, 2001, 2004). Ils écrivent en 2005 un document de travail sur la neuroéconomie, intitulé « *une défense de l'économie sans pensée* », qui fait l'objet d'une publication en 2008[100]. Cet article est très représentatif des réactions sceptiques exprimées le plus souvent par les économistes vis à vis de la neuroéconomie. Gul et Pesendorfer font valoir deux principaux arguments à l'encontre de l'approche mobilisée par les neuroéconomistes. Tout d'abord l'économie s'appuie selon Gul et Pesendorfer sur la théorie des préférence révélées et n'a pas vocation à traiter des mêmes problèmes théoriques que la psychologie et les neurosciences. A cette critique formelle portant sur la spécificité théorique de chaque discipline s'ajoute un argument visant plus spécifiquement le contenu des modèles de la neuroéconomie. En suggérant que les choix observés ne manifesteraient pas les préférences réelles des individus, les neuroéconomistes ne feraient que manifester leur propre point de vue que Gul et Pesendorfer qualifient de « *philosophique* », ce qui peut laisser craindre des dérives paternalistes.

L'idée d'un hiatus formel entre l'économie et toutes les disciplines qui traitent de la « *pensée* » et des processus psychologiques forme le noyau dur de la critique de Gul et Pesendorfer : « *l'économie et la psychologie répondent à des questions différentes, utilisent des concepts différents, et s'occupent de différents types de preuves empiriques* » (Gul et Pesendorfer, 2005, p.1). Cet écart concerne donc à la fois les méthodes, les théories ou les concepts, et les données. Gul et Pesendorfer montrent par exemple que la notion de « *signal* » (*cue*) telle qu'elle est définie par les neuroscientifiques, n'est pas pertinente en économie ((Gul et Pesendorfer, 2005, p.11-12). Les psychologues parlent de « *réponse déclenchée par un signal* » (*cue-triggered response*) lorsque, par exemple, la vision d'un hamburger « *déclenche* » un désir de consommation de frites, ou lorsque la présence de dealers dans un environnement urbain « *déclenche* » chez l'*addict* un désir de consommer de la drogue. Or, du

100L'article a été disponible dans sa version non-définitive dès 2005, bien avant sa publication, et a eu une large influence dans la littérature. C'est la raison pour laquelle j'ai choisi ici de m'appuyer sur la version de 2005, qui a été au centre des débats sur la neuroéconomie.

point de vue de l'économiste, le premier cas renvoie à la notion de biens complémentaire, alors que le second est compris comme un phénomène d'externalité (négative). En économie, ces deux concepts distincts fournissent des liens explicatifs différents avec les variables économiques, et en particulier avec les prix. Si les deux biens sont qualifiés de complémentaire, l'économiste s'attend à ce que leur demande varie de manière similaire à un changement de prix. La notion d'externalité suggère plutôt l'idée d'une défaillance de marché (*market failure*), et d'une déformation des prix (Gul et Pesendorfer, 2005, p.11-12).

Des phénomènes similaires du point de vue de leur déroulement psychologique peuvent donc renvoyer à de phénomènes économiques qui doivent être conçus de manière distincte. Inversement, un même type de phénomènes économiques peut s'appuyer sur des processus psychologiques hétérogènes. Cette observation a deux conséquences. Tout d'abord, l'économie ne dit rien des processus mentaux ou neuronaux sous-jacents aux choix qu'elle décrit. Surtout, les neurosciences et la psychologie n'ont absolument pas vocation à tester les hypothèses théoriques des économistes. Gul et Pesendorfer ne sont pas opposés à ces discipline en tant que telles, mais à la tentative visant à utiliser des arguments psychologiques en économie. La critique ne porte pas sur la neuroéconomie directement, mais constitue plutôt une réponse à la critique de l'économie dite « *standard* » par les neurosciences, telle qu'elle est exprimée notamment par Camerer, Loewenstein et Prelec (2005) (Gul et Pesendorfer, 2005, p.1).

Gul et Pesendorfer prônent donc une stricte séparation entre l'économie d'un côté, la psychologie et les neurosciences de l'autre. La démarche des neuroéconomistes est naïve, car elle suppose que les problèmes théoriques sont les mêmes de part et d'autre de cette frontière disciplinaire: « *les neuroéconomistes importent les questions et les concepts de la psychologie et ré-interprètent les modèles de l'économie comme si leur objectif consistait à répondre à ces questions. Le modèle du choix standard en économie est considéré comme un modèle du cerveau, qui est alors envisagé comme inadéquat. Soit l'économie est traitée comme une science du cerveau d' « amateurs » et est rejetée comme telle, soit des données neuronales sont utilisées comme des données économiques pour rejeter des modèles économiques* »(Gul et Pesendorfer, 2005, p.2).

Cet argument repose sur la définition, ou plutôt la réduction de l'objet de la science économique au seul domaine des choix observés et observables. L'approche de l'économie que Gul et Pesendorfer qualifient de « *standard* » et qu'ils se proposent de défendre renvoie finalement à l'acceptation de la théorie des préférences révélées: « *dans l'approche standard, les termes maximisation de l'utilité et choix sont synonymes […] Les données pertinentes sont*

les préférences révélées, c'est-à-dire les choix de consommation effectifs étant données les contraintes individuelles. Ces données sont utilisées pour calibrer des modèles (pour identifier des paramètres particuliers) et les modèles calibrés ainsi obtenus permettent de prédire les choix futurs mais aussi des variables d'équilibre comme les prix. Par conséquent, la théorie standard (positive) identifie les paramètres du choix au comportement passé et relie ce paramètres aux choix futurs et aux variables d'équilibres. L'économie standard se concentre sur les préférences révélées parce que les données économiques sont produites sous cette forme » (Gul et Pesendorfer, 2005, p.6).

Cette définition est bien sûr critiquable, puisque de nombreux économistes qui se considèrent eux-mêmes comme appartenant au courant « dominant » ou « standard » n'acceptent pas automatiquement la théorie des préférences révélées[101]. Par ailleurs, Gul et Pesendorfer font comme si le travail théorique de l'économiste consistait uniquement à paramétrer des modèles pré-existants: mais qui s'occupe, en amont, de construire ces modèles? Surtout, la restriction des données économiques aux seuls choix observés n'est justifiée que parce que « *les données économiques sont produites sous cette forme* ». Rien n'empêche de penser, cependant, que les données neuronales puissent permettre de faire également des prédictions sur les comportements, ou améliorer celles-ci: pourquoi décider alors de les rejeter *a priori*?

La seule véritable défense de l'approche « *standard* », qui selon Gul et Pesendorfer revient à considérer les choix observés comme la conséquence de la maximisation d'une fonction d'utilité (*cf. supra*), repose sur l'argument de la « flexibilité ». Ce modèle standard, quoique limité sur le plan de l'explication psychologique, est le meilleur car il peut s'accommoder de n'importe quel type de comportements (Gul et Pesendorfer, 2005, p.43). Or si un modèle est flexible et peut rendre compte de tous les comportements, sa capacité prédictive est quasi-nulle. Pourtant, Gul et Pesendorfer défendent précisément les modèles économiques en limitant leur fonction à la seule prédiction des comportements.

La critique de Gul et Pesendorfer n'est donc pas seulement formelle. Le reproche ne porte pas seulement sur la définition des concepts propres à chaque discipline. En effet, le fossé théorique n'est pas nécessairement infranchissable: il est tout à fait possible de

101 Voir par exemple Ross, 2005, pour un traitement pour approfondi de ces questions. Un développement sur le rôle de la théorie des préférences révélées serait bien sûr trop long et dépasserait le cadre établi de ce travail. Il est possible néanmoins de suggérer ici que les difficultés liés à la définition de ce qui est appelé « approche standard » ou « courant dominant » s'expliquent simplement par le fait que l'économie dite standard est en fait extrêmement hétérogène et recouvre des courants de recherche très différents. Ceci nous a conduit à rejeter, dans le cadre de cette étude d'histoire de la pensée économique, l'utilisation des termes de « courant standard » ou « dominant » (*cf.* introduction). Nous ne traiterons donc pas de la question -vaine selon nous- de savoir si l'appartenance à ce courant suppose, ou non, l'adoption de la théorie des préférences révélées.

concevoir que des processus psychiques, bien que définis dans les termes de l'analyse psychologique et/ou neuroscientifique, puissent aider l'économiste à prédire les comportements des individus, en lien avec des variables économiques. Pour Gul et Pesendorfer, la neuroéconomie pose problème, d'abord parce qu'elle mobilise des concepts économiques sans maîtriser leur définition, ou tente de faire passer pour économiques des concepts psychologiques, mais d'abord et surtout parce qu'elle tend à remettre en question le fondement normatif de l'approche standard, en suggérant que les agents peuvent agir à l'encontre de leur propre intérêt. C'est effectivement un thème assez présent en neuroéconomie, à travers notamment l'étude des comportements impulsifs. Il y a là selon Gul et Pesendorfer une attitude théorique profondément étrangère à l'économie, qu'ils qualifient d'ambition « *thérapeutique* »: « *la neuroéconomie poursuit des ambitions thérapeutiques: elle cherche à améliorer les objectifs de l'individu. Les questions centrales de la neuroéconomie sont: comment les individus effectuent-ils leurs choix? Ces processus de prise de décision permettent-ils effectivement d'augmenter leur bien-être? A l'inverse, les économiste analysent la manière avec laquelle les choix individuels interagissent au sein d'un ensemble institutionnel, étant donné les objectifs de ces individus* » (Gul et Pesendorfer, 2005, p.8).

L'accusation portant sur le caractère « *thérapeutique* » de la neuroéconomie et de ses ambitions normatives sera approfondie dans le chapitre 7, consacré au thème du paternalisme. Il importe ici de souligner cependant que l'argument « défensif » de Gul et Pesendorfer reposant sur les préférences révélées est en fait insuffisant pour justifier l'étanchéité des frontières disciplinaires. Ces derniers tendent alors de remettre en cause, au fil du développement de l'argumentation, le contenu même des théories psychologiques et neuroscientifiques. Gul et Pesendorfer font notamment valoir le fait que ces théories réduisent l'homme à l'animal: « *le fait que des substances addictives chez le rat le soient également chez l'homme n'a aucune pertinence pour l'économie* » (Gul et Pesendorfer, 2005, p.23). Surtout, les auteurs finissent par dénoncer le manque de sérieux dans les interprétations des expériences de neuroimagerie, qui engageraient selon eux des « *prises de position philosophiques* » de leurs auteurs: « *ce que les auteurs décrivent comme preuve n'est en fait bien souvent qu'un exposé relevant d'une prise de position philosophique. Ils décident que le cortex est associé à la planification de l'action (le choix rationnel), alors que d'autres processus (probablement dans les régions limbiques) représenteraient des influences physiologiques dans la décision* » (Gul et Pesendorfer, 2005, p.24).

L'article de Gul et Pesendorfer est très représentatif des critiques les plus communes adressées par les économistes à la neuroéconomie. Celles-ci partent au départ d'un reproche

formel: les neurosciences sont incontestablement intéressantes, mais l'économie traite des choix observés, en lien avec des variables économiques comme les prix, et n'a pas vocation à traiter de processus psychologiques. En développant cet argument, on s'aperçoit que ce qui pose problème pour Gul et Pesendorfer est l'idée selon laquelle les choix observés ne permettraient pas de révéler les préférences. Gul et Pesendorfer ne souhaitent pas au départ s'attaquer directement à la psychologie et aux neurosciences, mais à leur transformation en théorie économique. Mais ils finissent par remettre en cause la validité et la scientificité de ces disciplines prises isolément[102]. Or, la défense de l'« économie sans pensée » vaut aussi bien pour l'économie comportementale, l'économie expérimentale et la neuroéconomie. Elle oppose à la neuroéconomie des arguments qui ne sont pas spécifiques aux neurosciences, mais qui valent contre toute tentative visant à fournir des fondements psychologiques plus solides à l'économie.

2. Les neuroéconomistes sont-ils des économistes comportementalistes « gâtés »? La neuroimagerie comme jouet coûteux (Rubinstein, 2008)

Ariel Rubinstein, dans un article publié en 2008 dans le numéro spécial de la revue *Economics and Philosophy* consacré à la neuroéconomie (Rubinstein, 2008), remet en cause l'intérêt théorique des techniques de neuroimagerie pour l'économie à partir d'un argument très simple: les données neuronales ne sont pas pertinentes en économie car leur coût est beaucoup trop élevé. D'autres indicateurs, plus faciles à recueillir (temps de réponse, réaction électro-dermale) sont plus adaptés aux besoins de l'économie comportementale. Étant lui-même économiste comportementaliste, et donc favorable aux tentatives visant à donner des fondements psychologiques plus solides à l'économie, Rubinstein propose ici une critique pragmatique de la neuroéconomie qui ne se borne pas, à l'inverse de celle de Gul et Pesendorefer, à rejeter *a priori* l'utilisation d'indicateurs psychologiques dans l'étude du comportement économique.

Pour Rubinstein, l'approche dite « *standard* » élabore des modèles de la forme

102A la fin de l'article, Gul et Pesendorfer écrivent notamment: « *au bout du compte, les arguments scientifiques ne jouent qu'un rôle mineur dans la neuroéconomie: lorsqu'il s'agit de fournir une justification à la prise de position philosophique selon laquelle il y a une différence entre ce que les gens veulent et ce qu'il est bon pour eux, l'interprétation subjective du visage d'une souris fonctionne aussi bien qu'une expérience de neuroimagerie* » (Gul et Pesendorfer, 2005, p.24)

suivante (Rubinstein, 2008, p.492) :

$$c(A) = a$$

La fonction *c(x)* permet de prédire qu'un individu rationnel faisant face à l'alternative *A* choisira l'option *a*. L'économie comportementale propose des modèles du type :

$$c(A, f) = a$$

Le paramètre *f* renvoie à un effet de cadrage (*framing effect*): faisant face à l'alternative *A*, l'individu choisira *a* si la présentation de l'alternative A est « *lue subjectivement* » avec le cadrage *f*. Cela implique que, pour une même alternative, un individu effectuera des choix différents en fonction du cadrage retenu, qui dépend notamment du mode de présentation des options. Pour Rubinstein, les modèles en neuroéconomie expliquent les choix de la manière suivante :

$$c(A) = (a, x)$$

L'individu choisit l'option *a* dans l'alternative *A* et produit des indicateurs neuronaux (taux d'oxygénation du sang dans certaines zones du cerveau, activité neuronale, *etc.*) qui sont enregistrés pendant la prise de décision. Ces variables *x* pourraient servir, en économie comportementale, à expliquer le rôle des effets de cadrage *f* sur le choix. Les données neuronales pourraient ainsi permettre de comprendre et d'identifier les types comportementaux mis en évidence par les économistes comportementalistes (Rubisntein, 2008, p.495).

Rubinstein suggère donc des « *modèles dans lesquels la distribution des types serait une primitive* » (Rubinstein, 2008, p.493). En d'autres termes, il s'agirait de déterminer des liens de causalité entre des variables neuronales et des effets de cadrage, et, par suite, d'en déduire le choix retenu de l'individu :

$$i(x) = f$$
$$c(A, f) = a$$

Les données neuronales *x* permettent de prédire que l'individu adoptera le cadrage *f*, et donc qu'il choisira l'option *a* parmi *A*. Rubinstein propose ici un mode de collaboration entre neurosciences et économie qui s'appuie sur la notion d'effet de cadrage. Ce faisant, il considère implicitement que la neuroéconomie a naturellement vocation à s'intégrer au sein du programme de recherche kahnemanien sur les effets de cadrage, ce qui sera précisément remis en question dans notre seconde section. Ce qu'il importe de souligner ici est que cette approche, du point de vue de Rubinstein, soulève deux problèmes. Tout d'abord, les

neurosciences n'auraient ici pas vraiment de fonction théorique. Il s'agirait simplement de confirmer au niveau neuronal des types comportementaux préalablement identifiés par les économistes comportementalistes. Il n'y aurait pas à proprement parler production de « *nouvelles idées économiques* » (Rubinstein, 2008, p.489).

Or, pour Rubinstein, les indicateurs neuronaux utilisés par les neuroéconomistes ne sont en fait pas adaptés à cette fonction de détection des types comportementaux. En effet, les données neuronales sont extrêmement coûteuses à produire. Elles supposentun appareillage extrêmement lourd, en particulier dans le cas de la neuroimagerie. Selon Rubinstein, la plupart des travaux de neuroéconomie conduisent en effet, dans une perspective dualiste, à distinguer des choix « *instinctifs* » et « *cognitifs* », qui peuvent être compris comme deux cadrages différents d'un même problème décisionnel (Rubinstein, 2008, p.489). Du point de vue de Rubinstein, la prédiction de l'effet de cadrage sur l'individu peut être réalisée dans ce cas de façon beaucoup plus simple, sans nécessairement s'appuyer sur la neuroimagerie. Rubinstein a notamment réalisé une expérience, sur Internet, portant sur le paradoxe d'Allais. Les résultats indiquent que les temps de réaction fournissent une variable pertinente pour discriminer entre les choix instinctifs des choix cognitifs, qui respectivement violent ou respectent l'axiome d'indépendance des préférences (Rubinstein, 2007).

Rubisntein préconise donc l'utilisation en économie comportementale d'indicateurs comportementaux simples et peu couteux à recueillir, tels que les temps de réaction ou les mouvements de l'œil. Pour Rubinstein, le cout élevé des techniques de neuroimagerie conduit précisément les neuroéconomistes à surestimer la fiabilité de ces outils. Cela explique pourquoi les interprétations des expériences de neuroéconomie sont généralement trop rapides et superficielles, car: « *des diagrammes en couleurs […] sont présentées comme des preuves indiscutables* » (Rubinstein, 2008, p.486). Ces réflexions sur le caractère séduisant et trompeur de la neuroimagerie rejoignent des analyses plus communes de la « *rhétorique* » de la neuroéconomie, qui ont été abordées dans le premier chapitre (*cf.* Mäki, 2010). Rubinstein soulève ici cependant une question importante, en soulignant la nécessité de justifier l'intérêt théorique des techniques neuroscientifiques, au regard d'autres instruments plus commodes à manipuler pour les économistes comportementaux.

3. La neuroéconomie comme néo-phrénologie: une attaque directe contre les neurosciences (Harrison, 2008-a et 2008-b)

L'article publié en 2008 par Glenn Harrison constitue probablement la critique la plus aboutie de la neuroéconomie, à la fois par son exhaustivité, sa précision et sa vigueur (Harrison, 2008-a). Réalisant lui-même des expériences d'économie en laboratoire, Harrison, comme Ariel Rubinstein, n'est pas opposé en principe à l'idée de donner des fondements psychologiques et/ou neuroscientifiques plus solides à la théorie économique du choix rationnel. Il rejette ainsi les positions extrêmes exprimées par Gul et Pesendorfer, qu'il condamne pour leur « *isolationnisme* » (Harrison, 2008-a, p.339). Harrison va cependant plus loin que Rubinstein, puisqu'il fait valoir que les données neuronales sont non seulement plus couteuses à obtenir, mais soulèvent d'abord et surtout d'importantes difficultés d'interprétation. Les divers problèmes méthodologiques liés à l'inférence d'états mentaux à partir de variables physiologiques conduisent Harrison à condamner la neuroéconomie comme une « *néo-phrénologie* » ((Harrison, 2008-b, p.536).

Les reproches adressés par Harrison à la neuroéconomie portent d'abord sur la « *rhétorique révolutionnaire* » de certains neuroéconomistes, en référence notamment à l'article de Camerer, Loewenstein et Prelec (2005). Harrison reprend dans le détail les arguments que ces auteurs invoquent pour remettre en cause la validité de l'approche dite « standard » en économie. D'une manière similaire à Gul et Pesendorfer, Harrison montre que cette approche qualifiée de standard n'est nullement remise en cause, celle-ci postulant simplement que les choix observés révèlent les préférences: « *Non, ce n'est pas du tout ce que les économistes supposent. Nous disons que les choix révèlent les préférences, c'est tout* » (Harrison, 2008-a, p.308). Une telle hypothèse est parfaitement compatible avec les diverses observations mises en avant par Camerer, Loewenstein et Prelec, notamment avec l'idée de préférence dépendante du contexte ((Harrison, 2008-a, p.305), l'existence d'une utilité pour la monnaie (Harrison, 2008-a, p.306), la distinction entre un système du vouloir et un système des préférences (Harrison, 2008-a, p.308) ou la spécificité des domaines d'expertise (Harrison, 2008-a, p.309).

Pour Harrison, la critique de l'économie standard par Camerer, Loewenstein et Prelec (2005) repose donc sur une vision « *grossière, caricaturale* » et fait figure de « *coup bas* » (Harrison, 2008-a, p.306). Harrison rejoint ici les position exprimées par Gul et Pesendorfer en affirmant qu'une approche fondée sur les préférences révélées n'est pas nécessairement invalidée par les diverses données neuronales issues des expériences de neuroimagerie.

Harrison considère en outre que la critique des neuroéconomistes est non seulement infondée dans son principe, c'est-à-dire qu'elle n'invalide pas du tout l'approche par les préférences révélées, mais qu'elle est surtout extrêmement naïve, car les divers phénomènes psychologiques soulignés par les neuroéconomistes « *sont bien entendu connus des économistes depuis des décennies* » (Harrison, 2008-a, p.308). Ce qui semble au premier abord révolutionnaire ne constituerait en fait qu'un ensemble d'observations triviales.

Produisant des résultats théoriques extrêmement faibles, la neuroéconomie ne doit son succès, selon Harrison, qu'au talent extra-scientifique de ses principaux représentants, qui réussissent à « *bien vendre* », dans la sphère universitaire cette jeune discipline. Harrison dénonce le « *marketing académique* » qui entoure la neuroéconomie, ce qui, encore une fois, renvoie aux reproches courants concernant la « *rhétorique* » de la neuroéconomie (*cf.* chapitre 1, Mäki, 2011). Or en dehors de ces « *distractions* » ou du côté séduisant des images du cerveau en trois dimensions, force est de constater, selon Harrison, que les avancées théoriques sont pour l'instant très faibles (Harrison, 2008-a, p.303).

L'originalité de la critique de Harrison apparaît plus clairement dans la deuxième partie de son article, dans laquelle il se demande si les expériences de neuroimagerie peuvent, ou non, produire de la « *bonne économi*e » (Harrison, 2008-a, p.311). Harrison y examine dans le détail les théories et les interprétations proposées par les neuroéconomistes. Il identifie deux principaux problèmes méthodologiques. Harrison fait d'abord valoir que les neuroéconomistes manifestent une profonde méconnaissance des enjeux théoriques liés aux jeux économiques qu'ils utilisent dans leur expérience. Pour Harrison, les économistes comportementalistes et expérimentaux ont développé un ensemble important de réflexions relatives à l'interprétation des comportements observés dans de tels jeux ou protocoles de choix. Par exemple, un comportement coopératif dans le cadre du jeu de l'investisseur est souvent interprété par les neuroéconomistes comme l'indice de la « confiance » du participant envers ses partenaires. Toutefois, les économistes travaillant sur ce jeu ont depuis longtemps mis en évidence plusieurs artefacts ou confusions *(confounds*) possibles; un joueur coopératif n'est pas nécessairement confiant, et ses motivations peuvent être très diverses: il peut être motivé par l'envie d'infliger un maximum de pertes à l'expérimentateur, par de l'altruisme , il peut avoir un goût pour le risque plus prononcé, il peut envisager le jeu comme étant répété bien qu'il ne le soit pas, *etc.* (Harrison, 2008-a, p.319).

Pour Harrison, la plupart des expériences en neuroéconomie sont mal conçues et négligent ces distinctions. Il faudrait concevoir des protocoles permettant de contrôler

spécifiquement chacun de ces artefacts afin de s'assurer, dans le cas du jeu de l'investisseur par exemple, que la coopération soit bien l'indice d'une plus grande confiance. Harrison exprime des reproches similaires à propos des expériences portant sur le choix inter-temporel, en référence notamment à l'expérience de McClure *et al.* (2004) décrite dans le chapitre précédent (cf. chapitre 4, section II).

A ces difficultés liées à l'interprétation des comportements s'ajoutent des problèmes concernant les inférences statistiques utilisées par les neuroéconomistes. La production de données neuronales par les techniques de neuroimagerie s'appuie en effet sur une série de simplifications statistiques. Celles-ci ne font que trop rarement l'objet d'une exposition détaillée dans les compte-rendus d'expériences. Pour Harrison, les techniques d'échantillonnage auxquelles les neuroscientifiques font appel font figure de véritable « *usine à gaz* »[103] (Harrison, 2008-a, p.312). D'un point de vue statistique, les expériences de neuroimagerie soulèvent d'abord un premier problème relatif à la faible taille des échantillons d'individus observés, qui ne dépassent que très rarement la douzaine (Harrison, 2008-a, p.315). En économie, les expériences, plus simples à mettre en œuvre, peuvent faire appel à un nombre bien plus important de sujets, ce qui permet d'obtenir des résultats plus robustes.

Le deuxième problème concerne la manière avec laquelle les données individuelles fournies par l'imagerie sont agrégées pour produire une image du cerveau dit « moyen » (problème dit du *pooling accross brain*) (Harrison, 2008-a, p.314). Harrison affirme que les simplifications statistiques conduisent à surestimer les seuils de significativité. En effet, cette phase de normalisation des données, aboutissant à la construction abstraite d'une cerveau moyen, minimise les variations individuelles et permet d'identifier plus facilement des zones spécifiquement activées dans la tâche. Enfin, Harrison regrette la pratique, en neurosciences, qui n'oblige pas les chercheurs à publier l'intégralité de leurs données (Harrison, 2008-a, p.315). Les compte-rendus d'expérience, en neurosciences, ne sont que très rarement exhaustifs, et il faut aller voir dans des documents annexes, souvent disponibles en ligne, les informations concernant les simplifications statistiques utilisées.

Harrison insiste donc sur deux difficultés méthodologiques principales. D'une part, les neuroéconomistes s'appuient sur des simplifications statistiques pour mettre en évidence des zones cérébrales qui seraient spécifiquement activées dans une tâche expérimentale donnée. Le problème ici concerne la production des données neuronales (activation des différentes zones du cerveau). Le deuxième type de difficulté est de nature interprétative: ces données

103Pour être tout à fait exact, en anglais, l'expression littérale utilisée par Harrison est celle d'« *usine à saucisse* » (*sausage making factory*) (Harrison, 2008-a, p.312)

sont associées à des fonctions cognitives ou des états mentaux supposés être mis en œuvre dans la tâche; or cette tâche expérimentale est acceptée sans réserve comme étant révélatrice d'une fonction cognitive déterminée (actualisation temporelle, confiance, *etc.*), en négligeant les possibles confusions pouvant donner lieu l'interprétation comportementale de ces protocoles. Pour Harrison, la neuroéconomie n'aboutit ainsi qu'à « *une mixture de preuves statistiques ad hoc et de fiction interprétative, présentée comme vérité ou connaissance* » (Harrison, 2008-a, p.304). Pour Harrison, ces « *fictions interprétatives* » sont d'autant plus trompeuses qu'elles se présentent comme « *vérité ou connaissance* »: en effet, l'inférence du neuronal au cognitif semble corroborée dans les travaux de neurosciences par d'autres études portant sur des zones du cerveau similaires. Ce problème est connu sous le nom d'« *inférence inverse* » (Harrison, 2008-b, p.535; Poldrack, 2006, p.39)

4. Le problème de l'inférence inverse

Dans les expériences de neuroéconomie, les activités neuronales ou cérébrales sont associées à des fonctions cognitives ou à des états mentaux déterminés. Cette approche suppose d'abord que le protocole utilisé permette effectivement de simuler la fonction en question; en d'autres termes, la tâche ne doit pas générer d'artefacts (*cf. supra*, Harrison, 2008-a, p.304). En second lieu, il est nécessaire de s'assurer que l'activité obtenue soit bien liée à la fonction étudiée. La zone du cerveau qui s'active à l'IRM_f peut très bien n'avoir aucun rapport avec l'engagement du processus cognitif à l'œuvre dans l'accomplissement de la tâche. C'est donc l'inférence des états mentaux ou des fonctions cognitives à partir des activités neuronales qui pose problème.

Comme le souligne Harrison, cette difficulté est souvent contournée par les neuroéconomistes, au prix d'un raisonnement fallacieux connus sous le nom d'« *inférence inverse* » (Harrison, 2008-b, p.535). Celle-ci consiste à s'appuyer sur la mise en évidence, dans une ou plusieurs études plus anciennes, de l'activation d'une région du cerveau lors de l'engagement d'un processus cognitif donné pour inférer, « en sens inverse », que l'activité présente de cette même région du cerveau représente l'engagement du processus cognitif en question (Bourgeois-Gironde, 2010, p.232). Ce raisonnement peut être reconstruit sous la forme suivante (*cf.* Poldrack, 2006, p.39):

(1) Dans cette étude, la zone du cerveau Z était active dans la tâche A.

(2) Dans d'autres études, lorsque le processus cognitif X était engagé, la zone Z était active

(3) Par conséquent, l'activité de la zone Z dans cette étude démontre la mise en œuvre du processus cognitif X dans la tâche A

Pour que ce raisonnement soit valide, il faudrait, comme le souligne Bourgeois-Gironde (2010, p.235), que la proposition (2) affirme une stricte équivalence, du type: « dans d'autres études, le processus cognitif X était engagé si et seulement si la zone Z était active ». Toute activité de la zone Z étant équivalente à l'engagement du processus cognitif X, la conclusion de la proposition (3) serait alors valide. Cela ne correspond cependant pas à ce qui a été observé dans les autres études, qui montrent que « lorsque le processus X est engagé, la zone Z est active ». Par conséquent, l'activité de la zone Z dans la tâche A ne représente pas forcément l'engagement du processus X, mais peut être celui d'un processus X', ou X", sans que cela n'aille à l'encontre des précédentes études, puisqu'une même région peut très bien-être associée à plusieurs fonctions X, X', X"... C'est la raison pour laquelle l'inférence est dite inverse. Il y a une inversion de la démarche qui, en neurosciences, consiste à localiser des fonctions cognitives dans des zones du cerveau (proposition 2), pour déduire de l'activité de ces zones à l'imagerie l'engagement de ces fonctions préalablement définies.

Bourgeois-Gironde propose d'assimiler l'inférence inverse à une abduction, au sens proposé par Peirce (1878). Une abduction peut ce comprendre comme une tentative pour conclure A de l'observation de B, étant donné que A implique B (Bourgeois-Gironde, 2010, p. 235). Ce raisonnement peut être utile, dans les sciences, pour générer des hypothèses, mais il ne permet pas d'établir des propositions certaines. De la même manière, dans les inférences inverses, l'appui sur les autres études peut servir à proposer des hypothèses interprétatives, mais ces interprétations demandent à être confirmées. Il faut en particulier s'assurer du bon niveau de description de la fonction X. Celle-ci doit être suffisamment générale pour inclure l'ensemble des sous-fonctions X', X",... prises en charge par la région Z, tout en conservant une pertinence interprétative. Il doit y avoir, en d'autres termes, une correspondance « un pour un »: la zone Z doit spécifiquement s'activer lors de l'engagement de la fonction X.

En négligeant ce travail de définition de ce que Bourgeois-Gironde appelle les « *ontologies cognitives* » (Bourgeois-Gironde, 2010, p.229), la plupart des études de neuroéconomie reposent sur une stratégie dite de la « même zone » (Bourgeois-Gironde, 2010, p.245). Une fonction cognitive est inférée de l'activité cérébrale observée, à la lumière

des autres études qui portent sur la même zone. Or, cette approche représente une inversion de la démarche des neurobiologistes qui, au départ, essayent au contraire de comprendre comment une fonction cognitive donnée, définie dans le cadre d'un protocole précis, est « prise en charge » dans le cerveau. Pour Harrison, le raisonnement interprétatif mise en œuvre par la neuroéconomie, remontant du cérébral au cognitif, représente ainsi une utilisation dévoyée des neurosciences, faisant de celles-ci une « *néo-phrénologie* » dont les hypothèses de localisation sont aussi fantaisistes qu'insuffisamment fondées[104]

Les critiques de la neuroéconomie partent généralement de l'existence d'un écart entre les concepts de l'économie et ceux de la psychologie et des neurosciences. Gul et Pesendorfer (2005) défendent une stricte étanchéité entre les disciplines. D'autres économistes (Rubinstein, 2008; Harrison, 2008-a) considèrent au contraire que leur discipline -l'économie- peut se doter de fondements psychologiques plus solides. Toutefois, il est alors délicat de justifier l'utilisation des données neuronales. D'une part, il existe des indicateurs comportementaux plus simples, moins couteux et plus adaptés aux besoins des économistes (Rubinstein, 2008). Surtout, les variables neurophysiologiques ne semblent pouvoir être pertinentes d'un point de vue cognitif qu'au prix d'hypothèses de localisation des fonctions cognitives dans le cerveau, qui soulèvent de lourds problèmes interprétatifs et statistiques (Harrison, 2008-a). La théorie des marqueurs somatiques de Damasio, appuyée sur l'interprétation de l'*Iowa Gambling Task* fournit une bonne illustration de ces difficultés méthodologiques.

104« *les neuroéconomistes sont véritablement les nouveaux phrénologistes, puisqu'ils se reposent essentiellement sur des hypothèses portant sur la modularité des processus cognitifs dans le cerveau. Mise à part quelques modifications mineures de l'ancienne craniologie avec l'activité du cerveau, les hypothèses fondamentales concernant la modularité restent essentiellement les mêmes que dans la phrénologie* »(Harrison, 2008-b, p.534)

B. Les critiques de la théorie des marqueurs somatiques et de l'*Iowa Gambling Task*

Les critiques méthodologiques de la neuroéconomie ciblent d'abord les difficultés liées à l'articulation de l'économie comportementale et des techniques de neuroimagerie. Un deuxième courant de critiques porte spécifiquement sur les travaux d'un auteur qui a eu une large influence sur la littérature, Antonio Damasio. Toutefois, les problèmes méthodologiques concernant la théorie des marqueurs somatiques et l'*Iowa Gambling Task* peuvent aussi se comprendre comme une illustration des critiques de l'économie comportementale, en tant qu'ils témoignent des difficultés générales liées à l'application en neurosciences de paradigmes d'inspiration économique.

L'*Iowa Gambling Task* (IGT) et la théorie des marqueurs somatiques constituent respectivement un protocole de laboratoire et un modèle interprétatif qui offrent une théorie assez large du comportement à partir d'une hypothèse de localisation d'une fonction cognitive dans une région du cerveau, le cortex préfrontal (*cf.* chapitre 4). Les travaux de Damasio, associés à la notion d'émotion, rencontrent un large succès ; mais, à partir des années 2000, ce programme de recherche fait face à d'importantes critiques au sein des neurosciences. L'apparente simplicité de la théorie des marqueurs somatiques masque d'importants problèmes méthodologiques, relatifs en particulier à l'interprétation des comportements dans l'IGT. Ces difficultés peuvent valoir comme des exemples des « *fictions interprétatives* » dénoncées par Harrison (2008-a).

Après avoir été considéré comme un protocole de test de l'impulsivité fiable et innovant (cf. chapitre 4), l'IGT a peu à peu fait l'objet d'importantes critiques quant à la significativité des résultats obtenus[105]. Certains éléments du protocole apparaissent d'abord complètement arbitraires. Le choix d'utiliser 4 piles de cartes, plutôt qu'un nombre plus ou moins élevé, n'est pas justifié. Kerr et Zelazo (2004), dans une étude intitulée « le développement des fonctions exécutives : le pari chez l'enfant », proposent une version modifiée de l'IGT avec deux piles de cartes. Avec ce protocole, les auteurs parviennent à établir une différence dans le comportement des enfants de 3 ans et celui des enfants de 4 ans (Kerr et Zelazo, 2004). L'IGT, comme les autres tâches de machines à sous multi-jeux qui ont été abordées ici, pourraient donc très bien n'utiliser que deux machines (tas de cartes) .

Le principal problème de l'IGT concerne l'interprétation du comportement des joueurs

105Pour une synthèse complète des critiques concernant l'IGT, voir Ross *et al.*, 2008, p.97.

qui persistent à choisir des cartes parmi les deux piles désavantageuses. En effet, comme la distribution des gains est initialement inconnue et doit être découverte, une mauvaise performance à l'IGT peut refléter une incapacité à découvrir les probabilités de gains à long terme associées à chaque pile. Cette incapacité peut s'expliquer par exemple par des difficultés de mémorisation et d'apprentissage. Mais une mauvaise performance à l'IGT peut aussi s'interpréter comme un goût pour le risque plus élevé : l'individu, ayant compris que certaines piles étaient plus risquées que les autres, continue à choisir ces piles pour la simple raison qu'il préfère les jeux risqués.

La théorie des marqueurs somatiques favorise plutôt la première hypothèse : pour Damasio et ses co-auteurs, les patients atteints de lésions frontales sont incapables, non pas de ressentir des émotions liées au gain ou à la perte subie, mais de mémoriser et d'anticiper au moment du choix suivant l'occurrence de ces émotions (Bechara, Tranel et Damasio, 2000). Une telle explication suggère donc une interprétation en termes d'apprentissage et de mémorisation. Or cette hypothèse est invalidée par des travaux plus récents. Brand *et al.*, 2005 proposent en particulier une variante de l'IGT, appelée « jeu de dés », dans laquelle les probabilités de gains et de pertes sont connues à l'avance[106]. Les patients atteints de lésions préfrontales persistent également, dans ce jeu, à choisir les options défavorables (Brand *et al.*, 2005). Ce résultat est problématique car il suggère que les patient lésés ont soit des difficultés à raisonner de manière abstraite sur les probabilités, ce qui va à l'encontre de la thèse d'une incapacité affective postulée par la théorie des marqueurs somatiques (*cf.* Damasio, 1994), soit plus simplement un gout pour le risque plus élevé que les sujets normaux. Il ne s'agirait donc pas dans les deux cas d'un problème de planification temporelle des émotions, au sens retenu par Bechara, Tranel et Damasio (2000), c'est-à-dire d'une incapacité à prendre en compte les conséquences affectives futures des décisions présentes (*cf.* chapitre 4).

Une expérience de 2008 indique que la performance des sujets dans l'IGT est en fait déterminée par la fréquence d'obtention des récompenses. Chiu *et al.* (2008) ont construit une variante de l'IGT, appelée « Soochow Gambling Task », dans laquelle les gains et les pertes respectivement associées à chaque pile de carte sont identiques à ceux utilisés par Damasio et ses co-auteurs, mais les fréquences d'obtention de ces gains et pertes sont différentes :

106Au lieu d'utiliser des cartes, le jeu repose sur des dés. A chaque essai, le joueur doit deviner le chiffre du dé. Il peut proposer un chiffre, deux chiffres, trois chiffres, ou quatre chiffres. En ne proposant qu'un seul chiffre, le joueur choisit le pari le plus risqué : la probabilité de gagner est de 1/6. S'il gagne (perd), il obtient (perd) 1000. Les gains/pertes respectifs pour les paris moins risqués (deux, trois et quatre chiffres) sont de 500, 200, 100. Les gains/pertes espérées sont donc relativement simple à calculer, et un joueur souhaitant maximiser son espérance de gain choisira toujours le pari le moins risqué, qui est le seul à offrir une espérance de gain positive (*cf.* Brand *et al.*, 2005, p.97)

Iowa Gambling Task	Fréquence de gains (G) et de pertes (L) pour 10 cartes	Valeur espérée (EV) pour 10 cartes	Prédiction du choix à partir de la fréquence de gains et de pertes	Prédiction du choix à partir de la valeur espérée
A (Bad)	5 G 5 L	-$250		
B (Bad)	9 G 1 L	-$250	B	
C (Good)	6.25 G 1.25 L 2.5 S	+$250	C	C
D (Good)	9 G 1 L	+$250	D	D
Soochow Gambling Task *	(pour 5 cartes)	(pour 10 cartes)		
A (Bad)	4 G 1 L	-$250	A	
B (Bad)	4 G 1 L	-$250	B	
C (Good)	1 G 4 L	+$250		C
D (Good)	1 G 4 L	+$250		D

Chiu *et al.*, 2008, p.4

Dans l'IGT, parmi les deux piles désavantageuses (*A* et *B*), la pile *A* offre des gains moins fréquents (une fois sur deux) que toutes les autres piles. La pile *B* est désavantageuse mais la fréquence de gain est élevée. Pour Chiu *et al.* (2008), le protcole de l'IGT est mal conçu car on ne peut pas vraiment savoir si les sujets améliorent leur décision en prenant en compte la valeur espérée de gain à long terme associée à chacune des piles (*expected value*), ou si c'est la fréquence d'obtention des gains qui détermine leurs choix. Les joueurs pourraient très bien apprendre à choisir *C* et *D* simplement parce que ces options offrent plus souvent des récompenses. En outre, la pile *B* est ambiguë car elle est désavantageuse mais les gains associés y sont fréquents : si les joueurs sont sensibles aux fréquences plutôt qu'aux gains cumulés à long terme, un sujet normal peut persister à choisir une pile désavantageuse (la pile *B*). Cette hypothèse semble plausible, puisque certains joueurs se concentrent exclusivement sur la pile *B* (Chiu *et al.*, 2008, p.2).

Afin d'éliminer ces différents artefacts dans l'interprétation de l'IGT, Chiu *et al.* proposent une matrice de gains et de pertes légèrement modifiée. Dans le *Soochow Gambling Task* (SGT), les gains espérés à long terme associés à chaque pile sont les mêmes que dans l'IGT, mais les deux piles désavantageuses offrent des gains plus fréquents que les piles avantageuses. Dans cette version modifiée, les auteurs observent que les sujets normaux, ayant des bonnes performances à l'IGT, ne parviennent pas à identifier la stratégie gagnante à long-terme (Chiu *et al.*, 2008, p.4). En faisant référence aux travaux de Richard Herrnstein (1974) et Howard Rachlin (1991), Chiu *et al.* suggèrent que les joueurs suivent en fait des stratégies consistant à reproduire un choix ayant assuré un gain, et à changer d'option suite à une perte. Cette heuristique, qui renvoie à la notion de « mélioration » chez Herrnstein (*cf.*

chapitre 2) permet d'expliquer pourquoi la fréquence de gains fait varier la performance des sujets à l'IGT et au SGT.

Chiu *et al.* (2008) considèrent que leur étude invalide la théorie des marqueurs somatiques de Damasio :« *dans l'IGT, les sujets normaux se tournent progressivement vers les bonnes piles […]. Nous proposons que cet apprentissage n'est pas du à une espérance de gain à long-terme plus élevée mais est plutôt l'effet de gains plus fréquents que les pertes. La théorie des marqueurs somatiques souligne que les marqueurs somatiques […] prédisposent les sujets normaux à se comporter en accord avec les conséquences perçues de leurs actions à long terme. Cette étude montre que même les sujets ayant des marqueurs somatiques intacts ne parviennent pas à se comporter en accord avec la recherche d'un gain à long terme dans le SGT. Les récompenses immédiates dépassent l'effet du gain à long terme dans le Soochow Gambling Task* » (Chiu *et al.*, 2008, p.6).

L'IGT et le SGT sont pourtant des tâches type machines à sous multi-jeux, similaires à celles utilisées par la science quantitative de la motivation pour étudier les dynamiques d'apprentissage. Il est donc raisonnable de considérer que les *multi armed bandit problems* peuvent servir à établir un lien, entre la fréquence des gains, les comportements de mélioration et les capacités de mémorisation et d'apprentissage (*cf.* l'étude de Bogacz, McClure, Li, Cohen et Montague (2007) abordée dans le chapitre 4). Or, à la lumière des recherches néo-comportementales, qui ont précisément porté sur le choix répété, la motivation, et la planification temporelle des décisions, il apparait que la théorie des marqueurs somatiques constitue un schéma interprétatif de l'IGT beaucoup trop indéterminé. Les thèses de Damasio sur les émotions ne fournissent en fait pas d'hypothèse descriptive des comportements très précises. En particulier, il y a une ambiguïté quant au fait de savoir si la projection émotionnelle qui intervient au moment du choix porte sur le gain ou la perte immédiate ou à long terme. Les deux interprétations sont également plausibles parce que Damasio et ses co-auteurs ne proposent pas de formalisation des dynamiques d'apprentissage.

L'indétermination de la théorie des marqueurs somatiques renvoie aussi à une mauvaise conception de l'IGT lui-même. Cette tâche est en effet censée mesurer les capacités d'apprentissage émotionnel des individus, mais elle incite en fait à jouer au coup par coup, en fonction des gains et des pertes à court terme. Rien n'indique que les joueurs normaux jouent réellement en fonction des gains espérés à long terme (*cf.* Chiu et al., 2008). Par ailleurs, les résultats initiaux obtenus par Bechara *et al.* (1994) ne sont pas très robustes : dans l'étude de Glicksohn, Naor-Ziv et Leshem (2007), environ la moitié des sujets sains (46%) affichent de

mauvaises performances à l'IGT. Les auteurs suggèrent qu'il existe d'importantes variations dans les performances individuelles à l'IGT, et qu'il convient de regarder les données de chaque joueur plutôt que de parler de norme stable du comportement (Glicksohn, Naor-Ziv et Leshem, 2007, p;195)

Conçu initialement pour détecter certains troubles comportementaux, l'IGT constitue en outre un critère de diagnostic assez peu fiable. Les performances à l'IGT ne sont pas corrélées de manière claire avec les performances dans d'autres tests comportementaux, et en particulier avec la mesure des coefficients d'actualisation temporelle[107]. Cela explique pourquoi l'IGT a progressivement été abandonné comme jeu paradigmatique dans le cadre des théories neuroéconomiques de l'addiction (*cf.* section IV). Par ailleurs, Damasio a considéré que sa théorie associée à l'IGT permettait d'expliquer non seulement les troubles de l'impulsivité, mais, plus généralement, tous les déficits fonctionnels liés aux « émotions ». L'indétermination du terme d'émotion a donc autorisé Damasio à suggérer par exemple une application de la théorie des marqueurs somatiques à la psychopathie (Damasio, Tranel et Damasio 1990), qui demeure encore une fois beaucoup trop vague pour caractériser de manière pertinente sur le plan comportemental les psychopathes[108].

La théorie des marqueurs somatiques rencontre donc, comme l'économie comportementale dans le scanner, des problèmes méthodologiques liés à l'identification de fonctions cognitives dans le cerveau. En dénonçant l'imprécision des hypothèses explicatives utilisées, les critiques envisagées dans cette section remettent en cause les localisations

107La plupart des études qui visent à établir une corrélation entre la performance à l'IGT et l'actualisation temporelle débouchent sur des résultats contradictoires. Harmsen *et al.* (2005) identifient par exemple un lien de corrélation négatif en montrant que les fumeurs, qui se distinguent généralement par une plus forte actualisation temporelle des récompenses, ont des performances plus élevées à l'IGT ; mais Rotheram-Fuller *et al.* (2004) aboutissent à une conclusion inverse en observant des sujets sous méthadone. L'étude de Evans, Kemsih et Turnbull (2004) établit une corrélation inverse entre performance à l'IGT et le niveau d'éducation (ce qui pourrait paraître contre-intuitif) ; pourtant, Jaroni *et al.* (2004) montrent que chez les fumeurs (qui actualisent de manière plus importante les récompenses futures), un niveau d'éducation plus élevé accroit les capacités de *self control*. Enfin, si les performances à l'IGT s'améliorent avec l'âge dans l'expérience de Hooper *et al.* (2004), certains sujets voient leur performance à l'IGT diminuer en vieillissant (Denburg, Tranel et Bechara, 2005)

108Les marqueurs somatiques permettant de marquer un scénario comme bon ou mauvais, l'idée ici consiste à suggérer que les psychopathes manifestent une incapacité à ressentir les émotions négatives associées à l'accomplissement d'actes anti-sociaux (meurtres, viols, *etc.*) et qui, chez les sujets sains, remplissent une fonction d'inhibition de ces conduites. Encore une fois, cette hypothèse demeure beaucoup trop indéterminée. Elle ne rend pas compte de plusieurs traits distinctifs du comportement des psychopathes mis en évidence dans d'autres études plus approfondies, en particulier l'absence de distinction entre règles morales et conventionnelles. Par ailleurs, la théorie des marqueurs somatiques ne permet pas de dissocier l'agressivité instrumentale, spécifique aux psychopathes, de l'agressivité réactive, ni d'expliquer pourquoi les psychopathes expriment des réactions émotionnelles à certains types de stimuli aversifs. Pour une critique complète de la théorie des marqueurs somatiques appliquée à la psychopathie, voir Blair, 2005, pp.93-98.

cérébrales mises en évidence par les neuroéconomistes. Même s'ils se posent effectivement dans la pratique de la recherche, ces problèmes méthodologiques ne sont cependant pas insurmontables. En particulier, les résultats théoriques issus de la neuroimagerie reçoivent une confirmation dans l'étude de cas pathologiques.

II. La construction d'un paradigme autonome (1): la référence à la pathologie

En neurophysiologie, l'introduction de nouvelles techniques expérimentales s'appuie toujours sur des connaissances anatomiques antérieures[109]. Contrairement à ce qu'affirme Rubinstein par exemple, en neurosciences, « *les diagrammes en couleurs* » du cerveau ne sont pas nécessairement acceptées par elle-mêmes comme des « *preuves indiscutables* » (Rubinstein, 2008, p.486). C'est parce qu'elles sont corroborées par d'autres types de preuves convergentes que les données de la neuroimagerie permettent d'identifier des fonctions cognitives dans le cerveau. De la même manière, il est possible de faire faire valoir qu'un tel processus de « *triangulation* » des méthodes d'observation a permis au microscope électronique de bénéficier de la confiance et des connaissances produites à l'aide du microscope optique (Guala, 2005, p.125-126).

L'enjeu dans cette section consiste à montrer que l'observation de cas pathologique fournit aux neuroéconomistes un moyen effectif de « *triangulation* » des résultats de la neuroimagerie. L'enthousiasme initial suscité par le potentiel révolutionnaire de l'IRM fonctionnelle (*cf.* Camerer, Loewenstein et Prelec, 2005) doit être -et a été- tempéré, mais les critiques de cette nouvelle technique méritent également d'être nuancées[110]. L'interprétation comportementale des patients atteints de lésions du cortex préfrontal dans le cadre de l'IGT et de la théorie des marqueurs somatiques soulève de nombreuses difficultés. En revanche, le paradigme du *reward learning* fournit un outil de diagnostic précis et utile pour caractériser sur le plan neuronal diverses pathologies liées à des troubles de l'impulsivité, dans le cadre de ce que Rangel, Camerer et Montague (2008) appellent la « *neuropsychiatrie computationnelle* » (A). En retour, l'observation de cas pathologiques permet d'établir des

109Cela ne signifie pas pour autant que les nouvelles techniques d'observation servent uniquement à confirmer des théories neuro-anatomiques antérieures. L'introduction des micro-électrodes s'est bien appuyée sur un ensemble de connaissances anatomiques concernant le système dopaminergique; pour, autant, les travaux de Glimcher ou de Schultz ont bien été à l'origine d'une importante transformation théorique, en substituant au cadre théorique stimulus-réflexe le paradigme du *reward learning*.

110La littérature critique se distingue elle-aussi par une dimension rhétorique, au sens proposé par McCloskey (1983) (*cf.* introduction). L'article de Harrison (2008-a) en particulier abonde en artifices langagiers. Considérant par exemple les procédures statistiques utilisées dans le traitement des données de neuroimagerie, Harrison parle d' « *économétrie de MacGyver* », du nom d'un célèbre héros de série télévisé célèbre pour son ingéniosité à toute épreuve, pour critiquer l'absence de justification dans la résolution des problèmes de représentativité (Harrison, 2008-a, p.314). Harrison multiplie également les métaphores et les comparaisons, affirmant ainsi que « *les scanners du cerveau ne peuvent pas s'allumer comme des sapins de Noël sans faire de très nombreuses hypothèses de modélisation* » (Harrison, 2008-a, p.314).

correspondances plus solides entre les activités cérébrales et les fonctions cognitives. Y compris dans les études qui ne portent pas directement sur des sujets atteints de troubles comportementaux, la référence à la pathologie et à la connaissance clinique est implicite, et tempère les difficultés méthodologiques soulevées par les inférences inverses (B).

A. Le paradigme du *reward learning* comme outil de diagnostic

La fiabilité de l'IGT comme outil de diagnostic comportemental, pourtant défendue par Damasio (2008), a été remise en question au cours des dernières années. En effet, la performance individuelle à l'IGT peut faire l'objet de diverses interprétations; en outre, elle n'est pas corrélée avec des autres indicateurs comportementaux de l'impulsivité (*cf. supra*). En revanche, le paradigme du *reward learning*, tel qu'il a été mis en œuvre dans les premières expériences de neuroéconomie au début des années 2000 (*cf.* chapitre 4) fournit une méthode plus précise pour détecter diverses pathologies comportementales associées à des troubles de l'impulsivité. Deux études récentes (Pessiglione *et al.*, 2006; Voon *et al.*, 2010) sont analysées ici pour illustrer la manière avec laquelle le concept d'apprentissage de la récompense peut servir à de caractériser sur le plan médical le comportement de sujets impulsifs.

Rangel, Camerer et Montague (2008) considèrent que « *le domaine d'application le plus important pour la neuroéconomie est la psychiatrie* » (Rangel, Camerer et Montague, 2008, p.554). La connaissance des processus neuropsychologiques liés à l'apprentissage de la récompense autorise en particulier, selon les auteurs, la constitution d'une « *neuropsychiatrie computationnelle* » (Rangel, Camerer et Montague, 2008, p.554). Cette méthode expérimentale en psychiatrie consiste à formaliser des algorithmes d'apprentissage individuels. Les paramètres de ces algorithmes sont utilisés comme outils de diagnostic. La corrélation de ces variables comportementales avec les données de la neuroimagerie permettrait en outre d'identifier les régions du cerveau responsable de cet apprentissage, et, éventuellement, d'agir sur ces zones par voie médicamenteuse.

Dans une étude intitulée « les signaux dopaminergiques d'erreur de prédiction sous-tendent la recherche de la récompense chez l'homme », Pessiglione *et al.* (2006) montrent

ainsi que le signal d'erreur de prédiction de la récompense exprimé dans le striatum est susceptible d'être modulé par des médicaments qui stimulent (L-Dopa) la production de dopamine ou au contraire diminuent (haloperidol) la sensibilité des récepteurs à la dopamine (Pessiglione *et al*., 2006). Cette expérience porte sur un *multi-armed bandit problem* tout à fait classique. A chaque essai, les sujets doivent choisir entre deux machines à sous. La première est associée à une forte probabilité (80%) de gain (1 livre sterling) et à une faible probabilité (20%) de gain nul. Les probabilités associées à la seconde machine (option) sont inversées (20% de chance d'obtenir un gain, 80% de chances d'obtenir un gain nul). Les joueurs participent à trois blocs d'essais distincts. Dans la condition de contrôle, les gains sont toujours nuls. Dans la condition « gain », la structure des récompenses correspond à la description ci-dessus. Dans la condition « perte », les gains sont remplacés par des pertes symétriques (1 livre sterling) (Pessiglione *et al.*, 2006, p.1042).

Les probabilités de récompense et de punition associées à chaque option étant initialement inconnues, les sujet doivent apprendre à sélectionner toujours l'option à forte probabilité de gain dans la condition de gain, et à éviter l'option à forte probabilité de perte dans la condition de perte. Les résultats comportementaux montrent qu'il faut environ une trentaine d'essais en moyenne pour que les sujets découvrent la meilleure stratégie (toujours choisir l'option à forte probabilité de gain dans la condition de gain, toujours choisir l'option à faible probabilité de pertes dans la condition de perte) Les résultats sont similaires pour les deux conditions -gain ou perte- mais, les temps de réaction moyens sont plus longs dans la condition de perte (Pessiglione *et al.*, 2006, p.1043).

Dans la deuxième partie, « pharmacologique », de l'étude, les expérimentateurs administrent aux 39 sujets des médicaments agissant sur le système dopaminergique. Deux groupes de 13 sujets reçoivent d'abord un traitement au L-Dopa ou haloperidol. Le troisième groupe reçoit un traitement « placebo ». Les sujets ayant reçus le traitement au L-Dopa gagnent significativement plus d'argent (c'est-à-dire qu'ils découvrent plus vite la stratégie gagnante) que les sujets traités au haloperidol et que le groupe placebo dans la condition de gain. En revanche, les sujets traités au L-Dopa n'apprennent pas plus vite à choisir l'option à faible probabilité de perte (Pessiglione *et al.*, 2006, p.1043)

A partir de ces résultats comportementaux, les auteurs établissent des algorithmes permettant de rendre compte des dynamiques d'apprentissage pour chaque sujet. Les deux paramètres principaux de ces algorithmes -température et coefficient d'apprentissage- sont ensuite utilisés dans l'analyse des données de la neuroimagerie. L'activité du striatum est corrélée avec l'erreur de prédiction de la récompense, à la fois dans les conditions de gains et

de pertes[111](Pessiglione *et al.*, 2006, p.1043).

En croisant des techniques pharmacologiques, de neuroimagerie et d'analyse comportementale, cette étude illustre les possibles applications du paradigme du *reward learning* au domaine thérapeutique: « *les résultats de cette étude donnent une idée du type de troubles cliniques dans lesquels la dopamine est impliquée, et pour lesquels le L-Dopa et le haloperidol sont utilisés comme agents thérapeutiques, en particulier dans la maladie de Parkinson et la schizophrénie* » (Pessiglione *et al.*, 2006, p.1044). La tâche de machine à sous multi-jeux utilisée dans cette étude permet en effet de formaliser un algorithme d'apprentissage correspondant à la performance moyenne des sujets normaux, et, par suite, d'identifier les modifications possibles de cet algorithme par des substances agissant sur l'ampleur de l'erreur de prédiction. La neuroimagerie offre en outre la possibilité de mettre en évidence et de cibler précisément les zones du cerveau impliquées dans cette tâche cognitive d'apprentissage. D'un point de vue clinique, le concept d'apprentissage de la récompense constitue ainsi une grille d'interprétation commode et suffisamment général pour regrouper des troubles jusqu'alors considérés comme disparates. En effet, les sujets traités au L-Dopa (en particulier dans le cas de la maladie de Parkinson) ou au haloperidol (dans le cas de la schizophrénie par exemple) manifestent souvent des dérèglements de l'impulsivité, mais qui se donnent à voir dans des formes très diverses: hypersexualité, *shopping* compulsif, addiction aux jeux financiers pour le L-Dopa; abattement, apathie, dépression pour les sujets traités au haloperidol. La quantification des troubles de l'apprentissage dans ces protocoles permettrait d'améliorer le dosage des traitements médicamenteux agissant sur le système dopaminergique.

Les pistes ouvertes par Pessiglione et ses co-auteurs sont approfondies dans une étude plus récente (Voon *et al.*, 2010). L'expérience porte ici sur des sujets parkinsoniens, ayant développé ou non des troubles compulsifs (addiction aux jeux financiers, dépenses compulsives) suite à leur traitement. Leur comportement est comparé à celui de sujets sains (Voon *et al.*, 2010, p.136). La tâche est identique à Pessiglione *et al.*, 2006. Les auteurs étudient sur les participants l'effet d'une substance agissant spécifiquement sur les récepteurs dopaminergiques, similaire au haloperidol. Les résultats montrent que ce médicament améliore l'apprentissage aux gains pour les parkinsoniens ayant développé des troubles compulsifs, mais pas chez les parkinsoniens n'ayant pas développé de tels troubles (Voon *et*

111Dans la condition de perte cependant, les auteurs identifient également une activité de l'insula droite antérieure, pour les erreurs de prédiction négatives. Cette observation confirme la spécificité des erreurs de prédiction négative dans la condition de perte, puisque les temps de réactions sont en général plus long. Sur l'asymétrie entre les erreurs de prédiction négatives et positives, on se reportera au développement dans le chapitre 4 portant sur l'aversion aux pertes (chapitre 4, section II).

al., 2010, p.139).

Ces deux études (Pessiglione *et al.*, 2006; Voon *et al.*, 2010) illustrent ainsi la manière avec laquelle le paradigme du *reward learning* peut, de façon pertinente pour les cliniciens, dissocier des formes de comportement pathologiques de conduites « normales ». L'application des recherches en neurosciences portant sur l'apprentissage de la récompense dans le domaine thérapeutique est à l'origine de ce que Rangel, Camerer et Monatgue appellent la « *neuropsychiatrie computationnelle* »: « *les modèles computationnels de l'apprentissage de la récompense fournissent un nouveau langage pour comprendre la santé mentale et un point de départ pour relier des substrats neuronaux déterminés à des résultats comportementaux. Le modèle du conditionnement instrumental permet de prédire l'existence de dysfonctionnements dans l'évaluation des récompenses, dans lesquels un médicament, une maladie ou un trouble du développement perturbe la capacité du cerveau à assigner une valeur approprié à des actions ou des états mentaux* » (Rangel, Camerer et Monatgue, 2008, p.554). En retour, cette approche médicale fondée sur l'observation de cas manifestement déviants permet aussi de corroborer les résultats de la neuroimagerie.

B. La connaissance clinique des pathologies comme « ontologie cognitive » au service de la neuroimagerie

Les instruments de neuroimagerie peuvent être mis au service d'une approche médicale, visant à évaluer des déficits neurocognitifs (*cf. supra*). Tous les travaux de neuroimagerie ne poursuivent pas un tel objectif. Cependant, même dans les travaux ne mobilisant pas de sujets traités ou lésés, la référence à la connaissance clinique est implicite. Celle-ci fournit une source de connaissance incontournable pour élaborer des « *ontologies cognitives* » adéquates (Bourgeois-Gironde, 2010, p.229), c'est-à-dire des correspondances pertinentes entre les activités cérébrales et les fonctions cognitives. Cette section a pour but de montrer que la psychopathologie améliore la qualité des inférences qui peuvent être établies des données neuronales aux fonctions cognitives, et nuance ainsi les critiques soulevées par le problème de l'inférence inverse.

Les études de neuropsychiatrie computationnelle ont pour but de caractériser sur le

plan neuronal des troubles cognitifs. Elles peuvent aussi en sens inverse servir à diagnostiquer un trouble à parti d'une activité cérébrale (1). L'interprétation cognitive des données cérébrales qui est alors proposée reçoit une justification dans la connaissance clinique préalable (2). Cette dernière sert généralement de référence implicite à tous les travaux de neuroimagerie. Elle permet en outre d'apprécier la performance des décideurs dans les différentes tâches à partir des critères cliniques de distinction entre sujets sains et pathologiques (3). La neuroéconomie représente donc un compromis entre le langage clinique et l'approche fonctionnelle et quantitative des troubles. Ce moyen-terme théorique offre une justification aux inférences inverses communément pratiquées par les neuroéconomistes (4).

1. Du cognitif au cérébral, et du cérébral au cognitif: le double niveau de lecture des études de neuropsychiatrie computationnelle

Les expériences de neuroéconomie reposant sur la différenciation de deux groupes de sujet (traités ou non-traités, « sains » ou atteints de pathologies comportementales diverses, *cf.* Pessiglione *et al.*, 2006; Voon *et al.*, 2010) poursuivent des objectifs ambivalents. D'un coté, ces études visent à définir un dysfonctionnement cognitif donné en termes fonctionnels, c'est-à-dire à le traduire en variables neuronales et comportementales. Dans les travaux précédemment abordés (Pessiglione *et al.*, 2006; Voon *et al.*, 2010) par exemple, les divers troubles de l'impulsivité manifestés par des sujets « pathologiques » sont formalisés par un algorithme d'apprentissage commun, dont les paramètres représentent à la fois les choix observés (variables comportementales) et les activités cérébrales. Le protocole et les outils de neuroimagerie fonctionnent alors comme un dispositif de diagnostic fonctionnel (*cf. supra*).

Cette démarche, qui correspond à l'approche la plus commune en neurosciences, consiste à caractériser une fonction cognitive donnée au niveau cérébral. Elle donne lieu à des observations du type « lorsque le processus X est engagé, la zone Z est active » (*cf.* section I, Poldrack, 2006). Mais la neuropsychiatrie computationnelle (*cf.* Rangel, Camerer, Montague, 2008, p.554) peut aussi se lire en sens inverse, comme une inférence d'un trouble cognitif à partir d'une activité cérébrale. Une fois que le protocole de diagnostic est établi, il doit en effet être possible par la suite de conclure, à partir de l'activité de la zone Z dans la tâche, au fonctionnement normal ou pathologique de X. Dans l'étude de Pessiglione *et al.* (2006), l'activité du striatum permet ainsi d'évaluer le degré d'impulsivité.

L'affirmation selon laquelle « *les résultats apportent un soutien à l'idée selon laquelle il existe un lien fonctionnel entre la dopamine, l'activité du striatum et la recherche de récompense chez l'homme* » (Pessiglione *et al.*, 2006, p.1042) dans cette même étude est donc susceptible de deux niveaux de lecture différents, puisque l'inférence est construite au départ du cognitif (la recherche de récompense) au cérébral (l'activité du striatum), mais peut ensuite être établie en sens inverse. Ce type d'étude peut ainsi se comprendre de la manière suivante:

1) connaissance clinique initiale: la pathologie P affecte le fonctionnement du processus cognitif X.
   exemple: les sujets traités au L-Dopa manifestent des troubles de l'impulsivité.

2) caractérisation fonctionnelle de la pathologie: dans la tâche A, les sujets atteints de la pathologie P se caractérisent par une activité cérébrale $Z_{pathologique}$ et un comportement $C_{pathologique}$.
   exemple: les sujets traités au L-Dopa, dans la tâche d'apprentissage de la récompense A, se distinguent par des gains inférieurs à la moyenne (variable comportementale) et par une activité du striatum supérieure à la moyenne (variable cérébrale)

3) Utilisation de la neuroimagerie comme outil de diagnostic: toute activité $Z_{pathologique}$ dans la tâche A implique un dysfonctionnement de X.
   exemple: toute activité du striatum supérieure à la moyenne dans la tâche A implique un trouble de l'impulsivité.

Il y a donc bien ici une forme -bien spécifique- d'inférence inverse (*cf.* Poldrack, 2006), puisqu'une activité $Z_{pathologique}$ dans la tâche A pourra être diagnostiquée comme un dysfonctionnement de X (proposition 3), alors qu'il est seulement établi au départ que P, qui implique un dysfonctionnement de X (proposition 1), entraine une activité $Z_{pathologique}$ dans cette même tâche (proposition 2). Au final, un dysfonctionnement cognitif -le trouble de l'impulsivité- est bien inféré de l'activité d'une zone du cerveau -le striatum-, dans le sens de la démarche des neuroéconomistes (*cf.* section I). Comme les autres cas d'inférence inverse, ce raisonnement peut être fallacieux, car l'activité de la région en question -le striatum- peut ne pas être en lien avec X (trouble de l'impulsivité), mais peut être avec d'autres fonctions X', X'', X'''..., sans que cela ne remette en question la proposition (2). Dans l'étude de Pessiglione *et al.* (2006), le trouble de l'impulsivité peut entrainer dans la tâche A l'activité cérébrale $Z_{pathologique}$ et un comportement $C_{pathologique}$, mais cela n'implique pas nécessairement en sens inverse que tout individu caractérisé par $Z_{pathologique}$ et $C_{pathologique}$ dans la tâche soit impulsif.

2. La connaissance clinique des pathologies comme outil de définition des fonctions cognitives

Les études de neuropsychiatrie computationnelle peuvent servir à inférer un trouble cognitif d'une activité cérébrale (*cf. supra*). Pour que cette inférence soit valide, la définition du processus cognitif X doit être purement fonctionnelle, ce qui signifie que le dysfonctionnement doit être compris au sens qui en été donné dans cette tâche uniquement: tout $Z_{pathologique}$ dans la tâche A manifeste un trouble de l'impulsivité au sens fonctionnel, c'est-à-dire un déficit dans le test en question. Par exemple, dans les études sur le regret abordées dans le chapitre précédent (Camille *et al.*, 2004), il est logiquement possible d'inférer la mise en œuvre d'un processus appelé « regret » à partir de l'activité du cortex orbito-frontal, à la condition de définir le regret comme le jugement permettant, dans le protocole utilisé, d'évaluer la différence entre le gain obtenu et le gain qui aurait pu être obtenu en choisissant l'autre option (*cf.* chapitre 4, section III). En revanche, comme le souligne Bourgeois-Gironde (2010, p.233), une définition plus large de la notion de regret, incluant par exemple une référence aux émotions, introduit une hypothèse interprétative qui dépasse le cadre strict du protocole. Dans ce cas, il n'est alors pas nécessairement vrai que l'activité du cortex orbito-frontal implique un regret chez le sujet.

Une définition stricte du trouble cognitif est retenue, ne s'appliquant qu'au protocole de test proposé, a cependant une portée descriptive très faible. En effet, tout $Z_{pathologique}$ dans la tâche A signifie un dysfonctionnement cognitif X, qui implique $Z_{pathologique}$ dans la tâche A; mais X ne peut être défini en dehors de la tâche A. Cette proposition peut se comprendre comme une quasi-tautologie, car le diagnostic fonctionnel ne sert alors qu'à détecter des déficits dans la capacité à réussir le test. Le trouble X n'a pas d'autre signification en dehors du test, ce qui revient en pratique à renoncer à inférer une fonction cognitive du neuronal.

La fonction cognitive qui fait l'objet d'une évaluation doit donc être décrite dans des termes suffisamment généraux, non restreints à la tâche d'évaluation, pour avoir une pertinence clinique et descriptive. La définition doit être assez large pour inclure toutes les sous-fonctions X', X", X'"... susceptibles d'être mise en œuvre dans la tâche. Une définition trop étroite entraine le risque d'erreurs signalées plus haut: l'observation d'un $Z_{pathologique}$ dans la tâche n'est pas forcément lié à un dysfonctionnement de X, mais peut aussi s'expliquer par un dysfonctionnement de X', X", X", …De l'autre côté, la définition de la fonction X doit être suffisamment restreinte de manière à ne désigner que la capacité qui est spécifiquement atteinte chez les sujets qui « sous-performent » dans la tâche. Par exemple, l'étude de

Pessiglione *et al.* (2006) ne doit pas être comprise comme un dispositif d'évaluation des facultés d'apprentissage au sens large, car les individus impulsifs dans cette tâche sont bien capables d'apprendre certains contenus de connaissance. Il convient de préciser le type d'apprentissage qui est en jeu ici (apprentissage de la récompense).

L'enjeu consiste à élaborer le bon niveau de correspondance, dit « un pour un », ici entre un trouble neuronal (traitement au L-Dopa) et la caractérisation de ce trouble au niveau cognitif, ce qui nous ramène au problème posé par l'inférence inverse (*cf.* section I.C). C'est ici que, selon nous, la connaissance clinique joue un rôle essentiel. La caractérisation fonctionnelle des troubles (proposition 2) s'appuie en effet sur le présupposé selon lequel la populations d'individus atteints de la pathologie P, recrutés pour « calibrer » le test, est *spécifiquement* caractérisée par un dysfonctionnement de la fonction X. En d'autres termes, la longue observation des patients en clinique permet d'identifier le trait commun à un ensemble d'individus pathologiques, et sert ainsi d'outil indispensable dans l'élaboration de ce que Bourgeois-Gironde appelle les « *ontologies cognitives* » (Bourgeois-Gironde, 2010, p.229).

La description du trouble représente *in fine* un compromis entre la définition purement fonctionnelle et quantitative, selon laquelle la fonction atteinte ne désigne que la capacité à réussir le test, et la définition d'inspiration clinique, héritée d'une tradition d'analyse verbale des patients. Mais c'est la connaissance clinique, antérieure, qui permet toujours de justifier et de s'assurer que le niveau de description de la fonction soit le bon. Par exemple, dans l'étude de Pessiglione *et al.* (2006), l'expression de « trouble du *reward learning* » élargit la signification du déficit manifesté par les sujets traités au L-Dopa au delà du protocole de l'étude, puisqu'il peut s'appliquer à un ensemble de tâches relativement proches, et, de surcroît, conserve une pertinence pour caractériser finement la nature du trouble sur le plan clinique, comme le suggèrent Hare, Camerer et Rangel (2008).

L'analyse clinique des pathologies mentales constitue donc une source de connaissance indispensable pour appuyer toute démarche visant à localiser, dans le cerveau, des fonctions cognitives, que Harrison dénonce, à propos de la neuroéconomie, comme une tentative de « *néo-phrénologie* » (*cf. supra*, Harrison, 2008-b, p.534). La connaissance fournie par la clinique ne règle pas définitivement le problème posé par l'inférence inverse, mais elle améliore la qualité des inférences pouvant être établies du cérébral au cognitif. On rappelle que les raisonnements dits d'inférence inverse prennent la forme suivante (*cf.* section I, Poldrack, 2006, p.39):

(2) Dans cette étude, la zone du cerveau Z était active dans la tâche A.

(2) Dans d'autres études, lorsque le processus cognitif X était engagé, la zone Z était active

(3) Par conséquent, l'activité de la zone Z dans cette étude démontre la mise en œuvre du processus cognitif X dans la tâche A

Comme le suggère Bourgeois-Gironde, les inférences inverses peuvent être considérées comme plus ou moins certaines, dans « *un intervalle de confiance allant de la pure hypothèse à la conclusion définitive, en fonction des conditions* » (Bourgeois-Gironde, 2010, p.238). En effet, la conclusion (proposition 3) est plus ou moins probable en fonction de la sélectivité de la zone Z, c'est-à-dire selon la probabilité pour que Z soit *spécifiquement* activée par le processus X. Si la région prend en charge un nombre élevé de fonctions, cette probabilité sera faible, parce que la zone Z sera susceptible d'être activée par un autre processus que X[112]. Pour Bourgeois-Gironde, l'interprétation cognitive des activités cérébrales doit être réalisée en prenant en compte la sélectivité des zones étudiées (Bourgeois-Gironde, 2010, p.242). Plusieurs critères peuvent être utilisés. La connectivité fonctionnelle de la zone (avec d'autres zones) fournit un élément important d'appréciation[113]. Mais c'est surtout à partir des « méta-analyses » et des bases de données, c'est-à-dire de travaux de synthèse établissant des cartes cognitives du cerveau, avec la liste des fonctions associées à chaque zone, que les neurobiologistes peuvent évaluer la sélectivité des activités observées. Or, comme le souligne Bourgeois-Gironde, ces outils montrent qu'il existe un fort chevauchement des différentes régions. La complexité de la tâche de délimitation des « *ontologies cognitives* » explique ainsi sans doute que ce travail soit insuffisamment approfondi dans la plupart des études de neuroéconomie (Bourgeois-Gironde, 2010, p.242).

112La problème, ici, peut se comprendre comme une définition trop étroite des fonctions: c'est la raison pour laquelle plusieurs fonctions X, X', X'', X'''... peuvent être associées à Z. Inversement, si la définition des fonctions est trop large, la zone Z peut avoir été activée dans les études précédentes en raison de l'engagement d'une sous fonction de X. Par conséquent, l'engagement du processus X n'implique pas forcément à chaque fois l'activation de Z, et la conclusion apportée par l'inférence inverse peut être erronée.

113En effet, la connectivité fonctionnelle permet d'apprécier le rôle de la zone au sein d'une structure plus globale, voire eau sein du cerveau. Une zone fortement connectée sera susceptible d'être activée lors de la mise en œuvre d'un grand nombre de processus cognitifs différents.

3. La distinction clinique du normal et du pathologique: une référence implicite et sous-estimée des travaux de neuroimagerie

En s'appuyant sur leur propre expérience clinique (ou celle de leurs collègues psychiatres qui participent au travail de recherche), les neurobiologistes sont en mesure de contourner -sans les régler définitivement- les difficultés posées par l'inférence inverse, et de conceptualiser adéquatement les fonctions prises en charge par les différentes régions. Bien qu'ignorée par Bourgeois-Gironde (2010), l'importance des hypothèses de localisation suggérée par l'observation clinique est pourtant bien connue. Les neurologues ont toujours tiré parti de l'observation des sujets lésés ou atteints de troubles divers pour avancer des hypothèses de localisation, comme Broca avec l'aphasie par exemple. Cependant, dans le cas de la neuroéconomie, qui se comprend comme une approche fonctionnelle et quantitative, le rôle joué en amont par les concepts empruntés à la médecine mentale passe souvent inaperçu, pour deux raisons.

Tout d'abord, toutes les expériences de neuroimagerie ne s'inscrivent pas explicitement dans une démarche médicale de « neuropsychiatrie computationnelle », visant à caractériser un trouble cognitif sur le plan neuronal et comportemental. Par ailleurs, les hypothèses de localisation ne semblent pas mobiliser des preuves issues de la clinique. La plupart des expériences de neuroimagerie sont réalisées sur des sujets normaux et non sur des sujets pathologiques, patients de services de neurologie et/ou de psychiatrie. Pourtant, la connaissance du « normal » est toujours référée, en amont, à la connaissance du pathologique. Cela apparaît clairement dans une perspective plus globale sur l'ensemble des travaux ayant été publiés sur une zone ou une fonction cognitive.

L'exemple des études sur le regret décrites dans le chapitre précédent fournit une bonne illustration du rôle théorique joué par les cas pathologiques dans la recherche en neurosciences. Initialement, l'hypothèse selon laquelle le cortex orbitofrontal serait impliqué dans le regret provient d'une expérience de 2004 portant sur des sujets atteints précisément de lésions dans cette zone du cerveau (Camille *et al.*, 2004). Cette étude fondatrice a permis de montrer, grâce à l'enregistrement de l'activité électro-dermale, que les lésions du cortex orbitofrontal n'empêchent pas les patients de ressentir une déception lors d'une perte monétaire, mais bloquent l'apparition d'émotions plus complexes. Les chercheurs proposaient de qualifier ces jugements contrefactuels de « *regret* » (Camille *et al.*, 2004).

C'est à l'appui de cette première expérience que Giorgio Coricelli et ses co-auteurs ont pu par la suite conclure, dans le cadre d'une étude postérieure de neuroimagerie, que l'activité

du cortex orbitofrontal dans la même tâche expérimentale était associée, chez le sujet sain, à l'expérience du regret (Coricelli *et al.*, 2005). Dans un second temps, ce paradigme expérimental a pu servir d'outil de diagnostic. Ainsi, Larquet *et al.* établissent que certains patients schizophrènes manifestent des déficits fonctionnels dans cette tâche (Larquet *et al.*, 2010, p.266). La « *tâche du regret* » (*regret gambling task*) peut alors se comprendre comme un test neurocomportemental, offrant une une grille de compréhension et d'évaluation unitaire de troubles comportementaux observés dans diverses pathologies.

La référence à des travaux portant sur des sujets pathologiques doit être distinguée de celle qui peut être faite à d'autres études de neuroimagerie sur sujets normaux,. La première apporte, outre une information dite de « la même zone », la garantie que la fonction étudiée -ici, le regret- soit correctement décrite (ou au moins dans un sens pertinent d'un point de vue clinique). Or cette différence de nature dans les sources de « triangulation » des études de neuroimagerie n'est souvent pas spécifiée. Les faisceaux d'évidences antérieures avancés par les expérimentateurs pour appuyer leurs résultats peuvent effectivement laisser craindre souvent un phénomène d'inférence inverse. Par exemple, Coricelli et ses coauteurs écrivent au début du compte-rendu de leur étude:

> « *des faisceaux d'évidence indiquent que le cortex orbitofrontal encode la valeur relative des stimuli et met à jour la saillance des récompenses primaires et secondaires [Gottfried et Dolan, 2004; Rolls, 2000, Tremblay et Schultz, 1999] Des récompenses à la fois simples et abstraites telles que le gain ou la perte monétaire suscitent des émotions qui servent à guider le comportement, et le cortex orbitofrontal est un candidat naturel pour la production de telles émotions [Elliott et al., 2000; Kringelbach et Rolls, 2004; Breiter et al., 2001; O'Doherty et al., 2001. Nous voulons aussi suggérer ici que le cortex orbitofrontal module ces émotions par un processus de contrôle top-down, dans lequel un processus cognitif paradigmatique -le jugement contrefactuel- contribue à la réponse émotionnelle et module le comportement de choix. L'existence d'un substrat neuroanatomique spécifique dont dépend une émotion complexe comme le regret a été mis en évidence dans une étude montrant que cette émotion appuyée sur des processus cognitifs dépend de l'intégrité du cortex orbitofrontal (Camille et al., 2004)*» (Coricelli *et al.*, 2005, p.1255).

Une lecture rapide pourrait laisser croire ici que les auteurs tombent directement dans le piège dénoncé par Harrison (Harrison, 2008-b, p.535), puisque le faisceau d'évidence auquel il est fait référence semble renvoyer uniquement à d'autres études de neuroimagerie portant soit sur une zones similaire (le cortex orbitofrontal), soit sur une fonction cognitive pouvant être associée de près ou de loin à l'expérience du regret. Pourtant, on l'a vu, l'expérience de Camille *et al.* (2005) portant sur des patients lésés fonde bien, d'un point de

vue clinique, l'inférence qui est proposée entre l'activité du cortex orbitofrontal et le regret[114].

En second lieu, même dans les études comme celles de Pessiglione *et al.* (2006) remplissant des objectifs médicaux, l'importance de la connaissance clinique est également ignorée, voire minimisée par les chercheurs eux-même. Ces derniers conçoivent en effet les paradigmes expérimentaux issus du néo-comportementalisme comme des dispositifs d'évaluation fonctionnels visant précisément à se substituer aux catégories cliniques traditionnelles. Pourtant, ces catégories permettent bien d'élaborer des concepts pertinents pour décrire les fonctions au niveau cognitif (*cf. supra*). Les connaissances cliniques ont en outre une influence à un second niveau, qui n'a pas encore été abordé ici. C'est bien en effet à partir des critères cliniques empruntés à la médecine mentale « traditionnelle » qu'est possible, au départ, une discrimination entre un groupe de sujets sains ou normaux et un groupe de sujets pathologiques. Or l'évaluation quantitative d'une capacité ou d'un déficit dans un test neuro-cognitif suppose la définition du niveau de performance « normal » d'un sujet rationnel. Elle renvoie implicitement à un critère clinique de distinction entre les joueurs rationnels et pathologiques. La connaissance clinique permet donc non seulement de définir les termes descriptifs utilisés au niveau cognition, mais de statuer et de dire s'il y a fonctionnement « normal » ou « pathologique » du processus en question.

Cette interprétation n'apparait jamais explicitement dans la littérature. Souvent, les tâches de machine à sous multi-jeux sont prises comme des mécanismes d'évaluation purement quantitatifs de la performance des individus. Le gain obtenu par les sujets est pris sans réserve comme une mesure objective de son degré de rationalité. Camerer écrit par exemple que « *dans un jeu, il y a une critère simple pour évaluer la performance: qui gagne le plus d'argent?* » (Camerer, 2008, p.205). Il est également courant, dans le cadre de ces protocoles, de définir les comportements pathologiques comme des comportements « couteux »: « *les comportements inadaptés tels que l'addiction aux paris financiers ou au shopping sont, par définition, des problèmes de choix compulsifs […] impliquant des coûts négatifs* » (Voon *et al.*, 2010, p.135)

Cette définition du comportement pathologique pose d'importants problèmes, notamment parce que l'on peut en fait tout à fait concevoir des *multi armed bandit problems* dans lesquels les individus présentant des déficits fonctionnels obtiennent de meilleures performances que les sujets sains (Shiv *et al.*, 2005). En outre, dans l'étude de Voon *et al* (2010), les patients ayant des signaux d'erreur de prédiction amplifiés apprennent plus vite,

114Cela étant, la justification proposée ici par Coricelli *et al.* (2005) demeure critiquable, en tant qu'elle élargit significativement le contenu du concept de regret, en faisant notamment référence à la notion controversée d' « émotion » (*cf. supra*)

dans la condition de gain, à identifier la meilleure stratégie et obtiennent en moyenne des gains monétaires plus importants. La définition du meilleur comportement varie donc selon le protocole. Un sujet parkinsonien ayant développé une addiction aux paris peut, dans certaines tâches fonctionnelles, obtenir des performances supérieures à la normale, tout en étant manifestement handicapé dans de très nombreux autres comportements. C'est donc, au départ, en jugeant à partir de critères cliniques que les sujets parkinsoniens par exemple sont de type « pathologiques » que les expérimentateurs sont en mesure d'apprécier la rationalité des décideurs dans la tache considérée.

4. La neuroéconomie comme compromis entre le langage clinique et l'approche fonctionnelle

L'ancrage pathologique des hypothèses de localisation avancées par les neuroéconomistes offre un moyen-terme dans le problème de l'inférence inverse, à partir du raisonnement suivant:

(1) Dans cette étude, la zone du cerveau Z était active dans la tâche A_1.

(2) Dans d'autres études, portant sur des tâches similaires A_1, A_2, ...A_n, lorsque le processus cognitif X était engagé chez le sujet sain, la zone Z était active (activité $Z_{normale.}$). Chez le sujet atteint de la pathologie P affectant le processus X, la zone Z n'est pas active ou activité $Z_{pathologique}$

(3) Par conséquent, l'activité de la zone Z dans cette étude démontre la mise en œuvre (normale) du processus cognitif X dans la tâche A

La connaissance clinique antérieure a une influence à deux niveaux. Elle permet d'une part d'élaborer des « ontologies cognitives » adéquates (Bourgeois-Gironde, 2010, p.229), en fournissant un bon niveau de correspondance entre les activités cérébrales et la fonction X, et en s'assurant que cette dernière soit bien spécifiquement engagée dans la série de tâches A_1, A_2, ...A_n. A un second niveau, la connaissance clinique établit un critère de distinction entre sujets normaux et pathologiques, qui sert de principe d'évaluation dans la tâche.

Cela ne signifie pas pour autant que le rôle de la neuroimagerie et de l'approche fonctionnelle se borne à confirmer des connaissances cliniques antérieures. A partir d'une hypothèse de localisation suggérée par la clinique, l'intérêt de l'IRM_f consiste justement, d'une part, à valider cette intuition, et, d'autre part, à proposer une compréhension plus fine de la fonction cognitive envisagée. Celle-ci n'est pas seulement associée à une zone du cerveau,

mais à un schéma d'activité dans une zone du cerveau: dans l'étude de Pessiglione *et al.* (2006) par exemple, la fonction d'apprentissage de la récompense est réduite à un algorithme qui décrit l'activité (et non une simple activation du striatum) (*cf. supra*). Surtout, la fonction est traduite en variable quantitative, ce qui améliore l'appréciation pouvant être portée sur sa mise en œuvre.

La production des connaissances en neurobiologie fait donc apparaitre un progrès et une continuité entre des connaissances originaires fournies souvent par la clinique, et les données plus récentes fournies par le nouvel appareillage expérimental. Comme le souligne Guala, dans les disciplines expérimentales, « *la découverte causale est cumulative* » (Guala, 2005, p.78) et le progrès scientifique progresse par inductions particulières, plutôt que « *par les moyens d'une généralisation audacieuse* » (Guala, 2005, p.200).

Le défaut de la critique méthodologique avancée par Harrison (2008-b, p.535) consiste de ce point de vue à postuler que les découvertes scientifiques revendiquées par les neurobiologistes à l'aide des techniques de neuroimagerie font table rase des connaissances neuroanatomiques antérieures. Or, et c'est le cas pour le regret, mais plus généralement pour la plupart des travaux de la neuroéconomie, les résultats théoriques importants de la neuroimagerie ont toujours été dérivés d'intuitions relatives à des dysfonctionnements et des modes de comportement pathologiques. Dès le départ, les premières expériences de neuroéconomie portant sur l'apprentissage de la récompense chez l'homme s'appuyaient sur d'autres travaux de neuroimagerie mettant en évidence l'implication des mêmes régions, et du striatum en particulier, dans les phénomènes addictifs, chez des sujets consommateurs de cocaïne (Breiter *et al.*, 1997).

Harrison lui-même reconnaît que l'observation des sujets lésés constitue une réponse efficace à l'argument de l'inférence inverse: « *une autre perspective sur le problème de l'inférence inverse consiste à accorder davantage de poids à l'observation du comportement de patients lésés* » (Harrison, 2008-a, p.313). Cependant, Harrison se montre réservé quant au potentiel théorique des données fournies par les sujets lésés: « *malheureusement, si l'on se restreint aux seules lésions du cerveau, alors les échantillons* [de sujets] *vont être très petits et l'on aura de très nombreux artefacts; en outre cette approche en terme de lésion soulève des problèmes éthiques* » (Harrison, 2008-a, p.313). Il n'est cependant nullement nécessaire de se restreindre aux seules preuves fournies par l'observation du comportement des patients lésés mais qu'il s'agit précisément de croiser ces preuves avec celles produites par l'RM_f sur des sujets sains (*cf. supra*).

Les réserves « *éthiques* » exprimées par Harrison font référence à la possible

utilisation de connaissance à vocation prioritairement thérapeutiques ou médical dans un but purement théorique. Plus loin, Harrison affirme en effet qu'il existe indiscutablement une incompatibilité entre les objectifs médiaux et « *scientifiques* » de la recherche: « *les méthodes statistiques utilisées dans la plupart des cas sont celles qui sont en vigueur dans la littérature neuroscientifique, mais ces normes et pratiques répondent souvent plutôt à des besoins cliniques (par exemple, détecter une tumeur, faciliter la prévention des arrêts et des crises) qu'à des besoins scientifiques* » (Harrison, 2008-a, p.314)

Selon Harrison, il y aurait donc d'un côté la recherche pure, préoccupée uniquement de théorie, et de l'autre côté une recherche appliquée. Or, dans les faits, les deux objectifs sont poursuivis par la plupart des études de neuroimagerie. Par exemple, Pessiglione *et al.* (2006) visent à la fois à fournir une justification théorique à une hypothèse de localisation, et, dans le même temps, à proposer un protocole de test neurocomportemental pouvant servir d'outil de diagnostic. Le rôle théorique joué par des cas considérés comme pathologiques en clinique ne soulèvent pas de difficulté et n'est pas nécessairement un obstacle à la scientificité de ce programme de recherche. S'il est admis qu'un spectre de comportements, de lésions et de déficits fonctionnels peuvent être caractérisés comme manifestement inadaptés, alors il semble raisonnable de considérer que le comportement rationnel peut se définir théoriquement comme l'antithèse de ces cas pathologiques. Cet ancrage clinique permet ainsi à la neuroéconomie de se déployer comme une entreprise de quantification de la motivation et de ses troubles. Les apories suscitées par l'économie comportementale dans le scanner sont ainsi contournées par cette référence à la pathologie. La mise en avant de la notion de monnaie neuronale commune au détriment du cadre dualiste émotions/raisons contribue par ailleurs à distinguer la neuroéconomie des *behavioral economics.*

III. La construction d'un paradigme autonome (2): la mise en avant d'une monnaie neuronale commune à l'homme et à l'animal

L'économie comportementale dans le scanner repose sur le postulat d'une opposition entre deux types de processus cognitifs, les uns étant envisagés comme « rationnels », les autres, plus instinctifs, étant qualifiés d'« émotionnels » (*cf.* chapitre 4). Au fil du développement de la neuroéconomie, ce principe explicatif a perdu progressivement de sa pertinence pour interpréter les résultats expérimentaux. Ceci s'explique d'abord par l'ambivalence du rapport entre émotions et rationalité. Il est en effet possible de considérer, à la manière de Damasio, que les premières participent pleinement à la seconde; il existe cependant dans certains protocoles une incompatibilité entre émotions et rationalité (A). Le cas du choix inter-temporel montre que les neuroéconomistes ont progressivement abandonné le concept d'émotion pour revenir à l'hypothèse plus ancienne d'un système unitaire d'évaluation des récompenses. La thèse de la monnaie neuronale commune révèle ainsi l'ancrage biologique et évolutionnaire de la neuroéconomie (B).

A. L'ambivalence des émotions

La notion d'émotion, empruntée à Damasio, a été jugée commode au départ par certains neuroéconomistes, pour rendre compte de l'opposition, dans le cerveau, entre deux types de processus cognitifs. Cette distinction émotions/raison est également adoptée par Daniel Kahneman (Kahneman, 2003). Or cette apparente convergence entre le programme de recherche kahnemanien et les neurosciences masque une importante contradiction interne. En effet, dans la perspective de Damasio, les émotions participent à la raison . Pour Damasio, un patient atteint de lésions préfrontales ayant pour conséquence des déficits « émotionnels » manifeste clairement une incapacité à adopter des décisions conformes à ses intérêts (Damasio, 1994). A l'inverse, en économie comportementale, le mode de raisonnement appuyé sur les émotions est envisagé comme étant plus intuitif. Il permet une résolution plus rapide de problèmes décisionnels complexes, mais il est aussi une source d'approximations et

d'erreurs (*cf.* Kahneman, 2003). L'économie comportementale dans le scanner est ainsi traversée par une tension interne, puisque certaines études réalisées par des économistes comportementalistes vont à l'encontre de la thèse de Damasio selon laquelle les émotions sont un prérequis de la rationalité. Deux expériences sont ici mobilisées à titre d'exemple (Shiv *et al.*, 2005; Sokol-Hessner *et al.*, 2009).

Dans une étude intitulée « le comportement de l'investisseur et la face sombre des émotions », Shiv *et al.* montrent que des sujets lésés, caractérisés par des déficits émotionnels, obtiennent des performances supérieures à la normale dans un jeu de pari financier (Shiv *et al.*, 2005). Les expérimentateurs ont recruté pour cette expérience quarante et un sujets, dont quinze (les patients « cibles ») étaient atteints de lésions dans différentes zones du cortex préfrontal associées au traitement des émotions[115]. Sept autres sujets possèdent des lésions dans d'autres régions du cerveau, qui ne sont pas considérées comme étant impliquées dans les émotions; les dix-neuf participants restant étant des sujets « *normaux* » (Shiv *et al.*, 2005, p.435).

La tâche, très simple, repose sur un jeu de pari financier répété à l'identique vingt fois. Au début de l'expérience, les participants reçoivent 20 dollars de mise de départ. A chaque essai, le sujet doit choisir entre « investir » et « ne pas investir ». S'il investit, l'expérimentateur jette une pièce en l'air, et, en fonction du résultat, le sujet gagne 2,5 dollars (si la pièce tombe sur pile) ou perd un dollar (si la pièce tombe sur face). S'il n'investit pas, le joueur conserve ses gains, sans modification (Shiv *et al.,* 2005, p.436). Dans ce jeu, un joueur souhaitant maximiser l'espérance de ses gains doit investir à chaque essai, puisque l'espérance de gain associée à l'investissement est strictement positive. Or, la plupart des sujets s'abstiennent généralement d'investir au moins pour certains essais, le plus souvent après avoir subi une perte. Les résultats de cette expérience montrent que les sujets normaux investissent dans 57,6% des essais, alors que cette proportion s'élève à 83,7% pour les patients cibles (Shiv *et al.*, 2005, p.436).

Shiv *et al.* concluent donc que les émotions peuvent avoir une « *face sombre* » dans le comportement économique. A l'inverse, des choix rationnels impliquent dans bien des cas une « *neutralité émotionnelle* ». Par exemple, un conducteur dont le véhicule dérape sur une plaque de glace a souvent tendance, sous l'effet de la peur, à appuyer sur la pédale de frein. Cette réaction aggrave encore plus le dérapage. Un conducteur plus rationnel, c'est-à-dire n'agissant sous l'effet de la peur, se rappelle que le comportement adapté consiste à ne pas

115Les lésions de ces quinze patients cibles touchaient soit l'amygdale (trois patients), soit le cortex orbitofrontal (huit patients), soit le cortex insulaire sensori-moteur droit (quatre patients)

appuyer sur la pédale de frein. Cette absence de ressenti émotionnel lui permettra de prendre une meilleur décision (Shiv *et al.*, 2005, p.435). Les auteurs admettent cependant que ce genre d'expériences n'est pas généralisable à tous les domaines. Les émotions ne jouent pas nécessairement toujours un rôle négatif dans la prise de décision. Leur étude constitue un cas bien spécial, puisqu'il s'agit d'une décision risquée, mais non pas dans l'incertain: le résultat de chaque pari est inconnu *ex ante*, mais la distribution de probabilité des gains et des pertes dans ce jeu simple est facile à comprendre pour les sujets (Shiv *et al.*, 2005, p.438). Or, dans la plupart des expériences de neuroéconomie, il s'agit au contraire de « découvrir » des probabilités de gains et de pertes associées à diverses options de choix.

Pour Shiv *et al.*, il ne s'agit donc pas d'affirmer que les émotions sont toujours les adversaires de la raison, mais plutôt d'identifier les types de situation dans lesquelles celles-ci peuvent, ou non, favoriser la rationalité des décisions: « *la question n'est pas de savoir si les émotions peuvent conduire à de bonnes ou de mauvaises décisions. La recherche doit plutôt déterminer les circonstances dans lesquelles les émotions peuvent être utiles ou perturbantes, et ce lien entre les circonstances et les émotions peut fournir une explication intéressante du comportement humain* » (Shiv *et al.*, 2005, p.438). Une telle approche suppose donc de se concentrer sur l'environnement de la décision, plus que sur le type de processus cognitif engagé dans la décision.

Prise de manière autonome, la notion d'émotion perd donc de son pouvoir explicatif du point de vue de la rationalité, puisque des sujets caractérisés par des déficits dans le traitement des émotions ne prennent pas nécessairement des décisions défavorables. En outre, chez le sujet sain, comme le suggère l'exemple de la conduite automobile avancé par Shiv *et al.* (2005, p.435), l'amélioration des capacités dites de *self control* permet souvent d'améliorer la performance des décisions. Cette conclusion constitue notamment le résultat principal d'une expérience célèbre de Camerer et ses co-auteurs, intitulée « penser comme un trader réduit sélectivement l'aversion au risque » (Sokol-Hessner *et al*., 2009).

Sokol-Hessner *et al.* proposent aux participants de leur expérience une série de 140 choix binaires entre, à chaque fois, un gain certain ou une loterie, qui peut porter soit sur un gain ou une perte équiprobable (mais de montants différents) ou deux gains (de montants différents) équiprobables. Les gains et les pertes, à chaque essai, sont compris entre -24 dollars et +30 dollars (Sokol-Hessner *et al.*, 2009, p.5037). La plupart des sujets manifestent dans ces décisions de l'aversion au risque, c'est-à-dire qu'ils préfèrent obtenir de manière sûre un montant monétaire plutôt que de parier sur une loterie dont l'espérance mathématique de

gain est égale à ce gain certain. A partir des choix retenus par les participants, les expérimentateurs peuvent donc estimer la sensibilité au risque de chaque individu (Sokol-Hessner *et al.*, 2009, p.5037). Les sujets sont informés que, parmi l'ensemble de leurs décisions, l'une d'entre elles sera tirée au sort et permettra de déterminer leur gain ou leur perte effectif à la fin de l'expérience. A l'issue de chaque choix en faveur du pari, les sujets sont en outre informés immédiatement du résultat obtenu (montant du gain ou de la perte).

Pour mettre en évidence le rôle des émotions, deux conditions sont successivement utilisées. Les joueurs répondent donc successivement à deux blocs identiques de 140 choix. Dans la première condition, on présente aux sujets le texte suivant:

> « *Assistez. Lorsque vous voyez le signal « assistez », concentrez vous sur cette décision financière uniquement, sans prendre en compte les autres décisions à venir. Dites vous que seul le pari importe, et que celui-ci pourrait être celui pour lequel vous allez être payé. Ainsi, vous pourriez gagner ce montant positif, mais vous pourriez aussi perdre ce montant négatif et devoir rendre l'argent à l'expérimentateur. Approchez chaque essai comme si vous faisiez seulement ce choix particulier dans l'étude d'aujourd'hui. Concentrez vous sur les valeurs dans ce pari, et sur les gains certains. Demandez vous comment vous vous sentiriez si vous perdiez le montant négatif, et comment vous vous sentiriez si vous perdiez le montant négatif. Laissez toutes les pensées et toutes les émotions relatives à ce choix particulier se produire naturellement, sans essayer de les contrôler. Il est important que vous vous concentriez sur les décisions monétaires que vous affrontez au moment présent, en stricte isolation des autres décisions* » (Sokol-Hessner *et al.*, 2009, informations supplémentaires, p.3).

Après avoir répondu au premier bloc de questions, dans lequel le signal ci-dessus apparaît à chaque fois, les sujets doivent répondre à nouveau aux mêmes décisions financières dans un second bloc, mais cette fois ci dans une condition d' « auto-régulation »:

> « *Réévaluez. Lorsque vous voyez le signal « réévaluez », pensez à chacune de ces décisions financières dans le contexte de toutes vos décisions passées et futures. Considérez chaque décision comme si elle constituait un « portefeuille ». Rappelez vous que vous effectuez de très nombreuses décisions similaires. N'essayez pas d'évaluer vos gains totaux, mais approchez simplement ces paris en gardant en tête leur contexte. Imaginez que vous faites face à l'une de ces décisions maintenant. Une manière de suivre ces instructions consiste à vous imaginez être un trader. Vous prenez des risques financiers tous les jours, c'est votre quotidien. Imaginez que c'est votre métier et que l'argent en jeu n'est pas le votre, c'est celui de quelqu'un d'autre. Bien sûr, vous voulez toujours obtenir des gains importants (votre métier l'exige). Vous faites cela depuis longtemps, et vous continuerez de le faire à l'avenir. Tout ce qui importe est que vous parveniez à atteindre un sommet à la fin, une perte ici ou là n'a pas d'importance du point de vue de votre portefeuille global. En d'autres termes, vous gagnez parfois et vous perdez parfois. Il est important que vous vous concentriez sur chacune de ces*

décisions dans les contexte de toutes les autres décisions que vous allez prendre aujourd'hui pendant les essais d'autorégulation » (Sokol-Hessner *et al.*, 2009, information supplémentaire, p.3).

Les expérimentateurs obtiennent deux résultats importants. D'une part, l'aversion au risque est liée avec les réactions émotionnelles. Lorsque l'individu subit une perte dans un pari, l'ampleur de la réponse émotionnelle, mesurée par la réaction électrodermale, dépend en effet de la sensibilité individuelle au risque, mesurée par les choix effectifs dans l'ensemble du bloc d'essais. D'autre part, les niveaux initiaux d'aversion au risque dans la première condition diminuent significativement (ainsi que les réactions électro-dermales) dans la seconde condition (Sokol-Hessner *et al.*, 2009, p.5038).

Dans ce jeu financier, la diminution des émotions relatives aux pertes et au risque permet au sujet d'améliorer ses performances. En effet, celui-ci a alors une aversion au risque plus faible, et donc une préférence pour des gains certains moins marquée, ce qui lui permet d'élever son espérance de gain total. Pour Sokol-Hessner *et al.*, cette expérience révèle que, en matière d'investissement financier, les « *experts* » se distinguent des individus normaux par une capacité de mise en perspective et de recul par rapport à leur propre ressenti émotionnel (Sokol-Hessner *et al.*, 2009, p.5036).

Les auteurs, cependant, ne concluent pas à l'existence d'une relation d'incompatibilité générale entre émotions et rationalité, ce qui dépasserait le cadre de cette expérience. Ce type d'étude incite plutôt, comme le suggèrent Shiv *et al.*, à rechercher et identifier les circonstances dans lesquelles les émotions peuvent, ou non, favoriser les choix rationnels (Shiv *et al.*, 2005, p.438). S'il n'est pas invalidé par ces travaux expérimentaux, le concept d'émotion ne constitue donc pas un principe explicatif pertinent. Pour certains jeux ou paris financiers, comme par exemple l'IGT, les émotions sont les conditions nécessaires du comportement rationnel; dans d'autres protocoles, les meilleure décisions sont prises par les joueurs les moins émotifs. La notion d'émotion est en fait sous-déterminée, et renvoie à des processus cognitifs hétérogènes (apprentissage, mémoire, aversion aux pertes, *etc.*), ce qui explique ces difficultés. C'est sans doute la raison pour laquelle le terme d'émotion a progressivement été abandonné dans la littérature, au profit du concept plus ancien, hérité du programme néo-comportementaliste, de monnaie neuronale commune.

B. Abandon des modèles duaux et retour des interprétations biologiques et évolutionnistes: le cas du choix inter-temporel

Les neuroéconomistes s'inscrivent dans le prolongement d'une tradition théorique néo-comportementaliste, nettement distincte de la psychologie d'inspiration kahnemanienne. A l'inverse des modèles duaux de la décision, les approches néo-comportementalistes visent à faire apparaitre une continuité évolutive entre la rationalité et les émotions. Ces deux perspectives théoriques donnent donc lieu à des conflits interprétatifs. Ces tensions se donnent à voir en particulier dans le domaine du choix inter-temporel, qui illustre la manière avec laquelle l'économie comportementale dans le scanner s'est progressivement transformée en une neuroéconomie assumant plus explicitement son héritage néo-comportementaliste.

Deux travaux importants de l'économie comportementale dans le scanner (McClure *et al.*, 2004; McClure *et al.*, 2007) se proposaient d'établir une théorie de l'actualisation quasi-hyperbolique dans le cerveau. En montrant notamment que certaines régions étaient spécifiquement activées, dans des problèmes d'arbitrage inter-temporel, par les choix impliquant des récompenses immédiates, McClure *et al.* défendaient la thèse d'une dissociation entre deux systèmes d'évaluation dans le cerveau (*cf.* chapitre 4). Cette interprétation a cependant progressivement été remise en question, notamment par Kable et Glimcher (2007) et par Hare *et al.* (2009).

Dans une étude intitulée « les corrélats neuronaux de la valeur subjective dans le choix inter-temporel », Kable et Glimcher contestent d'abord l'idée selon laquelle des régions du cerveau seraient spécifiquement impliquées dans l'évaluation des récompenses immédiates. Kable et Glimcher utilisent un protocole similaire à l'étude de McClure *et al.* (2004) et proposent à leurs sujets des séries de choix entre une récompense monétaire immédiate et une récompense plus importante avec un délai. A la différence de McClure *et al.* cependant, Kable et Glimcher maintiennent à chaque fois le montant de la récompense immédiate à 20 dollars et font varier uniquement le montant et/ou l'échéance de la récompense avec délai (de 20,25 dollars à 110 dollars, et de 6h à 180 jours). A partir des choix observés, les expérimentateurs établissent pour chaque individu une fonction d'actualisation temporelle des utilités (monétaires) futures. Les paramètres de cette fonction sont ensuite utilisés dans l'analyse des données de la neuroimagerie, ce qui permet de mettre en évidence des régions qui encodent la somme corrélée des deux récompenses, leur différence, et leur importance relative (Kable et Glimcher, 2007, p.1626).

Les résultats de l'IRM$_f$ montrent que l'activité dans trois régions (striatum ventral, cortex médian préfrontal, cortex médian orbitofrontal) est corrélée positivement avec le montant de la récompense variable, et négativement avec son délai (Kable et Glimcher, 2007, p.1627). Or ces régions correspondent précisément à celles ayant été identifiées par McClure *et al.* comme les composantes d'un système d'évaluation « impulsif », spécialisé dans l'estimation des récompenses immédiates. Ce résultat permet donc de réfuter l'interprétation de McClure *et al.* Commentant l'expérience de McClure et al. (2007), Kable et Glimcher écrivent ainsi:

> « *leur conclusion était appuyée principalement sur le résultat selon lequel ces trois régions ont une activité beaucoup plus importante pour les choix impliquant une récompense immédiate que pour ceux impliquant seulement des récompenses avec délais. Cependant, ce résultat empirique, que notre étude ne contredit pas, est aussi compatible avec l'hypothèse selon laquelle ces trois régions encodent la valeur subjective de toutes les récompenses avec n'importe quel délai, puisque la valeur subjective d'une récompense immédiate est plus importante que celle d'une récompense avec délai. Le changement d'activité observé ici dans ces trois régions lorsque seule la récompense avec délai varie permet de rejeter l'hypothèse selon laquelle ces régions encoderaient uniquement les récompenses immédiates* » (Kable et Glimcher, 2007, p.1631).

Kable et Glimcher ne remettent donc pas en question le résultat empirique de McClure *et al.* (2007): il est logique que des régions spécialisées dans l'évaluation inter-temporelle des récompenses aient une activité beaucoup plus importante dans des choix impliquant des récompenses immédiates. Mais ce résultat n'est pas suffisant pour justifier l'interprétation selon laquelle il existerait une dissociation entre deux systèmes d'évaluation dans le cerveau.

Néanmoins, l'hypothèse théorique de McClure *et al.* n'est pas pour autant nécessairement invalidée. Elle pourrait être ici sauvée de la réfutation puisque, dans l'expérience de Kable et Glimcher, les choix impliquent toujours une option en faveur d'une récompense immédiate (20 dollars tout de suite). Par conséquent, il est tout à fait normal, dans la perspective de McClure et ses co-auteurs, que les régions appartenant au système impulsif soient ici activées différentiellement, car il y a à chaque fois un gain sans délai qui est en jeu[116]. En même temps, il serait possible d'envisager, dans l'étude de Kable et Glimcher, que la récompense immédiate implique en fait un délai implicitement, puisque le sujet doit au moins

116Kable et Glimcher reconnaissent par ailleurs qu'une limite principale de leur étude concerne l'absence de choix portant uniquement sur des récompenses avec délai : « *Certaines limites de cette étude suggèrent de la prudence dans les interprétations. Tout d'abord, à l'inverse de McClure et al. (2004), nous n'avons pas inclus de condition expérimentale impliquant des choix entre deux récompenses avec délai. Les études, à l'avenir, devront inclure de tels choix pour vérifier que ces régions n'encodent pas la valeur subjective d'une manière catégoriquement différente lorsqu'aucune récompense immédiate n'est disponible* » (Kable et Glimcher, 2007, p.1632)

attendre jusqu'à la fin de l'expérience pour obtenir sa gratification monétaire[117].

Un problème similaire se pose dans les choix entre deux récompenses avec délai, proposés par exemple dans l'expérience de McClure *et al.* (2004). Comme le souligne avec justesse Harrison, il y a une ambiguïté quant au fait de savoir si le sujet envisage le délai le plus court comme étant nul (parce qu'étant évalué en termes relatifs, comparativement avec le délai plus long associé à l'autre option), ou s'il envisage au contraire les montants absolus de ce même délai. Dans la première hypothèse, qui ne peut pas être complètement rejetée, tous les choix binaires impliqueraient en fait implicitement une récompense immédiate (*cf.* Harrison, 2008-a, p.316). Le délai d'obtention de la récompense, dans ces études de neuroimagerie sur le choix inter-temporel, pose donc d'importants problème d'interprétation. De ce point de vue, le protocole de l'étude de Kable et Glimcher est sans doute plus cohérent que celui de McClure *et al.* (2004, 2007), puisqu'il consiste à maintenir fixe l'option en faveur de la récompense immédiate et à faire varier les paramètres de la récompense avec délai, ce qui permet d'observer les effets directs d'une augmentation ou d'une diminution du délai et du montant des gratifications offertes. Toutefois, si elle ne conduit pas à valider l'hypothèse interprétative de McClure *et al*. l'expérience de Kable et Glimcher n'est pas non plus suffisante pour la réfuter.

La controverse entre McClure *et al.* et Kable et Glimcher met en évidence un conflit entre deux lectures possibles des expériences de neuroimagerie portant sur le choix inter-temporel. Dans une étude intitulée « le *self control* dans la prise de décision implique la modulation du système d'évaluation situé dans le cortex préfronatl ventromédial », Hare, Camerer et Rangel proposent un compromis théorique, en dissociant, au sein du système neuronal d'évaluation des récompenses, le rôle fonctionnel spécifique de différentes régions (Hare, Camerer et Rangel, 2009). L'expérience de Hare, Camerer et Rangel repose sur une discrimination comportementale nette entre deux groupes de sujets. Conformément à ce qui été proposé dans la section précédente, cette approche inspirée de la clinique permet de préciser le sens des interprétations cognitives pouvant être établies à partir des données de la neuroimagerie (*cf.* section II). Ici, l'inférence proposée favorise plus nettement l'hypothèse d'une monnaie neuronale commune, défendue par Kable et Glimcher.

Les participants à l'expérience de Hare, Camerer et Rangel sont trente-sept individus

117Ce problème, lié à un délai minimum dans l'obtention des récompenses monétaires, avait précisément conduit McClure à reproduire son étude initiale avec des récompenses gustatives (eau et jus de fruit). L'administration directe, dans le scanner, de volume d'eau ou de jus de fruit par l'intermédiaire d'une paille, permettait ainsi de contrôler beaucoup plus finement les délais d'obtention effectifs des récompenses (*cf.* McClure *et al.*, 2007).

obèses qui suivent un régime. Dans la première partie de l'expérience, les sujets doivent évaluer 50 aliments, en fonction du goût et de la qualité diététique (une note différente est attribuée pour ces deux critères). A partir de ces choix, les expérimentateurs identifient un aliment jugé neutre par l'individu, à la fois pour son goût et sa qualité diététique. Les participants doivent ensuite choisir, dans le scanner, entre les différents aliments et cet aliment neutre, dans une série de décisions binaires successives. Ils indiquent à chaque fois le degré d'attractivité de l'aliment considéré sur une échelle allant de 1 à 5 (cette mesure fournit ainsi une estimation de la valeur relative accordée par le sujet à cet aliment, comparativement à l'aliment neutre).

L'originalité de l'étude de Hare, Camerer et Rangel, par rapport aux expériences précédentes consiste ainsi, d'une part, à utiliser un délai « symbolique » dans l'obtention de la récompense, puisque c'est ici l'effet positif sur la santé à long-terme (caractère diététique de l'aliment) qui est envisagé ici comme une forme de récompense à long-terme, mais d'abord et surtout à tirer parti d'une discrimination comportementale entre des sujets sains et des sujets impulsifs. En effet, les choix observés permettent à Hare, Camerer et Rangel, 2009 de distinguer nettement deux types d'individus. Un premier groupe de 19 participants, que Hare, Camerer et Rangel, 2009 qualifient de « *self controllers* », effectuent leurs décisions en prenant en compte à la fois le caractère diététique et le goût, alors que les membres du second groupe (18 sujets) choisissent uniquement en fonction du goût (Hare, Camerer et Rangel, 2009 ; p.646)

Les résultats de la neuroimagerie montrent tout d'abord que l'activité dans le cortex ventromédian préfrontal (CVMP) reflète, pour tous les sujets, la valeur subjective attribuée à l'aliment considéré. Pour les sujets appartenant au groupe des *self controllers*, l'activité du CVMP est donc corrélée à la fois avec le goût et le caractère diététique, et seulement avec le goût pour les individus *non self controllers*. En outre, une région spécifique s'active -le cortex dorso-latéral préfrontal (CDLPF)- lors des choix impliquant l'exercice d'une capacité de *self control*, c'est-à-dire dans les choix conduisant à rejeter un aliment appétissant mais peu diététique. Cette activité s'observe pour tous les sujets, mais elle est relativement plus importante chez les sujets *self-controllers* (Hare, Camerer et Rangel, 2009, p.647).

Pour Hare, Camerer et Rangel, ces résultats montrent que le CVMP encode la valeur subjective des récompenses, et que cette valeur subjective est susceptible d'incorporer des éléments hétérogènes (propriétés gustatives, diététiques, *etc.*). L'activité du CDLPF serait quant à elle requise dans les situations impliquant des conflits possibles entre ces différents

facteurs (santé et goût par exemple). La modulation du signal de valeur par le CDLPF permet ainsi au CVMP d'incorporer des objectifs de long-terme dans l'évaluation des stimuli.

Cette interprétation peut sembler très proche de celle de McClure *et al.* (2004, 2007). Hare, Camerer et Rangel affirment en effet que la prise en compte d'objectifs de long-terme implique la modulation de l'activité du CVMPF par une région distincte (Hare, Camerer et Rangel, 2009, p.647), ce qui revient bien à dissocier le rôle fonctionnel de ces différentes régions selon l'horizon temporel. Cependant, Hare, Camerer et Rangel font valoir que, « *contrairement à leur hypothèse* [de McClure et al., 2004], *[...] ceci est le cas non pas parce qu'un signal d'évaluation distinct est encodé dans le CDLPF, ce qui, dans notre expérience, impliquerait une corrélation qui n'a pas été observée entre l'activité dans cette zone et les notes de santé. Le CDLPF exerce plutôt son influence en modulant le signal de valeur encodé dans le CVMPF* » (Hare, Camerer et Rangel, 2009, p.647-648). En d'autres termes, l'absence de corrélation entre les valeurs diététiques et l'activité du CDLPF remet en cause la distinction entre deux signaux d'évaluation des récompenses dans le cerveau.

La différence essentielle dans l'interprétation porte ici sur la compréhension des fonctions exécutives ou de planification temporelles associées au CDLPF. Hare, Camerer et Rangel parlent d'une « *modulation* » du CVMPF par l'activité du CDLPF, alors que McClure *et al.* conçoivent celles-ci plutôt en termes d'inhibition. En effet, chez McClure *et al.* (2004, 2007), les systèmes « patient » et « impulsif » sont tournés vers la réalisation d'objectifs opposés. Le premier a pour fonction de « freiner » l'activité du second. Dans un article plus récent, McClure affirme ainsi que le concept de « *fonction exécutive* » associée au CDLPF dans le choix inter-temporel est aux « *antipodes* » de la notion d'impulsivité (Bickel *et al.*, 2012). Cette approche dualiste, que Bickel et al. (2007) appellent « *théorie décisionnelle de la compétition neurocomportementale* » (*competing neurobehavioral decision theory*), a deux conséquences importantes, qui la distinguent nettement du cadre néo-comportementaliste.

L'interprétation de l'activité du CDLPF en terme d'inhibition vise d'abord à réduire directement un concept économique ou comportemental -l'actualisation temporelle- à une variable cérébrale. La plus ou moins grande oxygénation du sang observée dans cette zone est supposée rendre compte du coefficient de *discount* temporel appliqué aux récompenses avec délai. D'autre part, le système « patient » du CDLPF, dont le large développement est propre à l'homme, opère en sens inverse du système impulsif, localisé dans les régions limbiques et similaire à celui de l'animal. Par conséquent, dans la perspective de McClure, la capacité à

différer l'obtention d'une récompense est spécifiquement humaine[118].

A l'inverse, chez Hare, Camerer et Rangel, la fonction exécutive associée au CDLPF n'est pas conçue (seulement) comme concurrente ou opposée à celle du système d'avluation des récompenses immédiates, mais plutôt comme son prolongement évolutif. Le recours à l'explication en termes évolutionnistes permet de justifier la continuité entre l'évaluation des récompenses primaires et des formes de réflexions plus complexes, spécifiquement humaines : « *nous proposons que le CVMPF a évolué à partir de sa fonction originaire de prévision des valeurs à court-terme des stimuli, et que les êtres humains ont développé la capacité à incorporer des considérations à long-termes dans ces valeurs, en donnant à des structures telles que le CDLPF la faculté de moduler le signal de valeur primaire* » (Hare, Camerer et Rangel, 2009, p.647).

Dans la perspective de Hare, Camerer et Rangel, l'évaluation des récompenses secondaires ou symboliques s'appuie sur des circuits neuronaux qui, originairement, ont été conçus pour évaluer des récompenses primaires immédiates, et ont été progressivement détournés de leur fonction, en réponse aux problème environnementaux spécifiques posés par les sociétés humaines. Cette idée connue dans la littérature sous le nom d'« *hypothèse du recyclage neuronal* » postule ainsi que « *les inventions culturelles envahissent les réseaux cérébraux plus anciens d'un point de vue l'évolution et héritent de leurs contraintes structurelles* » (Dehaene et Cohen, 2007, p.390). En s'inscrivant dans ce cadre évolutionniste, Hare, Camerer et Rangel ne visent donc pas ici à réduire le concept économique et comportemental d'actualisation temporelle à une activité cérébrale, mais à suggérer que les processus cérébraux impliqués dans l'actualisation temporelle s'appuient sur des réseaux plus primitifs. Comme le propose Bourgeois-Gironde (2010, p.229), l'hypothèse du recyclage neuronal permet ainsi de mettre en évidence des contraintes évolutionnaires dans l'accomplissement de fonctions cognitives supérieures, à l'œuvre dans le comportement économique.

118« *il y a un large écart entre l'actualisation temporelle chez l'homme et chez les animaux. Les êtres humains effectuent de manière routinière des arbitrages entre des couts et des bénéfices immédiats, et des couts et bénéfices futurs, avec des délais de plusieurs dizaines d'année. A l'inverse, même les primates les plus avancés, dont la taille du cortex préfrontal est néanmoins largement inférieure, ne semblent pas supporter des périodes d'attente non-programmées dans la gratification supérieures à quelques minutes. Même si certains animaux sont capables d'effectuer des arbitrages temporels sur des longues périodes (par exemple, en stockant de la nourriture pour l'hiver), un tel comportement semble invariablement être instinctif, à l'inverse de la nature plus générale de la planification chez l'homme. Par ailleurs, l'ensemble des études portant sur les lésions du cerveau […] convergent pour montrer que les lésions du cortex préfrontal ont pour effet d'accentuer l'influence des récompenses immédiates, ainsi que de dégrader la capacité à planifier les décisions* » (McClure *et al.*, 2004, p.504).

Ici, dans le cas du choix inter-temporel, le recyclage neuronal implique que le CDLPF ait été conçu originairement pour évaluer des récompenses immédiates ou quasi-immédiates. Par conséquent, il n'y aurait pas d'un côté des circuits spécialisés dans les récompenses immédiates, et de l'autre des régions exécutives permettant, par inhibition, de prendre en compte des objectifs de long-terme. Au contraire, ces derniers doivent mobiliser les circuits de la récompense à court terme pour être pris en compte dans le processus de prise de décision. La contrainte évolutionnaire qui pèse sur l'actualisation temporelle est donc liée à la nécessité de fournir un contenu émotionnel aux récompenses à long-terme. En d'autres termes, il ne suffit pas, pour exercer son *self control*, de restreindre ses inclinations en faveur des aliments appétissants, mais il faut encore que la valeur diététique soit perçue comme appétissante, c'est-à-dire comme désirable, source de satisfaction, par le décideur.

Les interprétations évolutionnistes du choix inter-temporel aboutissent donc à remettre en question les interprétations du fonctionnement des zones pré-frontales en termes d'inhibition, et, plus généralement, les modèles duaux (*cf.* introduction à la première partie). Toutefois, si elle semble favoriser l'hypothèse d'une monnaie générale commune, l'étude de Hare, Camerer et Rangel (2009) ne débouche pas pour autant sur un rejet complet de l'étude de McClure *et al.* (2004). Les auteurs soulignent certains éléments de convergence avec les travaux de McClure[119]. La seule véritable objection empirique que Hare et ses co-auteurs adressent à McClure porte sur l'absence de corrélation observée entre l'activité du CDLPF et les valeurs à long-terme. Pour le reste, il s'agit de divergences relatives à l'interprétation des données, en termes de modulation ou d'inhibition (*cf. supra*).

D'une manière relativement proche, dans une étude récente sur les « couts décisionnels » en termes de délai et d'effort physique, Prévost *et al.* obtiennent des résultats convergents avec ceux de Kable et Glimcher, mais se contentent de se référer à un choix de « *modélisation computationnelle*» différent de McClure, sans pour autant envisager de remettre question ce paradigme interprétatif:

> « *Nos résultats sont difficiles à comparer à cette étude précédente [*McClure et al., 2007] *parce qu'un type différent de modélisation computationnelle a été utilisé pour rendre compte des données. Selon ces auteurs, il y a deux systèmes différents*

119Hare, Camerer et Rangel conçoivent leur interprétation plutôt comme un moyen-terme dans la controverse entre McClure *et al.* d'un côté, et Kable et Glimcher de l'autre. En effet, Hare, Camerer et Rangel suggèrent comme McClure *et al.* (2004) que les choix en faveur des récompenses à long terme impliquent l'activité du CDLPF; et, en outre, comme Kable et Glimcher que l'activité du CVMPF représente la valeur subjective de tous les choix: « c*omme Kable et Glimcher, nous trouvons des preuves soldes en faveur de l'existence d'un signal commun d'évaluation qui dirige les choix indépendamment du degré de self control exercé par les participants. Comme McClure et al., nos résultats suggèrent que le CDLPF joue un rôle critique dans le déploiement du self control* » (Hare, Camerer et Rangel, 2009, p.647).

d'évaluation dans le choix inter-temporel: le système impatient, qui réduit fortement la valeur des récompenses non-immédiates, et un système plus patient, actif à la fois dans les essais immédiats avec délais, et qui réduit moins fortement les récompenses avec délais. A l'inverse, notre étude, qui adopte l'approche proposée par Kable et Glimcher (2007), suggère qu'un unique système d'évaluation encode la valeur actualisée de manière hyperbolique des récompenses à la fois immédiates et avec délais. De plus, notre observation selon laquelle l'activité dans le striatum et le cortex ventromédian préfrontal varie lorsque seul le délai de la récompense fournit une preuve directe que ces régions n'évaluent pas exclusivement les récompenses immédiates » (Prévost *et al.*, 2010, p.14087)

L'unique observation empirique sur laquel l'expérience de McClure *et al.* (2004) peut être remise en cause est le résultat selon lequel les trois régions identifiées comme appartenant au système impulsif seraient spécifiquement et uniquement activées dans les choix impliquant des récompenses immédiates. Mais l'adoption d'un cadre néo-comportementaliste n'implique pas ici un rejet des interprétations proposées par McClure, parce que les données, produites dans un paradigme expérimental alternatif, seraient difficilement comparables. D'une manière générale, les travaux sur le choix inter-temporel qui, selon nous, s'inscrivent dans un cadre néo-comportementaliste, ne prônent donc pas nécessairement l'abandon des interprétations dualistes.

Par ailleurs, le débat entre les partisans de la monnaie neuronale commune et ceux des modèles duaux est loin d'être clos. Dans une étude de 2012, intitulée « les bases neuronales des différences culturelles dans l'actualisation temporelle », McClure apporte un nouveau soutien empirique à son hypothèse dualiste[120]. Le retour observé chez Kable et Glimcher (2007) ou chez Hare *et al.* (2009)à la notion de monnaie neuronale commune n'a donc pas éliminé complètement les interprétations dualistes qui, quoique désormais moins fréquentes, perdurent dans la littérature. Les neuroéconomistes continuent à utiliser et à opposer les

120Dans cette expérience, les auteurs comparent l'actualisation temporelle des récompenses monétaires chez des sujets occidentaux et asiatiques. Les résultats comportementaux indiquent que les sujets asiatiques sont relativement plus patients que les sujets occidentaux. Or, cette moindre diminution de la valeur des récompenses futures est associée à une activité plus faible du système impulsif (striatum ventral). En revanche, l'activité des zones exécutives (cortex post-pariétal et CDLPF) est similaire dans les deux groupes (Kim, Sung et McClure, 2012, p.655). Pour les auteurs, cette observation remet en cause l'une des conclusions de l'étude de Hare *et al.* (2009) selon laquelle l'exercice du *self control* impliquerait l'activité du CDLPF (*cf. supra*): ici, les sujets asiatiques relativement plus patients ne se distinguent pas par une activité plus forte dans cette zone (Kim, Sung et McClure, 2012, p.655). Cette affirmation pourrait être discutée. Pour Hare *et al.* (2009) l'activité du CDLPF permet certes d'élargir l'horizon temporel du choix et, éventuellement, dans les cas de conflit entre plusieurs valeurs, de trancher en faveur du long-terme. Mais cela n'implique pas que tous les individus plus patients soient caractérisés par une activité plus forte des régions préfrontales. En effet, des choix relativement moins impulsif peuvent s'expliquer non seulement par l'exercice de facultés de *self control*, mais aussi par une moindre attractivité des récompenses immédiates. Dans l'étude de Kim, Sung et McClure (2012), la plus grande patience manifestée par les sujets asiatiques est peut être liée à un goût moins prononcé pour les récompenses monétaires, sans lien avec l'actualisation temporelle. Quoiqu'il en soit, le débat reste ouvert. McClure continue de défendre les hypothèses interprétatives avancées dans ses travaux de 2004 et 2007 (McClure, 2004; 2007), en s'associant notamment au psychologue Bickel et à sa « *théorie décisionnelle de la compétition neurocomportementale «* (Bickel *et al.*, 2007; 2007)

concepts d'émotions ou d'impulsivité à celui de fonction exécutive. Cependant, l'hypothèse de la monnaie neuronale commune est de notre point de vue plus conforme au cadre néo-comportementaliste dont participe la neuroéconomie. Même s'il s'agit là encore de prospective, les évolutions et les avancées les plus récentes suggèrent que le noyau dur de la discipline tend à favoriser les interprétations évolutionnistes. L'influence de ce paradigme théorique hérité de la science quantitative de la motivation se donne à voir notamment dans la théorie neuroéconomique de l'addiction.

IV. La neuroéconomie comme projet de « psychiatrie économique » : le cas des comportements addictifs

Les travaux de neuroéconomie sur le choix inter-temporel offrent un éclairage original des comportements addictifs. L'addiction constitue également un objet d'étude pour les économistes comportementalistes. Il s'agira donc ici, encore une fois, de montrer dans ce domaine que l'analyse proposée par les neuroéconomistes se distingue de celle des *behavioral economics*. Le cas de l'addiction revêt cependant pour nous une importance particulière, car il manifeste le rôle déterminant joué par les critères cliniques dans la définition neuroéconomique de la rationalité. Celle-ci s'appuie non pas une norme économique de référence -stabilité des préférences, constance du taux d'actualisation- mais sur la caractérisation médicale de troubles addictifs, considérés comme irrationnels. La rationalité, dans cette approche, n'apparait que comme l'envers de l'irrationalité. Sa définition est renvoyée, par défaut, à la normalité des sujets « sains ».

La théorie neuroscientifique de l'addiction vaut donc à la fois comme illustration et comme aboutissement théorique du projet de psychiatrie économique dont la neuroéconomie est selon nous porteuse, et qui la différencie du programme kahnemanien. Plusieurs éléments de cette théorie font apparaître une convergence avec la tradition néo-comportementaliste qui a été étudiée dans le cadre de la première partie. La notion d'environnement addictif suggère tout d'abord une définition écologique de la rationalité, qui met en avant les dimensions biologiques et évolutionnaires du comportement (A). Cette approche brouille les rapports de la rationalité biologique et économique, qui sont toujours, en dernière analyse, définis en référence à la notion de pathologie (B). La synthèse des théories neurobiologiques de l'addiction proposée récemment par Ross *et al.* (2008) fait ainsi de la neuroéconomie un sous-domaine autonome de la théorie économique, qui tire profit d'une collaboration entre neuroscientifiques et économistes sans pour autant se confondre avec l'économie comportementale (C).

A. Une explication biologique et évolutionniste: la notion d'environnement addictif

Plusieurs auteurs ont récemment proposé d'utiliser les études néo-comportementales portant sur l'apprentissage de la récompense pour construire un modèle général des comportements addictifs, servant à la fois des objectifs thérapeutiques et/ou théoriques (Bernheim et Rangel, 2004; Bickel, 2007; Landreth et Bickel, 2008; Rowland *et al.*, 2008; Rangel, Camerer et Montague, 2008; Ross *et al.*, 2008). Cela peut apparaître paradoxal dans la mesure où la science quantitative de la motivation montre que l'impulsivité, au moins dans certaines circonstances, sert des intérêts biologiques, ou peut s'expliquer au moins par des considérations évolutionnistes (*cf.* chapitre 2). C'est ainsi que Don Ross et ses coauteurs retiennent des travaux de Richard Herrnstein l'idée selon laquelle l'incohérence temporelle est d'une certaine manière rationnelle:

> « *cette recherche dérivée de la loi d'égalisation des rendements a montré que l'irrationalité, au sens technique de l'économiste, est en fait normale. Ceci est très important et doit être répété: la recherche portant sur la loi d'égalisation révèle qu'à la fois les hommes et les autres animaux sont naturellement disposés à avoir des préférences incohérentes pour des récompenses futures au fur et à mesure de l'écoulement du temps , avec la diminution progressive des délais d'obtention* » (Ross *et al.*, 2008, p.57)

Pour Ross *et al.*, la perspective de Herrnstein renverse ainsi l'approche médicale traditionnelle, qui commence par supposer que les conduites incohérentes sont pathologiques, pour considérer au contraire que l'incohérence temporelle peut recevoir une justification *du point de vue de la rationalité économique*[121]. Toutefois, et c'est là une source importante de confusion (*cf.* chapitre 3), parmi l'ensemble des comportements impulsifs, certains ne sont pas optimaux *du point de vue de la théorie de l'apprentissage de la récompense*. En d'autres termes, il y a d'une part des conduites impulsives qui sont considérées comme rationnelles à la fois par la théorie économique et neuroscientifique, et, d'autre part, certaines conduites extrêmes, qui, bien qu'étant toujours compréhensibles comme rationnelles dans les termes de l'analyse économique, sont rejetées comme déviantes ou pathologiques.

La théorie neuroéconomique de l'addiction peut donc se comprendre comme une

121Nous rappelons que les comportements impulsifs et les renversements de préférence associés peuvent être compris comme rationnels dans la mesure où ils peuvent correspondre à la maximisation d'une fonction d'utilité inter-temporelle (hyperbolique). Un sujet impulsif qui regrette ses choix par la suite maximise bien son utilité dans cette perspective.

application particulière de la théorie du *reward learning* (*cf.* Ross *et al.*, 2008, p.135). Dans ce cadre théorique, l'addiction apparaît comme une « auto-stimulation » du système de la récompense. En effet, le système dopaminergique, au fur et à mesure de l'apprentissage, anticipe sur l'obtention de la récompense, si celle-ci est associée à un stimulus conditionnant, et réagit en conséquence à ce stimulus, et non plus à la gratification effective (*cf.* chapitre 3). Dans l'expérience de Schultz, Dayan et Montague (1997), une activité des neurones dopaminergiques est observée après conditionnement lorsque le stimulus sonore auquel est associé la récompense apparaît.

Le système de la récompense remplit donc une fonction essentiellement prédictive: « *au fur et à mesure de l'apprentissage, le système ne répond plus à l'obtention de la récompense elle-même mais à ses* [signaux] *prédicteurs* » (Ross *et al.*, 2008, p.138). Or, pour que cette anticipation se forme, deux éléments doivent être réunis. D'une part, il doit y avoir dans l'environnement des signaux (*cues*) associés de manière relativement stables à l'obtention d'une récompense. Mais, d'autre part, cette perspective de récompense doit conserver une degré minimal d'incertitude, pour pouvoir générer une activité dopaminergique. Si le récompense est obtenue à chaque fois que le stimulus de prédiction apparaît, avec une probabilité de 100%, il n'y a pas besoin de prédire quoi que ce soit. Dans les expériences sur le singe ou le rat, les travaux montrent ainsi que les signaux d'activité dopaminergique décroissent au cours du temps, par un effet d'érosion, si une récompense est distribuée à période régulière, de manière ininterrompue (*cf.* chapitre 3, section III).

Le sujet doit donc se représenter la récompense associée au signal prédicteur comme étant seulement potentielle, soit en délai et/ou en probabilité. C'est la raison pour laquelle il existe une séparation nette, sur le plan neuronal, entre la prédiction des récompenses et l'expérience subjective du plaisir. Le circuit dopaminergique encode le montant attendu de satisfaction, et l'étude de neuroimagerie de Berns *et al.* (2001) montre que la sensation de plaisir active d'autres régions du cerveau. Dans la littérature neurobiologique, il est ainsi fait référence à une différence entre le système des préférences (*liking system*) et le système du vouloir, qui fonctionne par anticipation, dans le but de motiver des actions (*wanting system*) (*cf.* Berridge, 2007).

Le caractère prédictif du système de la récompense constitue à la fois un atout et une source de défaillances, qui débouchent sur des addictions de nature diverses: addiction au jeu financier, à Internet, au sexe, troubles du comportement alimentaire, *etc.* Dans les conduites addictives, les individus parviennent en effet à identifier, pour reprendre les termes de Ross *et al.*, des « [signaux] *prédicteurs stables de surprise* » (Ross *et al.*, 2008, p.135). En d'autres

termes, ils trouvent dans leur environnement des signaux annonçant de manière certaine la perspective d'une récompense incertaine. Ceci permet de générer, à chaque apparition du stimulus prédicteur, une activité dopaminergique, source d'excitation. Par exemple, le système de la récompense d'un accroc aux paris financiers s'active à chaque fois que celui-ci voit la roulette tourner (ou la machine à sous tourner, le dé rouler, *etc.*). La principale source de plaisir de cet individu ne repose pas en fait dans la satisfaction associée aux éventuels gains financiers, mais dans la simple excitation liée la perspective d'un gain (mais aussi une perte). Le processus de choix devient plaisant en lui-même: « *ce qui conduit les individus à avoir un problème avec les paris financiers ets que cette activité est plaisante par elle-même* » (Ross *et al.*, 2008, p.165).

C'est donc le plaisir du jeu, qui, dans le cas du pari financier, explique les dérives addictives: « *les jeux financiers pris comme activité consistent à payer pour la possibilité d'être surpris* […]. *Le pari financier n'est simplement qu'un stimulation directe du système de la récompense* » (Ross *et al.*, 2008, 165). Pour Ross *et al.* le pari financier représente ainsi le prototype de l'activité addictive, parce qu'il constitue une dépendance sans objet, à l'inverse de l'addiction aux drogues, de l'alcoolisme, des troubles du comportement alimentaire, *etc.* Mais, plus généralement, cette manière de penser les addictions aux jeux financiers peut s'appliquer à tous les phénomènes addictifs. Dans le cadre théorique du *reward learning*, l'addiction s'explique par une dépendance à la surprise. L'administration régulière de surprise permet en effet à chaque fois de produire une erreur de prédiction, et ces « *erreurs de prédiction perpétuellement positives* » manifestent une « *auto-stimulation* » du système de la récompense (Ross *et al.*, 2008, p.170). La surprise fournit ainsi un équivalent aux substances addictives, comme la cocaïne ou l'héroïne, qui, par les propriétés chimiques agissent de manière similaire en produisant des pics d'activité dopaminergique artificiels.

Cela ne signifie pas que l'addiction ainsi définie implique un dysfonctionnement du système dopaminergique lui-même. Celui-ci continue, dans les conduites addictives, à évaluer les récompenses d'une manière cohérente avec la structure de son environnement. Si le pari financier peut déboucher sur un gain avec une probabilité non-nulle, il est logique que le déroulement du pari préalablement au résultat déclenche une activité dopaminergique. Le problème vient plutôt de la perception interne de l'environnement, qui tend à être exagérément biaisé en faveur d'un type de prédicteurs bien déterminés (les roulettes au casino par exemple), au détriment de tous les autres. Ces signaux ont l'avantage d'être associés de manière certaine à des récompenses incertaines. Ils mobilisent alors de manière croissante l'ensemble des ressources cognitives du décideur. En se concentrant sur ces prédicteurs, les

individus souffrant d'addictions font l'expérience d'une surprise récurrente, en dépit de ce qui se produit réellement. Ils envisagent alors l'environnement comme étant potentiellement une source importante de récompenses.

L'apport principal de la théorie neuroéconomique de l'addiction consiste à montrer que l'individu dépendant -l'*addict*- est toujours en un certain sens rationnel, dans la mesure où son comportement est toujours adapté à son environnement. Ce qui pose problème est l'environnement dans lequel cet individu évolue. Dans les conduites addictives, l'environnement (addictif) conduit à privilégier un certain type de récompenses au détriment de toutes les autres. La rationalité du décideur doit donc être appréciée de manière écologique, c'est-à-dire en en prenant pas seulement en compte les capacités cognitives internes de l'individu, mais plutôt la structure des récompenses offertes par l'environnement externe. Tous les individus sont donc rationnels et irrationnels en puissance, puisque l'irrationalité est liée à la nature addictive de l'environnement. Ces approches conduit donc à brouiller la frontière entre simple impulsivité et addiction, et donc entre rationalité et irrationalité.

B. Une définition pathologique de la rationalité économique: le rôle des critères cliniques

La théorie neuroéconomique de l'addiction tend à remettre en question le partage établi entre rationalité et irrationalité. En effet, l'impulsivité est la conséquence de l'actualisation hyperbolique des récompenses futures. Or celle-ci, comme le souligne Ross, est parfaitement rationnelle dans la perspective du néo-comportementalisme (*cf. supra*, Ross *et al.*, 2008, p.57). Pour Rachlin par exemple, il s'agit là d'une sorte de loi de la perception rendue nécessaire par les exigences de la vie biologique, qui oblige à favoriser l'environnement immédiat et proche (Rachlin, 2006, p.2). Tout l'enjeu consiste donc à distinguer ce qui relève d'une impulsivité « acceptable » du point de vue de l'organisme et ce qui relève de l'addiction, néfaste et morbide pour l'individu. Selon nous, ce problème est résolu en neuroéconomie par l'adoption implicite de critères cliniques de l'irrationalité: c'est parce que, en amont, certaines conduites, ou, plus précisément, certains environnements sont considérés comme potentiellement addictifs qu'il est possible ensuite de considérer que les individus impulsifs dans ces milieux particuliers sont irrationnels. Il convient de souligner que ce point de vue ne

correspond pas le plus souvent à celui des neuroéconomistes qui, comme Ross, envisagent précisément leur approche comme une alternative aux critères cliniques. La question de l'influence des critères cliniques dans les modèles neuroéconomiques de l'addiction sera cependant approfondie dans le dernier chapitre. L'enjeu ici, consistera plus simplement à montrer que le critère neuroéconomique de définition de l'addiction se distingue de celui qui est retenu en économie comportementale.

La démarche proposée par Ross *et al.* (*cf. supra*) peut apparaître paradoxale. Au départ, Ross et ses co-auteurs se posent un problème définitionnel; il s'agit de distinguer l'addiction des conduites impulsives mais rationnelles: « *pour la profession clinique, un problème scientifique important consiste à développer un cadre théorique permettant de tracer une frontière nette entre les formes de comportements répétitives saines ou normales et celles qui sont excessives et pathologiques* » (Ross *et al.*, p.1) Or, Ross *et al.* expliquent l'addiction à partir de l'hypothèse d'une actualisation hyperbolique des récompenses futures, et la recherche sur la loi d'égalisation des rendements a abouti à montrer que l'actualisation hyperbolique est toujours rationnelle (Ross *et al.*, 2008, p.57).

La question se pose donc de savoir ce qui distingue l'addiction de l'impulsivité. Le test de l'*Iowa Gambling Task* ne fournit pas d'évaluation solide en la matière. Des études montrent qu'un nombre important de sujets sains affichent des performances médiocres à l'IGT (*cf.* chapitre 5). Plus généralement, Ross *et al.* regrettent que la profession clinique se montre plutôt réticente à adopter une définition environnementale de l'addiction. Cela apparaît clairement dans le traitement médical de l'addiction au jeu, qui représente selon Ross et ses co-auteurs le prototype de l'addiction environnementale. Or s'il existe bien des tests comportementaux de l'addiction aux jeux financiers, comme comme le *South Oaks Gambling Screen* (Lesieur et Blume 1987), le jeu en lui-même n'apparaît pas en général comme problématique. La plupart des études portent plutôt sur la comorbidité avec d'autres troubles (Ross *et al.*, 2008, p.39).

Pour Ross *et al.*, la théorie neuroéconomique offre bien pourtant un critère objectif de définition de l'addiction au jeu, et, plus généralement, de toutes les addictions de type environnemental. A quoi peut-on reconnaître une comportement addictif dans une expérience de neuroéconomie? Il est d'abord évident que cette définition ne saurait reposer sur la nature addictive de la substance consommée, puisque Ross *et al.* proposent une explication environnementale de l'addiction dans laquelle « *les propriétés intrinsèques des drogues* [sont] *comme des cas particuliers de la structure plus globale et commune à toutes les addictions en général* » (Ross *et al.*, 2008, 164). Le gain monétaire obtenu dans les jeux de machines à sous

multi-jeux ne fournit pas non plus de critère objectif, puisqu'il est possible de concevoir des tâches dans lesquelles les sujets dépendants obtiennent des performances plus élevées que les sujets sains (*cf. supra*, section III).

La théorie proposée par Ross *et al.* suggère plutôt que le trait distinctif de l'addiction est d'abord et avant tout l'existence, sur le plan neuronal, d'une cascade d'erreurs de prédiction positives. Cette activité intense du système dopaminergique tend à inhiber les régions préfrontales qui permettent normalement de contrôler les réponses dopaminergiques. Ce phénomène neuronal est clairement et précisément mesurable. Au niveau comportemental, cette cascade d'erreurs de prédiction positives caractéristique des conduites addictives s'explique par l'interaction avec un « environnement addictif », dans lequel la perspective d'une récompense incertaine est associée de manière certaine avec un type de stimuli-prédicteurs. Sur le plan psychologique, les comportements addictifs sont donc source de désagréments et de souffrances de deux types. D'une part, la conduite devient obsessionnelle, et les individus manifestent les plus grandes difficultés à se concentrer que d'autres signaux prédicteurs. D'autre part, des phénomènes de manque apparaissent lorsque l'obtention régulière de surprise s'interrompt, en particulier lorsque l'individu n'a plus accès à l'environnement addictif.

Ces trois types de preuve -neuronales, comportementales, et psychologiques- sont considérées par Ross *et al.* comme des marqueurs objectifs de l'addiction. Pourtant, nous montrerons dans le dernier chapitre que ces critères sont en réalité insuffisants. Tout d'abord, l'existence, sur le plan neuronal, d'une cascade d'erreurs de prédiction positives est une condition nécessaire mais non suffisante pour qu'un comportement puisse être envisagé comme addictif, comme l'a reconnu Don Ross ultérieurement (Ross, 2011, p.39). Dans la mesure où les preuves psychologiques renvoient au seul vécu introspectif et à la souffrance ressentie par l'individu, la seule preuve « scientifique » repose sur la notion comportementale d'environnement addictif.

Toute la difficulté consiste alors, nous le verrons, à définir les environnements susceptibles d'être addictifs pour les agents. Par exemple, Ross *et al.* rejettent la possibilité que le *shopping* puisse constituer une conduite addictive (Ross *et al.*, 2008, 212-213), ce qui implique que les magasins ou les centres commerciaux ne puissent pas être regardés comme des environnements addictifs. Mais pourquoi ne pourrait-on pas être « accroc » à l'achat de vêtements de la même manière qu'un individu dépendant à l'héroïne? Même s'ils expriment des réticences à l'admettre, Ross *et al.* s'appuient en fait implicitement sur des critères cliniques pour effectuer le partage entre impulsivité « acceptable » et addiction. Si le

shopping, la navigation sur Internet ou le sexe ne peuvent être addictifs pour Ross *et al.*, c'est pour la simple raison qu'ils ne sont pas (encore) considérés comme tels par la communauté clinique[122].

Adopter des critères cliniques de définition de l'addiction signifie concrètement que l'étude neurobiologique s'appuie, en amont, sur la sélection d'une population de sujets ayant été diagnostiqués comme dépendants par les addictologues. Comme le reconnaissent Ross *et al.*, « *dans l'étude du jeu déréglé, les sujets utilisés dans les recherches sont généralement recrutés en utilisant des tests cliniques* » (Ross *et al.*, 2008), p.37). Étant donné le manque de précision de ces tests, les chercheurs utilisent le plus souvent des individus ayant développé des formes sévères d'addiction environnementale, pour lesquelles le diagnostic fait consensus. Il s'agit donc d'améliorer la précision des dispositifs de diagnostic clinique préexistants -grâce à de nouvelles tâches et de nouvelles techniques d'observation neurobiologique-, sans nécessairement remettre en cause radicalement leur principe.

L'influence des catégories cliniques joue donc à deux niveaux. Celles-ci permettent d'abord de définir les environnements ou les types de conduite pour lesquels un diagnostic d'addiction est susceptible d'être prononcé: cela sera le cas pour l'addiction au jeu par exemple, mais pas pour les achats compulsifs. Ensuite, pour ces environnements addictifs, les définitions médicales fournissent un principe de segmentation dans les populations de sujets utilisés, entre des individus sains et normaux. Ce calibrage comportemental sert par la suite à déterminer les variables neuronales caractéristiques d'une conduite addictive.

L'identification clinique des pathologies addictives ne joue donc pas seulement un rôle dans les études mobilisants des sujets dépendants. Comme cela a été envisagé dans la deuxième section, il est possible de faire valoir que la totalité des expériences en neuroéconomie fait référence implicitement aux critères cliniques, et ce même lorsqu'elles ne mobilisent pas de sujets « pathologiques ». En effet, le comportement rationnel dans de tels travaux est caractérisé comme le comportement moyen suivi par des sujets « normaux ». Plus précisément, les individus normaux sont ceux qui sont considérés comme « sains » dans le vocabulaire des neurobiologiste, c'est-à-dire ceux qui ont été déclarés comme tels par des médecins ou après avoir passé des tests cliniques avec succès. C'est donc encore une fois parce que les neuroéconomistes acceptent des définitions médicales comme données, qu'il est possible ensuite de sélectionner une population de participants « normaux ». C'est à partir de

122Cela ne signifie pas par ailleurs que ces conduites ne puissent être un jour être considérées comme potentiellement addictives, en fonction de l'évolution des pratiques cliniques. Ce point sera approfondi dans la suite de cette section.

ce premier tri que la rationalité des décideurs peut être évaluée quantitativement dans des tâches type machines à sous multi-jeux.

Cette approche introduit donc une confusion tout à fait révélatrice, selon nous, de l'importance des catégories médicales, entre rationalité et normalité. Est rationnel l'individu qui est normal, c'est-à-dire qui ne souffre d'aucune pathologie. Ce paradigme que nous avons qualifié de pathologique (Vallois, 2011) offre une compréhension à la fois limitée et étendue du choix rationnel. Il est limité parce qu'il s'applique à un nombre restreints de comportements ou d'environnements, pour lesquels l'impulsivité est susceptible de se transformer en addiction. La neuroéconomie des conduites addictives porte essentiellement sur l'addiction aux jeux financiers, les troubles du comportement alimentaire, et le problème de l'épargne des ménages. De surcroît, cette approche débouche sur une définition négative de la rationalité: celle-ci repose d'abord sur l'identification des comportements addictifs, et il n'y a pas à proprement parler de définition logique et axiomatique de la rationalité en tant que telle.

En même temps, d'un autre côté, il est manifeste que la définition neuroéconomique de l'addiction, et donc celle de la rationalité, peut faire l'objet d'une extension quasi-illimitée, suivant l'interprétation plus ou moins stricte retenue par les cliniciens. L'achat compulsif pourrait bien après tout représenter une forme d'addiction, puisque les individus accrocs au shopping font également états aussi de symptômes psychologiques associés à la dépendance, comme le manque par exemple. Tout laisse à penser que des comportements tels que l'addiction au shopping, à Internet ou au sexe soient prochainement reconnues comme addictives. La notion d'addiction comportementale a été élargie au cours des dernières années. Le centre d'addictologie de l'hôpital Brousse en France par exemple reconnaît comme pathologiques et traite désormais les addictions sexuelles et les achats compulsifs[123].

Il est significatif selon nous de ce point de vue de constater que Don Ross lui-même a progressivement suivi cette tendance. Dans son libre publié en 2008, il rejette le shopping, le sexe et Internet en dehors du champs des conduites potentiellement addictives (Ross *et al.*, 2008, p.66), mais a changé d'avis en la matière dans un article plus récent. Ross y considère que l'intégration prochaine des addictions comportementales dans la 5ème édition du Manuel diagnostique et statistique des troubles mentaux (DSM 5 - *Diagnostic and Statistical Manual of Mental Disorder),* édité par la *American Psychiatric Association* représente une avancée considérable. Cette nouvelle classification sera d'abord limitée à l'addiction aux jeux, mais, selon Ross, « *il y a un soutien de la communauté des utilisateurs du DSM en faveur de l'intégration additionnelle de l'addiction à Internet. La politique relativement conservative,*

123Voir http://www.centredesaddictions.org/

pour le moment, consistant à ne retenir que l'addiction aux jeux ne reflète que l'état plus limité des études empiriques en la matière » (Ross, 2011, p.39). Il ne fait donc pas de doutes, pour Ross, que la navigation sur Internet puisse être à l'origine d'addictions, ce qui va à l'encontre de ses positions défendues dans son ouvrage (*cf. supra*).

Le champ d'application de la théorie neuroéconomique de l'addiction dépend donc de définitions cliniques. L'évolution de ces définitions modifie donc la nature des comportements susceptibles d'être engagés dans des dérives addictives, comme le manifeste les changements de position de Don Ross concernant les addictions à Internet. En associant la rationalité à la normalité, les neuroéconomistes considèrent que le comportement rationnel n'est plus celui qui satisfait les axiomes d'une théorie (économique) de référence, mais celui qui n'est pas de nature pathologique. L'introduction des catégories cliniques dans la théorie économique du choix rationnel implique donc l'introduction de jugements de valeur relatifs à la question de savoir ce qui constitue l'individu « normal ». Cette approche remet donc directement en question le principe kahnemanien d'une séparation entre le domaine du descriptif et du normatif.

C. L'affirmation d'une identité propre: distanciation critique et réappropriation des figures historiques de la discipline

La théorie de l'addiction est l'occasion d'une distanciation critique de la neuroéconomie avec les *behavioral economics*. En effet, là où les neuroéconomistes proposent une compréhension environnementale du comportement addictif, appuyé sur des critères cliniques, les économistes comportementalistes s'en remettent à une explication en termes de limites cognitives internes aux décideurs, liés à des effets de cadrage (*framing effects*), en postulant une stricte séparation des registres descriptifs et prescriptifs. En se démarquant de cette perspective kahnemanienne sur l'addiction, Ross et ses coauteurs (2008) mettent ainsi en avant l'identité théorique propre de la neuroéconomie, en lien avec son héritage néo-comportementaliste.

La théorie de l'addiction de Ross *et al.* (2008) s'oppose à plusieurs modèles qui ont été développés en économie comportementale. Ces modèles peuvent être distingués en deux types, selon qu'ils s'appuient, ou non, sur l'hypothèse d'une actualisation quasi-hyperbolique

proposée par Laibson (1997) pour rendre compte des comportements addictifs. En s'inspirant de Laibson (1997), Gruber et Köszegi (2001) construisent par exemple un modèle dans lequel l'addiction résulte d'un trop fort biais en faveur du présent. D'autres modèles font cependant l'économie de cette hypothèse d'actualisation temporelle quasi-hyperbolique et supposent que l'addiction résulte d'un autre type d'effet de cadrage. Dans la théorie des « *facteurs viscéraux* » (*visceral factors*) de Loewenstein notamment, l'addiction résulte de pulsions irrépressibles, quasi-analogues à des réflexes pavloviens, liés à la peur du manque et du retrait. Ces processus viscéraux représentent aussi d'une certaine manière un effet de cadrage, dans la mesure où ils conduisent à déformer le mode de présentation des options possibles. Or, la théorie neuroéconomique remet en cause ce type d'explication car, comme le soulignent Ross *et al.*, « *la peur du retrait n'est certainement pas le facteur le plus important dans le maintien de l'attention* » (Ross *et al.*, 2008, p.156). En effet, « *ce qui amène certaines personnes à avoir un problème avec les jeux financiers est que le jeu est plaisant en lui-même* » (Ross *et al.*, 2008, p.156): un individu accroc au jeu prend donc un réel plaisir dans le jeu, et il ne joue pas pour éviter les conséquences déplaisantes liées à l'abstinence. Par ailleurs, les observations cliniques vont à l'encontre de l'hypothèse proposée par Loewenstein, puisque la plupart des individus dépendants aux drogues dures ne subissent pas l'expérience du manque décrite par Loewenstein (*cf.* McAuliffe, 1982).

Le modèle de Daniel Read (2001) avance une explication de l'addiction en termes d'« *actualisation temporelle subadditive* ». Celle-ci signifie que « *l'actualisation d'une utilité future est plus importante lorsque le délai est divisé en sous-intervalles que lorsqu'il n'est pas divisé. Cela peut expliquer le résultat le plus important associé à l'actualisation hyperbolique: une impatience décroissante, ou une relation inverse entre le taus d'actualisation et le délai* » (Read, 2001, p.5). Il s'agit donc là d'un modèle qui permet de rendre compte des phénomènes de renversement des préférences associés à l'actualisation quasi-hyperbolique de Laibson en postulant simplement que les arbitrages inter-temporels des agents dépendent d'un effet de cadrage portant sur la représentation subjective du délai. Un délai d'une journée n'aura pas la même importance selon qu'il soit représenté subjectivement comme la succession de vingt-quatre périodes d'une heure, ou comme l'écoulement d'une période unique d'un jour. D'une manière certes différente mais relativement proche dans son principe, Gul et Pesnedorfer (2001) proposent également une explication du renversement des préférences à partir d'effets de cadrage. Le *framing* ne porte pas ici sur l'actualisation temporelle, mais sur la perception de certaines options, associées à une tentation, l'idée étant

que le renoncement à ces options est associé subjectivement à une diminution d'utilité pour les autres options.

Pour Ross *et al.*, l'explication de l'addiction en termes de *framing effects* n'est pas nécessairement contraire à la théorie neuroéconomique de l'addiction: « *il n' y a pas à choisir entre une explication en terme de cadrage et une autre, picoéconomique. Ce n'est certainement pas faux de dire que les individus qui souffrent d'impulsivité devraient essayer de modifier leurs cadres* » (Ross *et al.*, 2008, p.246). L'approche « *picoéconomique* » renvoie aux travaux de Georges Ainslie, pour qui le choix inter-temporel se comprend comme un conflit entre plusieurs mois inter-temporels (*cf.* Ainslie, 1991). Dans cette perspective, les individus peuvent devenir plus patients en envisageant les problèmes d'arbitrage inter-temporels non pas comme des choix discrets, mais comme s'ils engageaient la totalité des arbitrages futurs du même type que l'agent sera susceptible de réaliser: il faut « *grouper* » (*bundle*) ces problèmes d'arbitrage et considérer qu'ils ne portent pas sur la consommation d'utilités discrètes, mais sur les principes que l'individu s'engage à suivre au cours de son existence. Or, comme le soulignent Ross *et al.*, « *le bundling, après tout, est une sorte de cadrage* ».

La théorie neuroéconomique de l'addiction, d'inspiration « *picoéconomique* », postule donc elle aussi des effets de cadrage:« *pour être effective, une règle personnelle doit être gardée en mémoire, et la série de choix individuels non-impulsifs qu'elle propose doit permettre plus de récompense que la série de choix impulsifs […]. Cela requiert la construction et la permanence d'états mentaux motivationnels qui incluent des considérations indépendantes du contexte immédiat […]. D'une manière générale, à la fois la maintenance et l'échec des règles personnelles sont des effets de cadrage. Le handicap cognitif est ici une diminution de la capacité à constituer un cadrage sophistiqué* » (Ross *et al.*, 2008, p.111).

Cependant, les modèles de Read (2001) ou de Gul et Pesendorfer rencontrent une limite en en faisant pas référence à l'actualisation hyperbolique. En effet, « *l'avantage d'insister sur le choix inter-temporel comme actualisation hyperbolique est que cela permet de faire référence clairement à une « rationalité économique »* »(Ross *et al.*, 2008, p.112). Il convient de bien préciser ce qu'entendent ici Ross *et al.* par « *rationalité économique* ». De leur point de vue, il n'est pas faux de parler de *framing effects* à propos de l'addiction, mais ce concept d'effet de cadrage ne permet pas d'expliquer l'origine et la formation de cette erreur. Par exemple, Read postule que la représentation subjective du délai influe l'ampleur de l'actualisation (Read, 2001). Mais qu'est ce qui fait qu'un individu soit incliné à considérer

qu'une journée représente vingt-quatre fois une heure plutôt qu'une période unique? Évidemment, pour Read, le cadrage dépend de la manière avec laquelle la question est posée au sujet. Son étude, précisément, vise à afire apparaître des différences intra-individuelles dans les taux d'actualisation selon le mode de présentation des problèmes d'arbitrage inter-temporels. Cependant, cette approche demeure limitée à la simple association entre des environnements ou des modes présentation des options et des types de *framing effects*, sans réellement expliquer la dynamique de l'interaction entre l'individu et son environnement.

Les modèles de l'addiction reposant sur la notion kahenmanienne d'effet de cadrage sont donc condamnés du point de vue de Ross à identifier des « biais » ou des « erreurs » dans l'allocation inter-temporelle des utilités par l'individu. Ces effets de cadrage peuvent être liés à des facteurs environnementaux, et notamment à la manière avec laquelle les problèmes d'arbitrage sont présentés aux sujets. Mais pour Ross, ces modèles n'invitent « *à aucune élaboration économique* » (Ross *et al.*, 2008, p.112), dans la mesure où il n'y a pas, à proprement parler, d'explication de l'origine et de la formation de cette erreur de raisonnement. A l'inverse, le paradigme du *reward-learning* permet de rendre compte qu'un même individu puisse adopter des attitudes variées face à un même problème d'allocation des consommations inter-temporelles, en fonction des paramètres spécifiques de sa fonction d'apprentissage.

Ross *et al.* concentrent leurs critiques sur les modèles de Loewenstein (1999), Read (2001) et Gul et Pesendorfer (2001), mais celles-ci peuvent valoir également à l'encontre de celui de Rabin et Köszegi (2001). En effet, ces deux auteurs supposent que l'addiction est liée à l'actualisation quasi-hyperbolique du type Laibson (1997), ce qui revient donc, encore une fois, à expliquer les comportements addictifs par un effet de cadrage, ici lié au paramètre β de la fonction d'actualisation, qui, nous le rappelons, représente le degré de myopie temporelle du décideur.

L'ouvrage de Ross *et al.* (2008) marque une étape importante dans le développement de la neuroéconomie, en mettant en évidence clairement les différences conceptuelles entre l'approche proposée par les neuroéconomistes de celle des *behavioral economics*. Bien que limité au problème de l'addiction, le livre peut se lire comme un authentique manifeste de la neuroéconomie, affirmant son autonomie et son indépendance théorique. Ross *et al.* prennent leur distance avec l'économie comportementale d'inspiration kahnemanienne, mais ils cherchent également à réinscrire la discipline dans le prolongement de cette tradition théorique que nous avons appelée néo-comportementalisme ou science quantitative de la

motivation, et que Ross et ses co-auteurs désignent plutôt par le nom de « picoéconomie », en référence aux travaux de Georges Ainslie.

Conformément à notre hypothèse d'interprétation historique, Ross *et al.* voient dans les premiers travaux de Herrnstein le point de départ du questionnement neuroéconomique (Ross *et al.*, 2008, p.51). Par la suite, « *beaucoup plus que n'importe quel auteur* », l'économiste et psychiatre Georges Ainslie aurait contribué au développement de ce programme de recherche, en approfondissant les conséquences théoriques des travaux de Herrnstein (Ross *et al.*, 2008, p.66).

Ces éléments d'auto-histoire qui apparaissent aux détours de l'explication théorique, et qui convergent avec notre récit historique, ne doivent cependant pas faire perdre de vue que Don Ross est avant tout un neuroéconomiste, c'est-à-dire un chercheur participant directement au développement de la discipline. Même s'il se considère lui-même comme méthodologue, philosophe et historien de l'économie, Don Ross comprend ses réflexions comme des contributions pouvant avoir des implications théoriques directes pour l'économie et les neurosciences. Il est donc dans un position purement interne au champ lui-même, ce qui signifie pas que la portée de ses analyses soit nulle. Bien au contraire, son ouvrage sur l'addiction a permis selon nous de renforcer l'identité théorique propre des neuroéconomistes, et de la différencier de celle des économistes comportementalistes. Il est significatif de constater à cet égard que Don Ross a lui-même inventé l'expression d'« *économie comportementale dans le scanner* », afin d'inciter les neuroéconomistes à se démarquer des *behavioral economics* (Ross, 2008). Pour Ross, la véritable neuroéconomie ou « économie neurocellulaire » est celle qui, dans le sillage des travaux de Glimcher, réinvente les rapports de l'économie à la psychologie et à la psychologie (Ross, 2008).

Conclusion du chapitre 5

Entre la publication, en 2005, par Camerer, Rangel et Montague, de l'article intitulé « la neuroéconomie: comment les neurosciences peuvent guider l'économie » (Camerer, Rangel et Montague, 2005) et celle, en 2008, de l'ouvrage de Don Ross consacré aux addictions (Ross *et al.*, 2008), la neuroéconomie évolue considérablement. Au cours de cette période courte mais décisive, les prétentions de réforme de la théorie économique par les neurosciences ont été à l'origine de nombreuses critiques en économie. Au delà de l'opposition de principe, appuyée sur la défense des préférences révélées, selon laquelle l'économie n'aurait pas, par nature, à traiter de processus cognitifs sous-jacents aux choix qu'elle observe, les critiques de la neuroéconomie ont souligné les difficultés méthodologiques liées à l'approche réductionniste de l'économie comportementale dans le scanner.

Cette remise en question du potentiel théorique, pour l'économie, de la neuroimagerie a paradoxalement été salutaire. En effet, elle a conduit les neuroéconomistes à préciser le sens des concepts cognitifs et comportementaux utilisés dans le compte-rendu de leurs expériences. Ce travail de clarification conceptuel a ainsi fait apparaître un écart théorique entre les approches dualistes héritées de la psychologie kahnemanienne, et une conception unitaire de la cognition, référée à un cadre néo-comportementaliste. La controverse qui eu lieu, à propos du choix inter-temporel, entre Glimcher et Kable (2007) d'un côté et McClure *et al.* (2007) de l'autre, est de ce point de vue révélatrice, car elle oppose selon nous les deux phases du développement de la neuroéconomie. L'abandon progressif de la référence aux émotions et à la théorie des marqueurs somatiques de Damasio, au profit de la notion de monnaie neuronale commune, témoigne également dans ce domaine du divorce entre *behavioral economics* et neuroéconomie.

Par ailleurs, la valeur théorique et la fiabilité des instruments de neuroimagerie a pu être défendue en considérant que la fiabilité de ces instruments est renforcée par la psychopathologie. La référence à la connaissance clinique des pathologies offre ainsi une réponse à l'argument de l'inférence inverse. En mettant en avant les dimensions pathologiques, biologiques et évolutives de la décision, la neuroéconomie de la fin des années 2000 converge avec les éléments distinctifs du paradigme néo-comportementaliste. Elle s'affirme alors comme un projet de psychiatrie économique. Dans le domaine du choix inter-temporel, la théorie neuroéconomique de l'addiction témoigne clairement de ce croisement de l'économie et des préoccupations relatives à la maladie mentale.

Conclusion de la deuxième partie

Les premiers travaux de neuroéconomie, publiés au cours de l'année 2001, confirment notre interprétation historique. Ils se situent en effet dans le droit prolongement de ce que nous avons appelé, dans la première partie, la préhistoire théorique de la neuroéconomie. Les trois expériences qui ont ici été analysées (Berns, McClure, Pagnoni et Montague, 2001; Knutson *et al.*, 2001; Delgado *et al.*, 2001) visent à mettre en évidence, comme chez le singe, un signal d'anticipation de la récompense. Les recherches de Wolfram Schultz sont notamment citées comme une référence importante.

Alors que l'on aurait pu ainsi concevoir un progrès continu des recherches, de l'animal au sujet humain, grâce aux nouvelles technologies, les neuroéconomistes se sont néanmoins très rapidement démarqués de leurs prédécesseurs travaillant sur le primate et le pigeon. Le passage à l'homme entraîne deux changements importants pour la discipline. Le premier est lié à l'étude des régions pré-frontales. Celles-ci sont impliquées dans des facultés de contrôle et de planification, conçues généralement comme spécifiques à la cognition humaine. Pour un chercheur comme McClure par exemple, le fort développement chez l'homme du cortex pré-frontal oblige le chercheur à introduire une nette séparation entre l'animal et l'humain, celui-ci étant caractérisé par une opposition entre la « raison » ou les capacités de contrôle et les émotions ou instincts partagées avec l'animal (McClure, Laibson, Loewenstein et Cohen, 2004, p.504).

Les fonctions de contrôle, mémorisation et de planification temporelle peuvent néanmoins s'intégrer dans le cadre théorique du *reward learning* (Bogacz *et al.*, 2007). De manière plus décisive, c'est l'arrivée des économistes comportementalistes qui est à l'origine des bouleversements les plus significatifs pour la discipline. A partir de 2002-2003, un basculement important s'opère. La tendance à opposer raison et émotions se généralise dans les travaux de neuroéconomie. Cette perspective dualiste s'inspire largement des travaux d'Antonio Damasio, pionnier de la recherche sur les émotions et le cerveau.

Au cours de la deuxième partie des années 2000, la neuroéconomie se distingue progressivement de l'économie comportementale. L'économie comportementale dans le scanner suscite de nombreuses contestations internes aux *behavioral economics*, sur l'usage des nouveaux instruments et en particulier de la neuroimagerie. Les neuroéconomistes se replient alors vers des interprétations s'inscrivant plutôt le paradigme néo-comportementaliste,

au détriment du dualisme défendu par Kahneman. Cette voie de recherche s'avère plus féconde, puisqu'elle se dégage de toute visée réductionniste. Nous avons pu ainsi dégager, dans le domaine du choix inter-temporel, un noyau dur théorique de la neuroéconomie. La théorie neuroéconomique des addictions rentre compte ainsi de la manière avec laquelle l'idéologie scientifique de psychiatrie économique peut approcher les problèmes d'arbitrage temporel.

Il convient de souligner que cette interprétation n'est qu'une reconstruction *a posteriori* du champ. Elle ne coïncide pas forcément avec l'émergence historique constatée de la discipline. D'une part, les interprétations dualistes, ainsi que les références à la notion d'émotion, subsistent encore aujourd'hui dans de nombreux travaux de neuroéconomie. D'autre part, la neuroéconomie s'est principalement développée au départ à partir de l'étude des interactions sociales, plutôt que celle du choix inter-temporel. Les travaux sur les addictions et le *reward learning* renvoient à un domaine plus tardif. Néanmoins, le noyau dur théorique que nous avons identifié dans cette seconde partie permet, rétrospectivement, d'envisager les travaux sur le choix inter-personnel comme relevant également d'un paradigme néo-comportementaliste.

TROISIEME PARTIE – LA CONQUETE DE NOUVEAUX OBJETS : LE CHOIX INTER-PERSONNEL ET L'ANALYSE DU BIEN-ETRE

Le traitement du choix inter-temporel par les neurosciences fait apparaître, au cours des années 2000, un retour progressif à la notion néo-comportementaliste de « monnaie neuronale commune » (Berns et Montague, 2002). Les interprétations dualistes ne disparaissent pas, mais se distinguent alors plus nettement du cadre théorique propre au *reward learning*. Après une période de tension marquée par la controverse entre Kable et Glimcher (2007) d'un côté et McClure *et al.* (2007) de l'autre, les chercheurs néo-comportementalistes prennent conscience de travailler à l'intérieur d'un paradigme différent de celui de l'économie comportementale (*cf.* Voon *et al.*, 2010).

L'enjeu de cette troisième partie est de montrer que ce noyau théorique néo-comportementaliste que nous avons dégagé des travaux sur le choix inter-temporel, peut s'étendre et s'appliquer à deux autres domaines du comportement: en particulier, le choix inter-personnel ou les interactions sociales, et le bien-être ou l'analyse normative de la décision. Il ne s'agit pas à proprement parler d'une « application » d'une théorie plus ancienne à de nouveaux objets, puisque ces trois domaines (choix inter-temporel, interpersonnel et analyse normative) se sont développés simultanément. Dans une perspective rétrospective, notre étude a pour but ici de reconstruire une cohérence aux recherches qui ont été entreprises, en montrant, comme pour le choix inter-temporel, l'évolution d'une approche type « économie comportementale dans le scanner » vers un paradigme néo-comportementaliste.

L'émancipation de la neuroéconomie et sa distanciation critique vis à vis des *behavioral economics* sont cependant moins nettes dans le cas des interactions sociales. L'absence de controverses explicites entre les partisans des deux approches rend notamment plus difficile, et dans une certaine mesure plus artificielle, leur distinction. Cependant, il est possible de suggérer là aussi l'émergence de travaux visant à appliquer la notion d'apprentissage de la récompense « sociale » au choix inter-personnel (chapitre 6. L'analyse du choix inter-personnel: la neuroéconomie entre neurosciences sociales, neuro-éthique et économie comportementale). La psychiatrie économique peut ainsi s'appliquer non seulement aux addictions mais aussi aux troubles de la cognition sociale.

L'analyse normative du comportement et du bien-être constitue un cas spécial. Celle-ci

peut se concevoir, comme l'étude des interactions sociales, comme une extension de la notion de *reward learning* à un nouvel objet. La neuroéconomie s'empare effectivement ici d'un nouveau domaine et lui applique un traitement original, à partir de la notion de pathologie. En même temps, l'analyse normative vaut plus qu'une simple application supplémentaire de la théorie du *reward learning*. Elle fait également figure d'aboutissement à notre enquête historique. La réflexion proposée par les neuroéconomistes sur le bien-être montre en effet que la neuroéconomie peut se comprendre littéralement comme un projet de psychiatrie économique, visant à appliquer un regard médical sur la définition du choix rationnel (chapitre 7. Paternalisme Libertarien et Psychiatrie Économique. L'apport de la neuroéconomie à l'analyse comportementale du bien-être).

Chapitre 6. L'analyse du choix inter-personnel: la neuroéconomie entre neurosciences sociales, neuro-éthique et économie comportementale

Dans la deuxième partie des années 2000, la neuroéconomie s'affirme progressivement comme une discipline autonome, en proposant notamment une modélisation originale du choix inter-temporel (*cf.* chapitre 5). Cependant, la neuroéconomie ne se limite pas à l'analyse de la décision individuelle. Les neuroéconomistes abordent également l'étude des choix inter-personnels, c'est-à-dire des choix dans lesquelles le comportement des autres joueurs constitue une variable stratégique pour le décideur. Au départ, entre 2001 et 2003, ce sont d'ailleurs des travaux de neuroimagerie portant sur le thème des interactions sociales, et non sur le choix individuel, qui popularisent le terme de neuroéconomie dans la littérature (McCabe *et al.*, 2001, Rilling *et al.*, 2002 ; Sanfey *et al.*, 2003).

Ce succès de la neuroéconomie que Fehr et Camerer qualifient de « *sociale* » (Fehr et Camerer, 2007) mérité cependant d'être nuancé. En effet, les travaux de neuroimagerie dans ce domaine se contentent le plus souvent de reproduire à la fois des protocoles et des interprétations préexistants, empruntés à l'économie comportementale. Le champ des interactions individuelles, tel qu'il est étudié par la neuroéconomie sociale, est ainsi restreint à trois types de jeu : le jeu de l'ultimatum et du dictateur, le jeu dit de la confiance ou de l'investisseur, et le dilemme du prisonnier. Le problème n'est pas que cette restriction offre une vision appauvrie des interactions sociales, mais que cette restriction soit directement reprise de l'économie comportementale, sans apport spécifique des neurosciences. En peinant à s'émanciper de l'économie comportementale, la neuroéconomie sociale fournit donc aux adversaires de l'économie comportementale dans le scanner ses meilleures critiques.

Ernst Fehr et Colin Camerer sont les principaux économistes-promoteurs de cette utilisation instrumentale de la neuroimagerie, dans l'analyse du choix interpersonnel. La forte tutelle exercée par l'économie comportementale sur la neuroéconomie sociale en fait un sous-domaine tout à fait particulier de la neuroéconomie (I. L'étude des interactions sociales: un sous-domaine particulier de la neuroéconomie). Dans la perspective de l'économie comportementale dans le scanner, les interactions individuelles et leurs mécanismes neuronaux sous-jacents sont interprétés en termes d'émotions et de préférences sociales (II.

Émotions et Préférences sociales le choix inter-personnel vu par l'économie comportementale dans le scanner). Cependant, l'enjeu consistera à montrer qu'une autre lecture de ces expériences est possible, en mettant celles-ci en rapport avec les travaux portant sur le *reward learning*. Ces interprétations conduisent notamment à substituer aux notions de préférences et d'émotions celles d'intelligence sociale (III. De l'apprentissage de la récompense au choix inter-personnel : la notion d'intelligence sociale). Les études de neuroéconomie sur les interactions individuelles peuvent ainsi être intégrées au sein du paradigme néo-comportementaliste. L'ancrage « pathologique » de ce programme de recherche se donne à voir dans ce domaine à travers les divers troubles et dysfonctionnement de l'intelligence ou cognition sociale (IV. Les implications prescriptives. Troubles et pathologies de l'intelligence sociale).

I. L'étude des interactions sociales: un sous-domaine particulier de la neuroéconomie

La neuroéconomie sociale est généralement considérée comme un sous-domaine à part de la neuroéconomie. Dans un article de 2003 visant à décrire les avancées principales de la neuroéconomie alors naissante, Colin Camerer divise ainsi la jeune discipline en trois sous-domaines principaux, avec les études portant sur les interactions individuelles d'un côté et celles concernant les thèmes de l'apprentissage de la récompense et du *self control* de l'autre (Camerer, 2003). La neuroéconomie sociale a pour principale spécificité de s'être construite à partir d'un usage purement instrumental de la neuroimagerie. Il s'agit en effet, dans le cadre de protocoles empruntés aux *behavioral economics*, de confirmer, au niveau cérébral, des théories explicatives du comportement préexistantes. Cette approche a évidemment un attrait immédiat pour les économistes comportementalistes, mais soulève aussi des critiques relatives au réductionnisme dont elle est porteuse.

Dans une étude de synthèse intitulée « la neuroéconomie sociale : les mécanismes neuronaux des préférences sociales », Ernst Fehr et Colin Camerer définissent la neuroéconomie sociale comme une nouvelle discipline qui « *combine les outils des neurosciences sociales avec des tâches bien structurées empruntés à la théorie économique. Ces protocoles sont associés à des prévisions théoriques concernant le jeu rationnel et l'efficacité sociale des équilibres, qui sont utiles pour interpréter les résultats et établir des régularités empiriques d'une étude à l'autre »* (Fehr et Camerer, 2007, p.419). Une telle définition assigne donc aux neurosciences une tâche plutôt limitée. Il s'agirait d'un apport purement technique, en enregistrant à l'aide des « outils » neuroscientifiques -la neuroimagerie essentiellement- des variables neurophysiologiques. La conception des protocoles et l'interprétation des comportements relèveraient quant à eux de l'économie.

Une telle démarche revient donc, littéralement, à faire de l'économie comportementale dans le scanner, puisque les théories et les expériences des *behavioral economics* sont directement reproduites dans le scanner. L'objectif est de réduire les concepts de l'économie comportementale à des activité cérébrales. Comme le souligne Bourgeois-Gironde (2010, p.229), cette neuroéconomie d'inspiration réductionniste « *inverse* » la démarche traditionnelle des neurosciences. Les économistes intéressés par un usage instrumental de la neuroimagerie cherchent en effet à confirmer des hypothèses explicatives du comportement par l'observation de l'activité cérébrale. L'économie comportementale fournit au départ une

tâche ou protocole A, ainsi qu'une théorie explicative de ce jeu affirmant que celui-ci engage le processus cognitif X. L'activité observée dans la zone Z lors de ce jeu permet de confirmer la mise en œuvre de X dans A, et donc la théorie initiale. A l'inverse, les neurobiologistes se donnent plutôt d'une fonction X, qu'il essayent de caractériser au niveau comportemental et cérébral. Ils conçoivent pour cela un protocole permettant d'observer la mise en œuvre de X. Ils observent alors une activité dans la zone Z, qu'ils associent à l'engagement de la fonction étudiée (Bourgeois-Gironde, 2010, p.239).

L'étude de de Quervain et al. (2004) fournit une bonne illustration de cette neuroéconomie sociale dans le scanner. Les auteurs y montrent, dans une variante du jeu de l'ultimatum avec une possibilité de sanctionner les joueurs égoïstes, que la punition dite altruiste active le striatum. Or cette région du cevreau est associée à la recherche de récompense. Cette observation est prise comme une confirmation de la théorie de la punition altruiste de Fehr (Fehr et Schmidt, 1999), selon laquelle la punition est source de plaisir (*cf. infra*, et de Quervain *et al.*, 2004).

Conformément à ce qui a été envisagé dans la première section du chapitre 5, l'économie comportementale dans le scanner rencontre néanmoins deux difficultés méthodologiques, exposées notamment par Harrison (2008-a, 2008-b). La première concerne l'inférence des processus cognitifs à partir des activités cérébrales. Lorsque les neuroéconomistes déduisent de l'activité Z l'engagement du processus X, ils s'appuient sur d'autres études ayant montré préalablement que l'engagement de ce même processus impliquait l'activité Z. Or cela n'entraine pas nécessairement que toute activité Z soit équivalente à l'engagement de X (*cf.* chapitre 5, section I).

Ce type de raisonnement, qualifié d'inférence inverse (Poldrack, 2006), est particulièrement fréquent dans la neuroéconomie sociale. Par exemple, dans leur étude célèbre portant sur le jeu de l'ultimatum, Sanfey *et al.* observent une activité du cortex dorsolatéral préfrontal (CDLPF), qu'ils interprètent de la manière suivante : « *le cortex dorsolatéral préfrontal (CDLPF) a été associé à d'autres processus cognitifs, comme la maintien et la poursuite d'objectif ou le contrôle exécutif (Miller et Cohen, 2001 ;Wagner et al., 2001). Par conséquent, l'activation observée de cette zone lors de la réponse à des offres inéquitables peut être associée à la représentation et au maintien actif des demandes cognitives de la tâche, qui consiste ici gagner autant d'argent que possible. Une offre inéquitable est plus difficile à accepter, comme l'indique le taux plus élevé de rejet de ces propositions, et il en résulte une charge cognitive plus importante pour que le participant puisse surmonter la forte*

impulsion émotionnelle à rejeter l'offre inéquitable » (Sanfey *et al.*, 2003,p.1757).

Cette citation semble faire apparaître, sur le plan logique, un cas exemplaire d'inférence inverse: l'explication proposée de l'activité du CDLPF n'est justifiée en dernière analyse qu'à partir d'autres études de neuroimagerie (Miller et Cohen, 2001 ;Wagner *et al.*, 2001). Le type d'explication proposé ici par Sanfey *et al.* apparaît très critiquable pour le neurobiologistes habitués à raisonner en sens inverse, du cognitif au cérébral (cf. supra), mais aussi du point de vue de l'économiste non-spécialiste des neurosciences. En effet, la neuroéconomie dans le scanner tend à accepter comme données des interprétations comportementales pré-établies, qui font pourtant l'objet de débats importants en économie comportementale. Ici, Sanfey *et al.*, admettent que « *le taux plus élevé de rejet de ces propositions* » reflète le fait qu'une offre inéquitable soit « *plus difficile à accepter* », et soit associée à une « *charge cognitive plus importante ».* Or, il existe de très importantes discussions théoriques au sein des *behavioral economics* concernant l'interprétation comportementale et cognitive des résultats du jeu de l'ultimatum. La lecture particulière qui est proposée ici repose sur une simplification de ces enjeux théoriques.

Les interprétations comportementales proposées ici par Sanfey *et al.* sont donc relativement grossières du point de vue d'un économiste comportementaliste spécialiste de ces question. Les neurobiologistes ignorent le plus souvent qu'il n'existe pas une seule et unique fonction cognitive pouvant être associé à un comportement donné dans une expérience, mais qu'il y a, selon la nature du protocole, plusieurs artefacts possibles (*cf.* chapitre 5). En ce qui concerne le jeu de l'investisseur, qui a fait l'objet de très nombreuses études de neuroimagerie, Glenn Harrison souligne par exemple que cette tâche « *offre un mélange prodigieux d'artefacts […]. Plusieurs choses différentes peuvent inciter quelqu'un à envoyer de l'argent, et à en renvoyer en retour, et ces choses sont bien connues. Chaque joueur peut être altruiste envers l'autre joueur. Ils peuvent aussi en vouloir à l'expérimentateur, et chercher à lui prendre autant d'argent que possible. La « confiance » peut aussi n'être que l e reflet d'un goût prononcé pour le risque. Les joueurs peuvent aussi envisager ce jeu unique comme s'il était répété. Par conséquent, à moins que l'on ne souhaite définir les termes de « confiance » et de « digne de confiance » par un amalgame de ces différentes motivations, il est nécessaire de concevoir des protocoles permettant de contrôler ces artefacts possibles* » (Harrison, 2008, p.320)[124].

124Il faudrait ajouter que l'altruisme lui-même comporte plusieurs artefacts possibles. Les économistes comportementalistes distinguent plusieurs motivations possibles pour les conduites altruistes: l'aversion à l'inéquité (Fehr et Schmidt, 1999), l'élan altruiste (*warm glow*, Andreoni, 1993), la réciprocité (Berg, Dickhaut et McCabe, 1995), *etc.* Un autre problème important concerne la question de savoir si ces choix altruistes portent sur les conséquences ou les résultats des actions, ou s'ils visent plutôt les intentions, comme

En reprenant directement les protocoles des *behavioral economics*, la neuroéconomie sociale est ainsi particulièrement vulnérable à la critique d'économistes comportementalistes soucieux d'exactitude dans l'interprétation cognitive des résultats. C'est la raison pour laquelle les attaques théoriques contre la neuroéconomie se concentrent généralement sur ce sous-domaine particulier (*cf.* Harrison, 2008-a).

La plus grande méfiance des économistes à l'égard de la neuroéconomie sociale, par rapport aux autres branches d'analyse de la neuroéconomie, s'explique peut être aussi par sa proximité avec des thèmes plus philosophiques. Un important programme de recherche, connu sous le nom de « neuroéthique » ou « philosophie expérimentale » (Cova, 2011; Ogien, 2011; Knobe et Nichols, 2007), s'est en effet développé au cours des dernières années autour des neurosciences et de la psychologie morale. Les travaux de neuroéconomie sociale portant sur la coopération ou l'altruisme sont en lien direct avec ceux de la neuroéthique, et convergent sur de nombreuses interprétations du choix éthique, en mettant en avant notamment une dimension émotionnelle ou affective (Greene, 2001; Greene et Haidt, 2002) ou intuitive (Haidt, 2002).

Quoique partageant des interrogations et des préoccupations similaires à la neuroéconomie sociale, les études de philosophie expérimentale ne seront pas abordées ici, parce qu'elles ne s'appuient pas sur des protocoles directement inspirés de l'économie mais plus généralement sur des dilemmes moraux assez variés. Cependant, cette proximité assumée par certains neurobiologistes avec la philosophie a souvent pour effet d'élargir l'interprétation des résultats expérimentaux à des discussions beaucoup plus générales, concernant par exemple la nature humaine. Paul Zak, chercheur influent dans ce domaine, offre sans doute la meilleure illustration de cette ambition philosophique. Il considère en effet que ses études portent sur « *la physiologie des sentiments moraux* » (Zak, 2010) et permettent de démontrer scientifiquement les thèses des philosophes et économistes du passé: « *mes découvertes pourraient être résumées de façon succincte en disant: Adam Smith avait raison. Les sentiments de fraternité et d'empathie nous rendent vertueux et nous détournent du vice. Mes études neuroéconomiques révèlent la physiologie sous-jacente à Adam Smith* » (Zak, 2010, p;10).

Zak défend par ailleurs des propositions méthodologiques assez naïves, en plaidant pour une approche strictement inductive: « *les modèles inductifs sont de la « bonne science* » »(Vercoe et Zak, 2010, p.124). D'une manière similaire aux premiers manifeste de

dans le modèle de Rabin (1993).

l'économie comportementale dans le scanner (*cf.* chapitre 4), Zak considère ainsi que les neurosciences permettent de révolutionner l'économie, laquelle aurait pour défaut de reposer sur « *une psychologie naïve du choix humain* » (Zak, 2010, p.2). Enfin, Zak n'hésite pas à vulgariser les résultats de la neuroéconomie, de façon parfois assez contestable. Il a ainsi écrit un article intitulé « huit leçons de la neuroéconomie pour les investisseurs financiers », dont la plupart des propositions se ramène au bon sens le plus élémentaire (contrôler son « appétit » pour le risque, ne pas suivre l'avis de tous, contrôler ses émotions, *etc.*) (Sapra et Zak, 2010).

Ce type de réflexions de « neurophilosophie » dépassent le cadre étroit du questionnement scientifique, et nuit parfois à la neuroéconomie sociale. Celle-ci apparaît donc comme un sous-domaine tout à fait particulier de la neuroéconomie. Elle exhibe en effet de façon exemplaire les principaux défauts de l'économie comportementale: réductionnisme neuronal, rhétorique révolutionnaire, positionnement méthodologique naïf. Sur le plan théorique, cette approche consistant à reproduire l'économie comportementale dans le scanner conduit à mettre l'accent, dans l'explication des interactions individuelles, sur les notions d'émotions et de préférences sociales.

II. Émotions et Préférences sociales: le choix inter-personnel vu par l'économie comportementale dans le scanner

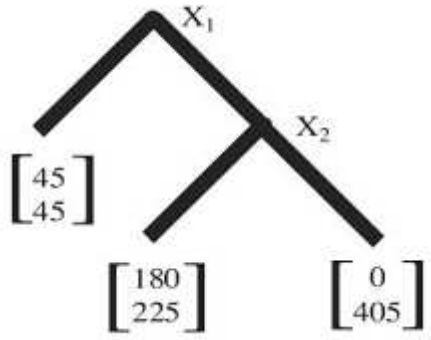
C'est au début des années 2000 que sont publiées trois études fondatrices pour la neuroéconomie sociale (McCabe *et al.*, 2001, Rilling *et al.*, 2002 ; Sanfey *et al.*, 2003), portant respectivement sur le jeu de l'investisseur, le dilemme du prisonnier et le jeu de l'ultimatum (A). Ces travaux suscitent un vif intérêt chez les économistes comportementalistes qui travaillent sur le choix inter-personnel. L'économiste Ernst Fehr a en particulier cherché à collaborer avec les neurobiologistes pour comprendre les mécanismes neuronaux sous-jacents aux processus cognitifs et comportementaux à l'œuvre dans les interactions sociales. Son programme de recherche, qui est appelé ici « théorie de la punition altruiste dans le cerveau » a fourni à la neuroéconomie sociale un modèle général pour l'interprétation des choix inter-personnels (B). Cette approche, qui repose sur une représentation duale de la cognition sociale, rencontre néanmoins des limites à la fois sur le plan descriptif et normatif (C).

A. Trois études fondatrices pour la neuroéconomie sociale (2001-2003)

Les premières études de neuroimagerie portant sur des protocoles inspirés des *behavioral economics* apparaissent entre 2001 et 2003. Il s'agit alors essentiellement de travaux concernant le choix inter-personnel. Le choix inter-temporel, tel qu'il est conçu en économie comportementale, ne fera l'objet d'expériences avec IRM_f qu'en 2004 (McClure *et al.*, 2004). Les trois premières études de neuroimagerie sur le jeu de l'investisseur, le dilemme du prisonnier et le jeu de l'ultimatum (McCabe *et al.*, 2001, Rilling *et al.*, 2002 ; Sanfey *et al.*, 2003) ont une valeur fondatrice pour la neuroéconomie sociale, et, plus généralement, pour l'ensemble de la neuroéconomie, en suggérant un possible enrichissement des interprétations de l'économie comportementale par les neurosciences.

Dans leur article publié en 2001, Kevin McCabe et ses co-auteurs proposent de rendre compte des comportements coopératifs dans le jeu de la confiance à l'aide d'une hypothèse

dite « du cerveau social » (McCabe *et al.*, 2001, p.11834). Douze sujets ont participé à cette étude, dans laquelle il s'agit de jouer soixante-douze fois d'affilée le même jeu, dit de l'investisseur, auquel correspond à la matrice de gain suivante:



(McCabe *et al.*, 2001, p.11834)

Les individus alternent la position du premier décideur (l'investisseur, représenté sur le graphe par X_1) et de second décideur (représenté par X_2). Ils peuvent affronter soit un autre joueur (humain), dont les décisions sont affichées à l'écran, soit un ordinateur, qui suit toujours la même stratégie. Cette information est divulguée par les expérimentateurs avant chaque séquence de jeu, et les individus savent s'ils jouent contre un un joueur humain ou contre un logiciel.

Les résultats de la neuroimagerie montrent qu'une zone spécifique située dans le cortex préfrontal médian s'active lorsque les joueurs décident d'accorder leur confiance en tant qu'investisseurs, ou de ne pas trahir la confiance qui leur a été accordée dans la position ud deuxième décideur. Cette activation n'atteint cependant un niveau significatif seulement lorsque le jeu oppose deux sujets humains. McCabe *et al.* associent donc à cette zone deux fonctions cognitives: d'une part, inhiber les inclinations en faveur de la satisfaction d'intérêts égoïstes (par exemple, trahir la confiance accordée par l'autre joueur et céder à la tentation de remporter un gain monétaire plus important); d'autre part, « *lire* » les états mentaux de l'adversaire et se mettre à sa place. Ces deux facultés, à la fois de contrôle et de « *mentalisation* » sont au fondement du « *cerveau social* » (Dunbar, 1998), dont le siège se trouverait donc dans le cortex préfrontal (McCabe *et al.*, 2001, p.11834).

Rilling *et al.* (2002) ont quant à eux étudié le dilemme du prisonnier. Comme dans l'étude de McCabe *et al.* (2001), les expérimentateurs indiquent au participant, avant chaque séquence de jeu, s'il affronte un autre participant humain ou un ordinateur. Le sujet choisit à chaque fois de coopérer ou d'adopter un comportement de *free-riding*. En comparant les deux

conditions, les auteurs observent que la coopération est associée à une activation à la fois de régions limbiques impliquées dans l'évaluation des récompenses et d'une région préfrontale -le cortex cingulaire antérieur (CCA)- impliquée dans la « *résolution des conflits* » (Rilling *et al.*, 2002,). Pour Rilling *et al.*, ces résultats suggèrent que la coopération est à la fois plaisante et constitue une sorte de récompense; mais qu'elle requiert aussi la résolution d'un conflit entre deux types de récompense liées à chaque décision (coopérer ou ne pas coopérer) en contrôlant la satisfaction égoïste.

L'expérience de Sanfey *et al.* (2001) est l'une des études de neuroéconomie la plus citée et la plus connue. Cette popularité s'explique sans doute par la simplicité de son protocole et du résultat obtenu. Les auteurs font jouer les participants au jeu de l'ultimatum. Conformément aux observations déjà relevées par les économistes comportementalistes, la plupart des sujets proposent des offres de partage relativement équitables, et rejettent les offres inéquitables, alors même que cette décision implique de renoncer un gain certain quoique faible. Sanfey *et al.* observent d'une part que les décisions de rejet chez les répondants sont associées à une activation de l'*insula* antérieure, une région impliquée dans les émotions « négatives », la peur en particulier. L'acceptation d'offres inéquitables, chez des sujets qui ne se préoccupent donc pas de sanctionner leur partenaire est est d'autre part liée à l'activation du cortex dorso-latéral préfrontal (CDLPF). Selon les auteurs, il y aurait chez la plupart des joueurs une impulsion affective et spontanée de rejet face aux offres inéquitables. Cette « *aversion à l'inéquité* » (*inequity aversion*) résulterait ainsi de l'activité de l'*insula* antérieure. Seule l'exercice de facultés de contrôle, associées au CDLPF, permettrait chez les joueurs capables de délibération de surmonter les réactions de rejet instinctives des offres inéquitables (Sanfey *et al.*, 2002, p.1757).

Les trois premières expérience de neuroimagerie sur les jeux « inter-personnels » de l'économie comportementale convergent donc sur l'idée d'une opposition entre ce qui relève des affects et de la raison. Cette représentation duale de la cognition offre un schéma interprétatif relativement général des interactions sociales, à partir notamment de la notion d'émotion, qui renvoie aux recherches d'Antonio Damasio (*cf.* chapitre 4),dont les travaux sont cités dans chacun des trois articles. Cette hypothèse dualiste a ainsi servir de plateforme théorique pour le déploiement de l'économie comportementale dans le scanner.

B. La théorie de la punition altruiste dans le cerveau: le programme de recherche de Ernst Fehr

Les premières expériences de neuroimagerie portant sur le choix inter-personnel ont suscité très rapidement un vif intérêt chez les économistes comportementalistes. L'idée d'une dualité entre deux types de processus cognitifs dans le cerveau a notamment beaucoup séduit les économistes qui, comme Ernst Fehr, souscrivaient déjà, avant de s'intéresser aux neurosciences, à une explication affective de la coopération et de l'altruisme. C'est à partir de cette hypothèse dualiste que s'est développée l'économie comportementale dans le scanner. Dans le domaine du choix inter-personnel, l'influent programme de recherche emmené par Ernst Fehr a permis de fédérer l'ensemble des travaux expérimentaux relevant de cette approche.

Ernst Fehr est un économiste comportementaliste qui a beaucoup travaillé sur les problème de l'altruisme, de la coopération et de la négociation. Il est connu pour avoir proposé, à la fin des années 1990, une théorie de la « *punition altruiste* » visant à expliquer l'origine des comportements altruistes et coopératifs (Fehr et Schmidt, 1999). Ce modèle s'appuie sur l'idée selon laquelle la simple possibilité d'une punition peut obliger, au sein d'une communauté, les individus égoïstes à coopérer ou à se comporter de manière altruiste, c'est-à-dire à prendre en compte le bien-être d'autrui. La coopération et l'altruisme peuvent donc constituer la norme d'un groupe, quand bien même la majorité d'individus qui le composent ne sont ni coopératifs ni altruistes, dès lors qu'il existe un nombre minimal d'individus prêts à sanctionner le non-respect des normes d'équité et de coopération (Fehr et Schmidt, 1999, p.855).

La théorie de la punition altruiste porte donc sur un trait comportemental tout à fait particulier -la « *réciprocité forte* »- que Fehr définit comme « *la prédisposition à coopérer avec les autres et à punir ceux qui violent les normes de coopération, avec un coût personnel, même lorsqu'il n'est pas probable que ces coûts puisse être remboursés par les autres ou à une date future* » (Gintis, Bowles, Boyd et Fehr, 2003, p.154). La réciprocité forte ne repose donc sur aucun calcul coût-avantage. Cela a deux conséquences importantes. Tout d'abord, les individus manifestant cette réciprocité forte n'évaluent pas les résultats des actes de leurs semblable, mais plutôt leurs intentions, puisqu'ils ne sont pas directement sensibles aux gains et aux pertes liées à la punition ou à la sanction (Fehr et Fischbacher, 2002; Fehr et

Fischbacher, 2003, p.788). Surtout, la réciprocité forte est inconditionnelle. Elle ne relève pas d'une forme de délibération et se distingue donc de la coopération motivée par des effets de réputation, et de l'altruisme réciproque, dans lequel l'altruisme est toujours une réponse à une favorable à une action altruiste qui lui est antérieure (Fehr et Schmidt, 1999, p.855).

La réciprocité forte peut se comprendre comme une tendance innée en faveur de l'altruisme, de la coopération et de la sanction des normes. Cette prédisposition résulte selon Fehr d'une évolution culturelle et naturelle de l'espèce humaine, l'idée étant que les communautés caractérisées par la présence d'individus ayant cette réciprocité forte dispose d'un avantage sur les groupes composés seulement d'individus égoïstes (Fehr et Fischbacher, 2003). Gintis, un auteur avec lequel Fehr a beaucoup collaboré, a ainsi proposé un modèle évolutionniste pour expliquer l'origine de la réciprocité forte dans les groupes (Gintis, 2003). Cependant, si Fehr et Gintis prennent en compte l'évolution phylogénétique, c'est-à-dire celle de l'espèce humaine, pour rendre de compte de l'altruisme, la réciprocité forte demeure pour eux un phénotype stable, qui produit chez un même individu toujours les mêmes effets en termes de comportements. En d'autres termes, cette perspective ne prend pas en compte l'évolution ontogénétique, qui se produit au cours de la vie de l'individu: celui-ci est ou non caractérisé par cette prédisposition innée à réciprocité. Le rapport de ces auteurs à la psychologie évolutionniste, quoique revendiqué, est ainsi plus compliqué qu'il n'y parait[125].

La théorie de la punition altruiste a fait l'objet d'assez nombreux tests empiriques. Ces tests portent d'abord sur le dilemme du prisonnier. Fehr et Gächter ont montré que, dans un dilemme du prisonnier répété mais sans effets de réputation, la simple possibilité de punir les *free-riders* permet d'augmenter de manière significative la coopération interindividuelle (Fehr et Gächter, 2000). Si la menace d'une sanction suffit donc à éliminer le free-riding, il reste à montrer que cette menace est bien réelle, c'est-à-dire qu'il existe bien des individus prêts à

125Ceci est notamment visible dans un article de 2006, dans lequel Gintis revendique l'héritage de la psychologie évolutionniste (Gintis, 2006). Pour Gintis, « *la principale affirmation de la psychologie évolutionnaire est que le cerveau humain a évolué ainsi parce que des cerveaux plus complexes et plu importants, en dépit de leurs coûts, permettent une meilleure adaptation* » (Gintis, 2006, p.2). Par conséquent, la formation du cerveau humain doit s'expliquer par l'évolution de l'espèce humaine dans son ensemble. Cependant, les contraintes évolutionnaires ont abouti à produire une « *nature humaine* », qui, depuis la fin du Paléolithique est demeurée essentiellement la même. Selon Gintis, la stabilité de cette nature humaine rend négligeable l'évolution ontogénétique (celle qui a lieu au cours de la vie de l'individu). Pour Gintis, « *les psychologues évolutionnistes sont nativistes et pensent que les capacités humaines élémentaires (par exemple, la grammaire universelle, la physique populaire, la psychologie populaire) sont présentes dès la naissance […] Les psychologues évolutionniste affirment par ailleurs que le cerveau humain est le produit de l'évolution pendant le Pléistocène, et qu'il n'y a pas eu de changements importants dans le cerveau humain depuis le développement de l'agriculture et de la vie urbaine à la fin du Holocène […]. Par conséquent, les hommes et les femmes modernes sont équipés de cerveaux qui remontent à l'age de pierre* » (Gintis, 2006, p.4)

punir les individus qui ne se préoccupent que de leur intérêt personnel. C'est précisément ce que visent les tests portant sur le jeu de l'ultimatum et du dictateur. Dans plusieurs études, Fehr a observé que la plupart des individus sont disposés à sanctionner les joueurs proposant des offres inéquitables au jeu du dictateur, et ce même lorsqu'ils ne participent pas directement au jeu en tant que répondant, mais qu'ils y assistent en tant que simple spectateur extérieur (Fehr et Gachter 2002; Fehr et Fischbacher, 2004)

La théorie de la punition altruiste rencontre cependant deux limites lorsqu'elle est restreinte à un niveau d'analyse seulement comportemental. Tout d'abord, elle s'applique assez mal au jeu de l'investisseur. La possibilité de punir les joueurs non-coopératifs a ici un effet négatif sur la coopération, car les sanctions n'apparaissent pas dans ce cadre comme légitimes, parce que motivées par la prise en compte de l'intérêt égoïste (Fehr et Rockenbach; 2003). Si la réciprocité forte constitue une explication possible de certains comportements coopératifs, d'autres hypothèses descriptives demeurent donc envisageables. La notion d'intelligence machiavélique proposée par les neurobiologistes a ainsi permis d'élargir le champ d'application du cadre interprétatif de Fehr, en rendant compte notamment du phénomène de confiance dans le jeu de l'investisseur.

Par ailleurs, la théorie de la punition altruiste postule l'existence d'une prédisposition innée à la réciprocité forte chez certains individus, qui peut être observée au niveau comportemental, mais dont l'origine, qui est supposée être liée à l'évolution du cerveau humain, reste inexpliquée. Comme le soulignent Fehr et Schmidt, la présence d'individus manifestant une réciprocité forte constitue dans le cadre de la théorie de la punition altruiste un bien public de second-rang (Fehr et Schmidt, 1999). Il reste à comprendre ce qui, originairement, peut motiver un individu à punir, indépendamment de tout calcul coût-avantage. L'enjeu consiste alors non seulement à montrer que les individus sont capables effectivement de sanctionner en assumant un cout personnel, mais d'expliquer les causes de la punition (Falk, Fehr et Fischbacher 2005, p.2018).

Les questionnaires utilisés dans les expériences sur le jeu de l'ultimatum fournissent un embryon de réponse à ce problème. Les sujets qui rejettent des offres inéquitables y indiquent généralement que leur décision a été motivée par une « *émotion négative* », de dégout, provoquée par le comportement égoïste de leur partenaire (Fehr et Fischbacher, 2004, p.85). L'expérience de Sanfey *et al.* (2003) citée plus haut confirme cette hypothèse, en associant cette aversion forte à l'inéquité à l'activité de l'*insula* antérieure (*cf. supra*).

Fehr a donc vu dans les neurosciences un moyen pour enraciner sa théorie de

l'altruisme et des comportements coopératifs dans une explication neurobiologique en termes d'émotions. L'idée selon laquelle la réciprocité forte aurait pour cause immédiate des mécanismes neuronaux affectifs a par la suite constitué le principe directeur du programme de recherche neuroéconomique d'Ernst Fehr. Un an après l'expérience de Sanfey *et al.* (2003), Fehr et ses coauteurs étendent au jeu de l'investisseur ce schéma interprétatif (de Quervain *et al.*, 2004). Le protocole de cette étude repose sur une version modifiée du jeu de la confiance, dans laquelle les sujets placés dans la position du premier décideur (l'investisseur) peuvent punir le deuxième joueur si celui-ci trahit la confiance qui lui a été accordée. Deux conditions sont distinguées. Dans la première, les punitions pouvant être infligées sont symboliques, c'est-à-dire qu'elles ne réduisent que le gain obtenu par le joueur qui punit, mais pas celui du joueur qui est puni. Dans la seconde condition, les punitions sont effectives et conduisent à amoindrir les gains des deux joueurs (de Quervain *et al.*, 2004, p.1254).

Les résultats indiquent que le punition effective est associée à une activation relativement plus importante du *striatum* dorsal (de Quervain *et al.*, 2004, p.1257). Cette région est associée au circuit de la récompense (*cf.* chapitre 4). Selon les auteurs, l'activité de cette zone représenterait une satisfaction émotionnelle liée à la punition (de Quervain *et al.*, 2004, p.1257). En d'autres termes, la punition altruiste s'expliquerait par un « plaisir de punir » originel qui inciterait les individus à sanctionner leurs partenaires non-coopératifs, et cela même lorsque cela diminue leur propre gain. Cette interprétation est justifiée par ailleurs par le fait que les sujets qui ont la propension à punir (effectivement) la plus élevée sont ceux dont l'activité du *striatum* dorsal est la plus forte (de Quervain *et al.*, 2004).

L'expérience de de Quervain *et al.* de 2004 à laquelle participe Fehr a une valeur fondatrice pour son programme de recherche. En effet, sur la base de ces résultats neurobiologiques, Fehr propose que l'apparente irrationalité de la punition altruiste soit en fait rationnelle, dans la mesure où la punition est source de plaisir, et peut se comprendre comme analogue à un bien économique (Fehr, Fischbacher et Kosfeld, 2005, p.347). Ce résultat est fondateur, parce qu'il permet d'expliquer les préférences sociales des individus par des mécanismes affectifs.

Les travaux de Fehr en neurosciences ont aussi porté, à un second niveau, sur le possible contrôle que les sujets peuvent exercer sur leurs réactions émotionnelles à l'équité et à l'inéquité. Ici, l'assimilation des résultats de la neuroimagerie au sein du cadre dualiste de Fehr est plus complexe, car les processus délibératifs supposés être opposés aux processus émotionnels remplissent, au niveau neuronal, deux fonctions distinctes, comme le suggèrent

McCabe et ses coauteurs dans leur expérience de 2001 (McCabe *et al*., 2001). Fehr reprend à son compte l'idée, proposée par Damasio notamment (*cf.* chapitre 4) selon laquelle les régions préfrontales exercent un contrôle des régions limbiques impliquées dans les émotions, mais cette fonction de contrôle, dans le domaine du choix inter-personnel peut se comprendre de deux manières différentes, comme inhibition d'une part, et comme mentalisation d'autre part.

Les études de neuroimagerie portant sur les interactions individuelles établissent d'abord que le cortex préfrontal permet d'inhiber les réactions instinctives et spontanées, associées aux régions limbiques. C'est le résultat principal de l'expérience de Sanfey *et al.* (2003) qui montre que l'activation du CDLPF permet aux joueurs capables de délibération de surmonter les réactions de rejeu instinctives des offres inéquitables (*cf. supra*). Fehr et ses collaborateurs ont approfondi ce résultat, en observant, toujours dans le cadre du jeu de l'ultimatum, que la stimulation magnétique transcrânienne (SMT) du CDLPF, qui conduit à inhiber l'activité de cette région, avait pour conséquence d'augmenter le taux de rejet des offres inéquitables (Knoch *et al.*, 2006). En d'autres termes, la désactivation relative du CDLPF rend les sujets moins capables d'exercer leurs facultés de raisonnement et de contrôle. Cette étude repose donc sur un technique -la stimulation magnétique transcrânienne- qui dépasse les limites de l'imagerie, en tant qu'elle permet d'établir des inférences causales directes sur le comportement.

Les régions préfrontales fonctionnent, dans le jeu de l'ultimatum par exemple, comme des mécanismes inhibiteurs. Ces mêmes régions exercent également des fonctions de contrôle d'un autre type, appelées dans la littérature facultés de « mentalisation » ou de processus de « théorie de l'esprit » (*Theory of Mind processes*, Singer et Fehr, 2005). Dans leur expérience de 2001, McCabe et ses co-auteurs montrent ainsi que dans le jeu de l'investisseur l'activité des zones préfrontales chez le premier décideur permettait de « *se mettre à la place* » du second joueur et d'imaginer ses intentions stratégiques (McCabe *et al.*, 2001). La reproduction de ce résultat dans une étude ultérieure a précisé la localisation de cette fonction, dans une zone particulière du cortex préfrontal, le cortex cingulaire antérieur (Gallagher *et al.,* 2002). Par ailleurs, Bhatt et Camerer ont montré, dans un jeu inspiré du concours de beauté de Keynes, que cette région est impliquée dans la formation de réflexions sur les croyances des autres jeurs, et que son activité augmentait avec le nombre d'itérations cognitives (les sujets pensent que je pense qu'ils pensent que je pense....) de la réflexion stratégique (Bhatt et Camerer, 2005).

Le cortex cingulaire antérieur serait donc responsable d'une capacité non pas

d'inhibition, mais plutôt de formation de stratégies complexes. Dans le domaine du choix inter-personnel, cette faculté est souvent comprise comme une intelligence machiavélique, c'est-à-dire comme une aptitude à comprendre et anticiper les stratégies des autres joueurs. Par exemple, dans le jeu de l'ultimatum, la possibilité de punir conduit les joueurs égoïstes à formuler des propositions plus équitables, par crainte de la sanction. Ici, l'équité n'est pas motivée par une réciprocité forte mais simplement par la menace d'une sanction, qui est anticipée par des sujets capables d'intelligence machiavélique. Fehr et ses co-auteurs approfondissent ce résultat à l'IRM_f, et montrent ainsi que, dans la version avec punition, la formulation d'offres plus équitables par des sujets ayant fait des propositions inéquitables dans la condition sans punition est associée à une plus forte activité du cortex cingulaire antérieur (Spitzer *et al.*, 2007, p.185)

Les travaux de Fehr et de son équipe en neurosciences aboutissent donc à opposer dans le domaine du choix inter-personnel des processus affectifs et délibératifs. Ces derniers remplissent une fonction de contrôle, qui comprend un mécanisme à la fois d'inhibition et de mentalisation. En assurant l'essor des modèles duaux en neuroéconomie sociale, le cadre interprétatif autour duquel le programme de recherche de Fehr a été construit a fourni un principe fédérateur pour l'ensemble des travaux dans ce domaine,.

C. Apports et limites des modèles duaux

Ernst Fehr et ses collaborateurs sont devenus assez rapidement l'une des équipes de recherche les plus actives dans le domaine de la neuroéconomie sociale. Les travaux réalisées par Fehr en neurosciences autour de la notion de punition altruiste et d'intelligence machiavélique ont permis de donner un principe de cohérence à un ensemble d'études assez disparates. D'autres équipes de recherche, en particulier celle de Paul Zak, se greffent au projet initial et développent des modèles duaux du choix inter-personnel. Cette dynamique débouche à partir de la deuxième moitié des années 2000 sur des articles de synthèse, pouvant être lus comme les premiers manifestes de la neuroéconomie sociale (Camerer et Fehr, 2006; Fehr et Camerer, 2007; Fehr, 2008). (1). Quoiqu'ayant fourni un schéma interprétatif commode pour les neurobiologistes et les économistes s'intéressant aux interactions

individuelles, les approches dualistes soulèvent néanmoins plusieurs difficultés théoriques (2).

1. Une approche fédératrice pour la neuroéconomie sociale

Les explications des interactions sociales proposées par Fehr s'appuient sur l'existence d'émotions négatives, liées au plaisir de punir (*cf.* de Quervain *et al.*, 2004) ou au dégout qu'inspirent les comportements égoïstes. Plus généralement, cette approche du choix inter-individuel s'appuie sur une représentation dualiste de la cognition (*cf. supra*). L'opposition émotions/raison a fourni à la neuroéconomie sociale un principe d'unification et de cohérence des principaux résultat empiriques. Le programme de recherche de Fehr a donc fonctionné comme un pôle d'attraction théorique. Son approche émotionnelle des interactions sociales a été approfondie du côté des émotions positives par Paul Zak, qui a travaillé sur les mécanismes affectifs à l'œuvre non pas dans la sanction des conduites intéressées, mais dans le don, la charité, l'altruisme et la coopération.

Zak n'est pas, comme Fehr, un économiste comportementaliste à l'origine. Il a cependant reçu une formation en économie, et ses travaux s'intègrent assez bien dans le paradigme général de Fehr, en associant de manière étroite émotions et préférences sociales. Ses premières recherches en neuroéconomie sont réalisées, comme Fehr, en 2004 (Zak, Kurzban, et Matzner, 2004). Zak est principalement connu pour ses travaux portant sur l'ocytocine. Ce neurotransmetteur, dont le rôle et l'existence ont été établis dès les années 1950, était alors associé, chez l'animal, à l'instinct de protection de la progéniture (Zak, 2008, p.3).

Comme dans d'autres domaines de la neuroéconomie, l'utilisation de l'IRM_f et de protocoles empruntés à l'économie comportementale a permis, au début des années 2000, d'étendre à l'homme des connaissances relatives au comportement animal. Avant l'apparition de la neuroimagerie, il était très difficile d'étudier le rôle de l'ocytocine sur des sujets humains. L'ocytocine a en effet été synthétisée dès les années 1950, et le fonctionnement du cerveau peut être modifié en faisant inhaler à ses sujets cette ocytocine de synthèse. Cependant, l'étude des effets de l'ocytocine sur le comportement apparaissait très incertaine car la concentration de cette neurohormone dans le sang est extrêmement faible et disparaît très rapidement (Zak, 2008, p.3). Grâce à l'IRM_f, il a été possible de mettre en évidence

directement l'influence de l'ocytocine sur l'activité du cerveau, et par la suite, de mieux connaître et calibrer ses effets à la fois sur le cerveau et le comportement humain (Zak, Kurzban, et Matzner, 2004).

Les principales découverts de Zak s'appuient néanmoins non pas sur la neuroimagerie, mais sur des comparaisons comportementales entre des conditions avec ou sans inhalation d'ocytocine. L'expérience fondatrice en la matière est celle que Zak a réalisé en collaboration avec Fehr en 2005 (Kosfeld *et al.*, 2005). Dans cette étude, les sujets jouent plusieurs séquences d'un jeu de l'investisseur classique, en alternant leur rôle (décideur 1 ou décideur deux) à chaque fois. Les sujets savent en outre qu'ils jouent avec de partenaires humains. Les cinquante-huit participants sont divisés en deux groupes. Les sujets, avant de jouer, inhalent une substance placebo (groupe « placebo ») ou de l'ocytocine (groupe « ocytocine »). Les résultats comportementaux indiquent que l'inhalation de l'ocytocine améliore considérablement la confiance accordée par l'investisseur dans le deuxième joueur. Sur 29 sujets du groupe placebo, 13 (ce qui représente 45% du groupe) transfèrent l'intégralité de leur capital initial, contre seulement 6 sujets dans le groupe placebo. Le transfert moyen de l'investisseur vers le second joueur est en moyenne 17% plus élevé dans le groupe de sujets ayant inhalé de l'ocytocine que dans le groupe placebo (Kosfeld *et al.*, 2005, p.674).

Les auteurs prennent soin de distinguer deux interprétations possibles des comportements coopératifs. Dans le jeu de l'investisseur, le transfert par le premier joueur de la totalité ou de la quasi-totalité de sa dotation initiale peut en effet être lu comme l'indice d'une grande confiance placée dans son partenaire, mais peut aussi être expliqué par une faible aversion pour le risque. Pour mettre en évidence l'effet spécifique de l'ocytocine sur la confiance, les expérimentateurs reproduisent le protocole initial, en remplaçant le deuxième décideur par un ordinateur prenant des choix aléatoires. La matrice du jeu est identique, mais les joueurs affrontent désormais une loterie, et non des partenaires humains. Dans ce jeu portant uniquement sur le risque (et non sur la confiance inter-personnelle), les transferts effectués par les sujets du groupe ayant inhalé de l'ocytocine sont de montant similaires à ceux du groupe placebo (Kosfeld *et al.*, 2005, p.674).

Dans une étude ultérieure, Zak, Stanton et Ahmadi (2007) montrent que l'ocytocine élève le montant des offres dans le jeu de l'ultimatum, mais reste sans effet sur les offres faites dans le jeu du dictateur. Cette expérience repose sur une méthode identique à celle de Kosfeld *et al.* (2005), en répartissant les sujets en deux groupes, « placebo » et « ocytocine ». Dans le jeu de l'ultimatum, les offres du groupe de sujets ayant inhalé de l'ocytocine sont d'un

montant 80% plus élevé que celui du groupe placebo. La différence entre es deux groupes n'est pas significative en revanche pour le jeu du dictateur, dans lequel le deuxième joueur n'a pas à accepter ou refuser l'offre qui lui est faite (Zak, Stanton et Ahmadi, 2007, p.1).

Cette étude précise donc le rôle de l'ocytocine dans les comportements altruistes. Pour les auteurs, l'ocytocine a un effet plus spécifique sur la « générosité », qui constitue un type d'altruisme bien particulier. L'altruisme, en général, peut se définir comme la disposition spontanée à aider les autres individus, ou d'améliorer leur bien-être, au détriment de son propre bien-être, sans nécessairement attendre de gain en retour, et donc sans prendre en compte d'éventuelle considérations stratégiques. Le montant des offres dans le jeu du dictateur donne une mesure de l'altruisme. La générosité, en tant que « *libéralité dans le don* », consiste à donner aux autres plus que ce à quoi ils pourraient s'attendre. Par exemple, donner vingt centimes à une personne sans domicile serait une manifestation de l'altruisme, et un don de 10 dollars à cette même personne illustrerait un comportement généreux (et aussi altruiste) Dans le jeu de l'ultimatum, les joueurs placés dans la position du répondant doivent s'attendre à un partage le plus inéquitable possible; et le montant des offres proposée par le premier jouer reflèterait donc le degré de générosité de cet individu (Zak, Stanton et Ahmadi, 2007, p.2).

Les auteurs approfondissent cette interprétation en expliquant que l'ocytocine produit des modifications du degré d'empathie des individus. Zak, Stanton et Ahmadi rappellent, en s'appuyant sur l'expérience précédemment citée de Sanfey *et al.* (2003), que, dans le jeu de l'ultimatum, les joueurs placés dans la situation du répondant expriment souvent des réactions émotionnelles négatives face à des offres inéquitables. Or, pour Zak, Stanton et Ahmadi, au moment de formuler leur proposition, les joueurs doués de faculté d'empathie seraient capables d'imaginer ces affects négatifs susceptibles d'être provoqués par une offre inéquitable, et le caractère désagréable de cette expérience les inciterait à être plus généreux (Zak, Stanton et Ahmadi, 2007, p.3).

L'empathie doit être distinguée des capacités de « mentalisation » ou des processus dits de « théorie de l'esprit ». Tous deux consistent à se mettre à la place d'un autre que soi, mais l'individu empathique imagine plus spécifiquement les émotions et les affects ressentis par l'autre, les facultés de mentalisation engageant la lecture non pas des états émotionnels, mais des intentions stratégiques des autres joueurs (*cf.* Fehr et Singer, 2006). L'empathie, qui constitue un prolongement direct des émotions primaires, peut être modifiée par l'ocytocine mais aussi par d'autres hormones. Zak a montré dans une étude ultérieure que l'administration de testostérone dans le jeu de l'ultimatum conduit en effet à diminuer le montant des offres

(Zak *et al.*, 2009).

A l'appui de ces résultats, Zak a développé un modèle explicatif de la générosité, appelé modèle HOME (*Human Oxytocin Mediated Empathy*). Ce modèle avance une explication empathique des comportements généreux, en suggérant que la détresse et la souffrance exprimée par autrui est, par le mécanisme de l'empathie, source de désagrément et nous incline spontanément à aider les autres individus (Zak, 2010). Ce modèle peut s'appliquer à la « prosocialité dyadique », c'est-à-dire aux relations entre deux individus, comme dans le cas du jeu de l'ultimatum, mais aussi, plus généralement, à toutes les relations inter-individuelles dans lesquelles la générosité peut être impliquée. Zak a également réalisé une expérience sur le don aux œuvres de charité notamment, dans laquelle il s'agit d'être généreux (ou non) envers non pas un autre joueur, mais un ensemble d'individus, qui ne sont pas présents dans le laboratoire (Barraza *et al.*, 2011). Dans cette étude, les sujets participent à plusieurs jeux (jeu de l'ultimatum, jeu de l'investisseur) à l'issue desquels ils obtiennent un gain financier. Les expérimentateurs demandent aux sujets, à le fin de la tâche, s'ils souhaitent donner une part de leurs gains à une œuvre de charité (la Croix Rouge ou le Croissant Rouge). Les résultats montrent que les dons dans le groupe de sujets ayant inhalé de l'ocytocine sont environ deux fois plus important que ceux du groupe placebo . L'ocytocine n'a cependant pas d'effet sur la proportion de donneurs dans le groupe (Barraza *et al.*, 2011, p.3).

Les travaux de Fehr et de Zak en neurosciences s'appuient donc sur la notion d'émotion, négative ou positive, pour rendre compte du choix inter-personnel. Cet intérêt commun est d'ailleurs à l'origine d'une collaboration entre les deux chercheurs (*cf.* Kosfeld *et al.,* 2005). Bien qu'omniprésente dans les comptes-rendus d'expérience et dans la discussion des résultats, la référence aux émotions est néanmoins susceptible d'une application variée et nuancée. La neuroéconomie sociale dans le scanner dont Fehr est l'une des figures importantes est sans doute moins naïve que ce que ses critiques laissent à penser. Harrison, par exemple, reprochait à McCabe *et al.* (2001) de négliger quatre artefacts possibles dans l'interprétation des résultats comportementaux du jeu de l'investisseur[126]. Or, Kosfeld *et al.* (2005) prennent soin de dissocier les effets de l'ocytocine sur la confiance et sur le risque. Par

126« *Plusieurs choses différentes peuvent inciter quelqu'un à envoyer de l'argent, et à en renvoyer en retour, et ces choses sont bien connues. Chaque joueur peut être altruiste envers l'autre joueur. Ils peuvent aussi en vouloir à l'expérimentateur, et chercher à lui prendre autant d'argent que possible. La « confiance » peut aussi n'être que l e reflet d'un goût prononcé pour le risque. Les joueurs peuvent aussi envisager ce jeu unique comme s'il était répété. Par conséquent, à moins que l'on ne souhaite définir les termes de « confiance » et de « digne de confiance » par un amalgame de ces différentes motivations, il est nécessaire de concevoir des protocoles permettant de contrôler ces artefacts possibles* » (Harrison, 2008, p.320)

ailleurs, cette interprétation est rendue encore plus précise par l'expérience de Zak, Stanton et Ahmadi, 2007, qui montre le rôle de l'ocytocine non sur l'altruisme, mais sur la générosité.

Si donc les premiers travaux de la neuroéconomie sociale font preuve d'une simplification excessive des enjeux interprétatifs au niveau comportemental et cognitif, il est indéniable que ce sous-domaine de recherche a progressivement affiné ses hypothèses explicatives, en dégageant peu à peu les résultats théoriques de leurs possibles artefacts. Glenn Harrison, dans sa critique des travaux de neuroimagerie portant sur le jeu de l'investisseur, reconnaît lui-même que ces deux expériences de Zak contrôlent plusieurs artefacts possibles de la confiance (aversion au risque plus faible, altruisme) (Harrison, 2008-a, p.320).

Ces expériences offrent en outre une réponse aux accusations d'« *inférence inverse* » de Harrison (Harrison, 2008-a, p.317). Ici, les interprétations proposées des résultats de l'IRM_f ne s'appuient pas nécessairement sur d'autres études ayant préalablement mis en évidence des activités cérébrales similaires. Des techniques telles que la stimulation magnétique transcrânienne permettent d'inférer directement une réponse comportementale (rejet plus fréquent des offres inéquitables) à partir d'une variable neuronale (inhibition de l'activité d'une région du cerveau). Pour Fehr, la stimulation magnétique transcranienne dépasse ainsi les limites de l'IRM_f (Eisenegger *et al.,* 2008; Knoch *et al.,* 2007). Par ailleurs, l'administration de neurotransmetteurs, dans les travaux de Zak, aboutit également à mettren évidence des liens de causalité directe entre variables cérébrales et comportementales. Comme le soulignent Fehr et Camerer, « *les données de la neuroimagerie ne permettent pas de faire des inférences causales, mais il est possible de se rapprocher de la causalité directe en comparant des groupes d'individus placés dans des conditions de traitement ou dans des conditions placebo* » (Fehr et Camerer, 2007, p.423). Plus généralement, le principe de la comparaison clinique, entre des individus sains ou placebo et des individus pathologiques ou ayant subi un traitement, offre un moyen efficace de « triangulation » des résultats de la neuroimagerie, en s'assurant d'un bon niveau de correspondance entre les régions du cerveau et leurs fonctions cognitives associées (*cf.* chapitre 5).

Enfin, la neuroéconomie sociale fait référence à des considérations phylogénétiques pour expliquer la formation des différentes régions du cerveau et de leur fonction. Les hypothèses de localisation proposées par les neurobiologistes peuvent donc être appuyées par les résultats de l'anthropologie évolutionniste. McCabe *et al.* (2001) justifient ainsi leur hypothèse selon laquelle les régions préfrontales seraient impliquées dans la coopération et les

interactions individuelles à partir d'une référence aux travaux de l'anthropologue Ronald Dunbar sur le « *cerveau social* ». Ici, l'interprétation des données de la neuroimagerie n'est pas (seulement) justifiée à partir d'autres travaux de neuroimagerie, mais à partir d'une théorie explicative de l'évolution du cerveau, appuyée sur des mesures et des comparaisons du crâne chez l'homme et d'autres espèces humaines (*cf.* Dunbar, 1998)[127].

Le problème de l'inférence inverse est moins sévère qu'il n'y paraît, et Fehr comme Zak cherchent à isoler les effets théoriques robustes de possibles artefacts comportementaux. Cependant, Harrison remarque, à propos du jeu de l'investisseur, que les travaux de Zak ne permettent pas d'éliminer un dernier type d'artefact possible, qui est beaucoup plus problématique pour l'ensemble de ce programme de recherche. Harrison souligne que les interprétations proposées par Zak sont toujours compatibles avec l'idée selon laquelle les sujets, dans ces expériences, imaginent ces séries de jeux non-répétés entre individus, comme s'ils étaient répétés. Cette tendance à envisager les interactions individuelles comme si elles étaient susceptibles de se reproduire, fournit aussi une explication plausible des transferts élevés dans le jeu de l'investisseur, puisque les considérations stratégiques dans un jeu répété sont de nature complètement différentes, et justifient la coopération notamment par des effets de réputation (Harrison, 2008-a, p.320). L'approche émotionnelle proposée par Fehr et Zak rencontre effectivement une difficulté majeure dans la question de savoir si les jeux étudiés acquièrent une signification en tant qu'interactions répétées ou non-répétées.

127Les travaux de Dunbar portent spécifiquement sur le néo-cortex ou cortex pré-frontal, dont le développement, dans l'évolution de l'homme, a été plus tardif. Dunbar suggère que l'accroissement de la taille du cerveau liée au développement du néo-cortex chez l'homme ou chez certaines espèces de primates ne peut s'expliquer par des contraintes du milieu « primaires », impliquant par exemple la maîtrise de certaines tâches motrices. En effet, le coût biologique du néo-cortex est extrêmement élevé, car il implique une dépense de fonctionnement métabolique très importante. Il est donc peu probable que les espèces ayant développé de larges régions pré-frontales soient dotées d'avantage adaptatifs significatifs. Pour Dunbar, la contrainte écologique auquel le néo-cortex permet de répondre est l'accroissement de la taille des communautés d'individu: le développement de cette région du cerveau serait apparu chez les espèces dans lesquelles le nombre plus important des individus composant les groupes aurait nécessité une complexification des fonctions cognitives impliquées dans la gestion des interactions individuelles. L'hypothèse du cerveau social de Dunbar suggère donc que « *les capacités de traitement d'information du néocortex émergèrent des contraintes liées à la taille des groupes, qui joua une rôle primordial dans son développement* » (Dunbar, 1988, p.184).

2. Émotions et préférences sociales: le problème de la répétition des interactions et de l'apprentissage

La « neuroéconomie sociale dans le scanner » converge sur l'idée d'une interprétation émotionnelle des interactions sociales. Les émotions fournissent un mécanisme d'explication commode des comportements observés dans le jeu de l'ultimatum, de l'investisseur ou dans le dilemme du prisonnier. Les sujets répondent en fonction des émotions qu'ils ressentent, et en fonction de leur capacité à maîtriser leurs émotions (intelligence machiavélique). Cette approche considère donc que les individus sont caractérisés par des préférences sociales, liées à leur affectivité particulière face aux autres individus. La modélisation théorique s'appuie sur ces préférences sociales individuelles pour prédire et rendre compte des comportements.

La notion de préférence sociale joue ainsi un rôle primordial dans la littérature de neuroéconomie sociale, qui, dans la lignée de Fehr ou de Zak, s'inscrit dans le prolongement direct de l'économie comportementale. Dans l'expérience fondatrice du programme de recherche de Fehr en neurosciences, de Quervain *et al.* supposent par exemple l'existence d'une « préférence pour la punition de la violation des normes » (de Quevain *et al.*, 2004,1p.257). Dans leur article manifeste de 2007, Fehr et Camerer mettent en avant l'« *importance des émotions et des préférences sociales* » (Fehr et Camerer, 2007, p.425). L'expression apparaît également dans le titre de certaines études; Tricomi, Rangel, Camerer et O'DOherty intitulent leur article « Preuves neuronales des préférences sociales averses à l'inéquité » (Tricomi, Rangel, Camerer et O'DOherty, 2010).

Si la notion de préférences sociales est si importante pour ce pan de la littérature en neuroéconomie sociale, c'est parce que, dans la perspective de Fehr, l'utilisation des neurosciences dans le champ des *behavioral economics* a précisément pour but d'expliquer et de mesurer les différences inter-individuelles (Fehr, 2008, p.230). Il s'agit, en d'autres termes, de fournir un fondement neuro-affectif aux préférences sociales des individus; ces données individuelles fournissant ensuite les paramètres du modèle théorique, à partir duquel il est possible de prédire les choix ultérieurs pris par ce même individu en présence d'autres sujets. L'idée consiste donc à « calibrer » un modèle comportemental -le modèle HOPE de Zak, ou la théorie de la punition altruiste de Fehr et Schmidt (1999)- à l'aide de paramètres individuels -par exemple, le degré d'empathie pour le modèle HOPE, la préférence pour la punition de la violation des normes pour la théorie de la punition altruiste.

Une telle approche repose sur une conception fixiste: chaque individu est censé être doté de préférences tables qui permettent de prédire son comportement. Cela a pour

conséquence importante de ne pas prendre en compte les effets d'apprentissage liés à la répétition des décisions, car un même individu adoptera toujours le même comportement dans des cas d'interactions similaires. Dans toutes les études portant sur les interactions sociales qui ont été citées jusqu'ici, les auteurs supposent ainsi que les sujets envisagent les différents jeux auxquels ils participent comme s'ils n'étaient pas répétés. Pourtant, les individus jouent bien dans ces expériences à plusieurs séquences répétées du même jeu. Mais les expérimentateurs indiquent que l'appariement des partenaires est aléatoire et varie à chaque essai, de telle manière que les interactions doivent être considérées à chaque fois comme uniques et non-répétées.

Comme le souligne Harrison, il n'est pas du tout sûr que les participants jouent dans ces expériences en considérant chaque interaction comme étant unique et indépendante des autres. Un problème tel que celui du jeu de l'investisseur se pose généralement, dans les interactions « réelles », hors du laboratoire, en supposant que le partenaire de l'échange sera à nouveau rencontré à l'avenir et fera l'objet de nouvelles interactions. La force de cette habitude peut très bien conduire les sujets à raisonner, dans le laboratoire, de la même façon, et à envisager « *le jeu non-répété avec les lunettes d'un jeu répété* » (Harrison, 2008-a, p.320).

Cette lecture « statique » des résultats de la neuroéconomie sociale est d'autant moins justifiée que la neuroéconomie dans son ensemble propose précisément une interprétation dynamique ou séquentielle de l'activité des mêmes zones du cerveau qui sont impliqués dans le choix individuel et inter-individuel, et qui sont associées au circuit de la récompense. Au sein du paradigme néo-comportementaliste, les réponses neuronales, cérébrales ou comportementales ne prennent de sens précisément qu'à l'intérieur d'une interaction répétée entre l'individu et son milieu. C'est à la lumière des contraintes spécifiques posés par l'environnement dans lequel l'individu interagit qu'un comportement peut être ou non qualifié de rationnel (et non à partir de caractéristiques individuelles du décideur).

Le programme de recherche de Fehr, auquel Zak a été ici associé, va donc assez clairement à l'encontre de l'approche évolutionniste du néo-comportementalisme. Zak comme Fehr font bien référence cependant à l'évolution, puisque selon eux les préférences sociales des individus ont un ancrage cérébral, qui résulte de l'évolution de l'espèce humaine. L'« équipement biologique » de notre moralité doit être compris comme une adaptation régressive aux contraintes de l'environnement dans lequel l'humanité s'est développée: « *la plupart des humains ont un sens moral intact qui nous guide dans nos décisions; Ce mécanisme utilise à la fois des régions du cerveau ayant une origine évolutionnaire ancienne, associée à la régulation corporelle et au mouvement, et des régions plus récentes, impliquées*

dans la délibération cognitive » (Zak, 2011, p.226). Il est clair néanmoins qu'une telle perspective laisse à penser que notre sens moral, quoique lié à l'évolution de l'espèce humaine, est inné. En d'autres termes, cette conception qui met en avant des explications phylogénétiques néglige l'ontogénèse et la progressive acquisition de ces facultés par l'individu (*cf.* section I).

Cela ne signifie pas pour autant que nos réactions soient mécaniquement déterminées par nos préférences sociales. Zak et Fehr insistent tous deux sur le rôle de l'environnement et des institutions, qui influencent fortement nos décisions. Zak précise ainsi qu'« *en tant que créatures sociales, nous nous acclimatons à l'environnement social dans lequel nous nous trouvons et nous régissons à la structure des incitations qui le caractérise. Le sens moral, comme toutes réponses comportementales, est flexible et ajustable* » (Zak, 2011, p.226). De la même façon, dans le cadre de la théorie de la punition altruiste, c'est la possibilité ou non de mettre en place des mécanismes de sanction jugés crédibles qui détermine les stratégies des joueurs. A la suite de Gintis (2003), Fehr considère donc que la coopération résulte de l'internalisation des normes extérieures. Selon Fehr, cette explication met l'accent sur le rôle des institutions et des règles collectives et se rapproche au moins autant de la sociologie, qui supposent que les acteurs obéissent à des normes, que de l'économie, pour laquelle le comportement ne résulte que l'intérêt personnel (Fehr et Gintis, 2007, p.43).

Cependant, si l'environnement peut affecter la manière avec laquelle les préférences sociales des individus s'expriment en termes de choix et de comportements réels, la dynamique de cette interaction reste en dehors de l'analyse. A un même environnement correspondra toujours la même réponse comportementale ou neuronale. Les données institutionnelles et individuelles constituent toutes deux des paramètres exogènes. Cette approche a donc pour limite principale de ne pas prendre en compte les effets d'apprentissage au niveau de l'individu. La limite est liée à la présupposition selon laquelle les émotions à l'œuvre des les interactions sociales sont des processus irrépressibles, analogues à des réflexes pavloviens. Zak résume ainsi l'apport principal de la neuroéconomie sociale de la manière suivante: « *un nombre importantes de recherches ont démontré que les représentations neuronales des valeurs morales sont automatiques et difficiles à supprimer, et utilisent souvent des régions du cerveau associées aux émotions. Le comportement social adapté requiert un équilibre entre l'approche et le retrait, la confiance et la méfiance* » (Zak, 2011, p.225)

Or, l'idée selon laquelle les émotions fonctionneraient à la manière de réflexes conditionnées est critiquable. Des expériences mettent en évidence un apprentissage de ces

émotions sociales, de la même façon que pour l'apprentissage de la récompense. En outre, cette approche dualiste, qui suppose des processus instinctifs automatiques pouvant ou non être contrôlés par des processus délibératifs, situés à un niveau cognitif supérieur, ne permet pas de comprendre comment ce deuxième étage de la cognition sociale émarge du premier. Par exemple, le CDLPF est supposé contrôler le rejet affectif des offres inéquitables dans le jeu de l'ultimatum, ou le cortex cingulaire antérieur est associé à la capacité de former des stratégies d'interactions plus élaborées, en prenant en compte les intentions stratégiques des autres joueurs (*cf. supra)*. Mais comment et pourquoi certains individus sont caractérisés par de telles aptitudes? Pour Zak et Fehr, la question est réglée puisque les préférences sociales sont des primitives des modèles, c'est-à-dire des données exogènes dont il n'est pas nécessaire de donner d'explications. L'objectif des recherches consiste simplement à discriminer entre des groupes d'individus « naïfs » et « sophistiqués », et à identifier les soubassements neuronaux de cette discrimination.

Fehr, dans son article originaire sur la punition altruiste, reconnaît que son modèle néglige les effets d'apprentissage: « *une limite évidente de notre modèle est qu'il ne peut pas expliquer l'évolution du jeu au cours du temps* » (Fehr et Schmidt, 1999, p.851). Cette limitation peut éventuellement se justifier sur le plan comportemental, en admettant par exemple que la plupart des interactions sociales, dans la vie moderne, se produisent entre individus anonymes. Mais cette perspective statique appuyée sur la notion d'émotion entre clairement en conflit en neurosciences avec les explications dynamiques ou séquentielles de la neuroéconomie.

Par ailleurs, le concept d'émotions en neurosciences, emprunté aux travaux de Damasio, soulève des problèmes du point de vue de l'interprétation normative (*cf.* chapitre 5, section I). Il est en effet délicat de définir la rationalité à partir des émotions, puisque celles-ci permettent dans certains cas de se comporter de manière adaptée aux circonstances, mais peuvent également se comprendre dans d'autres cas comme des biais ou des erreurs, affectant les facultés rationnelles des agents. Par exemple, Zak considère que la confiance et l'empathie sont des ressorts essentiels de l'activité économique, et que ces émotions sociales sont les conditions indispensables à la rationalité des échanges (Zak, 2010). Mais, dans le jeu de l'ultimatum par exemple, l'exercice de facultés de contrôle de soi et d'inhibition des réactions émotives constitue au contraire un moyen d'adopter des comportements plus sophistiqués: faut-il considérer que l'intelligence machiavélique est rationnelle, ou bien la neutralité émotionnelle doit-elle être envisagée comme une déviance? L'approche dualiste en neuroéconomie sociale rencontre donc des difficulté à la fois sur la plan normatif et descriptif.

III. De l'apprentissage de la récompense au choix inter-personnel: la notion d'intelligence sociale

Les travaux de Fehr et de Zak montrent que l'économie comportementale dans le scanner est souvent plus subtile que ce que ses détracteurs laissent à penser (*cf.* Harrison, 2008-a). Sa principale limite réside non pas dans leur confiance aveugle accordée aux nouvelles techniques de neuroimagerie, mais plutôt à l'appréhension des émotions à la manière de réponses automatiques ou des quasi-réflexes conditionnés. Une telle perspective néglige toute dynamique d'apprentissage des émotions sociales. L'enjeu ici consiste à montrer que le dualisme propre aux approches de Fehr et de Zak peut être replacé dans le paradigme plus général du néo-comportementalisme, en substituant à la notion d'émotion celle d'intelligence sociale. Des expériences mettent tout d'abord en évidence que les émotions sociales décrites par Fehr et Zak peuvent faire l'objet d'un apprentissage (et non seulement d'un contrôle) par des processus de *reward learning* similaires à ceux évoqués dans les chapitres précédents (A). L'opposition émotions/raison peut ainsi être remise en question en considèrant que le circuit de la récompense fournit un système d'évaluation fonctionnant de manière unitaire (B). Loin de constituer des paramètres fixes, les préférences sociales de chaque individu sont susceptibles de varier selon les circonstances, mais aussi selon l'expérience et l'apprentissage accumulé par l'individu: l'intelligence sociale se manifeste par la capacité à former des émotions sociales complexes (C).

A. *Reward Learning* et choix inter-personnel: une perspective dynamique sur les jeux inter-personnels

L'analyse du choix inter-personnel apparait assez éloignée des travaux de neuroimagerie sur l'apprentissage de la récompense et le choix inter-temporel qui ont été abordés dans les deux chapitres précédents. Pourtant, les régions activées dans ces études sont identiques et relèvent du circuit de la récompense. Par exemple, de Quervain *et al*. (2004) observent une activation du striatum dorsal dans leur étude, qu'ils associent à une satisfaction

liée à la punition altruiste (*cf.* section II). Or, les premières expériences de neuroéconomie ont également mis en évidence le rôle de cette région dans l'encodage d'erreurs de prédiction de la récompense. Si la neuroéconomie sociale appartient donc à la neuroéconomie, c'est donc aussi parce que toutes deux travaillent sur les mêmes objets, c'est-à-dire sur les mêmes zones du cerveau, en utilisant des méthodes comparables (IRM_f, comparaisons cliniques, *etc.*)

La plupart des analyses de la neuroéconomie qui ont été faites ici valent donc aussi pour la neuroéconomie sociale. En particulier, l'opposition entre une approche inspirée, ou « mise sous la tutelle » de l'économie comportementale, et une approche plus fidèle à l'héritage né-comportementaliste, structure également selon nous ce sous-domaine de la neuroéconomie consacré aux interactions sociales. Deux types d'interprétation peuvent ainsi être distinguées, l'un en terme de monnaie neuronale commune, et l'autre en terme de dualisme émotions/raisons. Le conflit interprétatif est cependant ici moins explicite que dans le cas du choix inter-temporel, dans lequel, une controverse était née entre Kable et Glimcher (2007) et McClure *et al*. (2007) (*cf.* chapitre 5, section III). L'absence de véritable controverse signifie aussi que plusieurs chercheurs -Colin Camerer notamment- ont participé tour à tour aux deux paradigmes. La lecture qui est proposée ici des résultats expérimentaux se comprend donc ici, à plus forte raison, comme une reconstruction historique *a posteriori* des enjeux théoriques.

Avant de justifier cette lecture particulière de la neuroéconomie sociale, il est nécessaire tout d'abord d'expliquer pourquoi les études de neuroimagerie portant sur le choix inter-temporel peuvent être comprises comme un prolongement du paradigme néo-comportementaliste. Ce dernier est consacré aux processus de *reward learning*. Les objets d'étude -les régions du cerveau activées- sont essentiellement les mêmes, mais il n'apparait pas évident, au premier abord, de saisir le lien entre choix inter-personnel et apprentissage de la récompense.

Un défaut important des premières études de « neuroéconomie sociale dans le scanner » réside précisément dans l'ignorance de ce lien entre *reward learning* et choix inter-temporel. Dans l'étude de de Quervain *et al.* (2004) par exemple, l'activité du *striatum* dorsal est associée à une satisfaction, liée au plaisir de punir les individus égoïstes. Or, cette zone, impliquée dans l'apprentissage de la récompense, n'encode pas en fait un plaisir effectivement ressenti, mais la valeur attendue ou espérée d'une récompense. Berns *et al.* (2001) mettent justement en évidence la séparation, au sein du cerveau, du traitement du plaisir hédonique et du plaisir attendu.

L'association de l'activité des régions limbiques à un plaisir ou une satisfaction, comme c'est le cas aussi dans l'étude de Rilling *et al.* (2002), est donc très discutable. Cette interprétation néglige une importante caractéristique de ces zones, qui fonctionnent par des processus d'essais et d'erreur. Le principal défaut des travaux de Fehr et de Zak est ainsi lié à leur caractère statique. En effet, en laissant de côté les dynamiques d'apprentissage à l'œuvre dans le « *reward* social », ces auteurs sont amenés à concevoir les réponses de ces régions comme des quasi-réflexes. Or, en dépit de leur automaticité, les zones du cerveau associées au circuit de la récompense sont effectivement capables d'apprentissage.

Une étude récente de Lin, Adolphs et Rangel (2012) met ainsi en évidence la similarité, dans le cerveau, des processus de *reward learning*, pour des récompenses monétaires et « sociales ». Cette expérience repose sur une tâche d'apprentissage de la récompense tout à fait classique, du type machine à sous multi-jeux, dans laquelle les sujets doivent choisir de manière répétée entre deux loteries, l'une étant plus avantageuse que l'autre sur le long-terme. L'objectif consiste pour les participants à découvrir la loterie la plus avantageuse et à continuer de choisir cette loterie jusqu'à la fin de la tâche. Dans la première partie de l'expérience, les récompenses sont monétaires: les sujets gagnent ou perdent, à chaque fois, des sommes comprises entre 0 et 1 dollar. Dans la deuxième partie, les récompenses sont dites « sociales »: un gain entraîne la visualisation d'un visage humain souriant, et l'audition, par des écouteurs, de mots à connotations positives tels que « *bravo* », « *excellent* », ou « *fantastique* ». A l'inverse, une perte est associée à un visage exprimant des émotions négatives (colère, peine) et à des mots à connotation négative tels que « *stupide* », « *idiot* », « *faux* » (Lin, Adlophs et Rangel, 2012, p.275).

Les résultats indiquent que ce sont exactement les mêmes régions qui encodent les deux types de processus de *reward learning*. Dans les deux conditions, l'activité du cortex ventromédian pré-frontal encode la valeur attendue de la récompense, au moment de la décision; et celle du *striatum* ventral encode l'erreur de prédiction de la récompense (Lin, Adlophs et Rangel, 2012, p.280). Cela laisse à penser que les différentes formes de satisfaction ou de plaisir évoqués par les premiers travaux de neuroéconomie sociale (plaisir de punir, plaisir de donner à une œuvre de charité, plaisir de coopérer, *etc.*) doivent en fait être compris comme des processus dynamiques de *reward learning*, constitutifs d'une seule et unique monnaie neuronale commune (*cf.* Berns et Montague, 2002). Une telle perspective remet ainsi en question l'approche dualiste proposée par Fehr et Zak.

B. Modèles duaux versus monnaie neuronale commune

Les problématiques relevant du choix inter-personnel en neurosciences s'intègrent dans l'analyse plus générale des processus du *reward learning*. La cognition sociale s'appuie en effet également sur des processus d'apprentissage. Une question déjà posée à propos du choix inter-temporel apparaît alors: faut-il considérer que ces choix sont le produit d'un arbitrage entre les émotions ou la raison, ou bien s'agit-il toujours d'un même signal encodant une « *monnaie neuronale commune* » (*cf.* Berns et Montague, 2002)? Dans le cas du choix inter-temporel, McClure *et al.* (2004, 2007), partisans du dualisme, soutenaient qu'un système « *impulsif* » traitait spécifiquement des récompenses immédiates ou disponibles à très court-terme, et qu'un système délibératif prenait en compte les objectifs à long-terme. Les défenseurs d'une approche unitaire rejetaient quant à eux l'idée selon laquelle deux signaux d'évaluation distincts seraient produits et mis en concurrence dans le cerveau (Kable et Glimcher, 2007, Hare et al., 2009).

Une telle controverse n'est pas apparue dans le domaine du choix inter-personnel. L'opposition entre les deux approches est sans doute moins tranchée que dans le cas du choix inter-temporel. Cependant, l'étude de Hare *et al.* (2009), qui remettait en question les interprétations dualistes, a été reproduite et étendue au *reward* « social » dans une expérience ultérieure, à laquelle a participé Colin Camerer (Hare *et al.*, 2010). Ce dernier travail expérimental ne constitue pas directement une critique des recherches réalisées par Fehr ou Zak mais permet néanmoins de pointer certains défauts des interprétations dualistes.

L'étude originale de Hare *et al.* (2009) portait sur des récompenses alimentaires: des sujets obèses, au régime, devaient successivement attribuer une valeur à 50 aliments, en fonction de leur goût et de leur qualité diététique, puis devaient placer des enchères sur ces items alimentaires (*cf.* chapitre 5). C'est un protocole similaire qui est utilisé dans l'expérience de 2010 (Hare *et al.*, 2010), mais les items alimentaires sont remplacés par des possibilités de dons à des œuvres de charité. Les sujets ne sont pas obèses mais « *normaux* ». A chaque « *item de charité* » correspond un descriptif de l'activité et des objectifs de l'œuvre en question. Les participants reçoivent 100 dollars au début de l'expérience. Ils doivent indiquer, pour chaque item, le montant du don, de 0 à 100 dollars, qu'ils seraient prêts à faire en faveur de cette œuvre de charité. A la fin de la tâche, l'un des items est tiré au sort, et le don est réalisé suivant le montant qui a été choisi par le sujet. La part restante du revenu initial (100 dollars) est versée au participant (Hare et al., 2010, p.583-584).

La neuroimagerie fournit deux résultats importants. Les auteurs observent d'abord, au moment de chaque décision, une activité du cortex ventro médian préfrontal, qui est corrélée avec le montant du don. D'autre part, cette activité du CVMPF est modulée, en amont, par l'activité de deux autres régions: le cortex supérieur temporal et l'*insula* antérieur. Dans l'expérience portant sur les récompenses alimentaires, Hare *et al.* montraient que l'activité du CVMPF était également susceptible d'être modulée par l'activité du CDLPF: cette région permettait de prendre en compte les objectifs de long-terme (qualité diététique) et d'émettre un signal encodant non pas la valeur à court terme (goût) mais à long terme. Ici, pour Hare *et al.*, l'activité du cortex supérieur temporal et de l'*insula* antérieure permettent aussi à l'individu de se détourner de la valeur immédiate, égoïste, liée au possible gain monétaire de 100 dollars, en faveur d'une valeur « supérieure » (donner à une œuvre de charité). Le choix inter-personnel présente donc selon les auteurs une analogie avec le choix inter-temporel dans la mesure où tous deux repose sur un arbitrage entre deux types de valeur (Hare *et al.,* 2010, p.588).

Dans le cas du choix inter-personnel, la prise en compte d'une valeur altruiste ne suppose cependant pas un élargissement de l'horizon temporel de la décision, comme dans le cas du choix inter-temporel, mais la transformation d'une perspective égoïste à un point de vue inter-individuel. Pour Hare *et al.*, l'activité de l'insula antérieure représente l'exercice de facultés d'empathie, qui permettent à l'individu d'imaginer le bien-être reçu par ses donataires. Le cortex supérieur temporal fonctionnr, comme cela a été montré dans d'autres études (*cf.* Fehr et Singer, 2006) comme une zone spécialisée dans la « mentalisation », permettant de concentrer son attention sur les pensées et les attentes des autres individus. La valeur accordée par l'individu à chaque œuvre de charité, et le don que celui-ci est prêt à effectuer, dépend donc en amont de la modulation de l'activité du CVMPF par ces deux régions. Les préférences sociales ne sont donc pas pour les auteurs données, mais varient en fonction des capacités d'empathie et de mentalisation de l'individu:

> « *Une hypothèse proposée en économie comportementale consiste à supposer que le montant donné aux œuvres de charité dépends seulement des préférences du donneur pour cette œuvre (Andreoni, 1990; Fehr et Schmidt, 1999; Fehr et Camerer, 2007). Les données concernant la connectivité fonctionnelle qui ont été présentées ici suggèrent que les capacités de cognition sociale peuvent aussi jouer un rôle dans la détermination de la taille du don, en influençant la manière avec laquelle la valeur du don, c'est-à-dire les préférences, est calculée au moment de la décision. Par exemple, un sujet qui n'active pas son insula au moment de la décision peut donner qu'un faible montant parce qu'il ne génère pas l'empathie nécessaire à la construction d'une préférence plus importante. De*

manière similaire, un sujet qui n'active pas pas son cortex supérieur temporal peut faire un don faible, non pas parce qu'il est indifférent aux bénéficiaires du don lorsqu'il est capable de prendre leur perspective, mais parce qu'il a des difficultés à concentrer son attention sur les autres » (Hare *et al.*, 2010, p.589).

La conclusion de cette expérience a été corroborée par une autre étude plus récente, portant sur la « *décision empathique* » (Janowski, Rangel et Camerer, 2012). Le protocole est encore une fois relativement similaire à celui de l'étude de Hare *et al.* (2009). Trente-deux sujets reçoivent un budget de 10 dollars et doivent placer des enchères sur des DVDs. Ces achats ne sont pas réalisés pour leur propre compte, mais pour le compte d'un sujet passif. Une courte biographie et une photo de ce sujet est fournie aux participants. Ces derniers doivent donc prendre des « *décisions empathiques* », c'est-à-dire choisir en s'imaginant être à la place d'un autre. Dans la deuxième partie de l'étude, les participants répètent cette tâche, mais pour leur propre compte. Les résultat indiquent que, dans les deux cas, l'activité du CVMPF est corrélée avec la valeur monétaire accordée par le sujet (l'enchère) pour chaque DVD. Toutefois, les décisions empathiques font spécifiquement apparaître une activité dans le lobe pariétal inférieur, une région connue pour son rôle dans les capacités d'empathie. Selon les auteurs, cette zone encode une « *variable qui mesure la distance entre ses propres préférences et celles de l'autre* » (Janowski, Rangel et Camerer, 2012, p.7). Pour Janowski, Rangel et Camerer, cette étude confirme les résultats de l'étude de Hare *et al.* (2010) dans la mesure où elle montre que l'activité du CVMPF, qui est corrélée avec les choix et les préférences, est susceptible d'être affectée, en amont, par l'exercice de facultés d'empathie.

Les trois études qui ont été décrites dans cette section (Lin, Adolphs et Rangel, 2012; Hare *et al.*, 2010; Janowski, Rangel et Camerer, 2012) aboutissent donc à remettre en question dans le domaine du choix inter-personnel les approches dualistes, à la fois dans leurs méthodes et dans leurs résultats. Tout d'abord, ces expériences ne se contentent pas de reproduire, à l'IRM_f, les expériences classiques des *behavioral economics* sur le choix inter-personnel (jeu de l'ultimatum, du dictateur, de l'investisseur ou dilemme du prisonnier), mais reposent sur un protocole spécifique aux neurosciences, conçu spécifiquement pour mettre en évidence un arbitrage entre deux types de valeur. Deuxièmement, les résultats de ces travaux montrent que les réseaux cérébraux impliqués dans le « *reward* social » fonctionnent de manière unitaire. Certes, des régions pré-frontales permettent d'exercer un contrôle sur l'activité des régions responsables directement de l'encodage de la valeur subjective, mais l'exercice de ces fonctions d'empathie ou de mentalisation n'aboutit pas à la production d'un signal d'évaluation alternatif. C'est bien la même région -le CVMPF dans ces trois études- qui

est corrélé aussi bien avec la récompense égoïste que la récompense altruiste ou empathique. En d'autres termes, la cognition sociale suppose toujours, même lorsqu'elle mobilise des objectifs « supérieurs », des processus de type affectif, associés ici à l'activité du CVMPF[128].

C. L'intelligence sociale comme capacité à former des émotions sociales complexes

Hare *et al.* (2010) proposent une interprétation unitaire des processus impliqués dans l'apprentissage de la récompense « sociale », dans laquelle la valeur subjective attribuée par l'individu aux différentes options du choix est susceptible d'être modulée par des « *capacités de cognition sociale* » ou d'« *intelligence sociale* » (Hare *et al.,* 2010, p.589), qui recoupent notamment des facultés d'empathie et de mentalisation. Ces fonctions sont également connues de Fehr, qui y a même consacré un article (*cf.* Fehr et Singer, 2006). Cependant, contrairement à ce que suppose le cadre dualiste de Fehr, ces capacités à l'intelligence sociale ne doivent pas être comprises comme l'envers délibératif des émotions sociales, mais plutôt, dans la perspective unitaire avancée par Hare *et al.* (2010) comme une aptitude à former des émotions adaptées.

Une étude de 2009 suggère ainsi que l'intelligence sociale repose sur la capacité à former des « *émotions sociales complexes* » (*cf.* Krajbich *et al.*, 2009, p.6). Cette expérience offre en outre un croisement intéressant des méthodes de l'économie comportementale dans le scanner et celles du néo-comportementalisme. En effet, les expérimentateurs observent le comportement de patients atteints de lésions du CVMPF – qui a été largement étudié dans le domaine du choix inter-temporel et du *reward learning* (*cf.* chapitre 4 et 5)- dans les jeux classiques utilisés en *behavioral economics* pour étudier le choix inter-personnel (jeu de l'ultimatum, du dictateur, de l'investisseur).

Les auteurs proposent une interprétation « émotionnelle » à chaque type de réponse dans les différents jeux. Ainsi le taux de rejet des offres dans le jeu de l'ultimatum et le

128Il serait possible d'objecter que l'idée selon laquelle la coopération, par exemple, serait source de satisfaction au même titre qu'une récompense monétaire apparaissait déjà dans les premières études de neuroéconomie sociale (Rilling *et al.,* 2002; Sanfey *et al.*, 2003). Toutefois, cette activité était interprétée -de façon erronée selon nous- comme l'indice d'un plaisir ou d'une satisfaction effective, alors que ce signal, replacé dans la perspective plus générale des travaux sur le *reward learning*, doit plutôt être compris comme un signal de prédiction de la récompense.

montant des transferts par l'investisseur dans le jeu de la confiance sont pris comme des mesures d'un « *coefficient d'envie* », c'est-à-dire de l'aversion spontanée qu'expriment les individus à voir leur partenaire s'enrichir à leur détriment. Le niveau des offres dans le jeu de l'ultimatum et du dictateur, et celui des transferts réalisés en retour par le second joueur dans le jeu de l'investisseur fournissent une estimation d'un « *coefficient de culpabilité* », lié à l'aversion spontanée qu'ont les individus à s'enrichir au détriment de leurs partenaires (Krajbich *et al.*, 2009, p.3).

La comparaison entre les sujets lésés et les sujets sains montre que le taux de rejet des offres dans le jeu de l'ultimatum et le montant des transferts par l'investisseur dans le jeu de la confiance sont similaires dans les deux groupes. En revanche, le niveau des offres dans le jeu de l'ultimatum et du dictateur, et celui des transferts effectués en retour par le second joueur dans le jeu de l'investisseur est significativement plus faible pour les sujets lésés. Ces derniers ressentent donc moins de culpabilité, mais leurs coefficients d'envie sont proches des sujets sains. Les patients atteints de lésions du CVMPF sont en outre parfaitement capables d'exercer des capacités de mentalisation, puisqu'ils attribuent aux autres joueurs des croyances à leur égard tout à fait similaires à celles des sujets sains (Krajbich *et al.*, 2009, p.5)

Les lésions au CVMPF n'empêchent donc nullement le fonctionnement des capacités dites de théorie de l'esprit. Par ailleurs, les sujets lésés ressentent également des émotions sociales, qui, comme l'envie, sont suscitées par le comportement des autres individus. Les troubles de l'intelligence sociale manifestés par ces sujets ne s'explique donc pas par un déficit des émotions sociales ou du raisonnement stratégique, mais plutôt d'une déconnexion entre les émotions sociales et les capacités de mentalisation. Pour les auteurs, les sujets lésés sont incapables de former des types plus complexes d'émotion sociale, qui supposent, comme la culpabilité, l'expression d'un affect dirigé vers l'attention de l'autre. Comme dans le choix inter-temporel, la cognition sophistiquée dans le choix inter-personnel, ou les capacités d'intelligence sociale, peuvent donc se comprendre comme une aptitude à diriger son ressenti émotionnel vers des formes plus abstraites de récompenses ou de valeur: pour Krajbich *et al.*, « *le CVMPF est impliqué dans la planification future des actions pour la même raison qu'il contribue à la négociation des interactions individuelles:* [cette région] *génère des émotions liées à des résultats qui ne font pas l'objet d'une expérience directe, mais qui sont seulement imaginés* » (Krajbich *et al.*, 2009, p.6)

Loin de se limiter à un simple mécanisme inhibiteur, l'intelligence sociale renvoie à une capacité positive à former des émotions complexes. Cette interprétation remet en cause la

dichotomie entre émotions et raison, puisque ces capacités impliquent à la fois des processus émotionnels et délibératifs. Plus généralement, l'ensemble des travaux -plus récents- qui ont été abordés dans cette section tendent à réinscrire les études de neuroéconomie sociale dans le cadre du néo-comportementalisme. Cela a une conséquence importante sur le plan interprétatif, car les comportements altruistes ou coopératifs ne sont plus compris comme le résultat de préférences sociales, mais sont justifiés par l'exercice de compétences ou de capacités à la cognition sociale.

La notion d'intelligence sociale résous ainsi une difficulté rencontrée par les travaux de Fehr. La prise en compte des dynamiques d'apprentissage permet en effet d'expliquer la formation des préférences sociales des individus. Les préférences sociales ne sont plus prises comme des paramètres fixes, mais au contraire comme des variables endogènes, susceptibles de variations chez un même individu, selon que celui-ci exerce ou non ses capacités de « *cognition sociale* » (*cf.* Hare *et al.*, 2010). Comme le suggère Camerer, les préférences ne sont plus dans cette perspective les points de départ ou primitives de l'analyse mais au contraire les points d'arrivée (Camerer, 2007, p.28).

Cette explication de l'origine des préférences sociales ne fait que déplacer le problème vers une autre question. Pourquoi en effet, les agents décident-ils d'exercer ou non leur intelligence sociale? Qu'est-ce qui, en d'autres-termes permet d'expliquer que certains sujets soient capables de se comporter de manière altruiste et coopérative, et d'autres ne le soient pas? Les neurosciences proposent deux réponses à ces questions. D'une part, le signal d'évaluation encodé par le CVMPF est, comme dans le cas du choix inter-temporel, sensible à l'environnement du choix, et il est notamment manipulable par des facteurs attentionnels (Hare, Malmaud et Rangel, 2011). Par conséquent, certains environnements seront naturellement plus favorables à l'altruisme ou à la coopération. D'autre part, l'exercice des capacités de cognition sociale présuppose la possession de ces aptitudes par le sujet. Dans l'étude de Krajbich *et al.* (2009, p.1), les individus sains sont ainsi déclarés « *aptes à la cognition sociale* », à l'inverse des sujets lésés.

La réflexion s'élargit ainsi dans deux directions. Il s'agit d'abord de modéliser l'influence de l'environnement sur le choix inter-personnel. Ces recherches portant sur « l'ingénierie environnementale de la décision » seront abordées dans le prochain chapitre. Surtout, les recherches sur l'intelligence sociale appellent, en amont, une définition neuronale et/ou comportementale de ce que constitue cette aptitude. Or, quand les expérimentateurs parlent de sujets capables de se « *comporter de manière adaptée avec leurs semblables* »

Krajbich *et al.*, 2009, p.5), ils entendent par là des individus n'ayant pas de troubles de la cognition sociale. La caractérisation des déviances et des pathologies de l'intelligence sociale occupe ainsi un rôle important dans la neuroéconomie sociale.

IV. Les implications prescriptives. Troubles et pathologies de l'intelligence sociale

Les approches dualistes de Fehr et de Zak rencontrent un problème normatif, lié au caractère ambigu des émotions, et à leurs implications équivoques pour la théorie du choix rationnel. Dans certains jeux, comme le jeu de l'ultimatum par exemple, l'expression de pulsions de rejet suites à un partage inéquitable empêche la réflexion rationnelle sur les coûts et les avantages. Les émotions s'opposent à l'exercice d'une intelligence qualifiée de machiavélique. D'un autre côté, les émotions empathiques dans le jeu de l'investisseur favorisent la confiance mutuelle, la coopération et élèvent le bien-être global. Faut-il, dans le domaine du choix inter-personnel, chercher à favoriser les émotions dites sociales, ou au contraire inciter les individus à exercer davantage de contrôle sur leur ressenti émotionnel?

Cette question ne soulève cependant pas, en pratique, de réels problèmes. Dans l'étude précédemment décrite de Krajbich *et al.* (2009) par exemple, il apparaît évident, à lire le compte-rendu de l'expérience, que le comportement rationnel est celui qui est suivi par les sujets sains ou normaux, dans la mesure précisément où les sujets lésés sont caractérisés par des troubles de la cognition sociale. Encore une fois, comme pour le choix inter-temporel, le présupposé selon lequel un certain nombre de comportements qualifiés de pathologiques sont irrationnels offre un fondement normatif à l'équivalence qui est introduite entre rationalité et normalité (*cf.* chapitre 5). Ce que *doivent* faire les individus en matière de choix inter-personnel consiste d'abord à ne pas connaître de tels troubles de la cognition sociale.

Les remarques similaires qui ont été faites dans le chapitre précédent à propos de ce paradigme « *pathologique* » (Vallois, 2011) s'appliquent donc aussi au domaine des interactions sociales. Comme pour le choix inter-temporel, les neuroéconomistes cherchent ici à utiliser des jeux économiques comme des protocoles de diagnostic. Cette tentative s'appuie cependant en amont sur la sélection de joueurs considérés comme « sains » ou « pathologiques », selon les critères cliniques en vigueur. C'est donc sur la base d'une identification médicale des sujets connaissant des troubles de la cognition sociale (autistes, sujets lésés, psychopathes, *etc.*) que les expérimentateurs peuvent par la suite estimer quantitativement, dans le cadre de protocoles standardisés, les variables neuronales caractéristiques de ces dysfonctionnements. L'influence des critères cliniques ne se limite pas aux seules études mobilisant des sujets pathologiques, mais plus généralement, à toutes les

expériences de neuroéconomie car celles-ci définissent toujours la rationalité comme celle du comportement sain.

Les implications normatives de cette compréhension pathologique des interactions sociales seront approfondies dans le chapitre suivant. Il convient de souligner que l'importance des préoccupations médicales dans la définition de la rationalité apparaît sans doute plus clairement ici que pour le choix inter-temporel. En effet, la neuroéconomie sociale ne peut s'appuyer sur une norme de rationalité clairement définie dans les problèmes de décision qu'elle étudie. Dans un jeu tel que celui de l'investisseur, l'identification du choix rationnel des deux joueurs dépend de l'objectif qui est visé, selon qu'il s'agisse de favoriser la rationalité collective (maximiser le gain total) ou individuelle (maximiser le gain de chaque joueur). A l'inverse, dans le cas du choix inter-temporel, un principe tel que la constance du taux d'actualisation, postulé par le modèle exponentiel de Samuelson, peut servir à nourrir l'illusion selon laquelle il existerait une norme économique de l'actualisation inter-temporelle.

La visée « *thérapeutique* », pour reprendre les termes de Gul et Pesendorefer (2005, p.8), de la neuroéconomie sociale est donc manifeste. Le projet de psychiatrie économique, visant à utiliser les protocoles de jeux comme des dispositifs de diagnostic, est cependant pour l'instant relativement moins développé que dans le domaine du choix inter-temporel et des comportements addictifs. L'évaluation des aptitudes à l'intelligence sociale peut remplir plusieurs objectifs. Le diagnostic peut d'abord porter sur les capacités dites de mentalisation ou de « *théorie de l'esprit* ». Les autistes sont notamment réputés pour souffrir de déficits fonctionnels dans ce domaine (Baron-Cohen, 1995). Vernon Smith, qui a participé à l'expérience fondatrice de McCabe *et al.* (2001) sur la coopération attache une grande importance à ces capacités de mentalisation, étant lui-même diagnostiqué d'une forme légère du syndrome d'Arnsperger (*cf.* chapitre 1, section I). Il propose ainsi d'utiliser les études de neuroimagerie sur le jeu de l'investisseur comme des instruments d'évaluation de ce qu'il appelle les « *compétences sociales* » des individus (Smith, 2007, p.316).

Il y a eu effectivement certaines tentatives allant dans le sens de la proposition de Smith. Hill et Sally (2002) ont notamment réalisé une étude célèbre de neuroimagerie sur des sujets autistes, en utilisant le jeu de l'investisseur et celui de l'ultimatum. Cependant, cette expérience suggère que ces jeux ne peuvent fournir de protocoles diagnostics fonctionnels pour ces troubles, car les sujets autistes se montrent capables, en général, de pallier leurs déficits en « mentalisation » par des processus de compensation leur permettant de prendre des décisions relativement semblables à celles des sujets sains (Hill et Sally, 2002).

Plus généralement, les troubles de l'intelligence sociale doivent être replacée dans la perspective écologique d'une interaction entre l'individu et son environnement. Par conséquent, il est contestable en principe d'utiliser des versions non-répétées[129] du jeu de l'investisseur, de l'ultimatum ou dilemme du prisonnier comme dispositifs de diagnostic de ces troubles. Il est certes possible de mettre en évidence, par exemple, que les individus souffrant de ce que Zak appelle des troubles liés à un déficit d'ocytocine (*oxytocin deficit disorder* (ODD) (Zak, 2005)), trahissent en général la confiance qui leur est accordée dans le jeu de l'investisseur. Ces troubles peuvent être à l'origine de graves troubles dans la vie réelle, puisque les individus concernés « *ont des difficultés à former des liens d'amitié et d'attachement amoureux* », qui sont souvent liées à la sociopathie (Zak, 2008, p.2). Cependant, comme le reconnaît Zak lui-même, cela ne signifie pas que, de façon mécanique, un déficit en ocytocine implique nécessairement une personnalité sociopathe, car « *l'histoire de l'individu et de son environnement jouent un grand rôle* » (Zak, 2010, p.545).

La portée d'une étude telle que celle de Krueger *et al.* (2012) par exemple, qui montre qu'une variation génétique dans le gène codant pour le récepteur de l'ocytocine entraine une diminution de la confiance accordée dans le jeu de l'investisseur, mérite donc d'être relativisée. En effet, il ne suffit pas d'avoir un déficit en ocytocine pour se transformer en sociopathe; de la même façon, les sujets ayant des lésions au CVMPF ne développent pas tous des troubles de la cognition sociale. De tels individus ont un terrain favorable pour l'apparition de ces troubles, mais il n'y a pas de lien nécessaire de cause à effet entre ces dysfonctionnements organiques et les symptômes comportementaux d'une intelligence sociale amoindrie.

Si par conséquent, comme le propose Fehr, les études de neuroéconomie sociale peuvent s'appliquer à la détection de la psychopathie et de la sociopathie (Fehr, 2008, p.225), il est nécessaire de prendre en compte la dimension environnementale de ces troubles. Le diagnostic doit s'appuyer non pas sur des études « *one shot* », mais plutôt sur des protocoles portant sur l'apprentissage. Les travaux de ce type, visant à détecter des troubles de l'intelligence sociale, restent pour l'instant peu développés en neuroéconomie sociale, car celle-ci s'est principalement appuyée sur la reprise de jeux non-répétés empruntés aux *behavioral economics*. Cependant, à la fin des années 2000, des chercheurs ont commencé à appliquer le paradigme du *reward learning* à la récompense sociale, comme dans l'expérience

129Généralement, dans les expériences, les jeux sont répétés, mais à chaque fois avec des partenaires différents et anonymes, ce qui revient, en pratique, à annuler toute effet d'apprentissage et de réputation. Ces tâche bien que répétées demeurent donc purement statiques; le terme anglais est plus parlant: il s'agit de *one shot studies*.

de Lin, Adolphs et Rangel (2012, *cf. supra*). Dans le même temps, des travaux en neurosciences sur la psychopathie, réalisés notamment par James Blair (Blair, 2005) ont avancé des explications assez proches pour ces troubles, en termes d'apprentissage des « *émotions sociales* ».

Comme l'a souligné Fehr (Fehr, 2008, p.225), la neuroéconomie sociale offre une explication originale de la psychopathie. Ce trouble représente selon nous l'innovation théorique principale associée à la notion d'intelligence sociale et peut se concevoir comme l'équivalent, dans le domaine du choix inter-personnel, du concept d'addiction environnementale, lequel correspond à un trouble non pas de l'apprentissage social mais de la planification inter-temporelle des récompenses. Les recherches réalisées par James Blair en neurosciences cognitives sur la psychopathie convergent avec ce cadre interprétatif du « *social learning* ». Nous quittons ici le terrain de la neuroéconomie à strictement parler, puisque Blair ne fait pas de référence explicité à la neuroéconomie, mais il est possible néanmoins de mettre en évidence des liens conceptuels.

Blair souligne que la psychopathie n'est pas un trouble du raisonnement moral. Les tests classiques, tels que celui proposé par Kohlberg notamment (Colby, Kohlberg *et al.*, 1983), consistant à évaluer le degré de complexité de la réflexion morale des sujets, échouent à mettre en évidence un écart significatif entre sujets normaux et sujets psychopathes (Lee et Prentice, 1988). Pour Blair, la psychopathie est plutôt liée à un déficit de nature émotionnelle, mais ce déficit doit être compris comme un trouble de l'apprentissage émotionnel (Blair, 2005, p.111). Il est significatif, selon nous, de constater de ce point de vue que Blair prend soin de distinguer son explication de la psychopathie de celle de Damasio. En effet, la référence damasienne renvoie pour nous à des schémas d'interprétation dualiste, ne prenant pas assez en compte la dimension évolutive des décisions (*cf. supra*). Comme le souligne Blair, la théorie des marqueurs somatiques de Damasio n'est certes pas directement une théorie de la psychopathie, mais Damasio a suggéré que la psychopathie pouvait se comprendre à l'aide du concept de marqueur somatique (Blair, 2005, p.93; Damasio 1994, Damasio, Tranel et Damasio 1990). Pour Damasio, ces marqueurs émotionnels permettent de marquer un scénario comme bon ou mauvais et seraient doc à l'œuvre également dans la cognition sociale. Les troubles de la sociopathie et de la psychopathie seraient ainsi liés à une absence de réponse émotionnelle face aux scènes de désastre social, de souffrance infligée à autrui, *etc.* (Damasio, Tranel et Damasio, 1990), et les individus caractérisés par de tels troubles seraient susceptibles d'obtenir des performances médiocres à l'IGT (Bechara *et al.*, 1994).

Blair observe effectivement que la performance des psychopathes à l'IGT est plus faible que celle des sujets sains (Blair *et al.*, 2001). Cependant, ces individus n'ont pas de réponses émotionnelles amoindries face à des stimuli représentant un événement social menaçant, comme une tempête ou un incendie entrainant la mort de plusieurs personnes, contrairement à ce que suppose Damasio (Blair, 1999; Blair *et al.*, 1997). Cette observation permet aussi de rejeter les explications classiques de la psychopathie en termes de peur, selon lesquelles les psychopathes seraient incapable d'anticiper et de craindre les stimuli associés à des expériences déplaisantes (Lykken, 1957). Les travaux de Blair ont permis de montrer que la psychopathie a plutôt pour origine un « *déficit dans le learning émotionnel* » (Blair, 2005, p.110), ce qui signifie que les psychopathes ont des difficultés non pas à exprimer de l'aversion, de la peur ou de la crainte, mais à utiliser ces émotions primaires dans la réflexion portant sur le choix inter-personnel. En utilisant un protocole relativement similaire à celui de l'expérience de Lin, Adolphs et Rangel (2012) sur le *social reward learning*, Blair observe que les psychopathes se distinguent non pas par une insensibilité face à des stimuli sociaux négatifs (visage exprimant une douleur par exemple), mais à élaborer des stratégies d'apprentissage destinées, dans ces tâches, à éviter l'occurrence de tels stimuli (Blair *et al.*, 2002; Blair et Cipolotti, 2000).

Conclusion du chapitre 6

L'étude du choix inter-personnel représente un sous-domaine tout à fait particulier de la neuroéconomie. En effet, la neuroéconomie dite sociale a rencontré très rapidement une large diffusion dans la littérature scientifique. Historiquement, ce sont au départ les expériences de neuroimagerie portant sur le choix inter-temporel qui ont suscité l'intérêt des économistes pour les neurosciences. Le succès de la neuroéconomie sociale est cependant à double tranchant, car il invite souvent les neuroéconomistes, dans ce domaine, à déborder du cadre interprétatif autorisé par la pratique scientifique, pour s'engager dans des réflexions de neuro-philosophie ou de neuro-éthique, dont le contenu théorique est particulièrement vulnérable à la critique méthodologique.

La neuroéconomie des interactions sociales s'est développée au départ, comme pour le choix inter-temporel, comme une économie comportementale dans le scanner. Cette approche conduit à mettre en avant le rôle des émotions dites sociales dans les interactions individuelles. La neuroéconomie fournit ici une représentation dualiste de la cognition bien moins naïve que ce qui lui est reproché par ses principales critiques méthodologiques. Les principaux promoteurs de la neuroéconomie sociale, formés et issus des *behavioral economics*, proposent des interprétations comportementales précises. Ils cherchent à contrôler les différents artefacts possibles des fonctions cognitives étudiées. Le principal écueil de cette approche se situe plutôt au niveau neuronal, car le cadre d'analyse des données neuronales est purement statique. Ce schéma interprétatif s'oppose à la perspective dynamique ou séquentielle du néo-comportementalisme.

Il est possible de proposer une autre compréhension des processus neuronaux à l'œuvre dans la cognition sociale, à partir du paradigme du *reward learning*. Il s'agit alors d'un apprentissage de la récompense dite sociale, mais les mécanismes d'anticipation, de prédiction et de correction de ces prédictions restent similaires. Cette perspective d'analyse débouche sur la notion d'intelligence sociale, comprise comme la capacité à former des émotions sociales complexes, comme la culpabilité par exemple.

Il convient de souligner néanmoins que l'opposition que nous avons introduite entre ces deux approches en neuroéconomie sociale est moins tranchée que pour le choix inter-temporel. Il n'y a pas eu, comme cela a été le cas à propos du problème de l'actualisation

temporelle des récompenses (*cf.* chapitre 5), de controverse explicite entre les chercheurs. Un même auteur comme Colin Camerer a même participé à des travaux que nous avons considérés comme relevant de l'économie comportementale dans le scanner (Fehr et Camerer, 2006), ou du néo-comportementalisme (Krajbich *et al.*, 2009). La porosité de la frontière entre ces deux approches ne constitue cependant pas une difficulté pour notre étude historique, précisément parce que la neuroscience du *reward learning* représente de notre point de vue un stade plus avancé, par rapport à l'économie comportementale dans le scanner. Il n'est donc pas nécessairement contradictoire qu'un même chercheur se soit successivement inscrit dans les deux démarches.

Il reste donc possible d'effectuer une reconstruction historique du champ de manière à faire apparaître celui-ci dans le prolongement du courant néo-comportementaliste. En effet, nous avons vu que la mise en avant d'une dimension biologique et évolutive du choix inter-personnel, à travers la notion de monnaie neuronale commune, permettait de résoudre les apories suscitées par la notion d'émotion sociale. En outre, les critères de classification clinique des troubles de la cognition sociale fournissent à ce programme de recherche un ancrage médical, et qui témoigne de l'importance des préoccupations liées à la pathologie. La neuroéconomie sociale peut ainsi être comprise elle aussi comme un projet de psychiatrie économique.

Chapitre 7. Paternalisme Libertarien et Psychiatrie Économique. L'apport de la neuroéconomie à l'analyse comportementale du bien-être.

Bien qu'apparaissant au départ comme un simple prolongement de l'économie comportementale -une « économie comportementale dans le scanner » (Ross, 2008)-, la neuroéconomie s'émancipe progressivement de la tutelle des *behavioral economics*. Elle s'affirme comme une discipline autonome à la fin des années 2000. Notre étude historique a donc abouti à distinguer la neuroéconomie de l'économie comportementale, et, plus généralement, des travaux de Daniel Kahneman (*cf.* introduction à la première partie).

D'un point de vue de l'histoire de la pensée économique, l'équivalence que nous avons introduit ici entre *behavioral economics* et le programme de recherche de Daniel Kahneman pourrait être discutée. Les économistes comportementalistes font preuve d'opportunisme et n'hésitent pas à s'affranchir du cadre kahnemanien. Nombre d'entre eux, comme Colin Camerer par exemple, opèrent à la frontière d'autres programmes de recherche, et s'inspirent de paradigmes divers en psychologie expérimentale. Comme le souligne Heukelom, les économistes comportementalistes ont manifesté la capacité à « *explorer une série de différentes disciplines et méthodes scientifiques* » pour enrichir leur description du comportement humain (Heukelom, 2009, p.123).

Cependant, même si les *behavioral economics* constituent un domaine scientifique ouvert à des approches diverses et variées, les historiens de la pensée considèrent en général que le noyau dur de ce programme de recherche -le principe d'une séparation entre le descriptif et le normatif, largement défendu par Kahneman (voir par exemple Kahneman, 2003)- est indiscutablement d'inspiration kahnemanienne. Heukelom écrit ainsi par exemple:

> « *les économistes comportementalistes restèrent toujours fidèles et revinrent toujours au cadre normatif-descriptif qui avait été originairement introduit par Kahneman et Tversky. Ce cœur conceptuel détermina la manière avec laquelle les économistes comportementalistes comprirent le monde, cela détermina les implications pour le bien-être qu'ils pouvaient défendre, cela détermina leur retour en arrière lorsque leurs explorations théoriques s'écartèrent trop loin de ce cœur conceptuel. La nouvelle terminologie de la rationalité fournit les fondation à la discussion sur le paternalisme en économie comportementale qui apparait au début des années 2000* » (Heukelom, 2009, p.123)

Selon nous, la neuroéconomie remet en cause ce principe d'une séparation entre le normatif et le descriptif, et lui oppose la notion de pathologie. Là où la psychologie kahnemanienne décrit des « erreurs » ou des biais au regard d'une norme théorique considérée comme donnée, les neuroéconomistes rejettent certaines conduites comme irrationnelles au nom de l'évidence de leur caractère pathologique. Or, comme le souligne Heukelom, la distinction du normatif et du descriptif (et sa remise en question) ont des conséquences sur le plan des « *implications pour le bien-être* », c'est-à-dire pour les recommandations ou prescriptions qui peuvent être proposées en économie publique à partir des résultats expérimentaux. En effet, selon Kahneman, l'économie comportementale et la psychologie ont simplement pour but de décrire le comportement humain. Cette description aboutit souvent à montrer que la norme de rationalité, telle qu'elle est définie par la théorie économique[130], n'est pas respectée par les individus « réels ». Par conséquent, des contributions en économie comportementale ont suggéré une possible régularisation comportementale du bien-être, en incitant les individus à « restaurer » la rationalité de leurs décisions. Cette « *économie comportementale du bien-être* » (*cf.* Bernheim et Rangel, 2009) ou « *économie publique comportementale* » (Bernheim et Rangel, 2008) prend donc acte du divorce entre ce que l'économie prescrit d'une part, et les choix réels des individus d'autre part; cette perspective kahnemanienne manifesterait ainsi ce que Wade Hands appelle le « *tournant normatif* » de la théorie économique contemporaine, dans laquelle la rationalité n'est plus qu'une norme à atteindre (Hands, 2009).

La neuroéconomie se donne ici, dans le domaine de l'analyse du bien-être au premier abord comme un simple prolongement des *behavioral economics*, en tant qu'elle débouche sur des recommandations normatives relativement similaires, visant à développer une « ingénierie comportementale » de la rationalité. Néanmoins, l'enjeu consistera, comme dans les chapitres précédents, à distinguer les propositions théoriques de la neuroéconomie de celles des *behavioral economics*. La neuroéconomie est en effet porteuse d'une remise en cause bien plus radicale de la théorie économique, puisque la validité de cette dernière est remise en cause à la fois sur le plan descriptif et normatif. En d'autres termes, il ne s'agit pas, pour les neuroéconomistes, de rétablir un idéal de comportement défini par les économistes, mais de proposer à la fois une nouvelle façon de décrire les comportements et de les réguler.

Dans cette perspective, l'analyse comportementale du bien-être ne se fonde alors plus,

130Il reste bien sûr à définir ce qu'est cette norme, ou à préciser la théorie économique qui doit servir de référence. Le problème de la définition de la norme, dans le paternalisme libertarien, sera abordé plus en détail dans la première section.

sur le plan normatif, en référence à la théorie économique, mais à des catégories cliniques. La prescription ne s'appuie pas sur une norme à atteindre, mais sur la détection de divers comportements pathologiques, qu'il s'agit littéralement de guérir. Ce chapitre étend donc au domaine de l'économie du bien-être les analyses qui ont été menées pour le choix inter-temporel et le choix inter-personnel, mais il a également pour but de caractériser la neuroéconomie comme un projet de psychiatrie économique, conformément à notre hypothèse interprétative (*cf.* introduction).

L'analyse du bien-être en économie comportementale se propose de corriger les erreurs qui apparaissent dans les choix réels des individus, en agissant la structure des incitations et l'environnement. Cette approche fait face cependant à de nombreuses critiques, portant sur l'arbitraire de la norme (I. L'économie comportementale du *bien-être* : la portée normative des *behavioral economics*). Un économiste comportementaliste et un neurobiologiste (respectivement Douglas Bernheim et Antonio Rangel) ont récemment proposé, en s'appuyant sur les neurosciences contemporaines, un modèle de régulation du bien-être, qui, selon nous, représente une perspective normative intermédiaire entre les *behavioral economics* et la neuroéconomie (II). Une perspective intermédiaire : la collaboration entre le neurobiologiste Antonio Rangel et l'économiste Douglas Bernheim). L'approche spécifique de la neuroéconomie en la matière sera enfin distinguée de celle de l'économie comportementale et caractérisée comme un projet de psychiatrie économique (III. Neuroéconomie et Psychiatrie économique).

I. L'économie comportementale du *bien-être* : la portée normative des *behavioral economics*

A partir du début des années 2000, un important débat théorique se développe en économie comportementale autour de la question du paternalisme. Plusieurs économistes ont en effet essayé de défendre l'idée d'une régulation des choix individuels, à l'appui de résultats théoriques des *behavioral economics*. Ces discussions ont donné naissance à un sous-domaine de l'économie comportementale à part entière, que Bernheim a plus récemment baptisé du nom d'« *économie comportementale du* bien-être » (*behavioral welfare economics*) (Bernheim, 2008). La réflexion normative en économie comportementale concerne principalement le choix inter-personnel et inter-temporel. Cette analyse comportementale du bien-être défend l'idée d'une régulation comportementale de la décision, en considérant que les « erreurs » observées en laboratoire représentent une violation de la rationalité économique (A). Une telle approche rencontre cependant une difficulté majeure, liée à l'absence de justification de la norme de comportement proposée (B).

A. De la description des erreurs à la prescription des normes

L'analyse comportementale du bien-être part du principe selon lequel les individus, dans la vie « réelle », se montrent souvent incapables de respecter la norme de rationalité prescrite par les économistes. Il s'agirait même là du résultat théorique principal des *behavioral economics* : pour Bernheim et Rangel, « *l'intérêt croissant dans le champ de la psychologie et de l'économie au cours des dernières années a été largement stimulé par l'accumulation de preuves indiquant que le modèle néoclassique de la décision fournit une description inadéquate du comportement humain dans de nombreuses situations économiques* » (Bernheim et Rangel, 2008, p.1). Or, si ce modèle économique est pris comme une norme à atteindre, les comportements humains apparaissent alors comme des erreurs. La mesure d'un écart par rapport à cet idéal de choix rationnel justifie alors une intervention.

Dans le domaine du choix inter-temporel, de nombreuses études empiriques dans les

années 1980 et 1990 ont montré en effet que l'actualisation temporelle ne se conformait pas en règle générale, chez l'homme, au modèle d'actualisation exponentielle de Samuelson (1938) (*cf.* chapitre 2, section II). Ce phénomène se manifeste notamment par un renversement des préférences au cours du temps, et recevait une explication dans le cadre de modèles d'actualisation quasi-hyperbolique, comme celui de Laibson (1997).

Or, bien que constituant une tendance relativement commune, la non-constance du facteur d'actualisation était interprétée comme une anomalie ou une erreur. Dans cette perspective, il est donc souhaitable de prendre des mesures visant à restaurer la cohérence temporelle des décisions. Laibson suggère par exemple que l'investissement dans des actifs illiquides permet d'empêcher toute modification des intentions initiales. Ce même auteur déplore également les innovations financières récentes, qui facilitent excessivement l'accès à la liquidité, et amplifient ainsi les tendances dépensières des individus (Laibson, 1997).

La plupart des travaux sur le modèle d'actualisation quasi-hyperbolique débouchent ainsi sur des recommandations pratiques, dont l'objectif vise à limiter la myopie temporelle des agents, celle-ci ayant pour conséquence un biais excessifs en faveur du présent (représenté dans le modèle par le facteur β). Ces recommandations peuvent être de deux types. Il peut s'agir tout d'abord de mécanismes d'engagement, qui offrent la possibilité de « se lier les mains » et de verrouiller les plans de consommation inter-temporels décidés *ex ante*. Dès les années 1950, Strotz évoquait ce genre de dispositifs, dans son article fondateur sur le choix inter-temporel (Strotz, 1956). Plus récemment, Thaler et Benartzi (2004) ont défendu la mise en place de programmes d'épargne automatique, dans lequel le salarié s'engage au départ à ce qu'un montant fixe soit prélevé de chacun de ses salaires et soit alloué à son épargne. A tout moment, le salarié peut décider de sortir du plan ; mais s'il ne ré-examine pas son choix, le prélèvement continue. Ces propositions ont fait l'objet de tests empiriques concluants, et le *Save More Tomorrow Program* conçu par Richard Thaler a permis effectivement d'élever le taux d'épargne des salarié dans les entreprises dans lesquelles ce dispositif a été mis en place (Thaler et Benartzi, 2004).

Ce type de mécanisme d'engagement encourage toute tentative de planification des choix futurs. Leur champ d'application est extrêmement large. Il ne se limite pas seulement au comportement d'épargne, mais concerne tous les comportements susceptibles d'être affectés par une éventuelle impulsivité du décideur. Par exemple, un individu cherchant à contrôler sa consommation alimentaire pourra « s'engager » en définissant ses menus pour la semaine ou pour le mois ; un individu dépensier pourra décider de plafonner le montant de découvert autorisé sur son compte ; un individu dépendant aux drogues dures pourra décider de suivre

une cure de désintoxication jusqu'à son terme, *etc.* (*cf.* Laibson, Chabris et Schuldt, 2008). L'idée, à chaque fois, consiste à empêcher tout renversement des préférences tout en respectant la liberté du décideur : c'est parce que celui-ci s'est montré inflexible *ex ante* qu'il est logique de lui refuser toute modification de son plan initial *ex post*.

L'analyse comportementale du bien-être propose également de modifier l'environnement de la décision, de façon à rendre les individus plus patients. Le biais en faveur du présent, dans le modèle d'actualisation quasi-hyperbolique, peut se comprendre en effet comme un effet de cadrage (*framing effect*), lié au mode de présentation du choix. Par exemple, à la cafétéria, je peut choisir un dessert appétissant mais peu diététique -une glace- au lieu de l'aliment peu appétissant mais diététique -une pomme- simplement parce que la glace est placée en première position dans la file. Un ordre inversé, ou, plus généralement, toute présentation contribuant à rendre plus visible l'option faiblement perceptible, permettrait à l'individu d'adopter une conduite plus « *rationnelle* ». Cette illustration célèbre, empruntée à Thaler et Sunstein (Thaler et Sunstein, 2003), ouvre ainsi un champ de régulation très large, puisqu'il couvre l'ensemble des stimuli environnementaux -économiques ou non-économiques- susceptibles d'affecter le *framing* du problème par l'agent.

Cet interventionnisme comportemental d'un genre nouveau ne se limite donc pas à la manipulation des variables économiques plus familières pour un économiste, comme les prix ou les quantités produites. Celles-ci relève plutôt de l'interventionnisme économique qualifié de « *standard* » ou de « *traditionnel* » par Bernheim et Rangel, qui n'appelle pas de traitement particulier par les économistes comportementalistes (Bernheim et Rangel, 2008, p.55). Certes, O'Donoghue et Rabin (2003) défendent la mise en place de taxes spécifiques sur les produits addictifs (*sin taxes*), mais cette proposition est assez critiquable dans la perspective d'une théorie comportementale de l'addiction (*cf.* section II). L'analyse comportementale du bien-être est plutôt une ingénierie de l'environnement du choix, qui se donne pour instruments d'action les règles et mécanismes d'engagement, et l'ensemble de de ce que Bernheim et Rangel (2009) appellent « *conditions auxiliaires de la décision* », qui se définissent comme des « *caractéristiques de l'environnement du choix qui peut affecter le comportement […]. Les illustrations typiques de ce conditions auxiliaires incluent notamment le moment auquel le choix est pris, la manière avec laquelle l'information ou les alternatives sont présentées, la caractérisation d'une option comme « status quo »* [option dont la valeur est présentée comme nulle ou neutre pour le décideur]*, le degré de saillance de l'option qui sera choisie par défaut, ou l'exposition à un effet d'ancrage* » (Bernheim et Rangel, 2008, p.55).

Toute la difficulté consiste à distinguer ce qui relève de ces conditions auxiliaires des

caractéristiques du bien proprement dites. Ce problème sera approfondi dans la deuxième section. Il importe de souligner ici que le planificateur, c'est-à-dire l'individu qui s'occupe de la conception de ces conditions, traite de variables qui ne sont pas prises comme pertinentes dans l'évaluation dite standard du bien-être (Bernheim et Rangel, 2008, p.55). Sans modifier l'éventail de choix possibles, l'intervention peut porter notamment, comme le suggèrent Bernheim et Rangel, sur l'information concernant les produits, leur mode de présentation, la publicité, le signalement d'effets secondaires ou indésirables, *etc.*

Ce type de régulation rentre dans le cadre de ce qui est connu sous le nom de paternalisme « *asymétrique* » (Camerer *et al.*, 2003) ou « *libertarien* » (Thaler et Sunstein, 2003, 2008; Sunstein et Thaler, 2003; Jolls et Sunstein, 2006; Gruber et Koszegi, 2001, O'Donoghue et Rabin, 2003, Ariely; 2008)). Nous préférerons ici le terme d'économie comportementale publique ou du bien-être, pour y inclure également les recommandations portant sur le choix inter-personnel. Les implications normatives de l'économie comportementale sont dans ce domaine peut être moins évidentes que dans le cas du choix inter-temporel. En effet, le partage entre les conduites rationnelles et irrationnelles est moins évident que pour le choix inter-temporel; et il est moins aisé de qualifier ici certains comportements -comme le *free-riding* par exemple- comme des erreurs.

Cependant, les *behavioral economics* peuvent aussi déboucher sur des recommandations pratiques dans le cas du choix inter-personnel, de deux manières différentes. Un premier type de raisonnement consiste à supposer que la coopération, et/ou la confiance entre les individus, est un objectif souhaitable en soi[131]. Ce point de vue est notamment défendu par Paul Zak. Dans un article de 2001, celui-ci montre, en s'appuyant sur un index national de confiance inter-personnelle, que la confiance favorise la croissance et l'investissement. Il suggère par ailleurs l'existence d'une « *trappe de confiance* » dans les pays pauvres (Zak, 2001). Si, par conséquent, la confiance et la coopération sont indispensables au bien-être et au développement, les comportements méfiants et de *free riding* peuvent être considérés comme des erreurs. Pour Zak, les politiques économiques doivent donc chercher à promouvoir la confiance. Sans défendre de propositions très précises, Zak suggère que la confiance est positivement corrélée au degré de proximité sociale, et à la robustesse des institutions, en particulier des institutions formelles (Zak, 2001, p.2).

Or, un planificateur cherchant à élever la confiance et la coopération entre les individus peut tout à fait tirer parti des expériences comportementales. Les très nombreuses

131Évidemment, on va le voir, ce présupposé est critiquable. Nous envisagerons ces critiques dans a prochaine sous-section. Nous nous limitons ici à un simple exposé des arguments en faveur de la régulation des choix inter-individuels.

études sur le dilemme du prisonnier ou le jeu de l'ultimatum permettent en effet de déterminer des facteurs susceptibles d'élever la participation au bien public, ou d'éliminer -au moins en partie- les comportements de *free riding*. Francesco Guala interprète ainsi l'utilisation des expériences en économie à la manière d'une ingénierie, qui aurait pour but de modeler le réel à l'image de la théorie économique. Ici, l'objectif consiste non à restaurer un équilibre de concurrence pure et parfaite, mais à élever le niveau de coopération entre individus[132].

Les nombreuses expériences sur la théorie de le punition altruiste suggèrent par exemple que les individus adoptent des conduites coopératives lorsqu'une punition des comportements de *free riding* est envisageable. Pour Fehr et Gintis, c'est ainsi la possibilité, ou non, de sanctionner qui détermine l'ordre social : si les sanctions sont possibles, l'homme « *social* » domine l'« *homme économique* », intéressé par son seul intérêt personnel. En l'absence d'une telle menace, l'égoïsme et l'intérêt personnel constituent la règle du comportement (Fehr et Gintis, 2007, p.53). La théorie de la punition altruiste s'applique donc aux situations d'interactions réelles, comme le soulignent Fehr et Gächter : « *Dans notre perspective, la punition du free riding joue aussi un rôle important dans la vie réelle. Il semble, par exemple, plus que probable […] que le licenciement de travailleurs soit fortement désapprouvé par leurs collègues, et que les casseurs de grèves font face à l'hostilité spontanée des employés en grève. L'impact énorme des opportunités de punition sur les contributions* [au bien public] *dans notre expérience suggère que la négligence de cette propension largement partagée à punir les free riders est susceptible de fournir de mauvaises prédictions, et, par conséquent, de mauvaises recommandations normatives. Les structures institutionnelles et sociales qui, théoriquement, sont supposées causer les mêmes comportements en l'absence de toute volonté de punir peuvent être à l'origine de comportements complètement différents si cette même volonté est prise en compte* » (Fehr et Gächter, 2000, p.993).

Pour Fehr, l'attention portée à la punition altruiste est susceptible d'avoir des implications dans l'économie réelle, en particulier dans le domaine des relations salariales (auquel Fehr a consacré de nombreux travaux dans les années 1990). Fehr ne propose pas vraiment de recommandation, mais se contente d'indiquer que les punitions altruistes jouent un rôle important dans la vie réelle. Il n'est pas du tout sûr, d'ailleurs, qu'il convienne à chaque

132Cette interprétation correspond plutôt à l'économie expérimentale, dans laquelle il s'agit effectivement de faire fonctionner des marchés simulés en laboratoire « à la manière » du marché concurrentiel. Cette approche en termes d'ingénierie est également défendue par Vernon Smith (2007). Sur le terrain, le chercheur le plus actif dans ce domaine est sans doute Alvin Roth, pionnier des *design economics*, qui a conçu notamment des dispositifs d'appariement pour l'allocation des stages d'étudiants en médecine, et pour le don d'organes.

fois de mettre ne place des sanctions pour favoriser la coopération et éliminer le *free riding*. Comme le souligne Zak, la plupart des règles définissant les sanctions sont informelles dans la vie réelle. La mise en place de nouvelles règles peut aussi avoir des effets contre-productifs sur la coopération (Zak, 2011, p.227)[133].

Dans le domaine du choix inter-personnel, la réflexion normative peut viser aussi, de façon plus large, non pas à promouvoir un type de comportement particulier -la coopération par exemple- mais un type de processus cognitifs, et notamment les processus délibératifs. Comme dans le cas du choix inter-temporel, l'approche dualiste conduit à favoriser le raisonnement et la délibération au détriment des émotions. Ce sont donc ces dernières qui seraient à l'origine d'erreurs dans la prise de décision. Pour Heukelom, les approches duales en économie comportementale seraient ainsi directement influencées par le partage normatif-descriptif établi par Kahneman[134], et reproduiraient, à l'intérieur même du décideur cette distinction entre ce qui doit être décidé et ce qui est décidé dans la plupart des cas[135]. Par exemple, dans le jeu de l'ultimatum, un joueur rationnel devrait, dans cette perspective, toujours accepter les offres inéquitables, dans la mesure où les rejets sont motivés par des processus affectifs irrationnels, qu'il est convient donc de contrôler.

133Zak s'appuie ici sur une étude réalisée par Gneezy et Rustichini en 2000, dans une crèche israélienne. Dans cet établissement, les parents sont tenus d'aller chercher leur enfant avant la fermeture de la crèche. Tout retard oblige le personnel à rester plus longtemps au travail. Gneezy et Rustichini observent le fonctionnement habituel de la crèche, dans lequel aucune sanction n'est prévue pour les retards, puis mettent en place un mécanisme de sanction financière. Les auteurs constatent que l'application de la règle a des effets contre-productifs, puisqu'elle augment le nombre de retards (Gneezy et Rustichini, 2000).

134Pour être tout à fait précis, Kahneman ne parle pas d'émotions et de raisons, mais plutôt d'intuitions et de raisonnement, qu'il définit ainsi:« *les opérations du premier système [intuitif] sont rapides, automatiques, associatives et souvent chargées d'émotions; elles sont gouvernées par l'habitude, et sont par conséquent difficile à ontrôler et à modifier. Les opérations du second système* [raisonnement] *sont plus lentes, nécessitent un effort et sont contrôlées par la délibération; elles sont également relativement flexibles et potentiellement dirigées par des règles* » (Kahneman, 2003, p.1451)

135Pour Heukelom, les modèles duaux représentent un prolongement direct de l'approche kahnemanienne: « *La littérature sur le choix inter-temporel et les systèmes duaux illustre l'incorporation, par l'économie comportementale, de la distinction normatif-descriptif* ». Les *behavioral economics* transforment cependant la distinction kahnemanienne, en faisant de la norme non plus un critère de rationalité externe à l'individu, mais en associant celle-ci à un type de processus cognitif: « *Les économistes comportementalistes en sont venus à considérer que le normatif non plus comme quelque chose d'externe à l'individu, mais comme un système rationnel fonctionnant en parallèle avec un système affectif, et avec lequel il lutte pour la domination. Comme Kahneman a contribué à cette nouvelle conception du normatif et du descriptif, cette évolution doit être comprise comme le reflet d'un développement dans la recherche comportementale au sens large, et non seulement dans l'économie comportementale* » (Heukelom, 2009, p.133). Il est intéressant de souligner que, selon Heukelom, la neuroéconomie se situe dans la même approche dualiste: « *certains économistes comportementalistes ont essayé de faire le lien entre cette approche en termes de système dual et la recherche en neuroscience et neurobiologie, ce qui a donné naissance à un nouveau sous-domaine appelé neuroéconomie. Cette littérature maintient la distinction normative-descriptive, mais la ré-interprète en supposant que ces deux éléments distincts représentent deux côtés du comportement humain. En d'autres termes, le normatif a changé de signification, en ne désignant plus quelque chose d'externe à l'individu, mais une des deux facultés propres à la nature humaine, qui luttent pour la domination* » (Heukelom, 2009, p.142). Dans notre point de vue, cette description n'est valable que pour la part de la neuroéconomie qui a été qualifiée d'économie comportementale dans le scanner. Une part importante des travaux s'émancipe de cette tutelle kahnemanienne.

L'exercice de facultés délibératives ne repose pas seulement sur l'inhibition des processus affectifs, mais s'appuie aussi sur des capacités d'empathie et de mentalisation (*cf.* chapitre 6). L'erreur, ici, consiste à ne pas exercer ces facultés. Elle se manifeste par divers comportements: transferts de montants faibles et/ou trahison de la confiance qui a été accordée dans le jeu de l'investisseur; propositions d'offres inéquitables dans le jeu de l'ultimatum, difficultés à agir pour le compte d'une autre personne (décision empathique), dons faibles ou nuls aux œuvres de charité, *etc.* Ce sont donc ici les processus empathiques et de mentalisation qui doivent être favorisés.

Dans le choix inter-temporel et inter-personnel, l'économie comportementale du bien-être repose sur l'identification d'erreurs dans les choix pris par les individus, pour prescrire des recommandations visant à rétablir une norme de comportement (respecter la cohérence temporelle, coopérer, agir de façon altruiste, contrôler ses pulsions, *etc.*). La conception des mécanismes effectifs par lesquels cette norme est susceptible d'être restaurée n'a pas fait l'objet ici de développements approfondis. La seule étude empirique visant à mesurer les effets de ce genre de régulation qui a été évoquée ici est celle de Taler et Benartzi (2004). Ce type de travaux relève en fait d'enquêtes de terrains (*field economics*) plus que de l'économie comportementale. L'économie comportementale du bien-être peut ainsi apparaître assez limitée sur le plan des recommandations concrètes. Le caractère peu opérationnel de ces réflexions est effectivement l'une des critiques importantes que l'on peut faire à ce type d'approche. Cependant, l'économie publique comportementale vise avant tout à fournir un fondement théorique à des dispositifs techniques de régulation du choix individuel. Nous essaierons ici de distinguer les critiques qui la concernent spécifiquement de celles qui portent plutôt sur ses applications, ou sur les disciplines appliquées qui se chargent de la traduire en actes. Or, la justification normative proposée par l'analyse comportementale, en dépit de l'apparente évidence selon laquelle les « erreurs » du choix inter-personnel et inter-temporel sont rejetées comme irrationnelles, soulève d'importantes difficultés théoriques.

B. Limites normatives

L'économie comportementale du bien-être a suscité un très vif débat chez les économistes à partir du début des années 2000. De nombreuses critiques ont été formulées quant à la faisabilité de ce projet. Un premier ensemble de critiques, dont nous ne traiterons pas ici, fait valoir que ce type de propositions, quand bien même celles-ci seraient fondées en droit, est simplement inapplicable dans les faits[136]. Notre analyse se limitera ici spécifiquement aux problèmes concernant la justification en droit de la régulation comportementale du bien-être, c'est-à-dire de son fondement normatif.

L'intervention est supposée rétablir une norme de comportement. Le problème est que cette norme est le plus souvent acceptée comme donnée. Or, la définition de ce que les individus doivent choisir ne va pas de soi et cache souvent des présupposés implicites, difficilement justifiables. Cela est particulièrement manifeste dans le cas du choix inter-personnel. Un premier ensemble de suggestions visent à favoriser les processus délibératifs au détriment des processus émotionnels. Or, la proposition selon laquelle les émotions devraient être contrôlées dans l'exercice de la cognition sociale débouche sur des recommandations contradictoires, dans la mesure où la notion d'émotion est ambivalente (*cf.* chapitre 6).

Les émotions peuvent effectivement être comprises d'abord comme des obstacles à la rationalité. Par exemple, dans le jeu de l'ultimatum, le répondant rationnel devrait dans cette perspective contrôler ses pulsions de rejet des partages inéquitables et accepter les offres de montant faible. Cette aptitude à la réflexion stratégique dans les choix inter-personnels, qui a été qualifié d'intelligence machiavélique, peut aussi se donner à voir dans le comportement du premier joueur, lorsque celui-ci, même égoïste, propose une offre équitable en anticipant une sanction possible de la part de son partenaire. Or, l'intelligence machiavélique ne saurait fournir une norme du comportement social, dans la mesure où elle suppose une inhibition des réactions émotionnelles, qui sont pourtant dans le cadre de la théorie de la punition altruiste, indispensables à l'application des normes de coopération. Si tous les individus apprenaient à contrôler leurs inclinations spontanées à punir les *free-riders*, il n'y aurait plus, selon Fehr, de

136En particulier, les décideurs publics, ou le gouvernement, ne possèdent pas nécessairement la connaissance nécessaire pour à la fois détecter les erreurs des agents, et les corriger (Rizzo et Whitman, 2009). Par ailleurs, l'intervention peut produire des effets contre-productifs si elle interfère avec les propres tentatives des individus de « débiaisement » (*debiasing*) (Kilck et Mitchell 2006; Whitman 2006). Surtout, le paternalisme libertarien repose sur la supposition que les décideurs ou gouvernant sont bien intentionnés. Or ceux-ci peuvent subir l'influence de lobbys, n'agir que pour leur propre intérêt (réélection), voire souffrir des biais identiques qu'ils sont supposés supprimer, ce qui remet en question leur capacité d'intervention (Glaeser, 2006),

coopération possible, en l'absence de toute possibilité de sanction.

Le problème, ici, est lié au fait que les processus délibératifs ne renvoient pas seulement à des mécanismes inhibiteurs des émotions, mais aussi à des capacités de mentalisation et d'empathie. Ceux-ci n'impliquent nullement une élimination des affects. Au contraire, un sujet empathique est capable de former des émotions sociales complexes. Chercher à promouvoir la délibération la réflexion dans le domaine des interactions sociales peut donc avoir deux significations bien différentes. Il peut s'agir d'abord, au sens strict, de contrôler ses émotions, afin de favoriser au maximum la réflexion stratégique sur l'issue des interactions, lorsque les deux types de processus -délibératifs et affectifs- sont effectivement en concurrence. Mais ce n'est pas toujours le cas et on peut très bien considérer par ailleurs que les émotions participent à la raison, et que l'expression de ces processus cognitifs de second niveau passe aussi par la manifestation d'un altruisme pur, dégagé de toute motivation lié à l'intérêt altruiste, comme c'est le cas dans les expériences sur le don aux œuvres de charité par exemple.

La promotion de la délibération ne saurait donc fournir un objectif en soi, parce qu'équivoque. Une solution consiste alors à choisir entre l'une des deux interprétations et à considérer, comme le fait Zak par exemple, que la coopération et la confiance sont par nature souhaitables (Zak, 2001). Un tel objectif semble difficile à admettre pour un économiste. Certes, dans le cadre d'un dilemme du prisonnier, il est évident que l'élimination des comportements de *free-riding* est souhaitable d'un point de vue agrégé, car elle élève le bien-être ou le gain total. Cependant, cela n'implique pas que les comportements coopératifs soient toujours rationnels, et que le *free riding* soit toujours une erreur. Ce qui est critiquable ici est l'affirmation selon laquelle la confiance constituerait une fin en soi, alors que, même d'un point de vue (neuro)-biologique, elle est plutôt un moyen de régulation du comportement, à la disposition d'autres fonctions organiques (comme la protection de la progéniture par exemple).

De la même façon, dans le cas du choix inter-temporel, les interprétations normatives du modèle d'actualisation quasi-hyperbolique partent toujours du présupposé selon lequel la stabilité du facteur d'actualisation, ou la prise en compte des intérêts à long-terme de l'agent, constitueraient des fins souhaitables par elles-mêmes. Cependant, rien, dans le modèle d'actualisation quasi-hyperbolique, ne permet de justifier de telles affirmations : pourquoi serait-il toujours rationnel de ne pas changer de plan? En outre, pourquoi faudrait-il favoriser les préférences à long-terme au détriment de celles à court-terme ?

Les propositions visant à « rectifier » la myopie temporelle des individus s'appuient ainsi sur deux présupposés distincts. Il est admis tout d'abord que la cohérence temporelle doit être un postulat minimal de rationalité, au moins sur le plan normatif. En économie comportementale, cette idée apparaît souvent sous la forme proposée par O'Donoghue et Rabin (1999), selon laquelle les individus dits « *sophistiqués* » seraient capables *ex ante*, de prévoir leur future incohérence et donc de souscrire à divers mécanismes d'engagement qui leur empêchent par la suite de modifier leurs plans initiaux. A l'inverse, les individus dits « *naïfs* » seraient incapables d'anticiper leurs futures défaillances (O'Donghue et Rabin, 1999). Il ne s'agit cependant nullement d'une justification : encore une fois, rien ne permet de considérer *a priori* que les individus sophistiqués sont nécessairement plus rationnels que les individus naïfs

O'Donoghue et Rabin considèrent donc comme allant de soi que l'incohérence temporelle soit une erreur qu'il convient de corriger. Plus généralement, en économie, la violation de la cohérence temporelle est effectivement considérée comme irrationalité manifeste, car des décideurs incohérents s'exposent à être « exploités » par des décideurs cohérents (voir par exemple Machina, 1989, p.1637-1638)[137]. Or, on l'a vu, les recherches sur la loi d'égalisation des rendements ont précisément montré que le renversement des préférences pouvait très bien être compris comme un comportement rationnel. En outre, quand bien même la cohérence temporelle *devrait* être respectée, il faudrait ensuite justifier la raison pour laquelle il serait toujours préférable, dans l'arbitrage inter-temporel, de favoriser l'utilité à long-terme au détriment de l'utilité à court-terme. Le présupposé, ici, est que les individus manifestent un biais excessif en faveur du présent, qu'il convient de rééquilibrer en imposant des facteurs (taux) d'actualisation plus (moins) élevés. Or, comme le soulignent Rizzo et Whitman, « *l'incohérence d'un individu qui actualise de manière hyperbolique peut être « rectifiée » en le forçant à actualiser de façon uniformément plus patiente -par exemple, en ayant toujours un facteur d'actualisation annuel égal à 0,9-, mais cette incohérence peut aussi être « rectifiée » en le forçant à actualiser de façon uniformément plus impatiente -en ayant toujours un facteur d'actualisation annuel égal à 0,8* » (Rizzo et Whitman, 2008, p.16).

137 Cet argument fait référence à ce qui est connu en théorie de la décision comme des cas de *Dutch book*, c'est-à-dire des paris qui assurent un gain certain quelle que soit l'issue du pari. De la même manière, un décideur cohérent peut réaliser un gain certain en échangeant avec un décideur dont les préférences se « renversent » au cours du temps. Supposons un tel individu A, avec une dotation initiale Y, et qui préfère, en t_0, X à Y, mais qui préfère ensuite, en t_1, Y à X. Par définition, il existe un montant minimal ε tel que (X – ε) soit préféré par ce même individu A à Y en t_0. Un décideur cohérent pourra donc facilement tirer parti des déficiences de A, en lui proposant d'échanger en t_0 (X – ε) contra sa dotation initiale Y. Il suffit ensuite d'attendre jusqu'en t_1, et de proposer à A d'échanger X contre sa dotation initiale Y. A n'ayant reçu dans l'échange précédent que (X – ε), il devra rendre plus que ce qu'il avait obtenu, et aura donc perdu au final ε. A l'inverse, le décideur cohérent aura « exploité » A et lui aura extorqué ε.

En considèrant que le choix inter-temporel représente un arbitrage entre différents « mois » temporels, rien ne permet de justifier une redistribution entre ces différents mois si cette allocation correspond à celle qui a été retenue par l'individu. Comme le reconnaissent eux-mêmes O'Donoghue et Rabin, il est seulement possible de plaider pour une distribution Pareto-optimale, c'est-à-dire que l'on peut éventuellement justifier une redistribution si elle augmente les utilités consommées à chaque période. (O'Donoghue et Rabin, 1999, p.105).

Quoique pouvant éventuellement être signalées par les auteurs concernés, ces difficultés sont généralement contournées en supposant, sans autres justifications, que les intérêts « *réels* » des individus sont ceux à long terme. O'Donghue et Rabin admettent par exemple que « *les comparaisons inter-temporelles de bien-être pour les individus ayant des préférences temporelles incohérentes sont par principe problématiques* », mais considèrent cependant que « *la perspective naturelle dans la plupart des situations est la perspective de long-terme, c'est-à-dire ce que vous souhaiteriez maintenant (si vous êtes parfaitement informé) pour votre comportement futur* » (O'Donoghue et Rabin, 1999, p.105). D'une manière similaire, Gruber et Köszegi « *prennent les préférences à long-terme des agents comme étant celles qui doivent être retenues pour la maximisation du* bien-être *social* » (Gruber et Koszegi, 2001, p.1297).

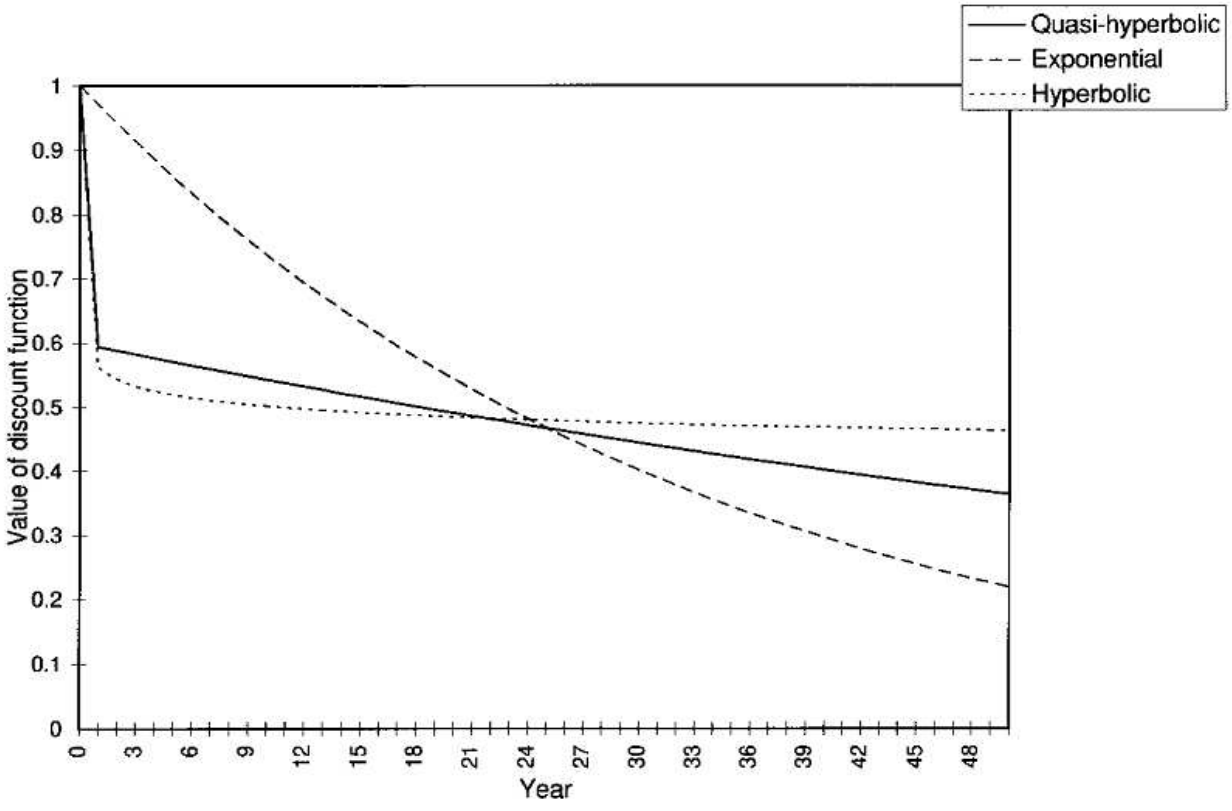
Dans ces propositions, le choix du long-terme au détriment du court-terme comme standard normatif est donc arbitraire. Un argument trompeur, qui apparaît fréquemment dans la littérature, consiste à affirmer qu'il s'agit là du standard normatif souhaité par l'agent lui-même (voir par exemple Gruber et Köszegi, 2001, p.1297). Or, l'actualisation hyperbolique implique précisément que les préférences de l'individu s'inversent au cours du temps: je peux préférer une salade à un hamburger dans certains environnements, mais pas dans d'autres circonstances. Comment définir alors mon standard normatif?

Par ailleurs, même s'il est admis que l'utilité à long-terme doit être privilégiée au détriment de l'utilité à court-terme, cela ne règle pas la question du « bon » facteur d'actualisation, qui doit être retenu comme critère normatif (*cf.* Rizzo et Whitman, 2008, p.18). Afin de mettre en place des politiques d'incitation à l'épargne, ou de lutte contre l'obésité, il est nécessaire de déterminer précisément le montant des taxes, des prélèvements, à partir d'un objectif quantifié de facteur d'actualisation. En d'autres termes, il ne suffit pas de dire que les individus doivent être plus patients, mais il faut spécifier le degré exact de patience (et d'impatience) qui doit être attendu dans les arbitrages inter-temporels.

C'est ici que le modèle d'actualisation quasi-hyperbolique offre un glissement normatif

tout à fait commode pour les défenseurs du paternalisme libertarien. Dans l'actualisation hyperbolique en effet, le taux d'actualisation n'est pas unique et constant comme dans l'actualisation exponentielle, mais diminue avec le délai. Il y a donc une infinité de taux d'actualisation chez un même individu, qui représentent, pour reprendre les termes de Rizzo et Whitman (2008, p.18), une « *continuité de possibilités normatives* ». Le modèle quasi-hyperbolique réduit ces possibilités à seulement deux options : le facteur d'actualisation applicable aux utilités immédiates égal à 1, et le facteur applicable aux utilités futures, égal au taux exponentiel diminué d'un paramètre *ad hoc* représentant le bais en faveur du présent (*cf.* section 2). Une telle représentation permet donc de considérer que le « bon » facteur d'actualisation est celui qui serait adopté si l'actualisation était exponentielle, sans fournir d'étalon précis pour ce paramètre.

Les économistes comportementalistes justifient généralement l'utilisation de fonctions quasi-hyperboliques par leur plus grande simplicité (Angeletos *et al.*, 2001). D'une manière similaire, Laibson (1997) considère que l'actualisation quasi-hyperbolique offre une approximation commode de l'actualisation hyperbolique. Dans son article de 1997, il fournit la représentation graphique suivante :



Laibson, 1997, p.450

Ici, la courbe d'actualisation exponentielle correspond à un taux d'actualisation $\delta = 0,97$; pour l'actualisation quasi-hyperbolique, $\beta = 0,6$ et $\delta = 0,97$; pour l'hyperbole généralisée[138], $\gamma = 10^5$ et $\delta = 5.10^3$. A l'appui de ce graphique, Laibson considère que la fonction hyperbolique représente donc une bonne approximation de l'hyperbole généralisée. Or cette simplification est trompeuse, car elle laisse à penser que l'actualisation *doit* être exponentielle. Une telle représentation n'est pas impliquée pas par les théories de l'actualisation hyperbolique issues de la psychologie évolutionniste.

La transformation de l'actualisation hyperbolique en quasi-hyperbolique permet donc de régler le problème lié à l'absence de facteur d'actualisation de référence, pouvant servir d'étalon normatif. La distinction entre actualisation hyperbolique et quasi-hyperbolique est souvent ignorée dans la littérature portant sur le choix inter-temporel pour deux raisons. Tout d'abord, les travaux de psychologie évolutionniste portant notamment sur le pigeon sont très souvent mobilisés, par les économistes comportementaux (voir par exemple Frederick, Loewenstein et O'Donghue, 2002; Chabris, Laibson et Schuldt, 2008), à titre de preuve empiriques supplémentaires de l'inadéquation du modèle d'actualisation exponentiel. L'actualisation quasi-hyperbolique est conçue comme une simple approximation des fonctions hyperboliques, choisie pour sa « *commodité* » (*cf.* Angeletos *et al.*, 2001).

Pourtant, les implications normatives ne sont clairement pas les mêmes. Mais ces distinctions passent souvent inaperçues parce qu'il est également possible de plaider, à l'appui des travaux de psychologie néo-comportementale, en faveur d'une régulation des choix individuels de type « environnemental », similaire aux propositions de l'économie comportementale du bien-être (*cf.* section III)[139].

De notre point de vue, les propositions de régulation fournies par la neuroéconomie et les approches néo-comportementalistes offrent cependant un fondement normatif bien plus solide que l'analyse comportementale du bien-être. Sa principale limite réside dans les arguments censés justifier la norme du comportement. Ses adversaires soulignent l'arbitraire des standards normatifs que ses défenseurs adoptent implicitement lorsqu'ils invoquent l'irrationalité ou les erreurs des individus (voir par exemple Mitchell, 2005). Dans le domaine

138Laibson (1997, p.450) propose une forme légèrement modifiée de l'hyperbole généralisée de Loewenstein et Prelec (1992), du type:

$$\lambda(t) = \frac{\gamma}{(1+\delta.t)}$$

139Si, en effet, dans la perspective de Ross *et al.* (2008) par exemple, l'actualisation hyperbolique est toujours « normale » au sens où elle peut toujours être comprise comme une réponse adaptée au milieu dans lequel l'individu interagit, cela ne signifie pas pour autant que les arbitrages inter-temporels appuyés sur ce principe soient toujours rationnels. Il y a en particulier des environnements propres à susciter des conduites irrationnelles (Ross *et al.,* 2008).

du choix inter-personnel, les émotions ne vont forcément aller à l'encontre de la rationalité individuelle -à moins de considérer que l'altruisme qui nait de l'empathie soit une forme d'irrationalité, ce qui ne correspond pas au point de vue des chercheurs concernés- et collective, dans la mesure où les processus affectifs sous-jacents à la punition altruiste permettent d'éliminer le *free-riding*. Pour le choix inter-temporel, l'absence de stabilité du facteur d'actualisation représente certes un écart par rapport au modèle de Samuelson, mais cet écart n'est pas pour autant forcément une erreur. Il peut souvent être justifié intuitivement : il n'apparaît pas illogique de préférer à la fois « une pomme aujourd'hui » à « deux pommes demain » et dans le même temps, « deux pommes dans un an et un jour » à « une pomme dans un an »[140]. Par ailleurs, le fait qu'un très grand nombre d'individus soient caractérisés par des fonctions d'actualisation quasi-hyperboliques suggère que cette « erreur » est extrêmement courante (*cf.* chapitre 2, section II).

Cette difficulté normative est sans doute liée, au fond, à une compréhension inadéquate des erreurs que les économistes comportementalistes prétendent mettre en évidence. Ces phénomènes sont caractérisés uniquement comme de écarts par rapport à une norme de référence. Cette approche ne fournit donc même pas, à proprement parler, de véritable description des choix réels des individus, dans la mesure où celle-ci supposerait une description des comportements irrationnels. Or, ces derniers sont présentés tautologiquement comme le produit de processus émotionnels qui sont eux même définis comme irrationnels. Ce qui manque donc est une véritable explication de l'irrationnel, ou en d'autres termes, une explication de ce qui fait que les individus se trompent. L'apport des neurosciences réside précisément dans une explication plus complète de cette logique de l'erreur.

140Le caractère « intuitif » et légitime, du point de vue de l'agent, de l'incohérence temporelle des décisions est parfois admis et a même fait l'objet de divers travaux expérimentaux (cf. chapitre 4). Certains économistes comportementalistes admettent donc que la constance du taux d'actualisation soit une règle qui ne doit pas toujours être suivie. Comme l'écrivent Frederick, Loewenstein et O'Donghue , « *la plupart des anomalies à l'actualisation* [exponentielle] *de l'utilité sont considérées comme des anomalies en référence à un modèle qui a été construit sans aucune attention portée à sa validité descriptive, et qui n'a aucun fondement normatif ou prescriptif* » (Frederick, Loewenstein et O'Donghue, 2002, p.21-22).

II. Une perspective intermédiaire: la collaboration entre le neurobiologiste Antonio Rangel et l'économiste Douglas Bernheim

Dans une série d'articles écrits en commun, l'économiste Douglas Bernheim et le neurobiologiste Antonio Rangel ont essayé d'approfondir les discussions récentes autour de l'économie publique comportementale à partir des avancées récentes des neurosciences (Bernheim et Rangel, 2004; 2005; 2008; 2009). Leurs premiers travaux (Bernheim et Rangel, 2004; 2005) portent plus spécifiquement sur la régulation des comportements addictifs. Ils y proposent une « *nouvelle théorie économique de l'addiction qui établit une passerelle entre les neurosciences et la politique publique* » (Berhniem et Rangel, 2005, p.11). Cette nouvelle compréhension de l'addiction à partir des neurosciences débouche sur un modèle pouvant s'appliquer plus généralement à tous les comportements susceptibles d'être régulés par l'économie publique comportementale (Bernheim et Rangel, 2009). La perspective développée par Rangel et Bernheim, à mi-chemin entre *behavioral economics* et neurosciences, peut se comprendre comme l'analogue, dans le domaine normatif, de l'économie comportementale dans le scanner. En effet, Bernheim et Rangel sollicitent l'appui des neurosciences afin de donner un fondement plus solide aux propositions de l'économie comportementale du bien-être (A), mais leur approche reste fidèle au principe kahnemanien d'un partage entre le descriptif et le normatif (B).

A. Un fondement neurobiologique à l'économie comportementale du bien-être

Bernheim et Rangel ont développé en collaboration un modèle des comportements addictifs (Bernheim et Rangel, 2004; 2005). Ce modèle peut s'appliquer par extension à tous les comportements dans lesquels l'impulsivité du décideur est susceptible de troubler la prise de décision, c'est-à-dire à l'ensemble du choix inter-temporel tel qu'il est étudié en économie comportementale. Bernheim et Rangel débouchent sur des propositions normatives relativement proches de celles qui ont été analysées dans la première section. Néanmoins, les

deux auteurs se distinguent en fournissant une justification neurobiologique au standard normatif retenu comme critère d'évaluation des comportements.

Bernheim et Rangel s'appuient en particulier sur les travaux, en neurobiologie, portant sur le système dopaminergique. Ils font référence notamment aux recherches de Wolfram Schultz (Bernheim et Rangel, 2004, p.1562). Pour Bernheim et Rangel, le circuit de la récompense fonctionne comme un « *mécanisme de prévision hédonique* » (*Hedonic Forecasting Mechanism*), qui a pour fonction de prédire la valeur des récompenses attendues. Ce système n'évalue pas la sensation hédonique à proprement parler, mais son anticipation; d'où une possible déconnexion entre ce que Bernheim et Rangel appellent le « *système des besoins* » et le « *système du désir* » ou de la motivation, représenté par l'activité du système dopaminergique, en référence aux travaux de Berridge (1996, 1999) (Bernheim et Rangel, 2004, p.1562-1563). En d'autres termes, un dysfonctionnement de ce mécanisme de précision hédonique motive l'organisme à poursuivre des objectifs qui n'ont plus de rapport avec ses besoins réels.

L'addiction illustre un tel phénomène. Pour Bernheim et Rangel, les comportements addictifs s'expliquent par une sensibilisation progressive à des stimuli associés à la substance addictive, et, qui, par la suite, lorsque l'individu est dépendant, provoquent une pulsion irrépressible de consommation: « *avec l'usage répétée d'une substance* [addictive]*, les stimuli associés à la consommation passée entraînent le mécanisme de prévision hédonique à surestimer le plaisir attendu, créant ainsi une forte et disproportionnée impulsion à consommer. Lorsque cela se produit, une part importante des processus de décision fonctionne en déformant l'information, ce qui conduit à des erreurs dans la prise de décision* » (Bernheim et Rangel, 2004, p.1559).

Cette analyse est proche de celle proposée par Ross *et al.* (2008), qui a été étudiée dans le chapitre 5, section IV. L'individu dépendant s'absorbe complètement dans un environnement addictif, composé d'un petit nombre de stimuli. Ces signaux, qui fonctionnent comme des prédicteurs stable de la surprise, entraînent une suractivité du système dopaminergique (*cf.* Ross *et al.*, 2008). Pour Bernheim et Rangel, la principale leçon de la recherche neuroscientifique sur les comportements addictifs est que ceux-ci ne sont pas liés à un plaisir ou une utilité trop importante, mais à une déformation du processus de choix (Bernheim et Rangel, 2005, p.19). Cette analyse vaut bien sûr pour les substances addictives, celles-ci ayant des effets directs sur le circuit de la récompense[141]. Cependant, dans la mesure

141Bernheim et Rangel proposent, en s'appuyant sur Gardner et David (1999) une liste de onze substances addictives, connues pour leur effet spécifique sur le circuit de la récompense: alcool, barbituriques, cocaïne, amphétamine, caféine, cannabis, hallucinogènes, nicotine, opioïdes, anesthésiants et dissolvants.

où elle met en avant une dimension environnementale, cette compréhension de l'addiction peut aussi s'appliquer plus généralement à tous les comportements compulsifs qui ne sont pas liés directement à l'une de ces substances addictives, comme l'addiction au shopping, au paris financiers, *etc.* Cette extension aux addictions dites comportementales, ne reposant pas sur l'usage d'une substance addictive (Bernheim et Rangel, 2004, p.1558), élargit donc la portée du modèle à l'impulsivité pris dans un sens très général. Celle-ci recoupe ainsi le domaine large des « anomalies » du choix inter-temporel étudiées par les *behavioral economics*.

Le modèle de Bernheim et Rangel se caractérise par deux résultats importants. Ces conclusions théoriques le distinguent de la littérature pré-existante et, pour Bernheim et Rangel, permettent de mieux rendre compte du phénomène de l'addiction, tel qu'il est observé en clinique. Tout d'abord, le modèle ne postule aucun effet cumulé de la consommation addictive. A l'inverse de la théorie de l'addiction rationnelle de Becker (Becker et Murphy, 1988), l'addiction n'est pas ici expliquée par des effets de complémentarité inter-temporelle, c'est-à-dire par le fait que la consommation présente soit d'autant plus plaisante que la consommation passée est importante (Bernheim et Rangel, p.1567-1568). Ces effets cumulés sont généralement invoqués pour expliquer le phénomène du manque et de la peur du manque, qui sont conçus comme les manifestations caractéristiques des conduites addictives. Or, pour Bernheim et Rangel, les observations cliniques montrent que le manque n'est pas un facteur décisif dans les troubles addictifs, et ce notamment pour la consommation de drogues dures[142].

En outre, le modèle de Bernheim et Rangel prédit une « *sophistication* » du comportement des individus dépendant, qui, encore une fois, est plus en accord avec la réalité clinique. Dans la perspective de Bernheim et Rangel, l'addiction s'explique par des stimuli qui déclenchent une envie irrépressible de consommer; mais cela n'empêche nullement les individus de comprendre que cette consommation est une erreur, qu'ils ne la souhaitent pas vraiment, et cela même au moment du choix. Un élément caractéristique de l'addiction est ainsi pour Bernheim et Rangel la volonté d'arrêter, qui s'illustre notamment par les nombreux mécanismes d'engagement auxquels les *addicts* ont recours. Ces derniers sont donc capables de sophistication au sens où, d'une part, ils comprennent que la pulsion addictive est une erreur, et, d'autre part, ils rationalisent cette pulsion en tentant de l'intégrer dans leux plans inter-temporels de consommation. Encore une fois, Bernheim et Rangel justifient ce principe

142Bernheim et Rangel s'appuient notamment sur l'étude clinique de McAuliffe (1982), qui montre que seulement 27,5% des individus dépendants dépendant à l'héroïne font l'expérience de ces symptômes de manque, et seulement 5% de ces mêmes individus envisagent la peur du manque comme un facteur explicatif de leur récidive.

en référence à l'observation clinique (Bernheim et Rangel, 2004, p.1571), qui suggère par exemple que les consommateurs d'héroïne sont constamment dans l'alternance de périodes d'abstinence et de rechute (Massing, 2000).

Pour Bernheim et Rangel, ce modèle offre une justification normative robuste pour défendre une régulation des choix individuels, ce qui permet de résoudre le problème central soulevé par l'économie publique comportementale. Leur article de 2009 s'ouvre ainsi sur le constat suivant: « *les modèles comportementaux sont utilisés de plus en plus dans l'évaluation des politiques publiques, ce qui soulève inévitablement des questions concernant le bien-être. Malheureusement, il n'y a pour le moment aucun consensus concernant les principes généraux qui sont censés gouverner de telles enquêtes normatives. Dans la plupart des cas, les économistes adoptent des critères ad hoc pour des modèles positifs particuliers, en proposant des justifications à partir d'intuitions vagues et sources d'inévitables controverses* » (Bernheim et Rangel, 2009, p.51).

Selon Bernheim et Rangel, les neurosciences fournissent un critère normatif plus solide -parce que fondé sur des considérations scientifiques- en montrant que les pulsions de consommation déclenchées par des stimuli addictifs (*cue-triggered process*) ne sont pas souhaitées par les individus. Bien qu'elle soit également critiquable, cette justification neurobiologique offre effectivement un principe non-ambigu pour prescrire des politiques de régulation des choix individuels. Dans leurs divers travaux, Bernheim et Rangel examinent ainsi à l'appui de leurs modèles les différents solutions possibles au problème de l'addiction, en envisageant leurs mérites et leurs inconvénients respectifs.

Le modèle permet tout d'abord de bien délimiter les comportements faisant l'objet de l'intervention. Dans cette analyse, les choix qui dégradent indiscutablement le bien-être sont ceux qui ne sont pas souhaités par l'individu. Ce dernier doit au moins regretter sa pulsion, et la volonté d'abstinence entre comme élément décisif. Par conséquent, Bernheim et Rangel excluent du champ d'intervention toutes les consommations -mêmes celles de substances addictives- qui ne font pas apparaître de projet d'abstinence future: « *le laissez-faire est […] la meilleure politique pour les consommateurs qui n'essayent pas sérieusement de s'abstenir (par exemple, les fumeurs ou les consommateurs de café qui se déclarent intégralement satisfaits de leur consommation* » (Bernheim et Rangel, 2004, p.1574). A l'inverse, une consommation même faible ou intermittente peut être considérée comme une erreur dès lors qu'elle est associée à un regret ou une volonté d'abstinence future.

En ce qui concerne les instruments de l'intervention, Bernheim et Rangel soulignent

l'efficacité limitée des taxes sur les substances addictives, qui sont notamment défendues par certains économistes comportementalistes (Gruber et Koszegi, 2001; O'Donoghue et Rabin, 2003). La simulation du modèle montre en effet que ces taxes peuvent avoir des effets positifs sur le bien-être dans certains cas bien précis, notamment pour des substances au prix peu élevé que les individus utilisent régulièrement (café, cigarettes par exemple). Mais en général, ces taxes conduisent surtout à pénaliser encore plus ceux qui ont la malchance de se trouver piégés par un environnement hautement addictif (Bernheim et Rangel, 2004, p.1577). Le modèle repose sur l'idée que les individus font face à des chocs aléatoire externes sur leur consommation, liés à la rencontre de stimuli addictifs, et suggère en fait plutôt de subventionner les individus qui ont le malheur de tomber dans la dépendance (Bernheim et Rangel, 2005, p.49-50). De façon similaire, pour Bernheim et Rangel, la criminalisation est critiquable, car elle aboutit à dégrader la situation fragile des individus dépendants[143].

Bernheim et Rangel plaident donc plutôt pour des politiques visant à compenser les conséquences négatives de l'addiction[144]. Le modèle prédit ainsi, pour les deux biens envisagés, qu'il est préférable de subventionner plutôt que de taxer (Bernheim et Rangel, 2005, p.49-50). Cette analyse cependant ne prend pas en compte les possibles externalités négatives liées à la consommation de drogues, et peut être inversée pour les rares substances addictives au prix faiblement prohibitif. Bernheim et Rangel considèrent plus généralement que les politiques ayant un effet seulement sur les prix, à la hausse comme à la baisse, ont au fond un impact limité parce que l'addiction est liée à des pulsions déclenchées par des stimuli (*cue-triggered process*), qui ne sont que faiblement influencées par le prix de la substance. Bernheim et Rangel défendent donc surtout des interventions visant à manipuler, précisément, cs processus quasi-automatiques déclenchés par des stimuli (Bernheim et Rangel, 2004, p.1580). Ce type de régulation renvoie donc à une action sur l'environnement de la décision, et s'appuie sur des propositions relativement proches de celles qui ont été envisagées dans la première section. Il s'agit en effet de diminuer l'influence néfaste des stimuli addictifs, en manipulant le mode de présentation des choix possibles[145].

143« *l'effet de la criminalisation sur le comportement est par conséquent potentiellement plus faible lorsque la consommation est irrationnelle, et plus forte lorsqu'elle est rationnelle -précisément les effets opposés d'une politique idéale. En outre, la criminalisation aboutit à exacerber les conséquences liées aux risques de l'addiction, contre lesquels il n'existe aucune assurance. Lorsque la société interdit une substance, ceux qui sont malheureusement suffisamment dépendants à cette substance subissent des couts disproportionnellement élevés. Des prix plus élevés peuvent les encourager à poursuivre des activités criminelles comme le vol et la prostitution, et la consommation de la substance par elle-même les définit comme criminels. Cela va exactement à l'inverse de principes d'une bonne assurance sociale des risques* » (Bernheim et Rangel, 2005, p.45)

144Ils préconisent ainsi toute une série de mesures d'aide et d'assistance en faveur des individus dépendants, comme par exemple la distribution gratuite de seringues (Bernheim et Rangel, 2005, p.49).

145Pour Bernheim et Rangel, il est ainsi souhaitable qu'un individu ayant développé une addiction dans un lieu

L'explication proposée par Bernheim et Rangel se distingue donc des modèles quasi-hyperboliques de l'addiction de l'économie comportementale (Gruber et Köszegi, 2001; O'Donoghue et Matthew Rabin, 2004) par la référence faite à des processus quasi-automatiques de consommation, déclenchés par des stimuli addictifs. Leur théorie offre ainsi un fondement normatif pour qualifier tout un ensemble de choix comme des erreurs, qui aboutissent à diminuer le bien-être l'individu. Cette justification, supposée être fondée par les résultats récents des neurosciences, s'appuie cependant sur une interprétation discutable des recherches sur le circuit de la récompense.

B. Émotions et processus réflexes – Un approfondissement insuffisant des concepts neurobiologiques

Quoique faisant référence aux recherches néo-comportementalistes sur le *reward learning*, le modèle de l'addiction de Bernheim et Rangel (2004, 2005), qui par extension, offre un cadre d'analyse pour l'économie publique comportementale (Bernheim et Rangel, 2008; 2009), reste fidèle à l'approche kahnemanienne de la décision. Bernheim et Rangel simplifient les résultats des neurosciences de manière à ce que ceux-ci respectent le partage normatif/descriptif proposé par Kahneman (2003). Bernheim et Rangel revendiquent explicitement l'adoption de cette distinction, en affirmant que « *le comportement et le bien-être doivent relever de modélisations différentes* » (Bernheim et Rangel , 2008, p.3).

Le partage normatif/descriptif est chez Bernheim et Rangel, comme plus généralement en économie comportementale, comme l'a souligné Heukelom[146], reproduit au niveau de

particulier -par exemple, pour une addiction aux drogues dures, le quartier dans lequel se déroule l'achat des substances- change radicalement son environnement quotidien (et déménage dans un autre quartier). Par ailleurs, l'intervention portant sur l'environnement peut viser à diminuer la visibilité des stimuli addictifs (pour le comportement alimentaire par exemple, interdire la publicité sur les produits gras et sucrés), ou renforcer la visibilité des stimuli aversifs -en rendant plus visibles par exemple les informations sur le caractère néfaste des substances addictives-, voire à en créer de toutes pièces (cas des visuels montrant des malades atteints de cancers du poumon sur les paquets de cigarettes).

146 *cf.* note 134 *supra*; « *certains économistes comportementalistes ont essayé de faire le lien entre cette approche en termes de système dual et la recherche en neuroscience et neurobiologie, ce qui a donné naissance à un nouveau sous-domaine appelé neuroéconomie. Cette littérature maintient la distinction normative-descriptive, mais la ré-interprète en supposant que ces deux éléments distincts représentent deux côtés du comportement humain. En d'autres termes, le normatif a changé de signification, en ne désignant plus quelque chose d'externe à l'individu, mais une des deux facultés propres à la nature humaine, qui luttent pour la domination* » (Heukelom, 2009, p.142)

l'individu lui-même. Bernheim et Rangel souscrivent en effet à une approche dualiste, dans laquelle deux modes de cognition - « *chauds* » et « *froids* »- sont distingués. Le mode chaud fait référence aux processus impulsifs déclenchés automatiquement par des stimuli, notamment addictifs. A l'inverse, la cognition froide est délibérée et rationnelle (Bernheim et Rangel, 2004, p.1559).

Or, l'interprétation dualiste des résultats de la neuroéconomie pose d'importants problèmes, liés d'abord à l'idée selon laquelle les processus cognitifs du premier niveau (chauds) seraient par nature irrationnels et s'apparenteraient à des quasi-réflexes pavloviens, c'est-à-dire à des réponses motrices (consommer la substance addictive) déclenchées par des stimuli sensoriels. La réflexion récente en neuroscience remet précisément en question le cadre théorique de la réfloxologie inspirée de Sherrington (*cf.* chapitre 3, section I), dans la mesure où le système nerveux ne se limite jamais à la simple transmission de commandes motrices, mais inclut toujours une dimension d'anticipation. Même dans les comportements les plus élémentaires, comme par exemple la salivation déclenchée par un stimulus associé à la nourriture, les expériences montrent que le système nerveux ne fonctionne pas comme un arc réflexe stimulus-réponse, mais évalue les différentes alternatives possibles. Contrairement à ce qu'affirment Bernheim et Rangel, les processus moteurs encodés par l'activité des neurones dopaminergiques n'est jamais déclenchée automatiquement par des stimuli, mais visent à estimer une valeur biologique perçue par l'organisme. Or, en réduisant ce signal de prédiction à un réflexe, Bernheim et Rangel nient la possibilité que ces réponses motrices -comme par exemple la consommation de drogue- puissent être contrôlées par l'individu, ou, plutôt ne peuvent expliquer ce qui fait qu'un individu, dans certaines circonstances, puisse être capables de surmonter ces *cue triggered process*.

Bernheim et Rangel sont donc contraints de considérer que, dans le cas des conduites addictives, cette prédiction est par nature biaisée, et que le système dopaminergique connait un dysfonctionnement. Or cela pose problème car le système dopaminergique ne connait pas à proprement parler de dysfonctionnements dans les conduites addictives. La théorie neuroéconomique de l'addiction proposée par Ross *et al.* (2008) suggère que, même dans les addictions les plus sévères, il n'y a jamais de dysfonctionnements du circuit de la récompense; celui-ci ne fait qu'accomplir sa fonction normale, qui est de prédire la satisfaction attendue, la cause du trouble étant que le processus de prise de décision devienne plaisant en lui-même (*cf.* chapitre 5, section IV).

La distinction entre mode froid et chaud, empruntée à Loewenstein (1996, 1999)[147], est donc critiquable du point de vue des résultats des neurosciences, sur lesquels Bernheim et Rangel prétendent paradoxalement s'appuyer. Or c'est précisément à partir de cette distinction que Bernheim et Rangel prétendent fournir un fondement normatif à l'économie comportementale du bien-être. En effet, ce sont les décisions prises dans le mode froid qui reflètent les véritables préférences individus, et tous les choix contradictoires pris dans le mode chaud représentent des erreurs. Le critère avancé revient, en pratique, à défendre un principe de cohérence des choix, en prenant comme référent normatif les choix pris à l'avance, lors de la planification *ex ante* des décisions de consommation. Pour Bernheim et Rangel, la régulation comportementale du bien-être dispose donc de données empiriques pour à la fois concevoir et mesurer les effets de ses interventions, en s'appuyant notamment l'écart entre les intentions déclarées des agents et leurs actions de consommation réelles : « *nous pouvons exploiter les données sur les relations entre les intentions et les actions suivies. En connaissant le taux de récidive à court terme (par exemple, à une semaine) parmi les individus qui essayent d'arrêter de fumer, on peut en déduire la fréquences des erreurs liées aux réponses réflexes, et fournir ainsi une justification pour faire des inférences raisonnables* [sur les politiques à suivre] » (Bernheim et Rangel, 2005, p.43).

Il serait donc possible d'inférer les préférences réelles des agents à partir des engagements futurs que ceux-ci sont disposés à prendre, et par suite, d'en déduire les couts implicites des consommations non-intentionnelles, c'est-à-dire non prévues dans le plan initial en termes de bien-être, (Bernheim et Rangel, 2004, p.1567). Pour Bernheim et Rangel, cette approche reste donc conforme au principe des préférences révélées (Bernheim et Rangel, 2009, p.53): tous les choix qui sont cohérents avec le plan initial révèlent les véritables préférences des individus. Seules les décisions prises de manière impulsives, sans respecter l'intention initiale, doivent être considérées comme non-pertinentes pour l'analyse du bien-être.

Cette défense du principe des préférences révélées conduit Bernheim et Rangel à rejeter les approches en termes de soi multiples, dans lesquelles l'individu est scindé en une série de sous-individus situés à chaque moment du temps, disposant chacun des ses propres

147A la différence de Loewenstein cependant, Bernheim et Rangel considèrent que les individus dépendants sont capables de comprendre l'irrationalité de leurs pulsions (Bernheim et Rangel, 2004, p.1559) . En d'autres termes, dans le mode froid, l'individu est tout à fait capable de comprendre que certains décision prises dans le mode chaud soient des erreurs. Le dualisme défendu par Bernheim et Rangel est donc moins strict que celui de Loewenstein, puisque les individus prenant des décisions irrationnelles peuvent néanmoins comprendre qu'il s'agit d'erreurs. Il n'en reste pas moins que Bernheim et Rangel, comme Loewenstein, envisagent les choix pris dans le mode chaud comme des pulsions irrationnelles.

préférences. Pour Bernheim et Rangel, cette perspective est inadéquate car chaque individu dispose d'un seul et unique ensemble de préférences. C'est la condition pour laquelle, précisément, il est possible de considérer que certains choix violent ces préférences. A l'inverse, dans les modèles d'actualisation quasi-hyperboliques interprétés en termes de soi multiple (Laibson, 1997; O'Donoghue et Rabin, 1999), rien en justifie pourquoi certains « mois » devraient être favorisés plutôt que d'autres. Le critère d'optimalité parétienne est dans ce cadre insuffisant (*cf.* chapitre 5, section II), et les auteurs recourent le plus souvent à un critère normatif *ad hoc*, en considérant que le bien-être est convenablement décrit par le facteur de *discount temporel* de long-terme (Bernheim et Rangel, 2009, p.69)

Le raisonnement en termes de soi multiple est donc fallacieux selon Bernheim et Rangel, car il implique que les individus ne commettent jamais d'erreurs à proprement parler, puisqu'ils ne font que contenter l'un de leur moi temporel. Or, pour Bernheim et Rangel, la possibilité de l'erreur est « *un prémisse central de* [leur] *analyse, qui est justifié par l'état actuel des connaissances concernant la neurobiologie de l'addiction* » (Bernheim et Rangel, 2009, p.69). Cette affirmation est critiquable, en tant qu'elle repose sur une simplification de la neurobiologie de l'addition. En outre, quand bien même les résultats des neurosciences favoriseraient cette interprétation dualiste, le critère normatif proposé pour différencier les intentions véritables des erreurs est, comme le reconnaissent eux même Bernheim et Rangel, beaucoup moins tranché qu'il n'y paraît.

Pour Bernheim et Rangel, le principe des préférences révélées est plus ou moins tenable selon que les choix soient pris en accord ou non avec les intentions initiales (*cf. supra*). C'est donc essentiellement la cohérence, ou l'absence de volonté d'abstinence qui définit un choix comme pertinent pour révéler les préférences de l'individu. Mais, dans leur analyse plus générale de l'économie comportementale du bien-être, Bernheim et Rangel suggèrent que l'applicabilité de ce critère est plus complexe qu'il n'y paraît en première analyse. Certains choix en apparence incohérents entre eux sont en réalité cohérents si l'on distingue les « *conditions auxiliaires du choix* ». Ces dernières sont définies comme des « *caractéristiques de l'environnement du choix qui peuvent affecter le comportement* » (Bernheim et Rangel, 2008, p.85). Par exemple, un individu peut préférer A à B, et aussi préférer B à A, sans que cela soit incohérent, si la première décision a été pris dans une condition C_1, distincte de la condition C_2 dans laquelle a été prise la seconde décision.

Cette notion de condition auxiliaire permet donc de légitimer, d'un point de vue normatif, certains choix qui peuvent apparaitre comme des erreurs. Par exemple, dans le cas du choix inter-temporel, si l'on considère que le moment du choix constitue une condition

auxiliaire, alors l'individu n'est contraint à aucun critère de cohérence temporelle dans ses choix. Il peut renverser à chaque fois ses préférences, sans que cela constitue une erreur, puisqu'à chaque condition auxiliaire, donc à chaque moment du temps, doit correspondre un optimum local (Bernheim et Rangel, 2009, p.83).

Évidemment, une telle approche conduit à démultiplier les préférences des individus, et l'enjeu consiste à délimiter un sous-ensemble de conditions auxiliaires qui sont pertinentes pour révéler le bien-être réel de l'agent. Le critère pour exclure certains choix peut être selon Bernheim et Rangel l'existence d'un mauvais traitement de l'information: par exemple, si les alternatives peuvent être présentées oralement ou écrites, il va de soi que, pour un individu sourd, le choix pertinent est celui qui est pris lorsque le problème est lu par le sujet lui-même. Dans le cas de l'addiction, les choix déclenchés automatiquement par des stimuli addictifs peuvent aussi être exclus **(**Bernheim et Rangel, 2009, p.84). Ici, des données relatives au contexte du choix (données neuronales, mode de présentation des alternatives), et non au choix au lui-même permettent de limiter le nombre de situations pertinentes du point du bien-être de l'individu.

Cette approche suppose donc l'adoption d'un critère de définition et d'exclusion des choix qui doivent respectivement être pris en compte ou exclus de l'analyse normative. L'exemple proposé par Bernheim et Rangel **(**Bernheim et Rangel, 2009, p.84), ainsi que leur théorie neuronale de l'addiction (Bernheim et Rangel, 2004) suggèrent que, dans le cas de l'addiction, tous les choix liés à des processus automatiques déclenchés par des stimuli addictifs doivent être rejetés comme non-pertinents. Cependant, dans leur article plus général portant sur l'analyse comportementale du bien-être et ses fondements normatifs, Bernheim et Rangel ne défendent l'utilisation d'aucun critère. Le cas des comportements addictifs est mobilisé à titre de pure illustration. Bernheim et Rangel présentent leur approche comme purement technique, et considèrent que celle-ci peut donner lieu à des résultats variés selon le critère normatif adopté:« *des analyses différentes aboutiront à différentes distinctions entre les caractéristiques de l'objet et les caractéristiques auxiliaires* […]. *Nous soulignons que les outils développés dans cet article fournissent une méthode cohérente pour réaliser une analyse du bien-être à partir des choix indépendamment de la manière avec laquelle on décide d'établir cette distinction* » **(**Bernheim et Rangel, 2009, p.85).

Les travaux de Bernheim et Rangel manifestent donc une tension entre la volonté de fournir un fondement normatif solide à l'analyse comportementale du bien-être à partir des neurosciences, et la réticence, caractéristique des économistes (comportementaliste) à proposer des critères d'évaluation normative des choix pris par les individus. Cette ambiguïté

est sans doute liée au fait que Bernheim et Rangel prennent la mesure des difficultés soulevés par l'économie comportementale du bien-être. Ces derniers comprennent que l'identification des situations de choix pertinentes du point de vue du bien-être, c'est-à-dire des situations dans lesquelles la théorie des préférences révélées doit être considérée comme valide, est associée à des risques de dérive, de « *pente glissante* », pou reprendre les termes de Rizzo et Whitman (2008). Bernheim et Rangel sont conscients des dangers liés à un abandon, même partiel, de la théorie des préférences révélées, au nom de principes paternalistes: « *le cas de l'orientation sexuelle est particulièrement instructif. Jusqu'à relativement récemment, la communauté clinique classait l'homosexualité dans la catégorie des troubles psychiatriques. Ce jugement était appuyé essentiellement sur un écart par rapport à une norme perçue, plutôt que sur des preuves scientifiques solides […]. Le traitement historique des homosexuels dans les sociétés « libres » illustre les dangers liés à la remise en question du principe des préférences révélées* » **(**Bernheim et Rangel, 2005, p.41).

Pour Bernheim et Rangel, il faut, afin d'éviter de telles dérives, accepter le paternalisme uniquement sur la base de preuves scientifiques solides et indiscutables: « *étant donnés les risques importants de dérives qui sont impliqués, si nous souhaitons relâcher le principe des préférences révélées pour évaluer des politiques publiques, il convient de fixer un standard scientifique élevé pour montrer, en s'appuyant sur des preuves objectives, que les préférences et les choix divergent systématiquement dans un contexte donné* » **(**Bernheim et Rangel, 2005, p.41). Évidemment, cet optimisme scientifique est illusoire, puisque, précisément, Bernheim et Rangel proposent un critère d'exclusion de certains choix sur la base d'une interprétation des neurosciences assez contestable. Par ailleurs, il est possible de faire valoir que le classement de l'homosexualité dans la catégorie des troubles psychiatriques revendiquait lui aussi l'autorité de la science.

L'approche proposée par Bernheim et Rangel rencontre donc deux principales limites. Tout d'abord, celle-ci s'appuie sur une lecture assez contestable des neurosciences, selon laquelle l'addiction serait liée à des processus analogues à des réflexes, qui ne seraient pas souhaités par les individus dépendants. En second lieu, Bernheim et Rangel n'établissent pas de façon définitive, selon nous, le lien entre médecine et analyse normative du bien-être en économie. Bernheim et Rangel mobilisent bien, comme cela a été envisagé dans cette section, des observations cliniques à l'appui de leurs analyses. Ils semblent donc saisir à de nombreux endroits l'intuition centrale de la psychiatrie économique, selon laquelle la théorie des préférences révélées ne peut être maintenue pour les cas pathologiques, par exemple lorsqu'ils écrivent que « *dans les cas extrêmes, presque toutes les situations de choix d'un individu*

peuvent devenir suspectes, et le choix devient alors une justification insuffisante à l'analyse du bien-être. De tels exemples sont fournis par les personnes souffrant de la maladie d'Alzheimer, ou d'autres formes de démence, ou de lésions dans le cerveau » (Bernheim et Rangel, 2009, p.86). Cependant, les critères normatifs qui peuvent être fournis par les classifications cliniques demeurent selon Bernheim et Rangel extérieurs à l'économie à proprement parler, qui doit, autant que possible, respecter le principe d'une stricte séparation entre le normatif et le descriptif.

III. Neuroéconomie et psychiatrie économique

Les propositions de régulation comportementale du bien-être qui ont été abordées jusqu'ici manquent de fondement normatif solide, permettant, pour certains choix, de défendre un relâchement de l'approche en termes de préférences révélées. La neuroéconomie offre selon nous une justification normative plus solide à ce paternalisme comportemental, en faisant référence à des critères de classification cliniques de choix et de comportements pathologiques.

Dans le cas du choix inter-personnel, les liens entre neuroéconomie et médecine sont assez clairs. Les jugements appréciatifs sur la rationalité des individus ne prennent sens, comme cela a été envisagé dans la première section, que dans une visée de détection des troubles de l'intelligence sociale. En effet, la théorie économique n'offre pas dans le domaine des interactions sociales de norme indiscutable du choix rationnel; l'altruisme, la coopération ou l'égoïsme peuvent être ou non envisagés comme des comportements optimaux selon les objectifs fixés (élever le bien-être global, maximiser le gain individuel), selon que l'on suppose par exemple, que l'altruisme soit en lui-même source d'utilité, *etc.* Par conséquent, même si Zak essaye de justifier « économiquement » sa démarche en prenant appui sur une corrélation entre confiance inter-personnelle et croissance économique, la rationalité des décideurs ne saurait ici être appréciée de façon aussi tranchée en considération seulement d'arguments économiques. C'est d'abord parce que les sujets ayant des difficultés dans l'exercice de la cognition sociale se distinguent par leurs choix dans les expériences que le comportement des sujets sains ou normaux peut être définis comme rationnel. Pour le choix inter-personnel, l'abandon de la théorie économique comme référent normatif est donc évident. Si ce projet visant à diagnostiquer des troubles de l'intelligence sociale, liés en particulier à la psychopathie, en neuroéconomie reste encore à développer, il est manifeste que les jeux économiques sont dans ce domaine au service d'une démarche médicale ou clinique.

Il convient de souligner cependant que cet ancrage médical reste implicite. Il apparaît par exemple à travers l'équivalence introduite entre la rationalité des décideurs et leur normalité. Mais la description de la neuroéconomie comme psychiatrie économique ne correspond pas nécessairement à la vision que les neuroéconomistes ont de leur propre travail. Kevin McCabe reconnaît ainsi un « *intérêt pour les pathologies* » en neuroéconomie, qui

« *s'explique parce que les neuroscientifiques sont payés pour étudier ces troubles* »[148]. Cependant, une telle explication suggère que cet intérêt médical n'est pas forcément partagé par les économistes, et que, par conséquent, la neuroéconomie n'est pas liée en principe, comme nous le défendons, à des préoccupations relatives à la médecine mentale. Par ailleurs, l'importance sous-jacente des critères de classification cliniques est beaucoup moins claire dans le cas du choix inter-temporel et de l'addiction que pour le choix inter-personnel. Si un chercheur comme Georges Ainslie admet volontiers l'influence et la dépendance à l'égard de la clinique, ce n'est généralement pas le cas des neuroéconomistes qui entendent développer une approche de neuropsychiatrie computationnelle, sur la proposition de Camerer, Rangel et Montague (2008). Un chercheur comme Don Ross affirme précisément la possibilité de construire une théorie purement quantitative de l'addiction, en indépendance complète à l'égard des critères cliniques plus anciens. Les travaux de cet auteur permettent de comprendre en quoi la neuroéconomie de l'addiction témoigne bien de l'apparition de considérations médicales en économie.

Nous rappelons que dans le cadre de la théorie de Ross *et al.* (2008), l'addiction peut se caractériser à partir de trois types de critères (*cf.* chapitre 5, section III):

i. Sur le plan neuronal, l'addiction se manifeste par une cascade d'erreurs de prédiction positives. Cette activité intense du système dopaminergique tend à inhiber les régions préfrontales qui permettent normalement de contrôler les réponses dopaminergiques. Ce phénomène neuronal est clairement et précisément mesurable.

ii. Au niveau comportemental, l'addiction s'explique par l'interaction avec un « environnement addictif », dans lequel la perspective d'une récompense incertaine est associée de manière certaine avec un type de stimuli-prédicteurs.

iii. Sur le plan psychologique, les comportements addictifs sont source de désagréments et de souffrances de deux types. D'une part, la conduite devient obsessionnelle, et les individus manifestent les plus grandes difficultés à se concentrer que d'autres signaux prédicteurs. D'autre part, des phénomènes de manque apparaissent lorsque l'obtention régulière de surprise s'interrompt, en particulier lorsque l'individu n'a plus accès à l'environnement addictif.

La nouveauté de cette théorie consiste à mettre en évidence des marqueurs neuronaux du comportement addictifs (la cascade d'erreurs de prédiction positives). Or, ce phénomène neuronal n'est pas toujours lié à une conduite addictive. Comme l'a reconnu Ross dans un article postérieur à son ouvrage de 2008, *« l'explication neurobiologique de la manière avec laquelle le circuit de la récompense devient absorbé par un réseau dominant de signaux attentionnel convergents sur une ou deux cibles (par exemple, l'alcool, la cocaïne, le pari*

148Correspondance personnelle avec l'auteur (21 mars 2012).

financier, le tabac, ou deux ou, plus rarement, trois de ces cibles), et gagne progressivement une influence grandissante sur les réponses motrices […] permet d'identifier les conditions nécessaires mais non suffisantes de la structure caractéristique des addictions » (Ross, 2011, p.39). En effet, une forte activité dopaminergique peut simplement impliquer que l'individu est fortement absorbé dans la tâche qu'il accomplit, sans que cela soit nécessairement néfaste ou pathologique.

Pour que la cascade d'erreurs de prédiction positives soit représentative d'un comportement addictif, il est donc nécessaire que l'individu fasse l'expérience de désagréments et d'inconforts, liés au manque et au caractère obsessionnel de sa conduite, et que le milieu dans lequel ce même évolue présente les caractéristiques d'un environnement addictif. Dans la mesure où la souffrance et le désagrément ne peuvent faire l'objet que d'une introspection de la part du sujet, le seul critère objectif qui reste est celui d'environnement addictif. Toute la difficulté consiste alors à définir les éléments distinctifs d'un tel environnement. Pour Ross *et al.* (2008), les environnements potentiellement addictifs sont des environnements dans lesquels la perspective des récompenses est « certainement incertaine », ce qui signifie que l'individu est en mesure d'identifier un ensemble restreint de stimuli associés à chaque fois à l'occurrence aléatoire d'une récompense (*cf.* chapitre 5, section IV; Ross *et al.*, 2008, p.213). Le risque devient alors source de plaisir en lui-même, et cette recherche de danger ou de perte (*loss chasing*) peut se comprendre comme un critère de définition comportemental de l'addiction.

Selon Ross *et al.*, cette condition comportementale impose une contrainte suffisante sur la définition neuroéconomique de l'addiction, et permet de d'exclure plusieurs comportements du champ de l'addiction environnementale. Toute conduite impulsive n'est pas nécessairement addictive au sens neuroéconomique. Par exemple, Ross *et al.* considèrent que le comportement sexuel ne peut faire l'objet d'une addiction, parce que la perspective de la récompense est soit trop incertaine (dans le cas de la séduction), soit trop certaine (dans le cas de la masturbation) (Ross *et al.*, 2008, p.213). La recherche du danger ou *loss chasing* doit précisément s'appuyer, comme dans le cas du jeu financier sur un aléa, mais qui doit être fortement codifié, afin que l'individu puisse y associer des stimuli et un ensemble de rituels: la surprise doit être « organisée ».

Cependant, ce critère comportemental du *loss chasing* est moins évident qu'il ne peut paraître à première vue. Le cas des troubles de la conduite alimentaire, qui est considéré comme un exemple classique d'addiction environnementale (*cf.* Rowland *et al.,* 2008) fournit une bonne illustration. En quoi le comportement alimentaire pourrait-il donner lieu à un

comportement de recherche du danger? Si un individu a la possibilité de consommer un hamburger gratuitement comme c'est le cas dans certaines expériences (voir par exemple Hare *et al.*, 2009), la perspective de la récompense apparaît certaine, immédiate, et sans risque. Le comportement alimentaire est peut être potentiellement addictif lorsque la nourriture est *perçue* par l'individu comme impliquant à la fois une gratification à court-terme et une perte à long-terme. Par conséquent, une personne obèse qui suit un régime peut être subjectivement récompensée simplement en décidant de prendre le hamburger (en addition du plaisir lié à sa consommation), parce que cette décision implique un danger potentiel. Dans cette perspective, l'incertitude porte sur la nature de la récompense représentée par la nourriture elle-même. Un anorexique peut de la même façon percevoir la nourriture comme fortement appétissante et ressentir en même temps du dégout, ce qui peut expliquer pourquoi les personnes souffrant de troubles du comportement alimentaire passent un temps disproportionné à choisir leur plat au restaurant ou à faire la cuisine. Ces individus deviennent totalement absorbés par les stimuli associés à la consommation alimentaire parce que celle-ci est perçue comme problématique, c'est-à-dire comme impliquant à la fois une récompense mais aussi des pertes possibles.

Cette hypothèse pourrait être généralisée à toutes les addictions environnementales. C'est parce que la consommation addictive acquière chez l'individu un statut subjectif ambivalent que la perspective de la récompense devient incertaine. L'explication que nous avançons n'apparait cependant pas chez Ross *et al.* (2008). Le but ici n'est pas d'approfondir le portrait psychologique de l'anorexique ou du boulimique, mais de montrer que le critère comportemental du *loss chasing* peut faire l'objet d'interprétation assez variables, selon la nature de l'aléa ou du risque recherché dans le comportement considéré. Pour qu'il y ait recherche de danger ou de perte au sens de Ross *et al.*, il est nécessaire dans tous les cas de s'appuyer sur des représentations subjectives de l'individu lui-même, et donc de mobiliser des preuves fournies par l'introspection. La notion d'environnement addictif est au bout du compte une construction purement psychologique, et la catégorie des addictions dites environnementales peut faire (et fait actuellement) l'objet d'une extension quasi-indéfinie (*cf.* chapitre 5, section IV). Par exemple, contrairement à ce qu'affirment Ross et ses co-auteurs (Ross *et al.*, 2008, p.213), il n'est pas si évident que les achats compulsifs ne puissent être envisagés comme des addictions comportementales. Après tout, les acheteurs compulsifs savent sans doute que leurs achats sont liés à des pertes (financières), et, comme pour les troubles du comportement alimentaire, la consommation addictive est perçue comme ambivalente. Ici comme pour les autres addictions environnementales, il est parfaitement

possible d'envisager que le « plaisir est dans le jeu »: ce n'est pas principalement l'achat de chaussures par exemple qui est source d'utilité, mais le processus d'achat en lui-même, avec ses rituels associés.

Encore une fois, la pertinence de ces explications pourrait être discutée, mais l'objectif ici est de souligner que le critère comportemental de définition proposé par Ross *et al.* n'est pas suffisant pour distinguer clairement la simple impulsivité de l'addiction. Il ne saurait être acceptable, en la matière, de se référer à l'usage commun du terme d'addiction, car, comme le soulignent Ross et ses co-auteurs, « *l'addiction au sens populaire du terme est un syndrome familier, culturel que certaines personnes adoptent comme une norme […] Cette norme peut en principe s'attacher à n'importe quel type de consommation* » (Ross *et al.*, 2008, p.161). Or si n'importe quel type de comportement peut être considéré comme une addiction, et si celle-ci doit être traitée, les risques de dérive paternalistes sont manifestes. Ross et ses co-auteurs sont conscients de ces risques de dérive et affirment:

> « *Nous ne souhaitons pas que les gens adoptent des préférences à l'égard des autres individus, en leur interdisant de tomber dans la dépendance à l'alcool, à l'héroïne, ou à la nicotine. Nous pouvons cependant admettre que ce type de préférences « autoritaires » trouvent une justification dans un fait objectif: les substances addictives en question ont des effets nocifs chez les individus qui en sont dépendants, y compris en référence aux standards normatifs des individus dépendants eux-mêmes. A l'inverse, il n'y a aucun consensus social, et aucune perspective sérieuse de pourvoir, un jour, en établir un, sur la quantité de shopping qui est excessive, ou sur le nombre d'heures qu'une personne doit passer chaque jour sur internet, ou à jouer aux jeux vidéos, ou sur le montant d'effort, de temps ou de cout d'opportunité émotionnel qu'un individu devrait consacrer à diverses formes de stimulation et de gratification sexuelle* » (Ross *et al.*, 2008, p.212-213).

Deux arguments sont avancés ici. Le premier est qu'il existe un « *fait objectif* », qui concerne effets « *nocifs* » de l'addiction « *en référence aux standards normatifs des individus dépendants eux-mêmes* ». Il convient de souligner qu'il ne s'agit pas ici du même argument que celui qui est proposé par Gruber et Köszegi (2001, cf. section I), selon lequel les « *standards normatifs* » de l'individu devraient spontanément favoriser les choix en faveur du long-terme. Pour Ross *et al*., les standards normatifs en question sont impersonnels et ne portent pas sur ce que l'individu désire accomplir à long-terme, mais plutôt sur la dangerosité « biologique » du comportement en question. Par exemple, une préférence pour la nage au milieu des crocodiles peut être dite « *mauvaise* » du point de vue des standards normatifs de l'individu ayant cette préférence, parce que ce comportement menace directement la vie de l'individu (Ross *et al.*, 2008, 212-213). En ce qui concernent les addictions, le conflit interne

qui est vécu au niveau psychologique est à l'origine de souffrance et de désagréments. C'est la raison pour laquelle les addictions peuvent être jugées nominativement comme « *mauvaises* »: « *en ce qui concerne la théorie du consommateur, la consommation addictive ne représente pas une différence notable à l'exception de l'intensité du conflit interne vécu par l'individu dépendant* »(Ross, 2011, p.26).

En second lieu, cette souffrance indiscutable est selon Ross *et al.* à l'origine d'une reconnaissance sociale. L'addiction doit donc faire l'objet d'un « *consensus social* », ou il doit au moins être possible d'en établir un. L'achat compulsif par exemple ne peut être classé dans le champs des addiction environnementales parce qu'« *il n'y a aucun consensus social, et aucune perspective sérieuse de pourvoir, un jour, en établir un, sur la quantité de shopping qui est excessive* ». Cependant, l'expérience psychosomatique de la souffrance est pour Ross et ses co-auteurs logiquement antérieure à sa reconnaissance sociale. Ce n'est pas à partir d'un consensus social qu'un type de comportement impulsif peut devenir potentiellement addictif; les individus souffrant d'addictions font l'expérience d'un conflit interne et sont par conséquent considérés comme tels par la société: « *tous les addicts considèrent leur état interne comme déplaisant, et par conséquent leurs amis, leurs proches et l'État peuvent à bon droit prendre part à cette négociation interne* » (Ross, 2011, p.36).

Pour Ross *et al.*, la preuve déterminante de l'addiction réside dans le pur fait psychologique de la souffrance ressentie par les individus dépendants. Encore une fois, cette expérience subjective ne peut recevoir de valeur scientifique et objective, comme le prétendent Ross *et al.,* que parce qu'elle a été reconnue comme telle en clinique. Comme elle ne constitue qu'une condition nécessaire mais non suffisante de l'addiction *(cf. supra*), la preuve neuronale -l'existence d'une cascade d'erreurs de précisions positives- vient simplement confirmer dans les expériences les critères cliniques utilisés en amont pour distinguer sujets sains et pathologiques. Par conséquent, il n'existe pas pour l'instant de preuves neuronales de l'addiction aux achats pour la simple raison qu'il n'y a pas encore eu d'études sur le sujet. Pourtant, certains acheteurs compulsifs évoquent des souffrances similaires à celles des accrocs au jeu financier, et certains médecins considèrent qu'il s'agit bien d'une addiction. Même si l'idée peut encore apparaître étrange pur beaucoup d'entre nous, peut-être que l'achat compulsif sera traité comme une addiction à l'avenir[149].

149Comme nous l'avons vu dans le chapitre 5, le centre d'addictologie de l'hôpital Brousse en France reconnaît et traite les addictions aux achats (*cf.* chapitre 5). Les critères cliniques semblent aussi dépendants, en amont, de l'évolution des normes sociales en la matière, et tout laisse à penser que la catégorie des addictions environnementales, qui a été intégrée dans le DSM 5, mais qui est limitée pour l'instant aux addictions aux jeux, fera l'objet à l'avenir d'un élargissement progressif. Par exemple, le fait qu'un nombre grandissant de joueurs en ligne se considèrent aux mêmes comme des *addicts* constitue probablement le premier pas vers la

L'identification des troubles et dysfonctionnements cognitifs, en neuroéconomie, est donc toujours relié selon nous de façon circulaire à des critères de classification cliniques. Les activités neuronales ou cérébrales ne fournissent jamais par elles-mêmes des preuves objectives, mais soulignent plutôt l'existence préalable de préoccupations médicales et sociales, relatives aux addictions et aux troubles de l'intelligence sociale. Cette affirmation ne doit pas du tout être comprise comme une remise en question du fondement normatif de la neuroéconomie et de son efficacité médicale. Au contraire, ces distinctions cliniques rendent possible, précisément, l'application et la mise en œuvre effective de diagnostics fonctionnels en neurosciences.

Les applications médicales de la neuroéconomie font ainsi l'objet d'un réel intérêt de la part des neurosciences et de la communauté clinique. Rowland *et al.* (2008) défendent par exemple l'utilisation des théories neuroéconomiques dans la détection des troubles du comportement alimentaire. D'une façon similaire, Takahashi (2010) propose de développer une neuroéconomie de l'obésité. Bickel *et al.* (2007) plaident pour une compréhension neuroéconomique de l'addiction aux drogues. Enfin, le paradigme du *reward learning* a aussi inspiré des protocoles de diagnostic de la psychopathie (*cf.* Blair *et al.*, 2001).

La neuroéconomie dispose donc d'importantes possibilités d'applications dans le domaine médical, mais principalement en tant qu'outil de diagnostic fonctionnel. Cette psychiatrie économique correspond donc bien à la description faite par Robert Castel de la médecine mentale moderne, dans laquelle les impératifs de soins sont séparés de la recherche fondamentale, consacrée plus spécifiquement au diagnostic et à la détection des troubles (*cf.* chapitre 2, section I). En effet, la question du traitement effectif des patients est relativement peu approfondie par les neuroéconomistes. Les expériences n'ont pas de réelles implications thérapeutiques. Les études pourraient au mieux servir à améliorer le dosage des traitement médicamenteux; mais la conception pharmacologique de ces médicaments relève de domaines de recherches assez éloignés. Les théories neuroéconomiques de l'addiction ou de la psychopathie n'apportent donc pas de solution de traitement innovantes, et la plupart des auteurs s'en remettent à des traitements assez classiques, comme les thérapies cognitives et comportementales (TCC) et les neuroleptiques (voir par exemple Bickel *et al.*, 2007, p.89).

reconnaissance officielle du jeu en ligne comme addiction potentielle (voir par exemple le site des « joueurs en ligne anonyme » http://www.olganon.org) .

Conclusion du chapitre 7

L'apparition d'un débat autour du paternalisme libertarien, ou d'une régulation comportementale du bien-être, au début des années 2000, joue un rôle déterminant dans notre récit historique. C'est en effet, selon nous, précisément à partir de ce moment que se développe, chez les économistes comportementalistes impliqués dans ces discussions, un intérêt tout particulier pour les neurosciences. La neuroéconomie est ainsi née autour de l'année 2003, lorsque la question des « troubles » et pathologies du comportement, et de leur nécessaire régulation devient un enjeu central en économie comportementale (*cf.* introduction).

Si les économistes comportementalistes en viennent à considérer que la neurobiologie et les techniques de neuroimagerie sont pertinentes pour leur discipline, ce n'est pas (seulement) parce que l'étude des processus neuronaux sous-jacents à la prise de décision offre un moyen de confirmer des hypothèses d'interprétation comportementale, mais d'abord et surtout parce les neurosciences, par leur ancrage médical et clinique, permettent de passer d'un registre descriptif à un registre normatif. Les *behavioral economics*, fidèles au principe kahnemanien d'une distinction entre le descriptif et le normatif, se destinent en effet à décrire le comportement « réel » des individus en laboratoire. Ces choix observés peuvent être dits plus ou moins rationnels à partir d'un optimum de comportement, censé être fourni non pas par la psychologie ou l'économie comportementale, mais par la théorie économique dite « *standard* » (Tversky et Kahneman, 1986, p.8252).

Tout le paradoxe de l'analyse comportementale du bien-être consiste ainsi à proposer une analyse normative en n'adoptant aucun jugement normatif sur ce que les individus doivent faire; les choix observés sont décrits comme des erreurs dès lors qu'ils violent des principes de rationalité que les économistes comportementalistes se défendent de choisir. Le problème de cette approche est que la théorie économique n'offre en fait pas de critères normatifs nets et indiscutables. Dans le choix inter-temporel, il n'apparait pas toujours nécessaire de respecter le principe d'une constance du taux d'actualisation avec le délai, tel qu'il est postulé par le modèle de Samuelson (1938). Le statut de cette norme est encore plus ambigu dans le cas du choix inter-temporel, selon que les objectifs concernent la rationalité individuelle ou collective des décisions.

La régulation comportementale du bien-être a donc soulevé d'importantes critiques qui mettent en évidence l'arbitraire de la norme de comportement retenue, ainsi que les risques de dérives autoritaires associées. Les neurosciences apportent de notre point de vue une solution théorique à ce débat, en justifiant l'abandon du principe des préférences révélées pour les comportements manifestement pathologiques. L'utilisation, en amont, de critères cliniques suffit alors à justifier les propositions de régulation avancées.

Le versant prescriptif de la neuroéconomie donne ainsi un sens définitif à ce que nous avons appelé « psychiatrie économique ». D'un autre côté, les réflexions qui ont été présentées ici peuvent aussi se concevoir comme un autre sous-domaine de la neuroéconomie. Comme pour le choix inter-personnel et inter-temporel, il est possible de distinguer une approche en terme d'économie comportementale dans le scanner, représentée par les travaux de Douglas Bernheim et Antonio Rangel, qui, tout en cherchant à ouvrir les *behavioral economics* aux neurosciences, reste fidèle au précepte kahnemanien d'une distinction entre le descriptif et le normatif. Une autre approche, plus fidèle selon nous à l'héritage néo-comportementaliste, revendique et assume plus explicitement ce que Gul et Pesendorefer appellent l'ambition « *thérapeutique* » (Gul et Pesendorfer, 2005, p.8) de la neuroéconomie.

Conclusion générale

La neuroéconomie s'inscrit dans le prolongement d'un courant de recherches plus ancien, en psychologie puis en neurosciences, portant spécifiquement sur le problème de la motivation et du *reward learning* chez l'animal. Les premières études dans ce domaine apparaissent dès les années 1960. A la suite des travaux pionniers de Richard Herrnstein sur la loi d'égalisation des rendements (Herrnstein, 1961), des psychologues cherchent à mettre en évidence une relation quantitative entre comportement et récompenses. Le néo-comportementalisme se comprend ainsi comme un conditionnement quantitatif ou comme une « science quantitative de la motivation », et approfondit de manière décisive l'étude du conditionnement chez Skinner.

Dans les années 1990, ce courant de recherche en psychologie expérimentale s'ouvre aux neurosciences et à l'électrophysiologie. Ce sont en fait, plus précisément, des neurobiologistes comme Paul Glimcher par exemple, qui prennent l'initiative de concevoir des tâches expérimentales inspirées de celles du néo-comportementalisme, du type machines à sous multi-jeux (*multi armed bandit tasks*). L'introduction des micro-électrodes, qui sont utilisées pour observer l'activité des neurones pariétaux chez le singe, est source d'innovations théoriques importantes pour l'étude quantitative de la motivation. La modélisation s'enrichit notamment par l'application d'algorithmes d'apprentissage empruntés au *machine learning* et à l'ingénierie (Sutton et Barto, 1998).

Les psychologues et neurobiologistes appartenant au courant néo-comportementaliste explorent des problèmes théoriques proches. Ils utilisent, chez le pigeon et le singe, des tâches expérimentales similaires. Pourtant, l'ensemble de ces travaux expérimentaux ne constitue pas, avant les années 2000, un programme de recherche à proprement parler. Les expressions de « néo-comportementalisme » ou de « science quantitative de la motivation » que nous avons utilisées pour le désigner ne sont qu'une reconstruction historique: des chercheurs comme Georges Ainslie ou Paul Glimcher par exemple, n'ont pas le sentiment d'appartenir à la même discipline et se pensent plutôt comme psychologue, psychiatre, ou neurobiologiste.

Une identité théorique commune se dégage néanmoins de la préhistoire théorique de la neuroéconomie. Même s'il ne bénéficie pas encore d'une reconnaissance académique comme nouvelle discipline, ni même comme nouvelle approche, ce domaine de recherche se

solidarise autour de ce que nous avons appelé un projet de « psychiatrie économique ». Les chercheurs -psychologues, psychiatres et/ou neurobiologistes- projettent en effet d'utiliser une modélisation inspirée de l'économie pour analyser des pathologies du comportement, principalement liés à des troubles de l'impulsivité.

La référence à l'économie est cependant trompeuse. Il s'agit en effet de représenter le comportement ou l'activité du système nerveux comme un processus d'optimisation sous contrainte. Les modèles utilisés n'ont cependant pas été empruntés à l'économie, mais ont été développés par des psychologues et des neurobiologistes. La confusion consiste à croire que ces travaux, parce qu'ils portent sur des mécanismes de maximisation de la valeur, jettent un pont entre la psychologie, les neurosciences, et le concept économique d'utilité espérée. L'incompréhension est d'autant plus fréquente que les chercheurs concernés mobilisent un vocabulaire théorique équivoque pour l'économiste, à l'image de Glimcher et de sa notion d'« *utilité espérée physiologique* » (Dorris, Bayer et Glimcher, 2005). Or les modèles d'espérance d'utilité en économie ne s'appliquent pas aux mécanismes du *reward learning* et ne peuvent être utilisées comme théories prédictives dans les tâches type machines à sous multi-jeux. Les recherches néo-comportementalistes n'ont donc pas vocation à tester, c'est-à-dire réfuter ou confirmer, des théories préalablement élaborées par les économistes.

L'« inspiration économique » revendiquée par les chercheurs renvoie donc plutôt à un effort de formalisation. Le projet commun à tous ces travaux expérimentaux consiste à caractériser des processus pathologiques comme des déficits fonctionnels, c'est-à-dire comme des variations quantitatives au dessus et en dessous d'une norme ou d'une moyenne. Cette approche implique l'utilisation non pas de modèles formels à l'économie mais de protocoles standardisés, fonctionnant comme des dispositifs de diagnostic fonctionnel, associés à un traitement quantitatif des variables observées. L'idéologie scientifique de « psychiatrie économique », qui informe de manière latente toutes ces recherches, possède en outre deux caractéristiques importantes: la quasi-totalité des travaux sont réalisés sur l'animal, et adoptent une perspective évolutionnaire et séquentielle. Le néo-comportementalisme se distingue ainsi nettement de la psychologie kahnemanienne. Celle-ci nourrit en effet des rapports conflictueux avec les psychologues évolutionnistes (*cf.* Tversky et Kahneman, 1996; Gigerenzer, 1996), et mobilise exclusivement des sujets humains.

Ce n'est que dans les années 2000 que le néo-comportementalisme se constitue véritablement comme une nouvelle discipline. A partir de 2002-2003, certains économistes comportementalistes manifestent un intérêt prononcé pour les neurosciences. Ils font alors de

la neuroéconomie un sous-domaine de l'analyse économique. Le passage de l'idéologie à la science, ou du néo-comportementalisme à la neuroéconomie, ne s'effectue pas sans changements théoriques importants. Les économistes comportementalistes tentent en effet au départ d'interpréter les recherches en neurosciences auxquelles ils participent comme des confirmations « cérébrales » de leurs propres explications du comportement. Cette « *économie comportementale dans le scanner* » s'avère cependant être une impasse théorique. L'application de schémas dualistes, supposant une opposition entre raison et émotions, inspirée de Kahneman (*cf.* Kahneman, 2003), s'oppose en effet à l'hypothèse unitaire d'une « *monnaie neuronale commune* » (Berns et Monatgue), plus conforme aux premiers travaux sur le *reward learning*.

Le retour au cadre théorique du néo-comportementalisme se révèle être une heuristique de recherche bien plus féconde en neuroéconomie. L'introduction de la neuroimagerie approfondit de manière décisive l'étude quantitative de la motivation. Ces nouvelles techniques expérimentales, non-intrusives, permettent de passer de l'animal au sujet humain. Chez l'homme, la compréhension des mécanismes du *reward learning* implique l'étude des facultés de planification et du cortex préfrontal. Par ailleurs, le champ d'application de la notion d'apprentissage de la récompense fait l'objet d'un élargissement, en portant désormais sur les choix interpersonnel et les troubles de la cognition sociale.

La neuroéconomie reste marquée par son héritage néo-comportementaliste, et par son projet de psychiatrie économique. La conclusion principale de notre travail est que le véritable apport de la neuroéconomie n'est pas d'instruire les économistes de « ce qui se passe dans le cerveau » du décideur, ou de fournir aux neurobiologistes une « théorie économique » du cerveau, mais bien plutôt d'introduire au sein de la théorie économique du choix rationnel des critères de classification clinique. C'est à l'appui de ces critères de délimitation entre le normal et le pathologique que les économistes comportementaux peuvent passer dans la même étude du registre descriptif au registre normatif.

Le cas du choix inter-temporel et des comportements addictifs fournit une bonne illustration à notre interprétation. Dans ce domaine, les avancées principales dans la modélisation ont été réalisées par des psychologues, qui ont proposé notamment l'hypothèse d'une actualisation inter-temporelle hyperbolique (*cf.* Chung et Herrnstein, 1967), que les économistes se sont contentés de reprendre, plutôt tardivement, dans leurs modèles (*cf.* Laibson, 1997). De l'autre côté, l'intérêt, pour les économistes comportementalistes, des travaux de neurosciences sur l'arbitrage inter-temporel ne saurait se limiter à une simple

confirmation de leurs propres modèles. Comme le souligne Rubinstein, les données comportementales, plus simples et moins couteuses à recueillir que les activités cérébrales, suffisent pour déterminer les fonctions d'actualisation de chaque sujet et leur degré d'impulsivité (Rubinstein, 2008). En revanche, pour mettre en place des politiques visant à diminuer l'influence du court-terme dans les choix des agents, comme le défendent par exemple Thaler et Benartzi (2004) en matière d'arbitrage consommation/épargne, il est nécessaire de se donner un objectif quantifié en termes de taux ou de coefficient d'actualisation temporelle: quelle proportion de son revenu un agent rationnel doit-il allouer à l'épargne plutôt qu'à la consommation immédiate? Cet objectif est donc la norme du comportement rationnel. Or la théorie économique ne permet pas de justifier le choix d'un taux plutôt qu'un autre, ni par ailleurs celui d'un écart à partir duquel un taux observé peut être considéré comme irrationnel (*cf.* Rizzo et Whitman, 2009). En introduisant des considérations relatives à la santé et à la médecine, des études de neurosciences comme celles de Hare *et al.* (2009) autorisent implicitement un partage normatif entre les individus: les sujets irrationnels sont ceux pour lesquels une pathologie a été diagnostiquée au départ de l'étude; les sujets rationnels sont les sujets sains ou normaux. La nouveauté décisive consiste à poser une équivalence entre normalité biologique et rationalité économique. C'est à cette extrémité normative que se situe selon nous le point de rencontre entre économie, théorie de la décision, psychologie évolutionniste et neurosciences.

Notre analyse met ainsi en avant un l'idée d'un paradigme pathologique. Elle débouche sur deux propositions principales, qui peuvent être opposées à d'autres interprétations de la neuroéconomie. D'une part, le rôle joué par l'innovation technologique, et en particulier par la neuroimagerie, est pour nous secondaire. Ces instruments ont bien été à l'origine de progrès théoriques (*cf. supra*), mais l'apport principal de la neuroéconomie réside dans l'adoption d'une distinction normal/pathologique propre aux sciences du vivant. Les expériences de neuroéconomie ne mobilisent d'ailleurs pas forcément l'IRM_f, et peuvent plus simplement s'appuyer sur la comparaison de sujets sains pathologiques (voir par exemple, Krajbich *et al.*, 2009). En outre, la préhistoire théorique de la neuroéconomie a vu le développement de concepts centraux pour la discipline, bien avant l'apparition de la neuroimagerie.

D'autre part, l'approche néo-comportementaliste dans laquelle s'inscrit la neuroéconomie se distingue radicalement des *behavioral economics*, influencées par la psychologie kahnemanienne (*cf.* Heukelom, 2009, p.121-149). L'étude quantitative de la motivation se transforme effectivement en nouveau sous-domaine de l'analyse économique à

partir du moment où des économistes comportementalistes commencent à s'intéresser aux neurosciences. Mais la tentative visant à confirmer par des données neuronales ou cérébrales des modèles issus de l'économie comportementale a rapidement soulevé des difficultés méthodologiques. Notre interprétation rejoint ici celle de Don Ross, selon laquelle l'économie comportementale dans le scanner représente une impasse théorique (*cf.* Ross, 2008). Conformément à ce que propose ce même auteur, le cœur théorique de la neuroéconomie se concentre plutôt du côté de l'« *économie neurocellulaire* », c'est-à-dire de l'analyse des mécanismes neuronaux du *reward learning* (*cf.* Ross, 2008). Plus généralement, la genèse théorique de la neuroéconomie montre que les économistes ont eu un rôle passif dans la constitution de la discipline, en reprenant à leur compte des données et des techniques expérimentales mais aussi des modèles théoriques développés par des psychologues et des neurobiologistes.

Ces deux propositions remettent en cause plusieurs interprétations figurant dans la littérature secondaire. Notre analyse va d'abord à l'encontre de l'idée, défendue par les promoteurs de l'économie comportementale dans le scanner (voir par exemple Camerer, Loewenstein et Prelec, 2005, ou Sanfey *et al.*, 2006), selon laquelle la neuroéconomie représenterait un approfondissement de l'économie comportementale au moyen des techniques neuroscientifiques. En effet, le néo-comportementalisme est de fait incompatible avec le cadre dualiste de la psychologie kahnemanienne. Cela remet par ailleurs en question l'appréhension des résultats expérimentaux, qui, en se référant souvent aux travaux de Damasio (Damasio, 1994), mettent l'accent sur le rôle des émotions (Bourgeois-Gironde, 2008, p.10; Camerer, 2003, p.1673). Inversement, la neuroéconomie ne doit pas être comprise, comme l'a parfois suggéré Glimcher, comme une « *théorie économique du cerveau* » (Glimcher, 2003, p.322), car la théorie en question, bien que faisant référence à une forme de maximisation de la valeur, n'a rien d'économique (*cf. supra*).

Les critiques méthodologiques (Rubinstein, 2008; Harrison, 2008-a et 2008-b; Poldrack, 2006; Gul et Pesendorfer, 2005) ciblent quant à elles les problèmes liés aux techniques de neuroimagerie. L'utilisation de l'IRM_f par des non-spécialistes a été à l'origine de difficultés méthodologiques, qui, bien que réelles, sont cependant largement surestimées. Les économistes comportementalistes sont en effet intéressés dans la possibilité de « lire » des états mentaux ou des processus cognitifs à partir des activités cérébrales, afin de favoriser une explication donnée du comportement. Or, cela n'est possible qu'au prix d'un renversement de l'approche classique en neurosciences, qui consiste au contraire à essayer de caractériser au

niveau cérébral une fonction cognitive donnée, plutôt que d'inférer celle-ci à partir de la première.

Le problème dit de l'inférence inverse (Poldrack, 2006) n'est cependant pas insurmontable. Pour des tâches standardisées et des fonctions cognitives bien délimitées, le nombre élevé de réplications et la connaissance clinique antérieure permettent d'assurer un degré de robustesse suffisant aux interprétations proposées par les neurobiologistes. Les principaux résultats de la neuroéconomie ont donc une validité essentiellement restreinte aux protocoles et aux troubles cognitifs associés au *reward learning*. De ce point de vue, tout l'enjeu des travaux sur le choix interpersonnel consiste par exemple à savoir si les mécanismes cognitifs impliqués peuvent être compris comme des processus d'apprentissage de la récompense « sociale ». Les neuroéconomistes réussissent donc à inférer des états mentaux à partir des activités cérébrales, mais dans un cadre interprétatif restreint, différent de celui de l'économie comportementale.

Le défaut principal de toutes ces analyses, qu'elles visent à défendre ou à critiquer la neuroéconomie, est de penser celle-ci comme la superposition de disciplines hétérogènes (neurosciences, économie, psychologie). Or la neuroéconomie n'est pas née, comme le suggèrent Camerer, Loewenstein et Prelec, de l'ouverture progressive de l'économie à la psychologie, puis aux neurosciences[150]. Il n'y a pas eu, on l'a vu, d'échange théorique d'une discipline à l'autre. La neuroéconomie est issue d'un domaine de recherches spécialisé, qui s'est progressivement constitué comme un nouveau domaine de savoir autonome.

D'un point de vue historique, la question pertinente n'est donc pas de savoir si ou comment les frontières disciplinaires peuvent, ou non, être franchies, mais plutôt de déterminer leur construction théorique. Comme le souligne Canguilhem, « *l'histoire des sciences est victime d'une classification qu'elle accepte comme un fait de savoir alors que le problème est de savoir de quel fait elle procède, alors qu'il faudrait entreprendre d'abord une histoire critique des classifications* » (Canguilhem, 1977, p.28-29). En suivant Canguilhem, les études méthodologiques sur la neuroéconomie peuvent être critiquées pour leur absence de mise en perspective historique. Les auteurs adoptent, pour chaque discipline, des définitions données de leurs objets, de leurs méthodes et de leurs frontières, pour se demander ensuite si

150 « *au cours des deux dernières décennies […], l'économie a commencé à importer des idées à la psychologie. L'économie comportementale joue maintenant un rôle de première importance dans le paysage intellectuel et a développé des nombreuses applications en économie […]. L'économie comportementale a été principalement influencée par une branche de la psychologie appelée « recherche comportementale sur la décision », mais d'autres sciences cognitives semblent mures pur la récolte. Des idées importantes vont sûrement venir des neurosciences, soit directement soit parce que les neurosciences vont bouleverser la psychologie qui à son tour va influencer l'économie* » (Camerer, Loewenstein et Prelec, 2005, p.9)

elles partagent, ou non, des intérêts théoriques communs. Par exemple, Gul et Pesnedorfer admettent que l'économie traite exclusivement de choix et de comportements observés, ce qui leur permet de rejeter toute forme d'explication psychologique ou neuroscientifique en économie (Gul et Pesendorfer, 2005). A l'inverse, Christian Schmidt considère que l'économie et les neurosciences peuvent converger autour du problème des « *décisions raisonnées* » (Schmidt, 2010, p.21).

Les positions de chaque auteur dans ces discussions dépendent évidemment de la délimitation retenue de chacune des disciplines. Le problème de l'interdisciplinarité tel qu'il est analysé dans la littérature secondaire est donc en grande partie artificiel. A la limite, dans cette perspective purement méthodologique, le problème de l'interdisciplinarité peut être appréhendé dans des termes logiques, indépendamment de toute considération relative aux neurosciences ou à l'économie, comme le fait Mäki (2009). Ce dernier propose des formes *a priori* de collaboration entre disciplines (trans-disciplinarité, multi-disciplinarité, impérialisme, interdisciplinarité verticale, horizontale, *etc.*, *cf.* Mäki, 2009, p.351), qui peuvent s'appliquer à n'importe quelle collaboration entre disciplines diverses.

L'absence de questionnement historique est solidaire d'une perspective interne à la discipline elle-même. Les analyses dites de méthodologie, d'épistémologie ou de philosophie portant sur la neuroéconomie se comprennent en effet non comme des études historiques, mais comme des contributions scientifiques. Il s'agit toujours d'apprécier le potentiel théorique de la neuroéconomie, et d'évaluer la pertinence d'un rapprochement entre économie et neurosciences. C'est la raison pour laquelle les discussions sont tournées vers la question du franchissement des frontières disciplinaires.

John Davis et Don Ross se situent cependant dans une perspective plus proche de la notre, en considérant la neuroéconomie comme un sous-domaine autonome et distinct de l'économie comportementale (Ross, 2008; Ross, *et al.*, 2008; Davis, 2010). Ross inscrit le développement de la discipline dans une tradition théorique très proche de celle que nous avons appelé néo-comportementalisme, associée en particulier aux travaux de Herrnstein et Ainslie (Ross *et al.*, 2008, p.51, et p.66). Ross différencie par ailleurs la neuroéconomie ou économie neurocellulaire de l'économie comportementale dans le scanner (Ross, 2008). Davis souscrit à l'interprétation de Ross et suggère que les avancées théoriques les plus significatives de la neuroéconomie n'ont pas été produites par un « *usage instrumental* » des techniques de neuroimagerie par les *behavioral economics* (Davis, 2010, p.574).

Les deux auteurs ne comprennent cependant pas leurs travaux comme des études

historiques. Ross se situe explicitement dans une visée évaluative. Sa distinction entre économie comportementale dans le scanner et économie neurocellulaire a vocation à mettre en évidence des « bonnes » et des « mauvaises » façons de faire de la neuroéconomie (Ross, 2008, p.473). Davis veut quant à lui dégager les implications théoriques des neurosciences, pour la conception de l'individu en économie (Davis, 2010). Dans les deux cas, l'analyse ne peut (et n'a d'ailleurs nullement l'intention) expliquer la raison pour laquelle les neurosciences, en premier lieu, pourraient intéresser l'économiste. En adoptant la séparation proposée par Ross et Davis entre neuroéconomie et *behavioral economics*, il est alors difficile de comprendre comment ce domaine de recherche en neurosciences puisse s'intégrer à l'analyse économique.

Dans l'un de ses travaux antérieurs, Davis suggère que la théorie économique contemporaine se caractérise par un « *pluralisme méthodologique* ». Celui-ci se manifesterait par la perte d'influence progressive du « *courant dominant néoclassique* » et par un éclatement croissant de l'analyse économique entre de multiples sous-domaines, influencés par des disciplines extérieures à l'économie (Davis, 2006, p.1). Mais, pour l'historien, il ne suffit pas de constater une ouverture de l'économie à d'autres branches scientifiques, et d'étudier les mécanismes par lesquels les économistes assimilent des concepts ou des modèles non-économiques. L'enjeu consiste à comprendre la manière avec laquelle ces sous-domaines se constituent. Dans le cas de la neuroéconomie, l'enquête historique montre que le néo-comportementalisme ne se transforme en nouvelle branche de l'analyse économique qu'au moment précis où apparaît, en économie comportementale, un débat théorique relatif aux pathologies et troubles du comportement. Invoquer de façon *a priori* des zones de rencontre entre économie et neurosciences aussi larges et indéterminées que l'« *individu* » (Davis, 2010) ou les « *décisions raisonnées* »(Schmidt, 2010, p.21) ne permet donc pas de rendre compte de l'émergence de la discipline, car, avant 2002-2003, les données, les techniques et les principaux résultats théoriques de la neuroéconomie étaient déjà disponibles, mais n'étaient pas alors considérées comme des contributions théoriques pour l'économie. Selon notre interprétation, s'il n'y avait pas eu l'apparition des débats autour du paternalisme libertarien au début des années 2000, le néo-comportementalisme n'aurait pas donné lieu à une nouvelle branche de l'analyse économique.

La genèse théorique de la neuroéconomie vaut donc comme une contribution critique, en remettant en cause les définitions acceptées des objets et des frontières de chaque discipline. A un second niveau, cette étude apporte une critique de nature plus « politique ».

Comme l'écrit Canguilhem, « *accepter sans critique la partition du savoir avant le procès historique où cet ensemble va se développer, c'est obéir à une idéologie* » (Canguilhem, 1977, p.28-29). Il ne s'agit pas ici de dénoncer une idéologie politique, au sens courant du terme, à laquelle participeraient les neuroéconomistes et leurs commentateurs. Dans le cas de la neuroéconomie, « *accepter sans critique la partition du savoir* » revient à considérer que les neurosciences traitent du cerveau, et peuvent avoir éventuellement un intérêt théorique pour des économistes, qui traitent de théorie mais aussi de politiques économiques. Notre analyse montre au contraire que tout ce domaine de recherche est solidaire d'un projet de régulation des comportements, et d'un nouvel interventionnisme économique de type environnemental, qu'il serait possible de rapprocher avec les analyses de Michel Foucault sur la biopolitique[151].

Les neuroéconomistes n'ont pourtant pour la plupart nullement l'intention d'élaborer une « psychiatrie économique », et ne s'identifient probablement pas à ce projet de régulation paternaliste. Mais, pour connaître l'idéologie scientifique dont est porteuse une discipline

151Ce serait néanmoins l'objet d'un autre travail de recherche. Nous nous contenterons ici simplement de suggérer que la neuroéconomie et les débats sur le paternalisme libertarien pourraient être interprétés comme des dispositifs caractéristiques du néolibéralisme, tel que l'étudie Michel Foucault dans la *Naissance de la Biopolitique*. Pour Foucault, l'État néolibéral, à l'inverse de l'État providence, vise une «*intervention qui ne serait pas de l'assujettissement interne des individus, mais une intervention de type environnemental*» (Foucault [1978-1979], 2004, p.265). Or, sans rentrer dans les détails, Foucault associe à ce nouvel interventionnisme une rupture théorique en économie, liée en particulier aux travaux de Gary Becker et de l'école de Chicago, à partir de laquelle l'économie se définit comme « *la science de la systématicité des réponses aux variations du milieu* » (Foucault [1978-1979], 2004, p.274). En suivant cette interprétation, Foucault suggère, dès les années 1970, que l'économie a vocation à incorporer des savoirs et des techniques à la psychologie et au comportementalisme: « *quand vous définissez l'objet de l'analyse économique comme ensemble des réponses systématiques d'un individu donné aux variables du milieu, vous voyez que vous pouvez parfaitement intégrer à l'économie toute une série de techniques, de ces techniques qui sont précisément en cours et en vogue actuellement aux États Unis et que l'on appelle les techniques comportementales. Toutes ces méthodes dont les formes les plus pures, les plus rigoureuses, les plus strictes ou les plus aberrantes, comme vous voudrez, vous les trouvez chez Skinner, et qui consistent précisément, non pas du tout à faire l'analyse de la significations des conduites, mais simplement à savoir comment un jeu donné de stimuli va pouvoir, par des mécanismes dits de renforcement, entraîner des réponses dont la systématicité pourra être notée* » (Foucault [1978-1979], 2004, p.274). Certes, Foucault n'a ici que l'intuition d'un rapprochement entre économie et psychologie, et il n'a sans doute pas pensé à l'éventualité d'une incorporation de la neurobiologie. Il n'a pu assister à l'extraordinaire développement des neurosciences au cours des vingt dernières années, et ses références aux « *techniques modernes comportementales* » (Skinner, Pavlov) apparaissent aujourd'hui un peu datées. Cela étant, rien n'empêche de penser que la neuroéconomie se situe dans le droit prolongement des analyses de Foucault sur la théorie économique contemporaine dans la mesure où, selon ce dernier, la psychologie comportementale sert de nouvelle « *technologie environnementale* » au néolibéralisme (Foucault [1978-1979], 2004, p.265). Les liens que nous avons établi entre neuroéconomie, paternalisme libertarien, et la régulation environnementale des pathologies comportementales, comme par exemple celle des addictions (*cf.* chapitre 7) convergent avec cette interprétation foucaldienne. Par ailleurs, toujours dans la *Naissance de la Biopolitique*, Foucault, en se référant aux travaux de Robert Castel, envisage ces techniques comportementales à la manière de ce que nous avons appelé « psychiatrie économique »:« *sur ces techniques comportementales, il y a un peu de littérature en France. Dans le dernier livre de Castel, la Société Psychiatrique avancée, vous avez un chapitre sur les techniques comportementales et vous verrez comment c'est, très exactement, la mise en œuvre, à l'intérieur d'une situation donnée -en l'occurrence un hôpital, une clinique psychiatrique- de méthodes qui sont à la fois des méthodes expérimentales et des méthodes impliquant une analyse proprement économique du comportement* » (Foucault [1978-1979], 2004, p.274).

scientifique, il ne suffit pas, et il n'est pas même nécessaire de sonder les opinions politiques des chercheurs. Souvent, les historiens de la pensée économique croient mettre à jour « l'idéologie politique » de courants de recherches en économie à partir des positions politiques prises par les auteurs qu'ils étudient. Par exemple, dans sa thèse consacrée à Kahneman et Tversky, Heukelom considère que les économistes comportementalistes sont général plus favorables que les économistes expérimentalistes à une intervention publique dans la sphère concurrentielle; par conséquent les *behavioral economics* seraient plus réformistes et moins libérales que le programme de recherche de Vernon Smith (Heukelom, 2009, p.152-153). Or l'idéologie scientifique n'a rien à voir avec les croyances personnelles des scientifiques, et il ne s'agit pas seulement de savoir si l'État doit ou non réguler les marchés. Les neuroéconomistes défendent sans doute à ce sujet des opinions diverses. Pourtant, s'il existe aujourd'hui des colloques consacrées à la neuroéconomie, si les neurosciences sont désormais enseignées à l'université dans des départements d'économie, c'est bien parce que la régulation médicale des troubles du comportement est devenue un enjeu de politique économique.

Annexes

Compte-rendus d'entretiens

-Entretien avec Sacha Bourgeois-Gironde (12/02/2009)

-Nicolas Vallois (NV): « Vous êtes chercheur à l'institut de sciences cognitives Jean Nicod. Mais votre parcours vous a en fait mené de la philosophie à la neuropsychologie expérimentale…

-Sacha Bourgeois-Gironde (SBG): « Oui, j'ai commencé par la philosophie traditionnelle, purement traditionnelle. Je suis rentré à l'institut Jean Nicod il y a 10 ans –le labo s'est crée il y a 10 ans, en 2000, et c'était d'abord un labo de philosophes, de philosophes analytiques- j'étais un philosophe analytique. En fait, j'ai été philosophe analytique assez jeune, cela remonte à mes études à l'ENS, plus précisément à la fin de mes études. Si on remonte plus tôt, j'étais pas du tout un philosophe analytique, j'étais un type paumé comme la plupart des gens qui font de la philo en France, c'est-à-dire dans la confusion mentale la plus totale, avec des paradigmes bizarres issus des années 1960-1970, un mélange de philosophie française obscure et de phénoménologie pas plus claire, et j'étais comme tout le monde dans ce bouillon culturel –c'est purement un bouillon culturel, qui s'impose à soi dans les classes prépas, et à l'ENS également. Je ne sais pas ce qui s'est passé, mais j'en suis sorti mentalement, ce qui n'était pas gagné, car j'ai quand même été jusqu'à faire un DEA avec Jacques Derrida. En fin, j'ai commencé, pendant 15 jours, un DEA avec Jacques Derrida. Heureusement, c'est un type très bien, il m'a dit : «il faut être rigoureux en philo». Venant de Derrida, c'était pas mal…. Et donc là je suis parti aux Etats-Unis, j'ai commencé la philo analytique. La philo analytique, j'en ai fait pendant pas mal d'années, j'ai publié… La philo analytique, c'est quelque chose qui marche toujours, mais qui est rentré dans des sciences particulières. Typiquement, au départ, c'était la philosophie du langage, puis la philosophie de l'esprit –essentiellement, les thèmes dominants sont l'esprit et le langage- et en fait aujourd'hui, beaucoup de personnes qui sont dans la philo analytiques sont rentrés dans les sciences elles-mêmes. Ici, dans ce labo, il y a des gens qui étaient des philosophes analytiques et qui sont des linguistes aujourd'hui, mais qui sont devenus des réels linguistiques, mondialement reconnus ; ou des gens qui faisaient de la philosophie de l'esprit, et qui font désormais des sciences cognitives, de la neuropsychologie expérimentale effectivement. Moi c'est un peu différent parce que j'avais des problèmes particuliers, des questions qui m'intéressaient sur la rationalité, et je me suis tourné plutôt vers l'économie, enfin la théorie de la décision au départ. Tenez, par exemple, ceci est un article qui va être publié dans *Theory and Decision*. En fait, il s'agit de questions philosophiques au départ, c'est pour cela que je vous le montre, pour que vous voyiez le cheminement. C'est une modélisation du type de violations –pour situer les effets de cadrage- au principe d'invariance. On regarde avec mon co-auteur le type de structures axiomatiques qu'il faut modifier pour accepter les effets de cadrage. Car ceux-ci ne sont pas des anomalies, mais des comportements rationnels, qui dépendent de facteurs pragmatiques. C'est vraiment quelque chose d'analytique au départ, philosophie du langage et théorie de la décision pure : sur quoi faut-il réfléchir dans une théorie axiomatique de la décision pour accepter les effets de cadrage et non pas les voir comme des anomalies ? Au début, il y a quelques années, je réfléchissais sur ces anomalies.

Cet article sort maintenant, mais la réflexion remonte assez loin. Qu'est-ce que l'irrationalité, ou que sont ces biais cognitifs, qui sont des violations à ce que l'on a encapsulé à un moment donné comme des axiomes de rationalité ? Je me suis intéressé au statut de ces choses là, philosophiquement parlant, mais en rentrant dans la technique : vous verrez, cet article est relativement technique. Finalement, je suis devenu beaucoup moins théoricien que si j'avais poursuivi dans cette voie, car j'ai fait de la logique étant jeune, et je me suis tourné vers des choses vraiment expérimentales. C'est comme cela que la neuroéconomie est arrivée. C'est aussi parce que j'avais de simples opportunités, de simples contacts, là où il faut, dans les hôpitaux…Après m'être lancé avec un certain enthousiasme naïf dans ce domaine, j'ai commencé une expérience sur les effets de cadrage en imagerie cérébrale, parce que j'avais discuté avec Kahneman, j'avais rencontré un certain nombre de personnes… J'ai passé beaucoup de temps à étudier ce qui se faisait en neurobiologie sur les effets de cadrage, les biais cognitifs, *etc.* C'est d'ailleurs le sujet de mon habilitation pour passer professeur, en 2006. En fait, les deux discours de Smith et Kahneman lors de la remise de leur prix Nobel tracent deux pistes : il y a la voie kahnemanienne, que j'ai explorée et la voie smithienne. Aujourd'hui, je suis plus convaincu par la seconde : j'ai changé de perspective. Mais suivant l'orientation que j'avais au départ, il était normal au départ de suivre la voie kahnemanienne. Je la suis toujours d'une certaine manière : envisager les biais, les anomalies, et de les présenter sous un angle plus précis. En fait, sur l'étude expérimentale des effets de cadrage, je n'ai pas réussi… Par contre, j'ai réussi d'autres choses, des choses très différentes, en neuroéconomie. Par exemple, même s'il faut encore reconstruire le discours, c'est le genre d'études que j'ai réalisées sur l'argent. Paradoxalement, c'est en effet beaucoup plus dur et moins direct d'articuler ces études à l'économie. A l'inverse, avec les rapports entre économie et psychologie, on est déjà dans quelque chose de bien installé, avec les travaux en théorie de la décision de Kahneman et Tversky, donc on bénéficie déjà d'une sorte de légitimité, qui à mon avis est factice. Même si ici on voit moins le lien direct entre ce que j'ai vu neurobiologiquement dans mon étude sur l'argent et l'économie, je pense que c'est malgré tout par là que l'on peut réussir. Et c'est donc ce que je vais essayer de faire dans les prochains mois et années. J'ai sauté quelques étapes, dans le récit, mais je vous ai posé quelques jalons.

-**NV:** On reviendra plus tard sur le problème de l'articulation avec l'économie. Dans votre ouvrage sur la neuroéconomie, on ressent en effet plutôt la voie smithienne. Votre parcours, qui vous a mené de la « confusion logique » des problèmes philosophiques traditionnels, à un traitement plus expérimental, plus précis, à l'aide des sciences cognitives, de problèmes relatifs à la rationalité et la décision, soulève deux questions : d'abord, est-ce que les neurosciences sont pour vous une « nouvelle logique » pour la philosophie analytique ? Ensuite, comment avez-vous réussi votre conversion : quelles sont vos compétences en neurobiologie ? Comment avez-vous appris à vous servir de l'appareillage technique des neurosciences, à lire et interpréter les résultats de l'imagerie cérébrale ?

-SBG: Cette formulation -les neurosciences comme « nouvelle logique » pour la philosophie du langage- est amusante : c'est exactement dans ces termes que j'ai moi-même formulé la chose. Pendant très longtemps, j'ai suivi le paradigme logique en philosophie analytique, je me suis accroché à la logique. Et je n'étais pas si bon que ça…En fait, cela ne fonctionnait pas. Du coup, j'ai effectué un tournant expérimental : j'ai interprété les contraintes que l'on affronte quand on conçoit un protocole comme des contraintes logiques d'un autre ordre. Je ne sais pas comment le qualifier davantage, mais j'ai considéré ma conversion à l'expérimentation par rapport à la logique, car celle-ci était le point de référence. Je considère effectivement l'expérimentation comme une série de contraintes formelles utilisées pour

traiter des problèmes philosophiques. Maintenant, beaucoup plus spontanément que ce que je faisais auparavant en logique, et peut être aussi parce que je m'y attarde plus, j'ai plus de visibilité –je publie plus, c'est aussi un critère- je me pose les problèmes philosophiques dans une dimension expérimentale, en termes de « que peut on tester ? ». On pourrait contester le fait que cela reste de la philosophie, mais je ne me pose pas cette question méta-philosophique. En procédant comme cela, j'ai abandonné une grande partie de ce qui fait un philosophe –en tout cas en France- c'est-à-dire qu'il est avant tout un méta-philosophe : il se pose beaucoup de questions sur ce qu'il est en train de faire. Au lieu de résoudre un problème, souvent, il se pose des questions sur la nature des problèmes. Moi, j'ai cessé d'être un méta-philosophe : je me pose des problèmes et j'essaie de les régler. Après, que l'on me dise que cela n'est plus de la philosophie, cela ne m'importe pas. Il y a des gens pour qui cela pose un problème et d'autres pas. J'en conviens, il y a des clivages, et l'on vit avec. Donc cette formulation est heureuse, car dans mon cas c'est exactement comme cela que j'ai vécu la transformation, en renonçant à une certaine vision de l'essence de la philosophie : la réflexivité, la méta-philo, etc. En ce qui concerne la question « ai-je réussi ma conversion ? », j'aimerais pouvoir répondre par l'affirmative, mais je n'en suis pas si sûr. Par exemple, si vous me demandez si je sais programmer une expérience ou si je comprends le fonctionnement des machines, des IRM, je serais forcé de vous répondre « non, pas très bien ». Je comprends quand même un peu, mais je ne suis pas un technicien, et peut être que pour être un véritable expérimentaliste, il faut maîtriser ces choses là. Ma limite est forte : je ne sais pas tout faire, loin de là. Mais c'est aussi une dimension du travail qui apparaît, c'est que l'on apprend à s'insérer dans des équipes –ce que l'on n'apprend pas du tout en philosophie. J'ai des collaborateurs, des co-auteurs : on écrit à plusieurs, et cela ne me gêne pas plus que cela. Je m'ancre donc dans des communautés compétentes. En même temps, j'ai une définition relativement lâche de la neuroéconomie du point de vue de la technologie. Cela a été popularisé grâce à l'imagerie cérébrale, au début des années 2000. L'IRM, c'est quelque chose de tout à fait récent, qui est apparu dans les hôpitaux, évidemment pour des usages médicaux avant tout. Mais la neuroéconomie est au-delà de l'imagerie cérébrale : à partir du moment où on corrèle un comportement d'intérêt pour l'économie avec une donnée physiologique, on est dans la neuroéconomie. Comme le dit fort justement Rubinstein dans son article critique sur la neuroéconomie, un temps de réaction à une question, dans la mesure où il s'agit d'un marqueur d'un processus neurocognitif, est en un certain sens une variable neuroéconomique. Si on regarde Damasio, ses travaux ne sont pas de l'imagerie, ou très minoritairement : ils s'appuient souvent sur la mesure de la réaction électro-dermale, envisagée comme marqueur physiologique de l'émotion. Il n'y a pas du tout d'imagerie cérébrale chez Damasio.

-**NV:** Il s'appuie également beaucoup sur la psychiatrie.

-SBG: Oui, il s'appuie sur la neuropsychologie, c'est-à-dire sur l'étude de patients avec ou sans des lésions cérébrales. A l'époque de la publication de ses travaux, dans les années 1990, Damasio ne présentait pas ses travaux comme de la neuroéconomie, parce que cela n'existait pas encore, mais c'est pourtant bien de la neuroéconomie. Damasio et Bechara ont d'ailleurs écrit un papier récemment pour finalement récupérer, ou re-situer leurs travaux dans une perspective qui est celle de la neuroéconomie. On peut donc avoir une définition extrêmement ouverte : l'essentiel, c'est d'avoir une corrélation entre un comportement d'intérêt pour l'économie et un marqueur physiologique. Mais une fois que l'on a cette corrélation, qu'est-ce qu'on en fait ? Et c'est là où les difficultés commencent…

-**NV:** Justement, dans votre ouvrage sur la neuroéconomie, vous attachez le plus grand soin à

montrer que celle-ci ne vise pas du tout à réduire les comportements à des données neurophysiologiques, à expliquer tout par les neurones, le cerveau. Les expériences ne permettent pas d'affirmer que tel processus neurophysiologique est effectivement requis pour l'accomplissement d'une tâche : on peut simplement établir des coïncidences, ou des corrélations, entre ces processus, et un protocole donné, qui doit correspondre à un « contexte social ».

-SBG: Oui, tout à fait, enfin le protocole encode de manière tout à fait minimaliste et abstraite un contexte social : c'est très peu socialisé en fait, au sens traditionnel du terme. On se contente d'abstraire certains traits d'un contexte social.

-NV: Cette position apparaît assez particulière. En effet, dans la littérature spécialisée, on débat souvent de l'intérêt de la neuroéconomie autour des questions suivantes : qu'est-ce qu'apporte un niveau de variables supplémentaires ? En quoi les données neurophysiologiques sont-elles intéressantes pour l'économiste ? *etc.* Dans ce genre de discussions, dans lesquelles s'insère justement Rubinstein, on affirme soit que ces variables permettent d'être plus réaliste, soit qu'elles sont largement inutiles pour l'économiste, et que celui-ci peut donc les négliger. Or vous semblez vous situer dans un tout autre débat, ou alors ce débat ne vous intéresse pas.

-SBG: Je ne pense pas que l'économie doive perdre son statut de science spéciale, au sens de Fodor. Peut être que le programme de la neuroéconomie consiste à rechercher des lois ponts entre les concepts de l'économie et les lois physiques : les neurosciences organiseraient le passage de l'un à l'autre. A un certain concept économique, prenons l'utilité par exemple , on associerait certaines activités neuronales qui elles-mêmes sont sous-tendues par des processus physiques de base. On a ainsi une réduction, au sens traditionnel de l'unité des sciences, de l'économie à des prédicats physiques. Il faut savoir si cela est d'abord possible, et je n'en sais rien. Je n'ai pas une position réductionniste en économie : je pense que le niveau d'explication des concepts de l'économie est autonome, je ne suis pas dans cette perspective là. Pourtant, tous les travaux de neuroéconomie, notamment ceux qui se font à Zurich, peuvent être interprétés dans une perspective très différente, qui conserverait le statut de science spéciale à l'économie, qui aurait ses concepts propres, ses fonctions propres, avec un niveau autonome d'explication du comportement humain. Seulement, il y a d'autres problèmes qui apparaissent : ici, on est pas sûr que ce que l'on appelle « économie » soit ce que les économistes considèrent comme tel. Cela ressemble davantage à une « anthropologie cognitive », avec une dimension évolutionniste. Dans une définition type Camerer –la neuroéconomie comme réduction de la décision économique à des activités cérébrales-, on perd l'économie comme science spéciale. Dans la voie smithienne, on perd peut être aussi l'économie, mais dans un autre sens cette fois : ce n'est plus seulement de l'économie. Cela reste à un niveau autonome où les concepts descripteurs et prescripteurs du comportement humain ne sont pas réduits à la physique. En fait, on a peut être fait des sciences humaines trop larges : c'est le problème des rapports entre économie et neuroéconomie. On n'arrive pas à cadrer.

-NV: Cette approche reconnaît donc un statut spécial à la fois à l'économie et à la neuroéconomie. Celle-ci serait animée par un projet particulier, celui d'une « anthropologie cognitive » comme vous l'avez dit, des artefacts économiques. Mais peut être peut on retrouver une certaine complémentarité entre ces deux sciences spéciales : l'économie serait une théorie, et la neuroéconomie travaillerait à un niveau intermédiaire, technologique, qui serait celui de l'artefact.

-SBG: Il y a une dimension technologique, c'est évident, mais au sens où l'archéologie a une dimension technologique. Il faut construire le passé, et cela ne peut se faire qu'avec des outils d'investigation adaptés. Quand j'étais plus jeune, je voulais être archéologue, et je me suis intéressé à l'épistémologie de l'archéologie. Il ne s'agit pas uniquement d'une lubie ou d'une métaphore légère : je pense être devenu archéologue, au sens où les résultats de mes travaux dépendent fortement de la technologie, des outils employés. Ce que vous dites est tout à fait correct : la neuroéconomie, je la conçois comme une archéologie cognitive. Certains pensent que cette appellation n'est qu'un *buzz* moderniste, mais en réalité, ce que l'on peut aller voir, c'est la survenance de mécanismes cognitifs dans des contextes artificiels, ou culturels, ancrés sur des fonctions neurobiologiques beaucoup plus anciennes. C'est un zoom, donc un outil d'investigation au sens de l'archéologie, sur cette interface entre ce qui est très ancien d'un point de vue de l'évolution et adaptation à des contextes économiques particuliers. Effectivement, il peut s'agir de l'apparition d'artefacts récents, intéressants pour l'économie, ou aussi de ce qui se passe lors de la prise de certaines décisions importantes pour l'économie. La neuroéconomie est une technologie d'investigation de cette interface qui s'est produite au cours de l'évolution.

-NV: En tant que technologie d'investigation des artefacts économiques, la neuroéconomie n'entretient-elle pas une différence essentielle par rapport à l'économie ? En travaillant au niveau de l'artificialité, les frontières sont toujours brouillées entre le descriptif et le prescriptif. N'y a-t-il pas toujours, de manière implicite ou non, dans les expérimentations, toute une série de recommandations, de conseils pratiques ? Dans votre livre, vous évoquez notamment des expériences sur le consentement au paiement des impôts : peut-on en tirer des recommandations et des méthodes pour concevoir le paiement des impôts, de manière à ce qu'il soit perçu comme plus « agréable » ? De même, votre expérience sur la monnaie pourrait avoir certaines implications sur la modélisation de cet « artefact économique » qu'est la monnaie.

-SBG: Il s'agit d'une question différente de la précédente. Auparavant, on parlait de la portée de la neuroéconomie en tant que science : que permet-elle d'observer, et comment ? A quel genre de grandes questions permet-elle de répondre ? On avait laissé en suspens la question de savoir si cela est intéressant pour les économistes. C'est une question à laquelle je ne suis pas sûr de pouvoir répondre encore. Mais ce que vous proposez ici est différent : peut-on utiliser ce que l'on observe pour mieux modéliser, mieux comprendre et aussi mieux appliquer. C'est une question qui m'est souvent posée, y compris par des gens qui sont directement intéressés par ces applications, que ce soit des financiers, des banquiers, des politiques, etc. J'ai déjà eu l'occasion d'essayer de répondre à cette question, et jamais de manière très conclusive, je dois l'avouer. En effet, je vois pas exactement les applications à retirer : on observe, on comprend mieux, mais est –ce qu'on peut pour autant mieux « cadrer » le paiement de l'impôt ? Pourtant, il y a des gens qui financent la neuroéconomie –en France, ce n'est pas trop le cas, mais il pourrait y avoir des fondations privées qui se mettent à financer, il y en a déjà en Suisse- qui sont exactement intéressés par ces applications futures. Sur la monnaie, cela aurait été peut être intéressant au moment de l'introduction de l'euro d'avoir mon étude à disposition. Peut être pour la prochaine réunification monétaire mondiale... *Design* des pièces de monnaie, je ne sais pas, parce que ce n'était que la première expérience, et je suis en train de me demander ce qu'il faut ensuite analyser pour mieux comprendre ce que j'ai fait : par exemple, le côté rond des pièces, *etc.* Il y a beaucoup de choses à aller voir. En ce moment, je suis en train de me demander ce qu'est la monnaie. Je fais référence à la thèse de Stanislas Dehaene sur le recyclage des circuits cérébraux dans certaines tâches culturelles comme la

lecture ou l'arithmétique. Je suis en train de me demander : la monnaie, est-ce plutôt du langage ou du nombre ? Je retrouve sans doute l'«eau tiède» qu'avaient trouvée des anthropologues de la monnaie il y a longtemps : la monnaie, c'est un symbole. C'est à la fois du nombre et du langage, mais pas seulement cela. Les activités que l'on observe lors du décodage nécessaire pour savoir si une pièce a de la valeur ou pas sont précoces, au sens où elles arrivent en même temps que ce que l'on observe pour les visages. Cela a donc l'air d'être du perceptuel, mais il ne s'agit pas de perceptuel, parce que l'on ne perçoit pas la valeur au sens littéral, on la traite. Cela ressemblerait plutôt au codage du sens d'un mot. Quand un sujet lit des mots comme « économie » ou « trapoutazou », il est capable de distinguer ceux qui appartiennent au langage de ceux qui n'en font pas partie. Il met 400 ms pour réaliser cela. Pour le visage, il met 150 ms. C'est donc aussi rapide que du perceptuel sans être du perceptuel, parce que c'est plutôt du sémantique. On pourrait parler de symbolique, c'est-à-dire du traitement quasi-perceptuel d'une donnée abstraite. Pour qu'un artefact économique comme la monnaie « réussisse », il faut peut être qu'il possède certaines propriétés fonctionnelles comme celles-ci : il ne faut pas que ça traîne, il faut que ça soit rapide, donc il faut que ça se greffe sur certaines fonctions perceptuelles, sans être soi-même du perceptuel, mais du sémantique…Les propriétés sont très subtiles, c'est très dur à observer et à décrire. Elles sont apparues par émergence, en survenant de propriétés plus anciennes. Reproduire de manière artificielle ces propriétés, cela me paraît difficile…

-NV: Sans doute, la neuroéconomie ne permet pas de modéliser *ex nihilo* une nouvelle forme de monnaie qui capturerait parfaitement l'ensemble de ces propriétés. Mais vos travaux peuvent néanmoins suggérer certaines pistes, certains traits, qui pourraient aider à la conception des monnaies déjà existantes.

-SBG: En tout cas, on réfléchit sur des propriétés inhabituelles de la monnaie. Il ne s'agit pas d'une analyse des fonctions traditionnelles de la monnaie. On est à un niveau de fonctionnalité très différent. Le défi, c'est d'articuler ces deux niveaux de fonctionnalité. C'est une question très intéressante, mais je ne sais pas encore y répondre. Mais effectivement, cela apparaît très séduisant, de réfléchir à la conception de la monnaie. Il y a un autre domaine qui pourrait être intéressant à cet égard, qui est celui de la répugnance morale. Cela m'intéresse de plus en plus. Cela renvoie à un débat entre Gary Becker et Alvin Roth sur le fait que tout n'est pas marchandisable, notamment dans le cas des marchés d'organes. Le dégoût, la répugnance morale, exercent semble-t-il des contraintes sur la libéralisation de certains marchés, ou plus précisément sur la mise en équivalence monétaire de certains types de transactions. Il y a là sans doute quelque chose sur lequel on peut agir de façon prospective, pour essayer de donner une piste positive à votre question. La répugnance morale est en effet quelque chose de très fluctuant. Je suis en train d'étudier les stratégies exactes de la répugnance morale, c'est-à-dire comment celle-ci se manifeste. L'économie s'intéresse à un très haut niveau de fonctionnalité, et moi je m'intéresse à quelque chose de beaucoup plus bas, plus petit, fin, basique. Roth a par exemple une théorie sur la répugnance morale : il fonctionne à ce niveau élevé. Moi, ce qui m'intéresse, c'est de comprendre les mécanismes de la répugnance morale. En comprenant ces mécanismes, peut être que l'on peut voir où et quand ces mécanismes ne jouent plus une contrainte sur ces marchés. Ce n'est pas très éloigné des questions liées à l'argent, d'une certaine manière. Ce que l'on appelle répugnance morale, c'est très souvent de l'indignation : « non, je ne pourrais jamais faire cela », ou « non, cela n'est pas acceptable ». Il faut comprendre la nature exacte de ces stratégies mentales. Par exemple, il y a un effet appelé « effet de contemplation », mis en évidence par certains psychologues : on ne peut pas contempler les choses répugnantes. Si l'on te propose un marché répugnant, et que tu es quelqu'un d'incorruptible, tu dis non tout de suite. Tout est

dans le « tout de suite ». Certains psychologues pensent que si on laisse un délai de contemplation, si on envisage donc l'évaluation répugnante, le simple fait d'évaluer cette option répugnante risque déjà d'inciter à sa propre corruption morale. Je suis donc en train de travailler sur l'effet de contemplation : les sujets craignent-ils vraiment d'envisager des options répugnantes ? Si on comprend mieux les mécanismes de la répugnance morale, notamment les stratégies comportementales qui visent à ne surtout pas considérer la chose, on pourrait peut être effectivement agir sur cette relation entre sphère de l'intégrité morale, exprimée par des biais de répugnance, et marché. On est toujours à la marge de l'économie, je ne suis pas un économiste, mais on peut travailler sur la libération des contraintes qui pèsent par exemple sur le marché d'organes. On peut prendre d'autres cas. Quand j'ai commencé à travailler là-dessus, des grands groupes de communication m'ont contacté, comme Publicis. Je n'ai pas donné suite, parce que je ne suis pas dans ce type d'orientation, mais ce que vous dites est intéressant, car ces gens voulaient effectivement travailler sur des stratégies de communication permettant de rendre certaines activités, comme la vente d'armes ou certains types de médecines, moins répugnantes. Cela allait même jusqu'à l'argent, souvent considéré comme répugnant –en France notamment, il y a un tabou sur l'argent, une culture anti-entrepreneuriale- donc jusqu'à des thèmes sociopolitiques intéressants. Sur les mécanismes de la répugnance morale, on peut donc espérer pouvoir modifier les stratégies de communication sur ces thèmes là. C'est intéressant, je ne suis pas simplement en mesure de le faire pour l'instant. Non pas parce que j'ai de la répugnance à le faire, mais parce que ce n'est pas mon business.

-NV: Mais en même temps, on sent qu'un lien est possible, entre cette étude expérimentale de mécanismes d'apprentissage économique, située à un niveau technologique, et la théorie économique proprement dite…

-SBG: Si vous me dites lesquelles, je serai ravi, parce que je pourrai réfléchir à ces liens possibles, qui ne seraient pas seulement des liens annexes, mais des liens plus structurels à l'économie. Quand la neuroéconomie est apparue, elle s'est posée comme quelque chose qui allait modifier de manière interne l'économie…

-NV: Peut être qu'une convergence peut apparaître justement sur le plan des applications pratiques. Vous avez par exemple évoqué Alvin Roth, qui justement est dans une démarche expérimentaliste très performative : il a conçu de réelles procédures de marchés, il voit son travail comme de l'ingénierie en laboratoire, etc. En même temps, toutes ces études expérimentales en neurosciences peuvent être appliquées au monde économique : des entreprises utilisent ainsi les études sur le mouvement oculaire, pour concevoir le design de leurs sites internet, *etc*. On observe donc toute une série de retombées pratiques, qui permettent d'établir un tel lien. Tout d'abord, qu'est-ce que vous pensez de toutes ces applications ?

-SBG: Cela ne me gêne pas du tout. Là aussi, il y a aussi souvent une réaction de répugnance ou de rejet concernant le fait que l'on monterait des stratégies de manipulation. Mais ce n'est pas du tout le genre de discours que j'aime tenir. Cela ressemble beaucoup trop à de la conspiration, à la théorie du complot…Ce n'est pas du tout mon idéologie, enfin mon ethos… Et en plus, surtout, je ne suis pas du tout convaincu que cela soit le cas. Je pense que quelqu'un qui voudrait se livrer à cela se planterait. On en est très loin en réalité. Oui d'accord, on peut mesurer des mouvements oculaire sur la lecture d'un texte sur un site Internet : mais est-ce que cela est de la manipulation ? Je ne sais pas…Oui, c'est de la manipulation au sens où le sujet n'est pas conscient de ces mouvements oculaires et l'on va le

guider davantage dans sa lecture. Mais cela n'atteint pas le libre-arbitre

.

-NV: C'est ce qui apparaît clairement dans votre ouvrage, et ce qui je pense est assez juste : les possibilités de manipulation du cerveau sont lointaines et très spéculatives.

-SBG: Oui, ce sont des craintes irrationnelles.

-NV: En revanche, il y a certaines applications qui peuvent toutefois être un peu plus contestables, et je voudrais avoir votre avis là-dessus. On voit apparaître quand même, en liaison avec les neurosciences, l'émergence de toute une littérature de *self help*, de développement personnel. Un journaliste américain a par exemple écrit un ouvrage pour expliquer comment la neuroéconomie peut nous aider à gagner en Bourse.

-SBG: Je trouve cela ridicule

-NV: Oui, mais toutes ces petites « recettes psychologiques » de développement personnel sont indiscutablement liées aux neurosciences.

-SBG: Oui, vous avez raison.

-NV: Qu'est-ce que vous pensez de ces applications, qui sont peut être plus réelles et plus contestables ? Je ne sais pas si vous avez un avis…

-SBG: Le journaliste en question est Jason Zweig, qui effectivement est un journaliste avant tout. Il a fait un vrai boulot de journaliste. Moi-même j'ai été contacté au moment où il écrivait son bouquin. Je ne suis pas cité dedans, parce que je n'avais pas d'expérience à proposer à ce moment là. Mais j'ai vu exactement ce qu'il faisait. Bon, c'est marrant, parce que c'est quelqu'un qui au départ est très convaincu par les travaux de Kahneman, et donc dans un sens, c'est dans le prolongement de ce que faisait Kahneman en termes de prescriptions. Beaucoup de domaines, notamment le domaine financier, se sont intéressés aux types de biais identifiés par Kahneman, parce qu'on pensait qu'on allait mieux comprendre ce qui se passe. Là c'est pareil avec la neuroéconomie : on est censé mieux comprendre le fonctionnement du trader, du consommateur… Cela ne va pas très loin à mon avis. C'est plus ridicule qu'autre chose. C'est plus un usage popularisé de la science. Oui, il y a des biais cognitifs, et on peut les voir dans la nature. Sauf qu'on dit parfois qu'une fois dans la nature, les biais cognitifs disparaissent, parce que dans le vrai marché, à terme, les biais s'effacent. C'est ce que montrent certains économistes. Par rapport à l'économie expérimentale, où l'on reproduit des anomalies en laboratoire, les études en *fields economics* vont sur le terrain –il y a un laboratoire spécialisé là dedans à Iéna, dans le centre d'économie expérimentale Max Planck, à Chicago- font disparaître les anomalies. Mais là on parle d'autres choses. Ici il y a simplement une récupération. C'est la décennie des neurosciences, c'est la décennie du cerveau, c'est de la folie. Toutes les semaines, il y a un truc sur le cerveau. Comme je le dis à la fin de mon livre, à la fin des années 1970, on a eu la lutte des classes, on a eu le sujet dans les années 1980, après je ne sais plus ce qu'il y a eu dans les années 1990, rien, sûrement, il ne s'est pas passé grand-chose dans les années 1990… Et maintenant, c'est le cerveau ! C'est le nouveau sujet à la mode, on ne sait pas trop pourquoi. Des fois, c'est complètement ridicule. L'autre fois, je me suis retrouvé comme ça dans un truc complètement ridicule, interviewé par la télévision suiss-romande à la FNAC. Ils ne voulaient pas m'interviewer dans mon bureau, ils voulaient que je sois à la FNAC. J'aurai du me douter du truc. Donc j'étais à la FNAC Saint-Lazarre, et ils me faisaient commenter en direct des comportements

de consommateurs. Vous voyez, lui, il paye en liquide et non en carte bleue ; cela ne vous rappelle rien ? Et là tu es déjà dans le piège, tu es filmé. Qu'est-ce que tu fais ? C'est bien fait pour toi, tu l'as cherché…Tu penses… « Effectivement il y a cette étude du MIT qui est sortie il y a un mois, sur les dépenses en carte bleue et en liquide, et qui montre que tu dépenses moins en liquide, parce que tu as le poids de l'argent dans la main, ce que tu n'as pas avec la carte bleue. Donc l'insula, toujours elle, s'active beaucoup plus quand tu payes en liquide ». Donc ensuite ils ont interviewé le type qui payait ses achats en liquides à la FNAC, c'était un type très louche, qui a expliqué sans le savoir l'article du MIT, que je leur avais expliqué juste avant… C'était très drôle ! Et là les journalistes : « ah oui, c'est vrai ce que vous dites ! ». Et là le type : « oui je suis tout à fait d'accord avec le monsieur, mais je dirais quelque chose de plus : tout ça c'est le lobby des banques ! ». Donc oui, la transition est très rapide et très fragile. Alors il y a des personnes comme Jason Zweig, qui font leur profession de la récupération de petites études qui ont été faites, comme celle que je viens de citer du MIT, et qui propose leurs recettes : « quand vous venez à la FNAC, payez en liquide, vous dépenserez moins ». Et c'est vrai, peut être, en plus. Le pire, c'est que c'est vrai. Par contre, le livre de Jason Zweig, je n'y crois pas. Vous l'avez lu ?

-**NV:** Je l'ai feuilleté rapidement.

-SBG: Cela vous paraît convaincant ?

-**NV:** Cela ressemble à des recettes psychologiques un peu naïves, un peu de bon sens.

-SBG: Oui, un peu de bon sens labellisé prix Nobel.

-**NV:** En même temps, ce livre a été préfacé par Olivier Oullier, il bénéficie d'une sorte de caution théorique.

-SBG: Oui, Olivier Oullier est beaucoup plus là dedans, c'est un bon ami, il l'a fait, comme moi je suis passé à la télévision suisse-romande. C'est le problème de toucher à une matière qui aujourd'hui est à la mode et qui dans dix ans ne le sera probablement plus. Il ne faut pas se leurrer quand même. Oullier est sans doute au moins aussi narcissique que moi. Quand on est sollicité parce qu'on fait quelque chose à la mode, c'est parfois dur de résister. Il faut faire attention à cela. Il ne faut pas exagérer, ce n'et pas de la starification, mais c'est quand même un domaine popularisé à un niveau parfois très faible. Toutes les neurosciences, d'ailleurs. Je pense qu'en fait, on est à la fin d'une certaine période. Il y a des gens qui ont fait des choses, bien avant Olivier ou moi, comme Camerer, Coricelli –d'ailleurs, c'est très bien ce qu'a fait Coricelli, je suis très convaincu, je trouve cela très intéressant, ses travaux sur le regret- et quel a été l'effet de cette masse d'articles ? Ils ont enlevé un tabou : on peut faire de l'économie expérimentale en imagerie cérébrale. Pourquoi est-ce si intéressant, si étonnant, au bout du compte ? Finalement, l'intérêt de l'article de Sanfey sur l*'ultimatum game*, c'est de dire « regardez, on a fait de l'imagerie cérébrale ». Après, il y a de l'interprétation dans cet article, mais cela ne va pas très loin. Une fois que l'on a levé ce verrou, qu'on a dit « on l'a fait et on peut le faire », au bout d'un moment, on arrête de le faire parce qu'on l'a fait. Donc on continue s'il y a véritablement quelque chose d'intéressant à faire. Je pense qu'on en est aujourd'hui à cette phase où soit on continue comme ça, et cela n'a plus aucun intérêt, parce que le message selon lequel « on pouvait le faire» est déjà passé, soit on trouve vraiment des pistes de recherche très précises. C'est tout à fait possible que la neuroéconomie ait lancé des pistes de recherche intéressante. Moi je m'inscris dans une piste particulière. Mais à un moment donné, cela ne va plus être la neuroéconomie, mais des pistes de recherche

différentes, qui ont pris leur source dans ce moment un peu bizarre, cette bulle médiatique, qui a enlevé le tabou. Et puis il y aura des gens qui feront des modèles neurocomputationnels de la dopamine dans le cas du regret -très intéressant, mais c'est des neurosciences- et puis d'autres, comme moi, peut être au départ parce que je suis philosophe, qui se demandent ce que nous apprend la neuroéconomie au sujet de l'évolution des comportements humains, des comportements économiques, de la nature de certains artefacts. Ce sont les questions que je pose ; mais qu'y a-t-il de commun entre les modèles de la dopamine chez le rat et mes questions ? Un certain niveau de données, peut être. Mais le positionnement est très divers. Pour vous c'est important, je pense –je ne sais pas si ce que je dis est vrai mais c'est comme cela que je perçois les choses- dans la mesure où vous vous lancez dans une épistémologie de cette chose là, si j'ai bien compris, et que vous êtes en plein contexte mouvant.

-NV: Ne peut on pas déjà envisager certaines voies d'approfondissement de la neuroéconomie ? Dans une perspective smithienne, la neuroéconomie travaille sur une « rationalité écologique », c'est-à-dire qu'elle corrèle des processus neurophysiologiques à certains contextes ou environnements. C'est une certaine vision de l'économie, bien particulière, mais n'est elle pas aussi appuyée sur un certaine vision des neurosciences, probablement réductrice ? Vous expliquez par exemple que les artefacts économiques sont toujours pris dans des contraintes perceptuelles liées à l'évolution de l'espèce, comme les mécanismes de reconnaissance des visages. L'idée selon laquelle notre cerveau construit des manières de voir et de représenter la réalité est omniprésente. De ce point de vue, est-ce que les neurosciences ne sont au fond utilisées uniquement comme une forme améliorée de psychologie de la forme ? Les neurosciences ne sont-elles pas ainsi réduites à n'être que l'étude des mécanismes élémentaires de la perception ? Peut on imaginer une neuroscience des fonctions cognitives supérieures ?

-SBG: Il y en a une. C'est vrai que les neurosciences se concentrent massivement dans l'étude des capacités de «base», motrices, perceptuelles, *etc.* Mais il y a quand même une neuroscience des fonctions cognitives plus complexes, comme la mémoire, le langage, *etc*. Et enfin il y a une neuroscience certes plus petite des fonctions supérieures, comme la décision et le raisonnement. Ici, pas dans ce laboratoire, mais dans un autre laboratoire du DEC (Département d'Etudes Cognitives), on a un des spécialistes mondiaux –Etienne Coquelin- des neurosciences de la décision. D'ailleurs, il écoute ce qui se dit en en neuroéconomie, même s'il ne veut pas dévier de son programme de recherche, en quoi il a raison. Et ce qu'a fait Dehaene, et qui est remarquable, ne concerne pas des fonctions de base : c'est le nombre, l'arithmétique, la lecture… C'est déjà très évolué, tout cela. Elle existe donc, cette neuroscience, et il y a des gens puissants, enfin compétents, dans le domaine. Maintenant, est-ce que la neuroéconomie se situe dans le prolongement de ces travaux ? Vous avez tout à fait raison de poser la question, parce que du coup, si ce n'est pas dans ce prolongement –votre question est très saine- cela voudrait dire que l'on a détourné finalement les neurosciences de leur programme incrémental de recherche. On les aurait aussi réduites, minimisées, en empruntant l'outil sans emprunter tout le programme de recherche qui est derrière. Je suis entièrement d'accord avec vous. Mais je pense que mon programme de recherches, en le situant précisément dans la continuité de ce qu'a fait par exemple Dehaene sur le nombre et la lecture, est respectueux des neurosciences, au sens où j'essaye d'articuler mon programme sur le programme interne aux neurosciences. Je ne fais pas juste un outil d'imagerie, un zoom. Tout à l'heure, je disais que c'était une technologie d'investigation –je le maintiens- mais c'est la question que je me pose qui peut être ancrée dans un programme neuroscientifique. Si on arrive à faire avancer les neurosciences vers la compréhension des mécanismes neurologiques qui sous-tendent le traitement des artefacts économiques, ou certains

comportements, dans la continuité de ce qui a déjà été fait pour certaines fonctions cognitives, comme la mémoire, le langage, le nombre, la reconnaissance des visages –ce sont des choses très étudiées en neurosciences- je pense qu'on a de fait intégré notre programme de neuroéconomie dans les neurosciences. Il y a un risque, effectivement, à la fois de désintéresser les économistes et les neuroscientifiques. Je suis entièrement d'accord avec la prudence que votre question suggère.

-**NV:** Oui, car les processus neurophysiologiques évoqués interviennent toujours à titre de contraintes : on explique l'apparition d'un artefact par une certaine pré-représentation donnée du cerveau.

-SBG: Oui, il dépend d'une certaine architecture…

-**NV:** …dont s'est progressivement doté le cerveau au cours de l'évolution.

-SBG: Mais c'est normal !

-**NV:** Mais y-a-t-il des explications de la manière avec laquelle de nouvelles pré-représentations, de nouveaux schèmes sensori-moteurs se forment ou apparaissent ?

-SBG: C'est possible que des gens se soient posés ce genre de questions, des gens comme Alain Berthoz par exemple. Mais je ne saurai pas répondre personnellement à cette question. Mais c'est une bonne question, et elle est étudiée à mon avis ».

-Entretien avec Mathias Pessiglione (12/11/2009)

-NV: «Vous êtes psychologue clinicien à la Pitié Salpêtrière. Pourriez vous tout d'abord présenter votre statut et votre métier?

-MP: Je suis avant tout chercheur à l'INSERM. J'ai un poste de chargé de recherche. J'ai passé le concours de recrutement des chercheurs. Il se trouve que j'ai aussi le titre de psychologue et j'ai donc une vacation comme psychologue clinicien au centre de Neuropsychologie et du Langage, qui doit s'appeler maintenant Centre des Maladies Cognitives et Comportementales, et qui est rattaché à la Fédération Neurologique.

-NV: Vous avez donc une activité exclusivement orientée vers la recherche?

-MP: Oui, à part une demi-journée par semaine où je fais le psychologue.

-NV: Vous n'êtes donc jamais en rapport avec des neurologues ou des psychiatres?

-MP: Pour mon activité de psychologue clinicien, je suis en relation directe avec des médecins. Et même lorsque je fais de la recherche, il y a souvent des collègues psychiatres. A chaque fois qu'on fait de la recherche chez les patients, il y a forcément dans le coup des neurologues ou des psychiatres.

-NV: Il y a donc dans votre activité une interaction avec le milieu médical, notamment au niveau de la mise au point et de la conception des protocoles. Je pense en particulier à l'utilisation de patients lésés dans certaines expériences.

-MP: Oui. On passe ici forcément par les médecins, ne serait-ce que pour le recrutement, les caractéristiques cliniques, et puis même pour le design de la manipulation. On discute avec les médecins en amont et en aval, au moins dans les projets qui impliquent des patients. Même pour les projets de neuroimagerie sur des sujets sains, il y a aussi des médecins impliqués.

-NV: Vous avez donc une formation de neurobiologie et de psychologie?

-MP: Oui. J'ai fait l'ENS en biologie. J'ai fait une maîtrise de neurosciences, puis un Master 2 de sciences cognitives. J'ai poursuivi en thèse, et au cours de la thèse j'ai repris les études en psychologie au niveau de la licence, jusqu'à un Master 2 de Neuropsychologie.

-NV: Quel était le sujet de votre thèse de sciences cognitives?

-MP: Il s'agissait d'un modèle de la maladie de Parkinson sur le singe.

-NV: A aucun moment vous n'avez eu de formation en économie.

-MP: Non.

-NV: Vous êtes néanmoins intervenu dans le cadre d'une université d'été sur la neuroéconomie. Même si cela peut vous faire sourire, vos articles de recherche peuvent être

qualifiés de « neuroéconomiques ».

-MP: Oui. Je ne sais pas exactement comment on pourrait les définir, car mes recherches sont peut être un petit peu à la limite du champ. En fait, il y a beaucoup de gens qui faisaient de la neuroéconomie « sans le savoir » ou sans avoir le mot pour le dire. Avant, tous les gens qui étudiaient les bases neuronales de la prise de décision faisaient ,et font encore, de la neuroéconomie « sans le savoir ». Après, il y a des ressemblances qui sont plus ou moins superficielles. Par exemple, quand je travaillais chez le singe, on utilisait des jus de fruits comme récompenses. Quand je suis passé chez l'homme, on a utilisé l'argent, pour différentes raisons, parce que c'est pratique, facile à quantifier, *etc.* Par conséquent, il y a là une ressemblance superficielle: on utilise de l'argent dans nos expériences pour motiver les sujets, alors on dit que c'est de la neuroéconomie.

-NV: Il y a toujours plusieurs définitions possibles de la neuroéconomie. Il y a une première définition assez naïve: la neuroéconomie serait l'étude des réactions du cerveau à certaines variables économiques, comme l'argent, le taux d'intérêt... Vous ne vous inscrivez pas vraiment dans cette approche, puisque vous n'avez pas de formation en économie. Un auteur [Don Ross] qualifie ce projet de « *behavioral economics in the scanner* »: cela consiste finalement à reproduire des résultats bien connus en économie expérimentale et les illustrer simplement par la neuroimagerie. La « vraie » neuroéconomie selon cet auteur serait l'«économie neurocellulaire». Elle serait un programme de recherche qui viserait à appliquer les théories de l'optimisation sous contraintes -toutes les mathématiques liées à l'analyse des processus d'équilibres- à l'apprentissage neuronal, donc à fournir des modèles mathématiques sur l'apprentissage neuronal. Cela concernerait notamment le système dopaminergique. Selon Don Ross, l'économie neurocellulaire serait au bout du compte une application de la théorie économique en neurosciences: le système dopaminergique est considéré comme un agent économique, c'est à dire comme un système d'optimisation sous contraintes... Dans ce sens là, vous sentez vous plus neuroéconomiste?

-MP: C'est exactement ce que j'ai fait. J'ai utilisé des algorithmes provenant d'autres disciplines pour essayer de modéliser ce qui se passait au niveau du système neuronal, à un niveau de description qui est en dessous de celui du cerveau dans son intégralité. C'est peut être ce que cet auteur entend par « cellulaire ». Mais les algorithmes, je ne les ai pas empruntés à l'économie. Je connaît très mal la littérature en économie. Ce sont des algorithmes du *machine learning,* conçus par des ingénieurs pour faire en sorte que les robots, les machines, apprennent à optimiser des gestes, des trajectoires, trouver des récompenses, *etc*. Ce sont peut être les mêmes algorithmes qui sont utilisés dans les problèmes d'optimisation en économie. C'est souvent le cas en effet que les mêmes mathématiques prennent des noms différents lorsqu'elles sont utilisées par des disciplines différentes.

-NV: Plus généralement, cela viendrait des sciences cognitives au sens large.

-MP: Oui; dans le domaine, le livre de Sutton et Barto fait autorité. C'est un livre qui balaie la plupart des algorithmes d'apprentissage. J'ai personnellement beaucoup utilisé cet ouvrage, et beaucoup de gens ont fait comme moi.

-NV: Dans tous ces articles qui sont publiés dans ce domaine, ce sont donc très souvent les mêmes algorithmes qui sont utilisés?

-MP: Oui. Toutefois, la dernière tendance consiste à utiliser des algorithmes bayésiens, qui

permettent de faire des divisions et des soustractions. En effet, en actualisant les variables dont on cherche à modéliser la dynamique, on peut soit soustraire, soit diviser, et, typiquement dans les inférences bayésiennes, on divise.

-NV: L'interprétation d'un auteur que vous devez connaître -Paul Glimcher...

-MP: Oui.

-NV: ... est la suivante: le grand changement dans les neurosciences contemporaines, c'est le passage à l'étude des actions motivées, orientées, par opposition avec les travaux plus traditionnels sur les actions ou comportements réflexes. Il y voit un changement de paradigme important en neurosciences. J'ai eu l'impression que vous suiviez un peu cette interprétation dans votre présentation à Aix-en-provences. Vous distinguiez également les actions réflexes du type pavlovien des *goal intended actions*. Dans un de vos articles publié dans *Brain* en 2003, vous essayiez de montrer que les troubles liés à la maladie de Parkinson, chez le singe, n'étaient pas tant des troubles moteurs mais plutôt des troubles du *learning*. Les singes atteints possèdent toujours un certain répertoire d'actions motrices, mais ils ne parvenaient plus à sélectionner le mouvement adapté, ils avaient également des difficultés dans la correction des erreurs, dans le *feedback*... On retrouve cette idée d'un niveau qui n'est ni du sensoriel ni du moteur, intermédiaire entre les deux. Dans votre article de 2005 paru dans le *Journal of Neurosciences*, vous montrez que l'intoxication des singes au MPTP n'a pas pour conséquence d'inhiber le cortex moteur, mais plutôt d'introduire une corrélation excessive dans l'activité neuronale, c'est à dire de diminuer la spécialisation fonctionnelle. Cela pourrait donc aussi correspondre à une difficulté à choisir un programme moteur parmi un répertoire d'actions motrices. J'ai l'impression qu'il s'agit là d'une idée récurrente dans ces études: on est dans un domaine qui n'est ni sensoriel ni moteur, qui concerne le choix entre plusieurs programmes possibles, dans la préparation à l'action. J'ai une première question: est-ce vraiment une nouveauté? Il s'agit d'une question assez naïve, je ne suis pas spécialiste en neurophysiologie, mais on pourrait dire, par exemple, que Broca avait déjà identité au XIXème siècle une fonction cérébrale associée au langage, et on ne se situe pas du tout ici dans le domaine des «réflexes». Est-ce-que l'affirmation de Glimcher selon laquelle «avant, les neurosciences, ce n'était que du réflexe» est-elle fondée?

-MP: Je ne comprends pas très bien cette affirmation. Pour moi, les réflexes, cela remonte au début de siècle, et aux travaux Sherrington. En revanche, quand vous avez reformulé en disant « on n'est plus au niveau moteur ou perceptif, mais dans cet espace entre les deux qui permet de faire des choix », cela m'a un peu plus parlé. C'est vrai si on considère l'imagerie cérébrale. Au fond, cela revient à dire que la vraie nouveauté, c'est l'imagerie cérébrale. Mais si l'on considère la neurologie, ce que vous disiez est vrai: les neurologues depuis longtemps s'intéressent à des troubles qui ne sont ni moteurs ni perceptifs et qui ont à voir avec des fonctions élevées, comme le raisonnement, la planification, le langage, la mémoire, etc. La neuropsychologie a un bon siècle, même plus maintenant. Cela fait très longtemps que les neurologues caractérisent les patients en termes de fonctions intellectuelles supérieures et pas en termes de fonctions sensorielles ou motrices. L'imagerie cérébrale permet aujourd'hui de rechercher les bases neuronales de fonctions supérieures chez l'homme.

-NV: Ce que vous dites, c'est que l'imagerie cérébrale a permis de localiser plus précisément ces fonctions.

-MP: On fait un peu mieux que de la localisation quand même. Si l'on ne faisait que localiser,

on ferait au fond, avec une technique différente, la même chose que les neurologues font avec les lésions: il y a une lésion ici, et il y a un déficit, par exemple un trouble du langage, alors cette région est impliquée dans le langage... Il s'agit, schématiquement, du raisonnement des neurologues. Traduit dans les termes de l'imagerie, cela donnerait: si je demande à mon sujet de faire une tâche de langage, de réciter un poème par exemple, et que cette région s'active, alors la fonction de cette région est de produire la parole ou le langage. En fait, l'imagerie ne se limite pas à la localisation des fonctions cérébrales: d'une part, on arrive à étudier les dynamiques au cours du temps; et, d'autre part, et à voir les mécanismes d'interaction d'une région avec une autre.

-NV: C'est aussi quelque chose que l'on peut faire avec des microélectrodes.

-MP: Oui. Effectivement, il faudrait parler d'imagerie au sens large du terme, et cela inclurait aussi les techniques d'enregistrement des activités unitaires. Dans ce cas, c'est beaucoup plus vieux que les années 1990. L'apparition de l'IRM fonctionnelle remonte aux années 1990. Par contre, l'électrophysiologie est beaucoup plus ancienne. Dans les années 1960, il y avait déjà des travaux qui ressemblent à ceux qu'on fait maintenant. Ce qu'on peut dire, et c'est peut être cela que Glimcher avait en tête, c'est que chez le singe, il y a certaines fonctions que l'on ne peut pas étudier, par exemple le langage ou la conscience, on n'est pas sûr que cela ait un sens chez le singe... En ce sens là, oui, on arrive à des fonctions plus élevées. Il me semble que les mécanismes de prise de décision que l'on étudiait chez le singe concernaient plutôt la sélection de l'action, ou alors des décisions perceptuelles -est ce que c'est un stimulus ou un autre- et pas le type de décisions qu'on utilise en neuroéconomie, où il n'y a pas de bonnes et de mauvaises réponses, mais où chacun va se déterminer en fonction de valeurs qui lui sont propres.

-NV: Ces études concernent principalement le système dopaminergique, en tout cas l'idée d'un espace intermédiaire entre le sensoriel et le moteur semble s'appliquer de manière privilégiée au circuit de la récompense.

-MP: Non, il y a également des études sur le cortex préfrontal. Le cortex préfrontal a pour caractéristique d'avoir des activités qui ne sont directement liés ni aux stimuli, donc à l'entrée sensorielle, ni à la sortie motrice. Par exemple, on a montré que le cortex préfrontal s'active pendant des délais où on doit retenir des informations. En l'absence de tout stimulus externe, on peut caractériser l'activité du cortex préfrontal dorsolatéral. Le cortex préfrontal ventral encode ,ou reflète, ce que l'on pourrait appeler les « valeurs »: à combien de récompense peut on s'attendre dans un contexte donné, combien un tableau plaît, combien on désire telle ou telle chose... Il s'agit donc de choses qui ont à voir avec les goûts personnels du sujet, et qui sont ni dans le stimulus, ni dans la motricité. Il ne faudrait donc pas se restreindre au système dopaminergique, mais y inclure tous les circuits frontaux-striataux-ventraux, que l'on appelle parfois limbiques, donc depuis le cortex orbito-frontal, striatum ventral, palidium ventral, et puis retour par le thalamus. Tout cela est innervé par les neurones dopaminergiques. Tous ces circuits concernent la neuroéconomie.

-NV: J'imagine qu'il devait y avoir déjà des connaissances physiologiques, anatomiques, sur ces circuits, antérieurement à l'apparition de l'électrophysiologie. Qu'est-ce qu'ont apporté précisément ces techniques? Est-ce que les connaissances anatomiques antérieures étaient si éloignées de la modélisation des dynamiques d'apprentissage?

-MP: Pour moi, l'anatomie est une sorte de connectique: cela permet de tracer de voies, de

voir quelle région est connectée avec quelles autres, donc d'identifier la structure, l'architecture des connexions neuronales. En revanche, l'anatomie ne renseigne pas sur la fonction. C'est pour cela que l'électrophysiologie a été utile, en étudiant les corrélations: est-ce que l'activité du neurone est davantage corrélée à tel ou tel paramètre de l'environnement, ou tel ou tel paramètre de la sortie motrice; et cela jusqu'à l'étude des récompenses.

-NV: Mais le rôle de la dopamine était-il déjà connu par exemple?

-MP: Il y a eu plusieurs étapes. Il y a eu d'abord la découverte du rôle des neurones dopaminergiques dans la maladie de Parkinson, par Friedrich Hassler en 1938. Il y a une très ancienne tradition d'anatomo-pathologie -l'étude des cerveaux *post-mortem*- en la matière. Ensuite, dans les années 1960 ou 1970, on a découvert les thérapies et médicaments dopaminergiques qui pouvaient améliorer les symptômes parkinsoniens. Ceci a développé l'idée selon laquelle la dopamine intervient dans les processus moteurs, parce que la maladie de Parkinson entraîne des pathologies motrices. Les malades sont rigides, ils tremblent, donc ils ont des symptômes moteurs. Maintenant on se rend compte que ce n'est pas vrai: les parkinsoniens ont aussi des troubles d'ordre affectif. Ils sont souvent apathiques, déprimés, *etc.* La découverte du lien entre dopamine et récompense a été fortuite. Il y avait aussi un autre faisceau d'évidence ou de preuves: toutes les drogues dites récréatives -je ne vais pas dire qui donne du plaisir- mais qui génèrent des expériences plaisantes, et qui jouent sur le système dopaminergique. Il y avait ces deux idées là pour la dopamine. D'une part, on disait, un peu rapidement, « c'est l'hormone du plaisir, parce qu'on savait que les drogues jouaient sur le système dopaminergique », et, d'autre part, la dopamine était associée la motricité, à cause de la maladie de Parkinson. Le lien entre les deux a été fait par l'équipe de Wolfram Schultz en 1983. Convaincu que la dopamine intervenait dans la motricité, il a été l'un des premiers à être capable d'enregistrer l'activité des neurones dopaminergiques. Il a cherché à corréler l'activité de ces neurones avec tous les paramètres du mouvements, avec l'amplitude, la vitesse, la direction du mouvement. Il n'a jamais rien trouvé, sauf que pour faire travailler les singes, il leur donnait des récompenses à la fin. Au bout d'un moment, il a remarqué qu'à chaque fois qu'il donnait la récompense au singe, le neurone s'activait. Il faut savoir que les électrophysiologistes se guident à l'oreille. Le train de *spikes,* qui est en fait une différence de potentiels, est converti en son. Les chercheurs installent un haut-parleur dans la pièce, et ils écoutent la décharge des neurones. Schultz a donc remarqué que le neurone s'activait à la récompense, et il a remarqué ensuite que le neurone s'activait à la récompense quand elle était inattendue. Il a alors élaboré le concept d'erreur de prédiction de la récompense, c'est à dire la différence entre ce que l'on obtient et ce que l'on attend. Il a fait ensuite le lien avec les algorithmes d'apprentissage, qui utilisaient justement ce signal -l'erreur de prédiction de la récompense- pour faire apprendre des systèmes artificiels. A partir de l'idée selon laquelle la dopamine code l'erreur de prédiction de la récompense, et que ce signal peut conduire l'apprentissage, y compris l'apprentissage de procédures motrices, il est possible de faire le lien entre dopamine impliquée dans le plaisir et dopamine impliquée dans la motricité. Je suis quand même d'accord avec votre constatation: il y a peu de systèmes qui sont aussi bien caractérisés, surtout d'un point de vue formel, que le système dopaminergique. De très nombreuses études enregistrent des activations, aussi bien en électrophysiologie qu'en IRM fonctionnelle, dans le cortex préfrontal. On sait que ces activations sont susceptibles de coder des valeurs, des probabilités, des délais... Mais les études ne sont pas toutes cohérentes et sont loins d'être aussi bien formalisées que le comportement des neurones dopaminergiques, parce que ce dernier est assez stéréotypé. Dans ces structures, tous les neurones réagissent à peu près de la même façon à la récompense, ils ont à peu près tous une décharge au même moment. Il s'agit finalement d'un système assez simple.

-NV: J'ai l'impression qu'il y a en fait deux types d'études. D'un côté celles sur les singes par exemple, ce qu'ont fait Schultz ou Glimcher, sur le mouvement oculaire, sur l'encodage par les neurones dopaminergiques de l'amplitude et la probabilité d'une récompense, puis l'erreur de prédiction. A partir de ça, on fait des algorithmes d'apprentissage... Vous avez travaillé là dessus: dans votre article paru en 2006 dans Nature, vous étudiez justement l'effet du L-Dopa et du Haloperidol sur l'erreur de prédiction, en faisant jouer des sujets à des loteries très simples. Il s'agit d'apprentissage dans un sens assez étroit du terme: ajuster des prévisions au fur et à mesure. De l'autre côté, des études sur l'apprentissage dans un sens plus complexe, que vous étudiez également, qui concerne le problème du choix dans un répertoire d'actions motrices. C'est ce que vous évoquiez à propos de la maladie de Parkinson: la difficulté à choisir parmi un répertoire d'actions motrices déterminées. Ces deux notions d'apprentissage sont-elles conciliables?

-MP: Les deux s'intègrent assez bien. Si vous vous donnez un choix binaire, le modèle économique le plus basique nous dit qu'il faut assigner une valeur à chaque objet, comparer les deux valeurs et choisir la plus haute. Après, pour savoir comment le cerveau apprend les valeurs de chaque option, on peut utiliser l'algorithme d'apprentissage basique. A chaque fois qu'une option est choisie -chaque option ayant une certaine valeur- à partir du résultat, le calcul de l'erreur de prédiction permet d'actualiser la valeur de l'option. Si le choix est répété, les valeurs seront donc modifiées. Les décisions sont ainsi améliorées d'un essai à l'autre.

-NV: Le choix de l'algorithme que l'on décide d'appliquer aux données est donc extrêmement important. Est-ce là que se joue l'innovation scientifique dans cette littérature? Ou que se fait la distinction entre un bon et un mauvais article?

-MP: Les modèles computationnels appliqués à l'analyse de données sont assez récents, et ont dix ans environ. Tous les chercheurs n'ont pas cette approche, et, en particulier, en électrophysiologie, il y a assez peu de groupes qui le font. En IRM, ces modèles commencent à se répandre. Effectivement, il peut y avoir un problème: le choix de l'algorithme peut conditionner le résultat. Typiquement, en IRM fonctionnelle, en utilisant un certain modèle, on peut trouver des activations différentes selon le modèle. La fonction du modèle consiste à générer des dynamiques. Certaines variables sont retenues, qui sont en fait des variables cachées de la décision. Le modèle est optimisé, c'est à dire que les paramètres libres du modèle sont ajustés pour que le modèle reproduise au mieux les données comportementales, et ensuite ce modèle optimal est utilisé pour générer des dynamiques. Ces dynamiques sont ensuite corrélées aux données d'imagerie. Par exemple, on peut prendre l'erreur de prédiction: le modèle, à chaque moment où tombe ou ne tombe pas la récompense, détermine l'erreur de prédiction théorique. Ce processus est répété au fil des essais. On obtient un vecteur, qui donne la succession des erreurs de prédiction au fil des essais. La dernière étape consiste à chercher dans les données d'imagerie les *voxels* dont le signal BOLD va corréler au signal théorique. Mais évidemment, à partir d'un autre modèle ou d'autres variables, d'autres activations seront trouvées. Le modèle sert donc essentiellement à générer des dynamiques, utilisées comme régresseurs pour caractériser les activations cérébrales. Tout ce que l'on peut dire, c'est que cette variable est suffisante ou non pour résoudre le problème. Alors, évidemment, s'il s'agit d'une tâche élémentaire, ce n'est pas compliqué. Pour résoudre une tâche d'apprentissage instrumental, il suffit d'avoir une règle de décision, qui peut être très simple: par exemple, choisir toujours la valeur la plus grande. Cela suffit. On cherche où est codée cette variable dans le cerveau; on trouve des régions qui effectivement expriment cette variable. On n'a pas démontré pour autant que le cerveau s'appuyait effectivement sur cette

variable pour prendre des décisions comme dans le modèle, mais on peut dire que cette région cérébrale exprime cette variable qui est suffisante en principe pour être la base de la décision.

-NV: Quelle est en général cette règle de décision extrêmement simple?

-MP: Cela varie selon les gens, mais il s'agit en général d'un *softmax*. C'est une règle qui permet d'ajuster ce que l'on appelle « température du choix » ou « balance exploration sur exploitation ». En effet, lorsque l'on estime les valeurs des deux options, on peut soit *exploiter* cette connaissance -c'est à dire choisir l'option avec la plus forte valeur- soit continuer à *explorer* un peu, c'est à dire prendre l'option qui a la valeur la plus faible, parce que les contingences ont peut être changé, ou parce que l'on s'est peut être trompé dans les estimations. Cela peut être bénéfique à long terme de continuer à explorer un petit peu, de ne pas se limiter à l'exploitation. C'est pour cette raison que l'on ne retient pas la règle la plus simple, mais une règle qui pondère avec une température la stratégie d'exploration ou d'exploitation.

-NV: On peut aussi intégrer le jugement contre-factuel, que certains appellent regret.

-MP: Oui. Mais cela dépend des protocoles, l'information n'est pas toujours disponible: pour produire un jugement contre-factuel, il faut avoir accès à ce qui se serait passé si l'autre option avait été choisie.

-NV: Finalement, il s'agit de neurosciences computationnelles, c'est à dire d'une approche en neurosciences fondée sur les sciences cognitives. Cela rejoint ce disait Paul Glimcher. Selon lui, le pionnier en ce domaine est David Marr. Êtes vous d'accord là dessus?

-MP: En neurosciences computationnelles, David Marr est une référence. Mais je pense aussi à d'autres études, notamment les modèles connexionistes de Marvin Minsky, qui doivent être plus vieux. Il y a fondamentalement deux classes de modèles: les modèles connexionistes, qui utilisent des réseaux de neurones, et les modèles computationnels, avec des boîtes et des flèches, et des opérations logiques entre les différents compartiments. En ce qui concerne les modèles computationnels, Paul Glimcher a probablement raison.

-NV: Quelle distinction faites vous entre modèles connexionistes et computationnels?

-MP: Dans un modèle connexioniste, on utilise un réseau de neurones. La tâche n'est pas décomposée en une succession d'opérations logiques, mais le réseau de neurones résous simplement la tâche. Il s'agit d'une forme d'apprentissage par modification des voies synaptiques, jusqu'à cela converge vers la solution qui est adaptée. Cela peut être efficace: il y a des réseaux de neurones partout, dans les distributeurs de billets des banques, qui retrouvent les codes. Ce sont des algorithmes qui sont assez efficaces, mais qui ne sont pas du tout opérationnels pour la compréhension scientifique parce qu'ils ne permettent pas du tout de décomposer les tâches en fonctions plus élémentaires. La référence pour les modèles computationnels est la machine de Türing, associée à la décomposition de tâches en séries d'opérations logiques.

-NV: Pourquoi la vision comme objet d'étude privilégié? Cela apparaît chez David Marr, chez Paul Glimcher aussi... Qu'y a-t-il de pratique dans le fait d'étudier la vision et pas une autre fonction?

-MP: La très grande majorité des stimuli auxquels nous sommes exposés sont des stimuli visuels. Il est probablement vrai que les primates se basent prioritairement sur la vision par rapport aux autres sens. Ceci s'explique peut être aussi par des contingences historiques, la vision étant le système qui a été le plus étudié, dont on connait le mieux les propriétés, et donc qui était le plus à même d'être modélisé.

-NV: Est ce que ces modèles neurocomputationnels ont une utilité dans le traitement des patients?

-MP: C'est une bonne question. On commence à utiliser les modèles cognitifs ou de neuroéconomie pour caractériser les déficits chez les patients, et les effets des traitements. Cela fonctionne pour certains troubles, comme l'impulsivité. Ainsi, un patient ayant une affection du cerveau déterminée a une variable du modèle modifiée: par exemple, le taux d'apprentissage, ou le taux d'exploration. Si l'on considère que l'impulsivité est la préférence pour les récompenses à court terme, alors avec un seul paramètre, qui est le *discount factor,* il est possible de caractériser l'impulsivité de ce patient. Ensuite, l'effet des traitements peut être amélioré: en donnant une certaine drogue à ce patient, les paramètres du modèle computationnel pourront être ajustés. Le patient sera moins impulsif...

-NV: Mais, en ce qui concerne par exemple l'impulsivité, on s'appuie sur un algorithme standard, qui reflète la moyenne des sujets...

-MP: Les études commencent tout juste. Il n'y a rien de tel qu'un modèle computationnel validé, trans-nosographique, qui permette de caractériser les déficits et les traitements. Ce que je suggère ici, c'est une sorte d'horizon. On n'en est pas encore là. Mais notre équipe s'inscrit dans cette approche: les algorithmes utilisés pour modéliser les activations en IRM servent également à caractériser le comportement des patients.

-NV: Il y a donc une possibilité d'application de ces études comme tests psychométriques

-MP: Oui. Cela peut être aussi utilisé comme marqueur pour le suivi des maladies ou des traitements par exemple.

-NV: La deuxième forme d'application concernerait une amélioration du dosage des traitements neuropharmacologiques?

-MP: Oui.

-NV: Cela renvoie à l'un de vos articles, où vous étudiez les effets du L-Dopa et du haloperidol. A la fin, vous concluez en disant que cette étude permettrait de mieux traiter les patients traités au L-Dopa...

-MP: J'ai du dire « mieux comprendre ». « Mieux traiter », je trouve cela un peu ambitieux. Je pensais précisément à une étude chez les patients que l'on a faite et qui vient de sortir dans *PNAS*. Effectivement, cela a fonctionné comme on le pensait. Les modèles ont pour vertu principale d'opérationnaliser les concepts. Très souvent, la neurologie et la psychiatrie sont limitées à des descriptions verbales un petit peu floues, un petit peu fleuries. Les concepts sont par conséquent un peu fluctuants, et l'on ne sait pas toujours de quoi excatement on parle. En faisant rentrer les concepts dans les équations, ceux-ci deviennent tout de suite beaucoup plus précis. Le deuxième avantage est la quantification. Cela permettrait donc d'améliorer

d'une part la description des maladies, et d'autre part la prescription des traitements.

-NV: Et il n'y aurait pas d'application pour les sujets sains?

-MP: En général, les journalistes posent la question de l'application dans le domaine de l'éducation. En général, je dis qu'il n'y a pas d'application. Je ne veux pas prendre de risque. Par ailleurs, y compris pour les patients, je réponds qu'il n'y en a pas. Parce que ce dont on vient de parler -utiliser un modèle computationnel pour un diagnostic ou prescrire un traitement- renvoie au très long terme. En ce qui concerne l'éducation, on pourrait imaginer également caractériser l'évolution des capacités cognitives des enfants avec des modèles, pourquoi certaines stratégies pédagogiques sont meilleures que d'autres, etc. Mais le même problème se pose: l'horizon est très lointain.

-NV: Les applications sur sujets sains semblent donc plutôt spéculative.

-MP: Oui.

-NV: Ces études débouchent souvent sur des recommandations de psychologie de sens commun. Je pense à un article récent de Drazen Prelec. C'est un jeu d'investissement, et il montrait que le fait de mieux gérer ses émotions, de se mettre dans la peau d'un trader, bref ce qu'il appelle « *emotional reappraisal* » permettait justement de mieux gérer ces problèmes d'apprentissage. J'ai l'impression qu'il s'agit du seul type de débouché «pratique», ce genre de recommandation psychologique assez floues.

-MP: Oui, mais je ne sais pas si cela est très spécifique à la neuroéconomie. Cela se généralise à tous les résultats de psychologie expérimentale, dans la mesure où celle-ci peut fournir des outils pour s'entraîner. Dans le passé, on a déjà vu cela avec la mémoire, la perception... Cela peut donner des stratégies de rééducation aussi chez les patients. Dans les cas de troubles de la mémoire, on peut faire des exercices, entraîner sa mémoire. En ce qui concerne l'orthophonie également, on peut entraîner la lecture, la prononciation, *etc.*

-NV: Par contre, ces applications dont on vient de parler, que Glimcher qualifie de « neuropsychiatrie computationnelle », seraient appuyées sur des algorithmes «moyens».

-MP: Oui, les différentes équipes convergeraient à terme vers un algorithme moyen, qui serait ensuite disponible comme outil dans la pratique médicale.

-NV: De ce point de vue, n'y-a-t-il pas une dépendance vis à vis des catégories médicales, dans la mesure où la définition du « bon » programme d'apprentissage suppose la mise en évidence de programmes « pathologiques »? On peut regarder les études dans le domaine d'une manière neutre, en considérant la mise au point d'algorithmes, mais d'un autre côté apparaît souvent l'idée selon laquelle il y aurait des « bons » et des « mauvais » programmes d'apprentissage. Y-a-t-il donc un assujettissement aux catégories du normal et du pathologique?

-MP: Vous posez le problème du normatif: faut-il considérer comme bon le comportement normal? Vous avez peut être raison En même temps, les algorithmes utilisés en économie permettent d'identifier un critère externe d'optimalité. On peut dire quelle est la meilleure règle d'apprentissage pour un ensemble de contextes donnés. Cela donne au moins une deuxième chance, disons une deuxième normativité. La norme peut donc être définie soit

comme le comportement moyen des sujets qui ne sont pas malades, ou qui n'ont pas été déclarés comme malades, soit comme comportement optimal, qui permet alors de s'évader un petit peu du comportement commun.

-NV: Mes dernières questions concernent les critiques méthodologiques qui apparaissent souvent dans la littérature. Un premier problème est soulevé par Harrison. Il affirme que la *voxel-based morphometry* est beaucoup moins fiable que les méthodes anatomiques traditionnelles. Cela renverrai au problème du « *pooling accross brains* ». J'ai deux citations que je souhaiterais vous faire lire.

-MP: La VBM est assez peu utilisée en fait. Cela consiste à faire des corrélations entre anatomie et structure du cerveau, c'est à dire, schématiquement, entre la quantité de matière de grise dans différentes régions du cerveau et des variables cognitives qui peuvent être dérivées de jeux économiques. Il y a beaucoup de problèmes méthodologiques qui sont relativement connus. Mais de toute façon, il est toujours possible de mettre en évidence un lien, c'est à dire une corrélation; par exemple, ceux qui ont le plus gros hippocampe vont donner le plus d'argent dans le jeu du dictateur, mais ce ne sont pas des corrélations qui vont permettre d'améliorer la compréhension des processus sous-jacents aux décisions telles qu'elles sont prises dans le jeu du dictateur. [lecture des citations] Harrison pointe ici le problème du cerveau moyen et du type de déformations qu'on utilise pour ramener le cerveau du sujet sur le cerveau moyen, avec toutes les approximations que cela comporte. La deuxième citation est liée. Effectivement, si l'on segmente le cerveau à la main -par exemple, on peut segmenter l'hippocampe- la précision sera plus grande qu'avec un algorithme qui le fait automatiquement. L'avantage de la segmentation automatique, c'est que cela va beaucoup plus vite. Des centaines de cerveau peuvent être traités assez rapidement. Et, d'autre part, le chercheur est à l'abri des hypothèses, d'une certaine manière: ce n'est pas lui qui essaie de démontrer une hypothèse, par exemple sur la taille de l'hippocampe, mais c'est l'ordinateur. Donc, s'il y a un biais, il n'est pas nécessairement en faveur de l'hypothèse. Ceci concerne l'IRM anatomique. Mais il y a aussi des critiques relatives à l'IRM fonctionnelle.

-NV: Oui, justement, en ce qui concerne le traitement statistique des données, il semble y avoir beaucoup de simplifications... [deuxième série de citations]

-MP: Ce sont des remarques de bon sens. C'est un piège général, mais il est vrai que les résultats d'imagerie cérébrale sont généralement surconsidérés et surinterprétés. Lorsqu'un *voxel* significatif est identifié, la tradition veut, bien que ce soit arbitraire, que le seuil de significativité soit à 5%. En d'autres termes, si un *voxel* est plus activé que ce qu'il aurait été par hasard, avec une chance de 5%, on considère qu'il est significatif. Bien sûr, comme il y a des milliards de *voxels* dans le cerveau, si l'on ne corrige pas pour les comparaisons multiples, c'est à dire pour le fait que le même test sera utilisé pour tous les *voxels*, il y aura nécessairement 5% de faux positifs. Il ne faut pas prendre cela pour des activations. Il s'agit cependant d'un biais dont tout le monde est conscient. Mais les gens corrigent plus ou moins bien pour les comparaisons multiples, ils ont des seuils plus ou moins astringents. Il faut donc effectivement faire attention. En plus, comme ces citations le soulignent, le résultat est dépendant de toute une série de traitements statistiques, et non-statistiques: il y a aussi les traitements temporels et spatiaux que l'on fait subir à tes données brutes. Les données sont balisées par exemple avec une gaussienne qui a une largeur plus ou moins grande, et cela va aussi influencer tes résultats. Je pense que le véritable critère de fiabilité est la réplication. Certains résultats sont indiscutablement robustes en IRM fonctionnelle.On peut faire des légères variations dans le paradigme, dans le traitement des données, mais ces mêmes

résultats seront toujours retrouvés. Toutefois, si l'on devait compter les résultats robustes en IRM fonctionnelle, deux mains suffiraient. Le temps fera le tri.

-NV: Le progrès de la connaissance dans ce domaine s'appuie donc sur la convergence des résultats expérimentaux. On pourrait donc répondre à ces critiques par le critère de réplicabilité des résultats. Il y a une autre critique qui est souvent faite, c'est la non-présentation des données statistiques, ou en tout cas leur difficulté d'accès.

-MP: Ce n'est pas tellement vrai, en tout cas ce n'est pas propre à l'imagerie. Il me semble que la communauté de l'imagerie a plutôt une bonne maturité, une bonne réflexion statistique. Comme de nombreux problèmes se posent, et qu'il y a clairement des abus, les chercheurs cherchent naturellement à mettre des garde-fous. Par rapport à d'autres disciplines, la réflexion est donc plutôt bonne. Maintenant, il y a évidemment des grosses différences entre les gens qui travaillent bien, et ceux qui travaillent pas bien. C'est en partie lié à l'existence de logiciels d'analyse relativement simples à utiliser, des « presse-boutons » aisément disponibles. Des gens mal informés, en tout cas qui ne sont pas au courant des opérations qui se cachent derrière les boutons, peuvent utiliser ces logiciels facilement, et ainsi obtenir et interpréter des images du cerveau.

-NV: C'est la critique la plus fréquemment adressée à la neuroéconomie: on met les sujets dans le scanner et on commente des activations, sans se préoccuper des simplifications en amont. Mais vous par exemple, possédez vous les compétences nécessaires pour comprendre intégralement le fonctionnement des machines et des logiciels?

-MP: Non, pas complètement. Je pense que personne ne comprend tout. Il n'est pas nécessaire d'être physicien nucléaire pour comprendre l'imagerie, même si c'est en fait la base du signal BOLD. C'est important de comprendre les principes de l'analyse, des corrections statistiques, sans forcément connaître tous les boutons du logiciel et ce qu'il y a derrière. De toutes façons, les logiciels sont tellement compliqués maintenant que personne ne peut tout connaître...Je pense que vous ne connaissez pas toutes les fonctionnalités de Word, mais vous pouvez quand même l'utiliser? C'est une question qui n'est pas facile à trancher: jusqu'où il faut connaître le logiciel pour pouvoir l'utiliser? Je pense qu'il faut distinguer entre des problèmes qui sont propres à l'imagerie et des problèmes qui ne le sont pas. Par exemple, le problème des comparaisons multiples n'est pas propre à l'imagerie. Celui de l'indépendance, qu'on a souvent pointé, l'est également. Cela consiste à sélectionner les données selon un critère qui est lié à l'effet que l'on veut montrer. Typiquement, on cherche une région qui est corrélée à un facteur personnalité, ou un déficit clinique, un score de dépression par exemple. Le chercheur scanne cinquante sujets -chaque sujet a son score de dépression- et identifie des corrélations. En d'autres termes, on cherche les *voxel* qui, d'un cerveau sur l'autre, vont corréler avec le score de dépression. On trouve un ensemble de *voxels*. Ensuite, on prend le coefficient de corrélation dans ces *voxels*. Et, en fait, de cette manière, on sélectionne des *voxels* dans lesquels le hasard, c'est à dire le bruit gaussien qu'il y a dans tes données, favorise la corrélation. Le coefficient de corrélation est ainsi surestimé. Il ne faut donc pas tenir ce coefficient obtenus dans les *voxels* sélectionnés comme une bonne estimation de la population en général. Quelles que fois, des résultats sortent ainsi, dans lesquels les auteurs affirment par exemple que telle région du cerveau explique 95% des traits psychotiques, laissant entendre qu'on a trouvé la cause aux comportements psychotiques. En tout cas, cela suggère que le volume d'une région explique 95% de la variance de la population en termes de personnalité psychotique. Ce qui n'est absolument pas le cas, parce que ce sont les *voxels* qui corrèlent qui ont été sélectionnés, donc la corrélation est surestimée. Il s'agit d'un biais typique, qui n'est

pas propre à l'imagerie. En électrophysiologie par exemple, les gens sélectionnent les neurones qui répondent à la tâche. C'est un cas typique de non-indépendance. Le critère de sélection doit être indépendant de ce qui est recherché. En revanche, ce qui est assez typique de l'imagerie, et de la neuroéconomie, c'est le problème de l'inférence inverse. Ce problème est assez lié au fait que les économistes surestiment les résultats des neurosciences. Ils pensent que, parce que telle région s'activait lors de telle tâche, cela veut dire que la fonction de la région est d'opérer cette tâche. Ils vont faire des inférences inverses: parce qu'ils ont lu quelque part que la région R était activée pendant le processus P, alors si dans leur manipulation la région R est activée, ils vont inférer que le sujet engage le processus P. C'est vrai si et seulement si la seule fonction de la région R est d'opérer le processus P. Mais ce n'est pas forcément vrai. Ainsi, en lisant certaines études, où il y a par exemple une activation de l'*insula*, certains économistes vont dire « le sujet est dégouté », parce qu'il se trouve que l'*insula* était activée quand on présentait des images dégoutantes aux sujets. Mais cela n'a jamais voulu dire que la seule et unique fonction de l'*insula* était de générer le dégoût.

-NV: Finalement, il ne doit pas y a voir beaucoup de chercheurs « sérieux » qui raisonnent ainsi?

-MP: Certaines études manifestent des cas flagrants d'inférence inverse. Mais il y a également des articles qui pointent ce genre de biais, comme ceux que vous avez cité pour les problèmes méthodologiques, et qui listent les bonnes pratiques et les erreurs à éviter.

-NV: Une dernière question: dans vos recherches, vous utilisez l'imagerie et les microélectrodes. Qu'est-ce qui guide votre choix entre ces deux techniques?

-MP: J'ai utilisé l'électrophysiologie pendant ma thèse, chez les singes. Chez les patients, on profite des électrodes implantées pour raisons thérapeutiques, pour la stimulation cérébrale profonde, pour enregistrer ce que l'on appelle des potentiels locaux. Ce sont des électrodes qui sont dans les structures profondes du cerveau. Ces potentiels locaux représentent schématiquement la somme des potentiels synaptiques. Pour l'instant, nous nous servons assez peu de l'électrophysiologie de surface, EEG ou MEG. Mais nous seront amenés à utiliser ces techniques, parce qu'elles sont complémentaires. L'IRM a une bonne résolution spatiale mais une très mauvaise résolution temporelle, alors que l'EEG et la MEG ont une très bonne résolution temporelle. Mais ce sont des potentiels que l'on enregistre en surface, et la résolution spatiale est par conséquent moins précise.

-NV: Un résultat validé par plusieurs techniques, c'est l'objectif?

-MP: Plusieurs laboratoires, plusieurs paradigmes, plusieurs techniques: je pense que cela constitue l'objectif pour avoir inductivement un effet qui soit robuste au travers de plusieurs de ces variations. Des résultats comme cela, qui tiennent et ne seront plus remis en question, sur des décennies, en IRM fonctionnelle, il n'y en a pas beaucoup ».

-Entretien avec Luc Mallet, 14/09/2010

-NV: « Vous êtes à la fois psychiatre et chercheur à l'Inserm. A ce titre, vous dirigez l'équipe de recherche « Comportement, Émotions et Ganglions de la base ». Vous partagez donc votre temps entre médecine clinique et recherche ?

-LM: On peut dire ça comme ça. Le thème de l'équipe consiste à essayer de comprendre comment un système cérébral –en l'occurrence, un système « profond », les ganglions de la base- traite des informations de type « émotionnel ». Les ganglions de la base sont impliqués dans la physiopathologie d'un certain nombre de troubles comportementaux, qui donnent lieu à des troubles psychiatriques. L'identification de ce traitement d'information, c'est-à-dire du fonctionnement et du dysfonctionnement de ces structures profondes en pathologie, a pour but de développer des traitements innovants, pour essayer d'agir sur le fonctionnement de ces structures, par le biais d'implantation d'électrodes –ce sont des techniques chirurgicales qui visent à implanter des systèmes qui vont moduler l'activité de ces structures et leur fonctionnement. Ce type d'intervention est réservé à des pathologies graves et résistantes. Ceci constitue le volant clinique de nos recherches : s'intéresser aux maladies psychiatriques qui résistent au traitement médical, à toutes les prises en charge, qui sont marquées par un certain nombre de désordres comportementaux, qu'on arrive à relier, à partir de données scientifiques, à un dysfonctionnement de ces structures. On arrive également à mettre en place des protocoles de recherche clinique pour voir si on a un intérêt à proposer ce type de traitement (implanter des électrodes). Ceci constitue la thématique de l'équipe. Dans cette thématique, on est amené à étudier chez l'homme des «modules élémentaires» du comportement, qui sont liés à des hypothèses fortes en termes de fonctionnement cérébral et de psychologie expérimentale. C'est ici que nos recherches convergent avec celles de Mathias Pessiglione : lui, il développe des concepts, des expériences de psychologie expérimentale qui sont liées à des hypothèses et des concepts forts sur le fonctionnement des mêmes structures. On est à la recherche de « paradigmes », que l'on va pouvoir tester chez l'homme, mais aussi chez l'animal, pour vérifier que cela soit bien lié au dysfonctionnement que l'on suppose. In fine, après avoir identifié tel dysfonctionnement dans telle pathologie –et si ce dysfonctionnement semble important, et est lié à cette structure- si on arrive à modifier l'activité de ces structures dans le bon sens par certaines interventions, alors on peut proposer ces interventions comme traitement. C'est un peu le cadre dans lequel on travaille. Là-dessus, la motivation, qui est l'aspect le plus proche de la neuroéconomie, est un concept central, parce qu'on peut «lire» un certain nombre de désordres psychiatriques par des anomalies de la motivation : soit des excès de motivation, soit des défauts de motivation, soit des déviances de motivation… C'est un concept très utile, clef, valise. A partir de là, évidemment, on est très intéressé par la caractérisation des systèmes cérébraux qui sont impliqués dans la motivation. C'est ici que l'on rejoint un peu la neuroéconomie, mais on est vraiment à l'extrémité.

-NV: Vous travaillez donc globalement sur les mêmes structures que Mathias Pessiglione.

-LM: On travaille sur des structures qui sont exactement les mêmes, qui sont notamment les ganglions de la base. Effectivement, on finit par retrouver vraiment des structures communes.

-NV: Vous partagez également cette approche, qui consiste à essayer de «quantifier» certains troubles neurologiques à partir de paradigmes de psychologie expérimentale, qui permettent d'isoler un trait de comportement.

-LM: Exactement.

-NV: Est-ce réellement une vraie nouveauté ? Je pense par exemple à une expérience de Mathias Pessiglione, dans laquelle on fait jouer des sujets à des jeux de paris financiers, pour modéliser par algorithmes leur comportement d'apprentissage. Ce genre de méthode est-il vraiment novateur ?

-LM: Le problème, c'est que l'on croie toujours que l'on connaît les choses, et puis on découvre après coup que l'on est très inculte : il y aura toujours des gens pour vous dire : « mais non, ça existe depuis très longtemps ». Mais, de cette manière, appliqué à la pathologie, cela me semble assez récent. Ce qui est nouveau, c'est de le faire avec la qualité que l'on a actuellement. L'idée d'isoler des composants élémentaires du comportement qui peuvent être pathologiques est incontestablement ancienne. Le fait de pouvoir le faire avec la possibilité de modéliser le comportement est plus récent.

-NV: On peut également aujourd'hui, avec des catégories et des concepts aussi larges que celui de « motivation », balayer assez large et saisir toute une série d'anomalies comportementales associées à des pathologies variées : Parkinson, Gilles de la Tourette, etc.

-LM: Disons que des caractéristiques du comportement de ces patients peuvent être lues en utilisant ce concept. Mais cela n'explique pas la totalité du patient, ni la totalité du trouble. Mais il est très pertinent d'envisager des particularités de leur comportement sous cet angle. Encore une fois, cette approche est liée à des hypothèses fortes sur le comportement et le fonctionnement de ces structures. Mais effectivement, ce qui est intéressant pour nous en tant que chercheurs, c'est que cela éclate complètement les catégories nosographiques : on ne se dit pas que l'on va étudier Parkinson, les TOC, Gilles de la Tourette, ou la dépression, mais, à partir d'une dimension qui est extrêmement bien caractérisée sur le plan comportemental, avec une modélisation possible, et qui en plus a l'intérêt d'être lié au fonctionnement de structures bien identifiées, on va s'apercevoir qu'il y a des anomalies et des différences de cette dimension comportementale, qui traversent le champ nosographique, et qui, soit font ressortir des éléments communs dans les dysfonctionnements, soit permettent de bien discerner certaines choses. Surtout, cela donne des pistes de travail, des pistes de thérapeutiques, c'est-à-dire sur des choses qui parfois ont des répercussions fonctionnelles assez importantes. Dans les troubles de la motivation, que vous avez évoqué à juste titre, Parkinson notamment, vous avez des phénomènes d'apathie qui peuvent être très handicapants pour les patients, et là, on a des pistes d'action intéressantes.

-NV: Cela correspond à ce que me disait Mathias Pessiglione : ces expériences permettent de remplacer les descriptions verbales, un peu floues, que l'on pouvait avoir en psychiatrie.

-LM: Oui, ces descriptions étaient verbales, mais avec des échelles qui ont été standardisées, qui ont fait l'objet de consensus entre les cliniciens du monde entier, avec des critères cliniques précis. Ces critères demeurent opérant en clinique, c'est-à-dire qu'ils permettent aux cliniciens de parler de la même chose avec les mêmes mots, et ils correspondent à une certaine réalité sur le terrain : les caractères évolutifs des patients, le fait que telle pathologie pourra répondre à tel traitement… Les travaux sur la motivation, ceux par exemple de Mathias Pessiglione, apportent un éclairage de plus, mais qui est orthogonal au précédent, très créatif sur le plan thérapeutique.

-NV: En même temps, Mathias Pessiglione se montrait assez prudent sur les possibles applications et débouchés thérapeutiques de ses recherches : cela permettrait éventuellement, par exemple, d'améliorer le dosage des médicaments.

-LM: Oui, il faut être très prudent, mais cela peut être un peu plus. On peut imaginer développer des techniques de réhabilitation cognitive, en utilisant des outils comme la réalité virtuelle. Si on identifie une composante qui est faible, mais pas non-existante, peut-être que l'on peut, par l'entraînement, restaurer un certain équilibre. Encore une fois, au-delà des médicaments, si on identifie des structures très précises dans ces systèmes profonds qui sont impliqués dans tel ou tel processus, cela donne des pistes pour aller cibler chirurgicalement, c'est-à-dire aller moduler directement l'activité de ces structures. Il y a quand même des pistes. Mathias a raison d'être prudent sur le délai d'application, qui ne sera pas sans doute très rapide.

-NV: De ce point de vue, vous ne voyez pas de décrochage entre les neurosciences fondamentales et la pratique médicale, entre les objectifs de recherche pure et les impératifs cliniques.

-LM: Si, bien sûr, il y a un décrochage, mais il a toujours existé. C'est amusant ce que vous dites parce que je dirais que la motivation et les concepts liés pourraient peut être –et j'insiste sur le «peut être»- remettre en cause ce clivage, en offrant un retour immédiat sur ce qu'on observe. Ces études ne sont pas complètement déconnectées : ce n'est pas comme ces études, où, par exemple, on met en évidence dans telle maladie neuro-dégénérative une anomalie moléculaire, avec l'accumulation d'une protéine anormale. Très souvent, ce genre de travaux est complètement coupé de la réalité clinique : il n'y a aucune possibilité de les raccrocher directement. De ce point de vue là, les travaux sur la motivation sont un petit peu différents.

-NV: Justement, il me semble que les travaux de neuroéconomie tendent à s'appliquer notamment au traitement des conduites addictives. J'ai ici par exemple un article associant neuroéconomie et addiction aux drogues [Bickel *et al.*, 2007], celui-ci avec les troubles du comportement alimentaire [Rowland *et al.*, 2008]; je ne sais pas si cela vous parle.

-LM: Je ne connais pas, mais cela ne m'étonne pas. Sur les addictions, bien sûr, et toutes les pathologies de l'impulsivité, il y a eu pas mal de choses, effectivement. Tout est lié en fait à la gestion du délai d'apparition d'une récompense : on peut voir l'impulsivité sous l'angle de l'incapacité à gérer ces processus d'appréciation d'une récompense attendue en fonction du temps. Il y a en effet un principe de base : plus une récompense est lointaine, plus elle perd de valeur. Mais on arrive à contrecarrer ce principe avec ce que l'on appelle «la capacité à voyager dans le temps» -time travel-, c'est-à-dire la capacité à se projeter dans le temps –une notion, je pense, empruntée également à la neuroéconomie- et qui permet de contrer ces effets de discounting de la récompense. Ceci est très intéressant, parce que cela va pouvoir sur le plan comportemental constituer une piste pour voir s'il n'y a pas des défauts de la capacité à voyager dans le temps et de se représenter les choses, qui permet justement d'annuler les effets du temps sur les récompenses attendues. Ce sont vraiment des pistes très créatives de recherche sur le comportement. Après, je ne suis pas spécialiste : je ne connais ces choses que de manière très superficielle.

-NV: J'ai l'impression, de l'extérieur, que le terme de «conduites addictives» peut s'appliquer à énormément de choses : le drogué, l'accroc aux jeux financiers, le boulimique,…Est-ce qu'on a pas une catégorie un peu trop large, très générique ?

-LM: Elle existait déjà. Cela fait longtemps que les psychiatres décrivent certaines choses par ce qu'ils appellent les « spectres ». Il y a les troubles et il y a les spectres : tel trouble est dans tel spectre, c'est une catégorie un peu plus large. Le spectre addictif décrit justement ce que vous venez de dire. Cela fait longtemps maintenant que l'on sait qu'il y a un *overlap* entre un certain type de troubles de conduites alimentaires, où il y a des conduites de boulimies, ce genre de choses, et les conduites addictives, les addictions aux drogues, aux jeux. Ce spectre addictif existait déjà cliniquement. Après, si des concepts neuroéconomiques viennent apporter de la consistance à ce spectre, cela peut être vraiment très intéressant, en apportant un pendant expérimental et scientifique à quelque chose qui avait quand même été un peu approché intuitivement par les cliniciens.

-NV: Cette capacité à voyager dans le temps que vous évoquiez me fait penser aux travaux de Damasio, et notamment à son étude des patients atteints de lésions du cortex préfrontal. J'ai envie de vous poser un peu la même question : le cas des lésions du cortex préfrontal, qui apparait systématiquement dans tous les ouvrages de neurosciences, avec l'histoire de Phineas Gage, n'est-il pas un idéal-type, ou plutôt une certaine manière assez frustre et simplifiée d'envisager les comportements intentionnels ? De la même façon, le terme d' «émotion», qui apparait dans l'intitulé de votre équipe de recherche, est une notion que Damasio et d'autres utilisent beaucoup. Mais comment définir ce concept ? N'est-ce pas un mot vide de signification ? Damasio affirme que la pensée est « incarnée », que l'émotion renvoie à une régulation homéostatique du corps : l'émotion est opposée systématiquement à la cognition. En même temps, la cognition est strictement incarnée, et il n'y a pas de division entre cognition et émotions. On peut donc se demander si finalement, ce n'est pas un concept vide, au sens contradictoire : les émotions sont ce qui s'opposent à la cognition «abstraite», notion à laquelle Damasio refuse pourtant toute réalité…

-LM: Oui, je suis assez d'accord. On a un peu de mal avec ces concepts. Il y a eu des choses très savantes qui ont été écrites là-dessus, des choses assez compliquées. Nous privilégions des approches assez opérantes : en utilisant des modèles pré-existants, le répertoire des émotions primaires par exemple, des affects… Mais cela reste assez vide, comme concept, par rapport à des concepts plus forts. C'est aussi ce que peut apporter la neuroéconomie, par le biais de la motivation, par exemple : des concepts plus remplis, plus solides, mieux déterminés. C'est une certitude ; historiquement, je pense que c'est ce qui va se passer. Après, Phineas Gage… C'est toute la théorie de Damasio, c'est un cas exemplaire… Je ne sais pas si cela rejoint ce que vous disiez tout à l'heure : là, on a affaire à un patient à qui il manquait un bout de cerveau. Il y a un lien structure-comportement extrêmement fort. C'est intéressant, c'est important, maintenant on est quand même passé à des études dans lesquelles on n'a pas besoin de patients à qui il manque des bouts de cerveau. Il y a toujours la tentative de la localisation : il manque un bout, donc ce bout s'occupait de ceci. Mais les règles ne sont pas si strictes, il y a des hiérarchies, des systèmes ; il y a des réseaux, des réseaux intriqués, des réseaux qui prennent le contrôle d'autres réseaux… C'est comme cela qu'il faut voir les choses, c'est une dynamique qu'on est loin d'avoir modélisé.

-NV: En même temps, de nombreux protocoles de neuroéconomie s'appuient sur la comparaison de sujets sains et lésés.

-LM: Oui, c'est toute la neuropsychologie qui se décline sous cet angle.

-NV: Justement, Mathias Pessiglione m'avait dit qu'il devait s'adresser à vous pour mettre en

place des expériences avec des patients. Pourriez-vous m'expliquer comment cela se passe en pratique ?

-LM: Pour les patients lésés, ce n'est pas moi qui m'en occupe. Si Mathias veut des patients avec des lésions, souvent, ce sont des patients qui sont suivis en neurologie. Après, en pratique, dans les centres de recherche comme ici, il faut bâtir un protocole de recherche clinique : il faut écrire un plan, demander de l'argent pour le faire… Les protocoles sont évalués par différentes personnes, des jurys, qui traitent plusieurs demandes, retiennent les meilleures, en fonction de l'importance scientifique. En suite, il y a un encadrement administratif, avec un promoteur d'essai, des grandes institutions, comme l'Inserm, qui «couvrent» en termes d'assurance et exigent de vous un certain nombre de choses, sous forme de cahier des charges, par rapport au fait de travailler avec des patients. Et enfin, vous mettez en place la structure qui permet d'assurer la fiabilité de vos observations.

-NV: Vous assurez l'encadrement des patients ?

-LM: Oui, on explique aux gens les tâches qu'ils devront réaliser. Il y a aussi toujours un consentement éclairé, une explication très approfondie de ce qu'ils vont faire, ils ont le droit de poser des questions, on leur conseille même d'en poser… Ensuite, ils signent un engagement pour pouvoir faire le protocole, qui stipule qu'ils ont bien compris de quoi il s'agissait. Il y a des comités d'éthique qui encadrent cela, qui contrôlent que ce que l'on fait correspond à toutes les règles qui ont été éditées en matière d'éthique. Après, c'est le travail des chercheurs cliniciens de faire ce genre de choses.

-NV: Pour vos recherches, vous utilisez l'imagerie ?

-LM: Oui.

-NV: Vous avez un accès facile aux scanners ? Faut-il réserver très longtemps à l'avance ?

-LM: Oui, cela peut être un problème, mais on a aussi certaines facilités. On bénéficie d'un environnement qui est quand même assez bien doté. C'est vrai qu'après, il y a beaucoup de recherches, donc il faut trouver des créneaux, se mettre dans la file d'attente…

-NV: Une dernière question : quelle est votre formation ?

-LM: J'ai fait médecine, après j'ai fait ma spécialité de psychiatrie à Paris. Assez tôt dans ma spécialité, j'ai fait de la recherche. J'ai fait un master de neurosciences. Je suis resté en neurosciences pendant mon internat, j'ai continué à travailler dans la recherche. Ensuite, j'ai pris un poste de chef de clinique, sorte de professeur assistant. J'ai donc eu des responsabilités cliniques plus importantes. En même temps, je continuais ma thèse de neurosciences, au CEA, je travaillais sur l'imagerie cérébrale. Après, j'ai fait un post-doc dans une unité CNRS pour faire de la psychologie expérimentale. J'ai commencé à travailler sur la stimulation au sein d'un centre d'investigations cliniques. J'ai passé le concours de l'Inserm et j'ai été recruté à l'Inserm, comme chercheur. Voilà, et j'ai monté une équipe, un labo…

-NV: La recherche occupe la majeure partie de votre temps ?

-LM: Oui, je suis directeur de recherche Inserm. Je fais de la clinique via mon activité de consultation dans ma spécialité, et j'ai une activité clinique au sens où je dirige des protocoles

de recherche clinique : je suis responsable et coordinateur. A un moment donné, il faut savoir faire les deux : un chercheur fondamental ne peut pas diriger des recherches cliniques, il faut être clinicien, ou médecin, pour pouvoir le faire ».

Bibliographie

-Adolphs, Ralph., Frederic Gosselin, Tony W. Buchanan, Daniel Tranel, Philippe Schyns, et Antonio R. Damasio. 2005. « A mechanism for impaired fear recognition after amygdala damage ». *Nature* 433 (7021): 68–72.

-Ainslie, George W. 1974. « Impulse control in pigeons ». *Journal of the Experimental Analysis of Behavior* 21(3) : 485-489.

-Ainslie, George W. 1975. « Specious reward: a behavioral theory of impulsiveness and impulse control ». *Psychological Bulletin*, 82(4) : 463-496.

-Ainslie, George W. 1992-a. « Derivation of « Rational » Economic Behavior from Hyperbolic Discount Curves ». *The American Economic Review* 81(2) : 334-340.

-Ainslie, George W. 1992-b. *Picoeconomics: The Strategic Interaction of Successive Motivational States within the Person.* Cambridge : Cambridge University Press.

-Ainslie, George W. 2001. *Breakdown of Will.* Cambridge : Cambridge University Press.

-Ainslie,George W., & John Monterosso 2002. « Will as Intertemporal Bargaining: Implications for Rationality ». *University of Pennsylvania Law Review* 151 : 825.

-Akerlof, George A. 1991. « Procrastination and Obedience ». *American Economic Review* 81(2) : 1-19.

-Andreoni, James. 1993. « An Experimental Test of the Public-Goods Crowding-Out Hypothesis ». *The American Economic Review* 83 (5): 1317–1327.

-Andrieu, Bernard. 2007. *La neurophilosophie*. 2e éd. Presses Universitaires de France – PUF.

-Angeletos, George-Marios, David Laibson, Andrea Repetto, Jeremy Tobacman, et Stephen Weinberg. 2001. « The hyperbolic consumption model: Calibration, simulation, and empirical evaluation ». *The Journal of Economic Perspectives* 15 (3): 47–68.

-Ariely, Dan. 2008. *Predictably Irrational: The Hidden Forces That Shape Our Decisions*. Harper.

-Aydinonat, N. Emrah. 2010. «Neuroeconomics: more than inspiration, less than revolution». *Journal of Economic Methodology* 17 (2): 159.

-Backhouse, Roger E., et Bradley W. Bateman. 2009. «Keynes and Capitalism». *History of Political Economy* 41 (4): 645-671.

-Baron-Cohen, Sacha. 1995. *Mindblindness: An essay on autism and theory of mind*. New York : Bradford Books.

-Barraza, Jorge. A., Michael. E. McCullough, Sheila Ahmadi, et Paul J. Zak. 2011. « Oxytocin

infusion increases charitable donations regardless of monetary resources ». *Hormones and behavior* 60 (2): 148.

-Barthes, Roland. 1984. *Le bruissement de la langue*. Seuil.

-Basso, Michelle A., et Robert H. Wurtz. 1997. « Modulation of neuronal activity by target uncertainty. » *Nature* 389 (6646): 66-69.

-Battalio, Raymond C., Leonard Green & John Kagel. 1981. « Income-Leisure Tradeoffs of Animal Workers ». *The American Economic Review* 71(4) : 621-632.

-Baum, William M. 1974. « On two types of deviation from the matching law: bias and undermatching ». *Journal of the Experimental Analysis of Behavior* 22(1) : 231-242.

-Bayer, Hannah. M., Brian. Lau, et Paul W. Glimcher. 2007. « Statistics of midbrain dopamine neuron spike trains in the awake primate ». *Journal of Neurophysiology* 98 (3): 1428–1439.

-Bechara, Antoine, Antonio Damasio, Hannah Damasio, et Steven W. Anderson. 1994. « Insensitivity to future consequences following damage to human prefrontal cortex. » *Cognition* 50 (1-3): 7-15.

-Bechara, Antoine, Daniel Tranel, et Hanna Damasio. 2000. « Characterization of the decision-making deficit of patients with ventromedial prefrontal cortex lesions ». *Brain* 123 (11): 2189 -2202.

-Becker, Gary. S. & Kevin M. Murphy. 1988. « A Theory of Rational Addiction ». *The Journal of Political Economy* 96(4) : 675-700.

-Berg, Joyce, John Dickhaut, et Kevin McCabe. 1995. « Trust, reciprocity, and social history ». *Games and economic behavior* 10 (1): 122–142.

-Bernheim, B. Douglas, et Antonio Rangel. 2005. «From Neuroscience to Public Policy : A New Economic View of Addiction ». *Swedish Economic Policy Review (article présenté à la « Conference on the Regulation of Unhealthy Consumption »).*

-Bernheim, B. Douglas, et Antonio Rangel. 2007. « Toward Choice-Theoretic Foundations for Behavioral Welfare Economics ». *The American Economic Review* 97 (2): 464-470.

-Bernheim, B. Douglas, et Antonio Rangel. 2009. « Beyond Revealed Preference: Choice-Theoretic Foundations for Behavioral Welfare Economics ». *Quarterly Journal of Economics* 124(1): 51-104.

-Bernheim, B. Douglas, et Antonio Rangel.. 2004. « Addiction and cue-triggered decision processes ». *American Economic Review*: 1558–1590.

-Bernheim, B. Douglas. 2008. *Behavioral welfare economics*. National Bureau of Economic Research.

-Berns, Gregory S., Samuel M. McClure, Giuseppe Pagnoni, et P. Read Montague. 2001.

« Predictability Modulates Human Brain Response to Reward ». *J. Neurosci.* 21 (8): 2793-2798.

-Berridge, Kent C. 2007. « The debate over dopamine's role in reward: the case for incentive salience ». *Psychopharmacology* 191 (3) : 391-431.

-Bhatt, Meghana, et Colin F. Camerer. 2005. «Self-referential thinking and equilibrium as states of mind in games: fMRI evidence». *Games and Economic Behavior* 52 (2): 424–459.

-Bhatt, Meghana, et Colin F. Camerer. 2005. « Self-referential thinking and equilibrium as states of mind in games: fMRI evidence ». *Games and Economic Behavior* 52 (2): 424-459.

-Bickel, Warren K., Michelle L. Miller, Richard Yi, Benjamin P. Kowal, Diana M. Lindquist, et-Jeffery A. Pitcock. 2007. « Behavioral and neuroeconomics of drug addiction: Competing neural systems and temporal discounting processes ». *Drug and Alcohol Dependence* 90 (1): 85-91.

-Binmore Ken. 2007. *Does Game Theory Work? The Bargaining Challenge.* Cambridge : The MIT Press.

-Blair, R. James. 2005. *The Psychopath: Emotion And The Brain*. Oxford: Blackwell Publishers.

-Blair, R. James R. 1996. « Brief Report: Morality in the Autistic Child. » *Journal of Autism and Developmental Disorders* 26 (5): 571-79.

-Blair, R. James R. 2001. « Neurocognitive models of aggression, the antisocial personality disorders, and psychopathy ». *Journal of Neurology, Neurosurgery & Psychiatry* 71 (6): 727–731.

-Blair, R. James R., Derek G. V. Mitchell, Rebecca A. Richell, Steve Kelly, Alan Leonard, Chris Newman, et Sophie K. Scott. 2002. « Turning a deaf ear to fear: impaired recognition of vocal affect in psychopathic individuals. » *Journal of Abnormal Psychology* 111 (4): 682.

-Blair, R. James R., Lawrence Jones, Fiona Clark, et Margaret Smith. 1997. « The psychopathic individual: A lack of responsiveness to distress cues? » *Psychophysiology* 34 (2): 192–198.

-Blair, R. James. R., et Lisa Cipolotti. 2000. « Impaired social response reversal ». *Brain* 123 (6): 1122–1141.

-Bloor, David. 1976. *Knowledge and Social Imagery*. Chicago: University Of Chicago Press.

-Bogacz, Rafal, Samuel M. McClure, Jian Li, Jonathan D. Cohen, et P. Read Montague. 2007. « Short-term memory traces for action bias in human reinforcement learning ». *Brain Research* 1153: 111–121.

-Born, Richard T., et David C. Bradley. 2005. « Structure and visual function of area MT ». *Annual Review of Neuroscience* 28 (21): 157-189.
-Bourgeois-Gironde, Sacha. 2008. *La neuroéconomie : Comment le cerveau gère mes intérêts*. Paris: Plon.
-Bourgeois-Gironde, Sacha. 2010. « Is neuroeconomics doomed by the reverse inference fallacy? » *Mind & Society* 9 (2): 229-249.
-Brand, Matthias, Elke Kalbe, Kirsten Labudda, Esther Fujiwara, Josef Kessler, et Hans J. Markowitsch. 2005. « Decision-making impairments in patients with pathological gambling ». *Psychiatry Research* 133 (1): 91–99.
-Braudel, Fernand. 1949. *La Méditerranée : L'espace et l'histoire*. Paris: Flammarion.
-Breiter, Hans C., Itzhak Aharon, Daniel Kahneman, Anders Dale, et Peter Shizgal. 2001. « Functional imaging of neural responses to expectancy and experience of monetary gains and losses ». *Neuron* 30 (2): 619-639.
-Breiter, Hans C., Randy L. Gollub, Robert M. Weisskoff, David N. Kennedy, Nikos Makris, Joshua D. Berke, Julie M. Goodman, Howard L. Kantor, David R. Gastfriend, Jonn P. Riorden, R.Thomas Mathew, Bruce R Rosen, et Steven E. Hyman. 1999. « Acute effect of cocaine on human brain activity and emotion ». *Neuron* 19: 591-611.
-Bush, Robert R., et Frederick Mosteller. 1951. « A mathematical model for simple learning ». *Psychological Review* 58 (5): 313–323.
-Camerer Colin F., George Loewenstein, et Drazen Prelec. 2005. «Neuroeconomics: Why Economics Needs Brains». *Scandinavian Journal of Economics* 106 (3): 555-579.
-Camerer, Colin F. 2003. « Strategizing in the brain ». *Science* 300 (5626): 1673.
-Camerer, Colin F. 2007. « Neuroeconomics: Using Neuroscience to Make Economic Predictions ». *The Economic Journal* 117 (519): 26-42.
-Camerer, Colin F., et Ernst Fehr. 2006. « When does" economic man" dominate social behavior? » *Science* 311 (5757): 47–52.
-Camerer, Colin F., Samuel. Issacharoff, Georges Loewenstein, Ted O'Donoghue, et Mathew Rabin. 2003. « Regulation for Conservatives: Behavioral Economics and the Case for" Asymmetric Paternalism" ». *University of Pennsylvania Law Review* 151 (3): 1211–1254.
-Camille, Nathalie, Giorgio Coricelli, Jerome Sallet, Pascale Pradat-Diehl, Jean-René Duhamel, et Angela Sirigu. 2004. « The Involvement of the Orbitofrontal Cortex in the Experience of Regret ». *Science* 304 (5674): 1167 -1170.
-Canguilhem, Georges. 1977. *Idéologie et rationalité dans l'histoire des sciences de la vie*. Paris: Vrin.

-Caplin, Andrew, et Mark Dean. 2008. « Axiomatic neuroeconomics ».in Glimcher, Paul W., Colin Camerer, Russell Alan Poldrack, et Ernst Fehr. 2008. *Neuroeconomics: Decision Making and the Brain*. Academic Press. 21–31.

-Capshew, James H. 1999. *Psychologists on the March: Science, Practice, and Professional Identity in America, 1929-1969*. Cambridge: Cambridge University Press.

-Castel Pierre-Henri. 2010. *L'Esprit malade. Cerveaux, folies, individus.* Paris : Les Editions d'Ithaque.

-Castel Robert. 1973. *Le psychanalysme*, Paris : Flamarion.

-Castel Robert. 1977. *L'Ordre psychiatrique : L'âge d'or de l'aliénisme.* Paris : Les Editions de Minuit.

-Castel Robert. 1981. *La Gestion des risques : De l'anti-psychiatrie à l'après-psychanalyse* Paris : Les Editions de Minuit.

-Castel, Françoise, Robert Castel & Anne Lovell. 1979. *La société psychiatrique avancée* Paris : Editions Bernard Grasset.

-Certeau, Michel de, Dominique Julia, et Jacques Revel. 1970. « La beauté du mort ». *Politique aujourd'hui* repris in M. de Certeau, 1993. *La culture au pluriel*, Seuil

-Chabris, Christopher F., David Laibson, et Jonathan P. Schuldt. 2006. « Intertemporal choice ». in S.N. Durlauf et L.E. Blume (eds), The New Palgrave Dictionary of Economics. Macmillan.

-Chaigneau, Nicolas. 2002. « Jevons, Edgeworth et les « sensations subtiles du cœur humain »: l'influence de la psychophysiologie sur l'économie marginaliste ». *Revue d'Histoire des Sciences Humaines* (2): 13–39.

-Chamberlin, Edward H. 1948. « An experimental imperfect market ». *The Journal of Political Economy* 56 (2): 95–108.

-Chiu, Yao-Chu, Ching-Hung Lin, Jong-Tsun Huang, Shuyeu Lin, PoLei Lee, et Jen-Chuen Hsieh. 2008. « Immediate gain is long-term loss: Are there foresighted decision makers in the Iowa Gambling Task ». *Behavioral and Brain Functions* 4 (1): 13.

-Chung, Shin-Ho & Richard J. Herrnstein. 1967. « Choice and delay of reinforcement ». *Journal of the Experimental Analysis of Behavior* 10(1) : 67-74.

-Cohen Michèle et Jean-Marc Tallon. 2000. « Décision dans le risque et l'incertain: l'apport des modèles non-additifs ». *Revue d'Économie Politique* 110 (5): 631-681.

-Colby, Anne, Lawrence Kohlberg, John Gibbs, et Marcus Lieberman. 1983. « A longitudinal study of moral judgment. » *Monographs of the Society for Research in Child Development*.

-Colby, Carol L., Jean-René Duhamel, et Michael E. Goldberg. 1996. « Visual, Presaccadic, and Cognitive Activation of Single Neurons in Monkey Lateral Intraparietal Area ». *Journal of Neurophysiology* 76 (5): 2841-2852.

-Commons, Michael. L. 2001. A short history of the Society for the Quantitative Analysis of Behavior. *Behavior Analyst Today* 2 (3): 275-279

-Coricelli, Giorgio. 2005. « Two-levels of mental states attribution: From automaticity to voluntariness ». *Neuropsychologia* 43 (2): 294–300.

-Coricelli, Giorgio, Raymond J. Dolan, et Angela Sirigu. 2007. « Brain, emotion and decision making: the paradigmatic example of regret ». *Trends in Cognitive Sciences* 11 (6): 258–265.

-Cova, Florian. 2011. *Qu'en pensez-vous ? Introduction à la philosophie expérimentale*. Germina.

-Cusset, François. 2003. *French Theory*. Paris: Editions La Découverte.

-Damasio, A. 2008. « Neuroscience and the Emergence of Neuroeconomics ». in Glimcher, Paul W., Colin Camerer, Russell Alan Poldrack, et Ernst Fehr. 2008. *Neuroeconomics: Decision Making and the Brain*. Academic Press. : 209-234.

-Damasio, Antonio. 1994. *Descartes' Error*. New York: Putnam Adult.

-Damasio, Antonio. 2003. *Looking for Spinoza: Joy, Sorrow, and the Feeling Brain*. 1er éd. Houghton Mifflin Harcourt.

-Davis, John B. 2007. « The turn in economics and the turn in economic methodology ». *Journal of Economic Methodology* 14 (3): 275-290.

-Davis, John B. 2010. « Neuroeconomics: Constructing identity ». *Journal of Economic Behavior & Organization* 76 (3): 574-583.

-Dehaene, Stanislas, et Laurent Cohen. 2007. « Cultural recycling of cortical maps ». *Neuron* 56 (2): 384–398.

-De Martino, Benedetto, Dharshan Kumaran, Ben Seymour, et Raymond J. Dolan. 2006. « Frames, biases, and rational decision-making in the human brain ». *Science* 313 (5787): 684–687.

-de Quervain, Dominique J.-F., Urs Fischbacher, Valerie Treyer, Melanie Schellhammer, Ulrich Schnyder, Alfred Buck, et Ernst Fehr. 2004. « The Neural Basis of Altruistic Punishment ». *Science* 305 (5688): 1254 -1258.

-Delgado, Mauricio R., Andrew Schotter, Erkut Y. Ozbay, et Elizabeth A. Phelps. 2008. « Understanding Overbidding: Using the Neural Circuitry of Reward to Design Economic Auctions ». *Science* 321 (5897): 1849 -1852.

-Delgado, Mauricio R., L. E. Nystrom, C. Fissell, D. C. Noll, et J. A. Fiez. 2000. « Tracking

the Hemodynamic Responses to Reward and Punishment in the Striatum ». *Journal of Neurophysiology* 84 (6): 3072 -3077.

-Denburg, Nathalie. L., Daniel Tranel, et Antoine Bechara. 2005. « The ability to decide advantageously declines prematurely in some normal older persons ». *Neuropsychologia* 43 (7): 1099–1106.

-Dimand, Robert W. 2007. «The Creation of Heroes and Villains as a Problem in the History of Economics». *History of Political Economy* 39 (Suppl 1): 76-95.

-diPetrio Mark. 2002. « The Goldilocks principles ». The flame 3 (1): 20-21.

-Dorris, Michael C., et Paul W. Glimcher. 2004. « Activity in posterior parietal cortex is correlated with the relative subjective desirability of action ». *Neuron* 44 (2): 365–378.

-Dosse, François. 2010-a. « Biographie, Prosopographie », *in* Delacroix, C. *et al. Historiographies : Concepts et débats*, 79-86. Gallimard.

-Dosse, François. 2010-b. « Histoire des mentalités », *in* Delacroix, C. *et al. Historiographies : Concepts et débats*, 220-232. Gallimard.

-Droulers, Olivier et Bernard Roullet. 2007, «Émergence du neuromarketing, apports et perspectives pour les praticiens et les chercheurs», *Décisions Marketing* 46: 9-22.

-Droulers, Olivier, et Bernard Roullet. 2010. *Neuromarketing - Le marketing revisité par les neurosciences du consommateur.* Paris: Dunod.

-Dunbar, Robin I. M. 1998. « The Social Brain Hypothesis ». *Evolutionary Anthropology: Issues, News, and Reviews* 6 (5): 178-190.

-Ehrenberg Alain. « Malaise dans l'évaluation de la santé mentale ». *Esprit* 324: 89-102.

-Ehrenberg Alain. 2008. « Le cerveau « social » : Chimère épistémologique et vérité sociologique ». *Esprit* 341: 79-103.

-Ehrenberg, Alain. 2010. *La société du malaise.* Paris : Odile Jacob.

-Elster, Jon. 1979. *Ulysses and the Sirens. Studies in Rationality and Irrationality.* Cambridge: Cambridge University Press.

-Elster, Jon, et Ole-Jørgen Skog. 1999. *Getting Hooked: Rationality and Addiction*. Cambridge: Cambridge University Press.

-Evans, Cathryn E. Y., Karen Kemish, et Oliver H. Turnbull. 2004. « Paradoxical effects of education on the Iowa Gambling Task ». *Brain and Cognition* 54 (3): 240–244.

-Falk, Armin, Ernst Fehr, et Urs Fischbacher. 2005. « Driving forces behind informal sanctions ». *Econometrica* 73 (6): 2017–2030.

-Fehr, Ernst. « Social preferences and the brain ». in Glimcher, Paul W., Colin Camerer, Russell Alan Poldrack, et Ernst Fehr. 2008. *Neuroeconomics: Decision Making and the Brain*.

Academic Press. 215–232.
-Fehr, Ernst et Urs Fischbacher. 2003. « The nature of human altruism ». *Nature* 425 (6960): 785–791.
-Fehr, Ernst, et Bettina Rockenbach. 2003. « Detrimental effects of sanctions on human altruism ». *Nature* 422 (6928): 137–140.
-Fehr, Ernst, et Colin F. Camerer. 2007. « Social neuroeconomics: the neural circuitry of social preferences ». *Trends in Cognitive Sciences* 11 (10): 419–427.
-Fehr, Ernst, et Klaus M. Schmidt. 1999. « A Theory Of Fairness, Competition, and Cooperation ». *Quarterly Journal of Economics* 114 (3): 817–868.
-Fehr, Ernst, et Simon Gächter. 2000. « Fairness and Retaliation: The Economics of Reciprocity ». *The Journal of Economic Perspectives* 14 (3): 159–181.
-Fehr, Ernst, et Simon Gachter. 2002. « Altruistic punishment in humans ». *Nature* 415 (6868) : 137-140.
-Fehr, Ernst, et Tania Singer. 2005. « The neuroeconomics of mind reading and empathy ». *CEPR Discussion Paper No. 5128*.
-Fehr, Ernst, et Urs Fischbacher. 2004. « Third-party punishment and social norms ». *Evolution and human behavior* 25 (2): 63–87.
-Fehr, Ernst, Urs Fischbacher, et Michael Kosfeld. 2005. « Neuroeconomic foundations of trust and social preferences: initial evidence ». *American Economic Review*: 346–351.
-Fibiger, Hans C., et Anthony G. Phillips. 2011. « Reward, Motivation, Cognition: Psychobiology of Mesotelencephalic Dopamine Systems ». In *Comprehensive Physiology*. John Wiley & Sons.
-Filler, Aaron. 2009. « The History, Development and Impact of Computed Imaging in Neurological Diagnosis and Neurosurgery: CT, MRI, and DTI ». *Neurosurgery* 65 (4): 29.
-Fontaine, Philippe, et Alain Marciano. 2007. «The Political Element in Economic Thought». *History of Political Economy* 39 (4): 567-570.
-Fontaine, Philippe. 2007. «From philanthropy to altruism: Incorporating unselfish behavior into economics, 1961-1975». *History of political economy* 39 (1): 1.
-Forget, Evelyn L., et Craufurd D. Goodwin. 2011. «Intellectual Communities in the History of Economics». *History of Political Economy* 43 (1): 1-23.
-Foucault Michel, 1969. *Qu'est-ce qu'un auteur ?* , *in Dits et Écrits*, Gallimard, 1994, t. I.
-Foucault, Michel. [1978-1979] 2004. *Naissance de la biopolitique : Cours au collège de France.* Paris : Seuil.
-Foucault, Michel. 2001. *Dits et Écrits, tome 1 : 1954-1975*. Paris: Gallimard.

-Foucault, Michel. 2001. *Dits et Écrits, tome 2 : 1976 - 1988*. Paris: Gallimard.
-Fox, Craig R., et Russell. A. Poldrack. 2008. « Prospect theory and the brain ». in Glimcher, Paul W., Colin Camerer, Russell Alan Poldrack, et Ernst Fehr. 2008. *Neuroeconomics: Decision Making and the Brain*. Academic Press. : 145–174.
-Frederick, Shane, George Loewenstein & Ted O'Donoghue. 2002. « Time Discounting and Time Preference: A Critical Review ». *Journal of Economic Literature* 40 : 351-401.
-Frederick, Shane, Georges Loewenstein, et Ted O'donoghue. 2002. « Time discounting and time preference: A critical review ». *Journal of economic literature* 40 (2): 351–401.
-Gallagher, Helen L., Anthony. I. Jack, Andreas Roepstorff, et Christopher D. Frith. 2002. « Imaging the intentional stance in a competitive game ». *Neuroimage* 16 (3): 814–821.
-Garcia, Patrick. 2010. « Histoire du temps présent », *in* Delacroix, C. *et al. Historiographies : Concepts et débats*, 282-295. Paris: Gallimard.
-Gardner, Eliot L., et James David. 1999. « The Nellfobiology of ». in Elster, Jon, et Ole-Jørgen Skog (eds). *Getting hooked: Rationality and addiction*. Cambridge University Press: 93.
-Gigerenzer, Gerd et Nathan Berg. 2010. « As-if behavioral economics: Neoclassical economics in disguise ». *History of Economic Ideas* 18 (1): 133–166.
-Gigerenzer, Gerd. 1993. « The bounded rationality of probabilistic mental models ». *In* K. I. Manktelow, & D.E. Over (Eds.), *Rationality: psychological and philosophical perspectives*, 284-313. Routledge.
-Gigerenzer, Gerd. 1996. « On narrow norms and vague heuristics: A reply to Kahneman and Tversky. ». *The Psychological Review* 103 (3): 592-593.
-Gigerenzer, Gerd. 2000. *Adaptive thinking: Rationality in the real world*. Oxford: Oxford University Press.
-Gintis, Herbert. 2006. « Adapting Minds and Evolutionary Psychology ». *Working Paper*
-Gintis, Herbert, Samuel Bowles, Robert Boyd, et Ernts Fehr. 2003. « Explaining altruistic behavior in humans ». *Evolution and Human Behavior* 24 (3): 153–172.
-Gintis, Herbert. 2003. « The Hitchhiker's Guide to Altruism: Gene-culture Coevolution, and the Internalization of Norms ». *Journal of Theoretical Biology* 220 (4): 407-418.
-Glaeser, Edward. 2006. « Paternalism and Psychology », *University of Chicago Law Review*, 73, p.133-156.
-Glicksohn, Joseph, Revital Naor-Ziv, et Rotem Leshem. 2007. « Impulsive decision-making: Learning to gamble wisely? » *Cognition* 105 (1): 195–205.
-Glimcher, Paul W. 2003. *Decisions, Uncertainty, and the Brain: The Science of*

Neuroeconomics. Cambridge: MIT Press.
-Glimcher, Paul W. 2010. *Foundations of Neuroeconomic Analysis*. Oxford University Press.
-Glimcher, Paul W., Colin Camerer, Russell Alan Poldrack, et Ernst Fehr. 2008. *Neuroeconomics: Decision Making and the Brain*. London: Academic Press.
-Glimcher, Paul W., et Aldo Rustichini. 2004. « Neuroeconomics: the consilience of brain and decision ». *Science* 306 (5695): 447.
-Glimcher, Paul W., Michael C. Dorris, et Hannah M. Bayer. 2005. « Physiological utility theory and the neuroeconomics of choice ». *Games and Economic Behavior* 52 (2): 213-256.
-Gnadt, James W., et Richard A. Andersen. 1988. « Memory related motor planning activity in posterior parietal cortex of macaque ». *Experimental Brain Research* 70 (1): 216–220.
-Gneezy, Uri, et A. Rustichini. 2000. « Pay enough or don't pay at all ». *The Quarterly Journal of Economics* 115 (3): 791–810.
-Goffman, Erwin. 1968. *Asiles; études sur la condition sociale des malades mentaux et autres reclus.* Paris : Les Editions de Minuit.
-Greene, Joshua, et Jonathan Haidt. 2002. « How (and where) does moral judgment work? » *Trends in Cognitive Sciences* 6 (12): 517-523.
-Greene, Joshua. 2001. « An fMRI Investigation of Emotional Engagement in Moral Judgment ». *Proceedings of the National Academy of Sciences* 97: 1143.
-Grenier, Jean-Yves & André Orléan. 2007. « Michel Foucault, l'économie politique et le libéralisme. *Annales, histoire, sciences sociales.* (5) : 1155-1182
-Gruber, Jonathan, et Botond Kőszegi. 2000. « Is Addiction "Rational"? Theory and Evidence ». *National Bureau of Economic Research Working Paper Series* No. 7507. http://www.nber.org/papers/w7507.
-Gruber, Jonathan, et Botond Kőszegi. 2004. « Tax incidence when individuals are time-inconsistent: the case of cigarette excise taxes ». *Journal of Public Economics* 88 (9): 1959–1987.
-Guala, Francesco, et Tim Hodgson. 2010. «The philosopher in the scanner (or: how can neuroscience contribute to social philosophy?)». *Journal of Economic Methodology* 17 (2): 147.
-Guala, Francesco. 2005. *The Methodology of Experimental Economics*. Cambridge: Cambridge University Press.
-Gul Faruk et Wolfgang Pesendorfer, 2008 «The Case for Mindless Economics» in Caplin, Andrew, et Andrew Schotter. 2008. *The Foundations of Positive and Normative Economics: A Handbook*. Oxford University Press.

-Gul, Faruk, et Wolfgang Pesendorfer. 2001. « Temptation and self-control ». *Econometrica* 69 (6): 1403–1435.

-Gul, Faruk, et Wolfgang Pesendorfer. 2004. « Self-control and the theory of consumption ». *Econometrica* 72 (1): 119–158.

-Gul, Faruk, et Wolfgang Pesendorfer. 2005. « The Case for Mindless Economics ». *Working Paper.*

-Hacking, Ian. 1975. *The Emergence of Probability: A Philosophical Study of Early Ideas About Probability, Induction and Statistical Inference*. Cambridge: Cambridge University Press.

-Hacohen, M. 2007. «Rediscovering Intellectual Biography and Its Limits». *History of Political Economy* 39 (Suppl 1): 9-29.

-Haidt, Jonathan. 2001. « The emotional dog and its rational tail: a social intuitionist approach to moral judgment. » *Psychological Review; Psychological Review* 108 (4): 814.

-Hands, Wade. 2009. The positive-Normative Dichotomy and Economics. *Working Paper.*

-Hare, Todd A., Colin F. Camerer, Daniel T. Knoepfle, John P. O'Doherty, et Antonio Rangel. 2010. « Value computations in ventral medial prefrontal cortex during charitable decision making incorporate input from regions involved in social cognition ». *The Journal of Neuroscience* 30 (2): 583–590.

-Hare, Todd A., Colin F. Camerer, et Antonio Rangel. 2009. « Self-Control in Decision-Making Involves Modulation of the vmPFC Valuation System ». *Science* 324 (5927): 646-648.

-Hare, Todd A., Jonathan Malmaud, et Antonio. Rangel. 2011. « Focusing attention on the health aspects of foods changes value signals in vmPFC and improves dietary choice ». *The Journal of Neuroscience* 31 (30): 11077–11087.

-Harmsen, H., G. Bischof, A. Brooks, F. Hohagen, et H.J. Rumpf. 2006. « The relationship between impaired decision-making, sensation seeking and readiness to change in cigarette smokers ». *Addictive behaviors* 31 (4): 581–592.

-Harrison, Glenn W. 2008-a. «Neuroeconomics: A Critical Reconsideration». *Economics and Philosophy* 24 (Special Issue 03): 303-344.

-Harrison, Glenn W. 2008-b. « Neuroeconomics: a rejoinder ». *Economics and Philosophy* 24 (03) : 533-549.

-Harrison, Glenn, et Don Ross. 2010. «The methodologies of neuroeconomics». *Journal of Economic Methodology* 17 (2): 185.

-Healy, David. 1998. *The Antidepressant Era*. Cambridge: Harvard University Press.

-Healy, David. 2002. *The Creation of Psychopharmacology*. Cambridge: Harvard University Press.

-Herman, Ellen. 1996. *The Romance of American Psychology: Political Culture in the Age of Experts*. Berkeley: University of California Press.

-Herrnstein, Richard J. & Drazen Prelec, D. 1991. « Melioration: A Theory of Distributed Choice ». *The Journal of Economic Perspectives* 5(3) : 137-156.

-Herrnstein, Richard J. & Heyman, Gene M. 1979. « Is matching compatible with reinforcement maximization on concurrent variable interval variable ratio? » *Journal of the Experimental Analysis of Behavior* 31(2) : 209-223.

-Herrnstein, Richard J. 1961. « Relative and absolute strength of response as a function of frequency of reinforcement ». *Journal of the Experimental Analysis of Behavior* 4(3) : 267-272.

-Herrnstein, Richard J. 1990. « Rational choice theory: Necessary but not sufficient ». *American Psychologist.* 45(3) : 356-367.

-Herrnstein, Richard J. 1991. « Experiments on Stable Suboptimality in Individual Behavior ». *The American Economic Review.* 81(2) : 360-364.

-Heukelom, Floris. 2009. Kahneman and Tversky and the making of behavioral economics. PhD dissertation. Amsterdam School of Economics.

-Heyman, Gene.M. & Richard J. Herrnstein. 1986. « More on concurrent interval-ratio schedules: a replication and review ». *Journal of the Experimental Analysis of Behavior.* 46(3) : 331-351.

-Horvitz, Jon C. 2000. « Mesolimbocortical and nigrostriatal dopamine responses to salient non-reward events ». *Neuroscience* 96 (4): 651–656.

-Huettel, Scott A. 2010. « Ten challenges for decision neuroscience ». *Frontiers in neuroscience* 4: 171-185.

-Ingvar, David H. 1977. L'idéogramme cerebral. *Encéphale* 3: 5–23, 1977.

-Janowski, Vanessa, Colin F. Camerer, et Antonio Rangel. 2012. « Empathic choice involves vmPFC value signals that are modulated by social processing implemented in IPL ». à paraître dans *Social Cognitive and Affective Neuroscience*.

-Jaroni, Jodie. L., Suzanne. M. Wright, Caryn. Lerman, et Leonard. H. Epstein. 2004. « Relationship between education and delay discounting in smokers ». *Addictive behaviors* 29 (6): 1171–1175.

-Jolls, Christine, et Cass R Sunstein. 2006. « Law of Implicit Bias, The ». *California Law Review* 94: 969.

-Jullien, Dorian et Nicolas Vallois. 2012. The Kahneman, Tversky and Gigerenzer controversy. *Working Paper.*

-Kable, Joseph W, et Paul W Glimcher. 2007. « The neural correlates of subjective value during intertemporal choice ». *Nature Neuroscience* 10 (12): 1625-1633.

-Kable, Joseph W. 2011. « The cognitive neuroscience toolkit for the neuroeconomist: A functional overview ». *Journal of neuroscience, psychology, and economics* 4 (2): 63-84.

-Kahneman, D. 2003. « Maps of bounded rationality: Psychology for behavioral economics ». *The American economic review* 93 (5): 1449–1475.

-Kahneman, D., et A. Tversky. 1996. « On the reality of cognitive illusions ». *The Psychological Review* 103 (3): 582-591.

-Kahneman, Daniel, et Amos Tversky. 1979. « Prospect Theory: An Analysis of Decision under Risk ». *Econometrica* 47 (2): 263-291.

-Kahneman, Daniel, Jack L. Knetsch, et Richard. H. Thaler. 1990. « Experimental Tests of the Endowment Effect and the Coase Theorem ». *Journal oof Political Economy* 98 (6).

-Kahneman, Daniel. & Amos Tversky. 1979. « Prospect Theory: An Analysis of Decision under Risk ». *Econometrica* 47(2) : 263-291.

-Kandel, Eric R., John H. Schwartz, et Thomas M. Jessell. 1991. *Principles of Neural Science*. New York: McGraw-Hill Medical.

-Kast, Robert. 2002. *La théorie de la décision*. Paris: La Découverte.

-Kerr, Aurora, et Philipp D. Zelazo. 2004. « Development of“ hot” executive function: The children’s gambling task. » *Brain and Cognition*.

-Klick, Jonathan, et Gregory Mitchell. 2005. « Government Regulation of Irrationality: Moral and Cognitive Hazards ». *Minnesota Law Review* 90: 1620.

-Knobe, Joshua, et Shaun Nichols. 2008. *Experimental Philosophy.* Oxford: Oxford University Press.

-Knoch, Daria, Alvaro Pascual-Leone, Kaspar Meyer, Valerie Treyer, et Ernt Fehr. 2006. « Diminishing reciprocal fairness by disrupting the right prefrontal cortex ». *Science* 314 (5800): 829.

-Knoch, Daria., Michael A. Nitsche, Urs Fischbacher, Christoph Eisenegger, Alavaro Pascual-Leone, et Ernst Fehr. 2008. « Studying the neurobiology of social interaction with transcranial direct current stimulation—the example of punishing unfairness ». *Cerebral Cortex* 18 (9): 1987–1990.

-Knutson, Brian, Charles M. Adams, Grace W. Fong, et Daniel Hommer. 2001. « Anticipation of increasing monetary reward selectively recruits nucleus accumbens ». *Journal of*

Neuroscience 21 (16): 1–5.

-Kosfeld, Michael, Markus. Heinrichs, Paul J. Zak, Urs Fischbacher, et Ernst Fehr. 2005. « Oxytocin increases trust in humans ». *Nature* 435 (7042): 673–676.

-Krajbich, Ian, Ralph Adolphs, Daniel Tranel, Nathalie L. Denburg, et Colin F. Camerer. 2009. « Economic games quantify diminished sense of guilt in patients with damage to the prefrontal cortex ». *The Journal of Neuroscience* 29 (7): 2188–2192.

-Krebs, John R., et Nicholas B. Davies. 1997. *Behavioral Ecology*. New York: Wiley-Blackwell.

-Kreps, David. M. 1990. *Microeconomic theory*. Princeton: Princeton University Press.

-Krueger Frank , Raja Parasuraman, Vijeth Iyengar, Matthew Thornburg, Jaap Weel, Mingkuan Lin, Ellen Clarke, Kevin McCabe, et Robert H. Lipsky. 2012. « Oxytocin Receptor Genetic Variation Promotes Human Trust Behavior ». *Frontiers in Human Neuroscience* 6.

-Laibson, David. 1997. « Golden Eggs and Hyperbolic Discounting ». *Quarterly Journal of Economics*. 112(2) : 443-477.

-Landreth, Anthony, et John Bickle. 2008. « Neuroeconomics, neurophysiology and the common neural currency ». *Economics and Philosophy* 24 (Special Issue 03): 419-429.

-Larquet, Marion, Giorgio Coricelli, Gaëlle Opolczynski, et Florence Thibaut. 2010. « Impaired decision making in schizophrenia and orbitofrontal cortex lesion patients ». *Schizophrenia research* 116 (2-3): 266–273.

-Latour, Bruno, Michel Biezunski et Steve Woolgar. 1979. *Laboratory Life: The Social Construction of Scientific Facts*. Paris: Sage Publications.

-Latour, Bruno. 1989. *La science en action : Introduction à la sociologie des sciences*. Paris: Éditions La Découverte.

-Lee, Ming, et Norman M. Prentice. 1988. « Interrelations of empathy, cognition, and moral reasoning with dimensions of juvenile delinquency ». *Journal of Abnormal Child Psychology* 16 (2): 127–139.

-Lenfant, Jean-Sébastien. 2010, « Between Axiomatics and Psychology. Probabilizing the Consumer in the 1950s », *Working Paper*. ESHET Conference, Amsterdam.

-Lesieur, Henry. R., et Sheila B. Blume. 1987. « The South Oaks Gambling Screen (SOGS): A new instrument for the identification of pathological gamblers ». *American Journal of Psychiatry* 144 (9).

-Levitt, Steven. D., et John A. List. 2007. « What do laboratory experiments measuring social preferences reveal about the real world? » *The Journal of Economic Perspectives* 21 (2): 153–174.

-Livet, Pierre. 2010. « Rational choice, neuroeconomy and mixed emotions ». *Philosophical Transactions of the Royal Society B: Biological Sciences* 365 (1538): 259–269.

-Lloyd, Georges. 1993. *Pour en finir avec l'histoire des mentalités*. Paris: Éditions La Découverte.

-Loewenstein, George. & Drazen Prelec. 1991. Negative Time Preference. *The American Economic Review.* 81(2) : 347-352.

-Loewenstein, George. & Drazen Prelec. 1992. « Anomalies in Intertemporal Choice: Evidence and an Interpretation ». *The Quarterly Journal of Economics* 107(2) : 573-597.

-Loewenstein, George. & Nachum Sicherman. 1991. « Do Workers Prefer Increasing Wage Profiles? ». *Journal of Labor Economics* 9(1) : 67-84.

-Loewenstein, George. & Richard H. Thaler. 1989. Intertemporal Choice. *Journal of Economic Perspectives.* 3(4) : 181-93.

-Loewenstein, George. 1987. « Anticipation and the Valuation of Delayed Consumption ». *The Economic Journal.* 97(387) : 666-684.

-Loewenstein, George. 1996. « Out of Control: Visceral Influences on Behavior ». *Organizational Behavior and Human Decision Processes* 65 (3): 272-292.

-Loewenstein, George. 2000. « Emotions in Economic Theory and Economic Behavior ». *The American Economic Review* 90(2) : 426-432.

-Loewenstein, Georges. 1999. « A visceral account of addiction ». *in Elster, Jon, et Ole-Jørgen Skog (eds). Getting hooked: Rationality and addiction. Cambridge:* Cambridge University Press: 235–264.

-Loomes, Graham, et Robert Sugden. 1982. « Regret Theory: An Alternative Theory of Rational Choice Under Uncertainty ». *The Economic Journal* 92 (368): 805-824.

-Lovell, Anne & Alain Ehrenberg. 2001. *La Maladie mentale en mutation.* Paris : Odile Jacob.

-Machina, Mark J. 1989. « Dynamic consistency and non-expected utility models of choice under uncertainty ». *Journal of Economic Literature* 27 (4): 1622–1668.

-Mäki, Uskali. 2009. Economics Imperialism. *Philosophy of the Social Sciences* 39 (3):351-380.

-Mäki, Uskali. 2010. « When economics meets neuroscience: hype and hope ». *Journal of Economic Methodology* 17 (2): 107.

-Mata, Tiago, et F. S. Lee. 2007. « The Role of Oral History in the Historiography of Heterodox Economics ». *History of Political Economy* 39 (Suppl 1): 154-171.

-McCabe, Kevin, Daniel Houser, Lee Ryan, Vernon Smith, et Theodore Trouard. 2001. « A functional imaging study of cooperation in two-person reciprocal exchange ». *Proceedings of*

the National Academy of Sciences of the United States of America 98 (20): 11832 -11835.

-McCloskey, Deirdre N. 1985. *The Rhetoric of Economics*. Madison : University of Wisconsin Press.

-McCloskey, Deirdre N. 1994-a. « How economists persuade ». *Journal of Economic Methodology* 1 (1): 15–32.

-McCloskey, Deirdre. 1994-a « How to Do a Rhetorical Analysis of Economics, and Why » *in* Roger Backhouse, ed., *Economic Methodology,* 319-342. Routledge.

-McClure Samuel M., et Kimberlee D'Ardenne. 2009. « Computational neuroimaging: monitoring reward learning with blood flow ». In Dreher JC, Tremblay L(Eds.). *Handbook of Reward and Decision Making.* Cambridge*:* Cambridge University Press.

-McClure, Samuel M., David I. Laibson, George Loewenstein, et Jonathan D. Cohen. 2004. « Separate Neural Systems Value Immediate and Delayed Monetary Rewards ». *Science* 306 (5695): 503 -507.

-McClure, Samuel M., et Kimberlee. D'Ardenne. 2009. « Computational neuroimaging: monitoring reward learning with blood flow ». in J.C. Dreher (eds). *Handbook of reward and decision making*: 229-270.

-McClure, Samuel M., Gregory S. Berns, et P. Read Montague. 2003. « Temporal prediction errors in a passive learning task activate human striatum ». *Neuron* 38 (2): 339–346.

-McClure, Samuel M., Jian Li, Damon Tomlin, Kim S. Cypert, Latané M. Montague, et P.Read Montague. 2004. «Neural Correlates of Behavioral Preference for Culturally Familiar Drinks». *Neuron* 44 (2): 379-387.

-McClure, Samuel M., Michelle K. York, et P. Read Montague. 2004. « The neural substrates of reward processing in humans: the modern role of FMRI ». *The Neuroscientist* 10 (3): 260–268.

-McClure, SamuelM, Keith M. Ericson, et David I. Laibson. 2007. « Time discounting for primary rewards ». *Journal of Neuroscience* 27 (21): 5796–5804.

-McCulloch, Warren S., et Walter Pitts. 1943. « A logical calculus of the ideas immanent in nervous activity ». *Bulletin of mathematical biology* 5 (4): 115–133.

-McFadden, Daniel. 1974. « Conditional Logit Analysis of Qualitative Choice Behavior ». In *Frontiers in econometrics*, 105-142. Academic Press.

-Milner, Brenda, Michael Petrides, et Michael L. Smith. 1985. « Frontal lobes and the temporal organization of memory. » *Human neurobiology* 4 (3): 137.

-Milner, Brenda. 1963. « Effects of Different Brain Lesions on Card Sorting: The Role of the Frontal Lobes ». *Arch Neurol* 9 (1): 90-100.

-Mirenowicz, Jacques, et W. Schultz. 1994. « Importance of unpredictability for reward responses in primate dopamine neurons ». *Journal of Neurophysiology* 72 (2): 1024 -1027.
-Mitchell, Gregory. « Libertarian Paternalism Is an Oxymoron ». *Northwestern University Law*
-Mongin, Philippe. 1984. « Modèle rationnel ou modèle économique de la rationalité? » *Revue Économique* 35 (1): 9–64.
-Mongin, Philippe. 2008. « Retour à Waterloo ». *Annales. Histoire, sciences sociales* 63: 139-169.
-Montague, P. Read., et Gregory. S. Berns. 2002. « Neural economics and the biological substrates of valuation ». *Neuron* 36 (2): 265–284.
-Moreno-Bote, Rubén, David C. Knill, et Alexandre Pouget. 2011. « Bayesian Sampling in Visual Perception ». *Proceedings of the National Academy of Sciences* 108 (30): 12491-12496.
-Mountcastle, Vernon. B., James C. Lynch, Apostolos. Georgopoulos, Hideo. Sakata, et Claudio Acuna. 1975. « Posterior parietal association cortex of the monkey: command functions for operations within extrapersonal space ». *Journal of Neurophysiology* 38 (4): 871–908.
-Nakahara, Hiroyuki, Hideaki Itoh, Reiko Kawagoe, Yoriko Takikawa, et Okihide Hikosaka. 2004. « Dopamine neurons can represent context-dependent prediction error ». *Neuron* 41 (2): 269–280.
-Newsome, William T., Kenneth H. Britten, et J. Anthony Movshon. 1989. « Neuronal correlates of a perceptual decision ». *Nature* 341 (6237): 52-54.
-Nik-Khah Edward, 2009. «Getting Hooked on Drugs: the Chicago School, the Pharmaceutical Project, and the Construction of Medical Neoliberalism», *Working Paper.*
-O'Donoghue Ted & Matthew Rabin. 1999. « Doing It Now or Later ». *The American Economic Review.* 89(1) :103-124.
-O'Donoghue, Ted, et Mathew Rabin. 2003. « Studying optimal paternalism, illustrated by a model of sin taxes ». *The American economic review* 93 (2): 186–191.
-Ogien, Ruwen. 2004. *La panique morale*. Paris: Grasset.
-Ogien, Ruwen. 2011. *L'influence de l'odeur des croissants chauds sur la bonté humaine*. Paris: Grasset.
-Palminteri, Stefano, Maël Lebreton, Yulia Worbe, David Grabli, Andreas Hartmann, et Mathias Pessiglione. 2009. «Pharmacological modulation of subliminal learning in

Parkinson's and Tourette's syndromes». *Proceedings of the National Academy of Sciences* 106 (45): 19179 -19184.

-Parker, Andrew J., et William T. Newsome. 1998. « Sense and the single neuron: probing the physiology of perception ». *Annual review of neuroscience* 21 (1): 227–277.

-Platt, Michael L, et Scott A. Huettel. 2008. « Risky business: the neuroeconomics of decision making under uncertainty ». *Nature neuroscience* 11 (4): 398-403.

-Pessiglione, Mathias, Ben Seymour, Guillaume Flandin, Raymond J. Dolan, et Chris D. Frith. 2006. « Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans ». *Nature* 442 (7106): 1042–1045.

-Petrides, Michael, et Brenda Milner. 1982. « Deficits on subject-ordered tasks after frontal- and temporal-lobe lesions in man ». *Neuropsychologia* 20 (3): 249–262.

-Phelps, Edmund S. & Robert A. Pollak. 1968. « On Second-Best National Saving and Game-Equilibrium Growth ». *The Review of Economic Studies* 35(2) : 185-199.

-Platt, Michael L., et Paul W. Glimcher. 1999. « Neural correlates of decision variables in parietal cortex ». *Nature* 400 (6741): 233-238.

-Poldrack, R. 2006. « Can cognitive processes be inferred from neuroimaging data? » *Trends in Cognitive Sciences* 10 (2): 59–63.

-Prévost, Charlotte, Mathias Pessiglione, Elise Météreau, Marie-Laure Cléry-Melin, et Jean-Claude Dreher. 2010. « Separate valuation subsystems for delay and effort decision costs ». *The Journal of Neuroscience* 30 (42): 14080–14090.

-Prost, Antoine. 1996. *Douze leçons sur l'histoire*. Paris: Seuil.

-Quartz, Steven R. 2008. « From cognitive science to cognitive neuroscience to neuroeconomics ». *Economics and Philosophy* 24 (3): 459–472.

-Rabin, Mathew. 1993.« Incorporating Fairness into Game Theory and Economics ». *American Economic Review* 83 (5): 1281–1302.

-Rabin, Matthew et Botond Kőszegi. 2006. « A Model of Reference-Dependent Preferences ». *The Quarterly Journal of Economics* 121 (4): 1133-1165.

-Rachlin, Howard & Leonard Green. 1972. « Commitment, choice and self-control ». *Journal of the Experimental Analysis of Behavior.* 17(1) : 15-22.

-Rachlin, Howard, 1995. « Behavioral economics without anomalies ». *Journal of the Experimental Analysis of Behavior* 64(3): 397-404.

-Rachlin, Howard, John Kagel & Raymond Battalio.1982. « Maximization theory in behavioral psychology ». *Behavioral and Brain Sciences* 4(3):371.

-Rachlin, Howard, Leonard Green & Barbara Tormey, B. 1988. « Is there a decisive test

between matching and maximizing? » *Journal of the Experimental Analysis of Behavior* 50(2) : 113-123.

-Rachlin, Howard. 1983. « How to decide between matching and maximizing: A reply to Prelec ». *Psychological Review* 90(4) : 376-379.

-Rachlin, Howard. 1991. *Introduction to Modern Behaviorism.* San Francisco: W.H. Freeman & Company.

-Rangel, Antonio, Colin Camerer, et P. Read Montague. 2008. « A framework for studying the neurobiology of value-based decision making ». *Nat Rev Neurosci* 9 (7): 545-556.

-Read, Daniel 2001. « Is time-discounting hyperbolic or subadditive? » *Journal of risk and uncertainty* 23 (1): 5–32.

-Rescorla, Robert, et Alan Wagner. 1972. « Variations in the Effectiveness of Reinforcement and Nonreinforcement ». *New York: Classical Conditioning II: Current Research and Theory, Appleton-Century-Crofts*.

-Rilling, James K., David A. Gutman, Thorsten R. Zeh, Giuseppe Pagnoni, Gregory S. Berns, et Clinton D. Kilts. 2002. « A neural basis for social cooperation ». *Neuron* 35 (2): 395–405.

-Rizzo, Mario J., et Glen Whitman. 2009. « Little brother is watching you: New paternalism on the slippery slopes ». *Arizona Law Review* 51 (3): 685–739.

-Robinson, David L., Michael E. Goldberg, et Gregory B. Stanton. 1978. « Parietal association cortex in the primate: sensory mechanisms and behavioral modulations. » *Journal of Neurophysiology* 41 (4): 910–932.

-Roitman, Jamie D., et Michael N. Shadlen. 2002. « Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task ». *The Journal of Neuroscience* 22 (21): 9475–9489.

-Rose, Nikolas. 1999. *Governing the Soul: The Shaping of the Private Self*. New York: Free Association Books.

-Rosenblatt, Franck. 1958. « The perceptron: A probabilistic model for information storage and organization in the brain. » *Psychological review* 65 (6): 386.

-Ross, Don, Carla Sharp, Rudy E. Vuchinich, et David Spurrett. 2008. *Midbrain Mutiny: The Picoeconomics and Neuroeconomics of Disordered Gambling: Economic Theory and Cognitive Science*. New York: Bradford Books.

-Ross, Don. 2005. *Economic Theory and Cognitive Science: Microexplanation.* Cambridge: The MIT Press.

-Ross, Don. 2008. « Two Styles of Neuroeconomics ». *Economics and Philosophy* 24 (Special

Issue 03): 473-483.

-Ross, Don, 2011. « Addictive, Impulsive and other Counter-Normative Consumption ». à paraître dans Victoria Wells and Gordon Foxall. *New developments in Consumer Behaviour*. Edward Elgar.

-Rotheram-Fuller, Erin, Steven Shoptaw, Steven M. Berman, et Edythe D. London. 2004. « Impaired performance in a test of decision-making by opiate-dependent tobacco smokers ». *Drug and alcohol dependence* 73 (1): 79–86.

-Rowland, Neil E., Cheryl H. Vaughan, Clare M. Mathes, et Anaya Mitra. 2008. « Feeding behavior, obesity, and neuroeconomics ». *Physiology & behavior* 93 (1-2): 97-109.

-Rubinstein, Ariel. 2003. « "Economics and Psychology"? The Case of Hyperbolic Discounting », *International Economic Review* 44(4): 1207-1216.

-Rubinstein, Ariel. 2007. « Instinctive and Cognitive Reasoning: A Study of Response Times* ». *The Economic Journal* 117 (523): 1243–1259.

-Rubinstein, Ariel. 2008. «Comments on Neuroeconomics». *Economics and Philosophy* 24 (Special Issue 03): 485-494.

-Sally, David, et Elisabeth Hill. 2006. « The development of interpersonal strategy: Autism, theory-of-mind, cooperation and fairness ». *Journal of Economic Psychology* 27 (1) (février): 73-97.

-Sally, David. 2000. « A general theory of sympathy, mind-reading, and social interaction, with an application to the prisoners' dilemma ». *Social science information* 39 (4): 567–634.

-Sally, David. 2001. « On sympathy and games ». *Journal of Economic Behavior & Organization* 44 (1): 1–30.

-Samuelson, Paul A. 1938. « A Note on the Pure Theory of Consumer's Behaviour ». *Economica* 5(17): 61-71.

-Sanfey, Alan G., James K. Rilling, Jessica A. Aronson, Leigh E. Nystrom, et Jonathan D. Cohen. 2003. « The Neural Basis of Economic Decision-Making in the Ultimatum Game ». *Science* 300 (5626) (juin 13): 1755-1758.

-Sanfey, Alan. G., Georges Loewenstein, Samuel M. McClure, et Jonathan D. Cohen. 2006. « Neuroeconomics: cross-currents in research on decision-making ». *Trends in cognitive sciences* 10 (3): 108–116.

-Sapra, Steven G., et Zak, Paul. 2010. Eight lessons from neuroeconomics for money managers. *CFA Institute Research Foundation Publications, Behavioral Finance and Investment Management* 2: 63-76.

-Saver, Jeffrey L., et Antonio R. Damasio. 1991. « Preserved access and processing of social knowledge in a patient with acquired sociopathy due to ventromedial frontal damage ». *Neuropsychologia* 29 (12): 1241–1249.

-Schelling, Thomas C. 1984. « Self-command in practice, in policy, and in a theory of rational choice ». *The American Economic Review* 74 (2): 1–11.

-Schmidt, Christian. 2010. *Neuroéconomie : Comment les neurosciences transforment l'analyse économique*. Paris: Odile Jacob.

-Schultz, Wolfram, Paul Apicella, Eugenio Scarnati, et Tomas Ljungberg. 1992. « Neuronal activity in monkey ventral striatum related to the expectation of reward ». *The Journal of Neuroscience* 12 (12): 4595–4610.

-Schultz, Wolfram, Peter Dayan, et P. Read. Montague. 1997. « A neural substrate of prediction and reward ». *Science* 275 (5306): 1593.

-Schultz, Wolfram. 2002. « Getting formal with dopamine and reward ». *Neuron* 36 (2): 241–263.

-Schultz, Wolfram. 2007. « Behavioral dopamine signals ». *Trends in neurosciences* 30 (5): 203–210.

-Selten, Reinhart. 1975. « Reexamination of the perfectness concept for equilibrium points in extensive games ». *International journal of game theory* 4 (1): 25–55.

-Sent, Esther-Mirjam. 2004. « Behavioral Economics: How Psychology Made Its (Limited) Way Back Into Economics ». *History of Political Economy* 36 (4): 735-760.

-Sent, Esther-Mirjam. 2005. « Simplifying Herbert Simon ». *History of political economy* 37 (2): 227–232.

-Shadlen, Michael N., et William T. Newsome. 1996. « Motion perception: seeing and deciding ». *Proceedings of the National Academy of Sciences* 93 (2): 628 -633.

-Shallice, Tim., & Evans, M.E. 1978. « The involvement of the frontal lobes in cognitive estimation ». *Cortex* 14: 294-303.

-Shiv, B., G. Loewenstein, et A. Bechara. 2005. « The dark side of emotion in decision-making: When individuals with decreased emotional reactions make more advantageous decisions ». *Cognitive Brain Research* 23 (1): 85–92.

-Simon, Herbert. 1957. *Models of man; social and rational*. New York: John Wiley.

-Simon, Herbert. 1976. *Administrative behavior.* Vol. 3. Cambridge: Cambridge Univ Press.

-Slovic, Paul, et Sarah Lichtenstein. 1968. « Relative importance of probabilities and payoffs in risk taking. » *Journal of Experimental Psychology; Journal of Experimental Psychology* 78 (32): 1-18

-Smith, John Maynard. 1982. *Evolution and the Theory of Games*. Cambridge: Cambridge University Press.

-Smith, Kip, John Dickhaut, Kevin McCabe, et John V. Pardo. 2002. « Neuronal substrates for choice under ambiguity, risk, gains, and losses ». *Management Science*: 711–718.

-Smith, Vernon. 2007. *Rationality in Economics: Constructivist and Ecological Forms*. Cambridge University Press.

-Sokol-Hessner, Peter, Ming Hsu, Nina G. Curley, Mauricio R. Delgado, Colin F. Camerer, et Elizabeth A. Phelps. 2009. « Thinking like a trader selectively reduces individuals' loss aversion ». *Proceedings of the National Academy of Sciences* 106 (13): 5035 -5040.

-Spitzer, Manfred, Urs Fischbacher, Bärbel Herrnberger, Georg Grön, et Ernst Fehr. 2007. « The neural signature of social norm compliance ». *Neuron* 56 (1): 185-196.

-Stephens, David W., et J. R. Krebs. 1987. *Foraging Theory*. Princeton: Princeton University Press.

-Strotz, Robert H. 1955. « Myopia and Inconsistency in Dynamic Utility Maximization ». *The Review of Economic Studies* 23(3): 165-180.

-Sunstein, Cass. R. et Richard H. Thaler. 2003. « Libertarian Paternalism Is Not An Oxymoron », *The University of Chicago Law Review* 70 (4): 1159-1202

-Sutton, Richard S., et Andrew G. Barto. 1998. *Reinforcement Learning: An Introduction*. Cambridge: MIT Press.

-Takahashi, Taiki. 2010. « Toward molecular neuroeconomics of obesity ». *Medical hypotheses* 75 (4): 393–396.

-Tanji, Jun, et Edward V. Evarts. 1976. « Anticipatory activity of motor cortex neurons in relation to direction of an intended movement ». *Journal of Neurophysiology* 39 (5): 1062 -1068.

-Terrall, Mary. 2006. «Biography as Cultural History of Science». *Isis* 97 (2): 306-313.

-Thaler, Richard H. & Hersh M. Shefrin. 1981. « An Economic Theory of Self-Control ». *Journal of Political Economy* 89(2): 392-406.

-Thaler, Richard H. 1981. « Some empirical evidence on dynamic inconsistency ». *Economics Letters* 8(3): 201-207.

-Thaler, Richard H., et Cass R. Sunstein. 2003. « Libertarian Paternalism ». *The American Economic Review* 93 (2): 175-179.

-Thaler, Richard H., et Cass R. Sunstein. 2008. *Nudge: Improving Decisions About Health, Wealth, and Happiness*. New York: Penguin.

-Thaler, Richard H., et Shlomo Benartzi. 2004. « Save more tomorrow^{TM}: Using behavioral

economics to increase employee saving ». *Journal of political Economy* 112 (S1): S164–S187.

-Tricomi, Elizabeth, Antonio Rangel, Colin F. Camerer, et John. P. O'Doherty. 2010. « Neural evidence for inequality-averse social preferences ». *Nature* 463 (7284): 1089–1091.

-Tversky, Amos, et Daniel Kahneman. 1986. « Rational choice and the framing of decisions ». *Journal of business*: 251–278.

-Tversky, Amos, Paul Slovic, et Daniel Kahneman. 1990. « The Causes of Preference Reversal ». *The American Economic Review* 80 (1): 204-217.

-Vallois, Nicolas. 2012. « The Pathological Paradigm of Neuroeconomics ». *Oeconomia* 1(4): 525-556.

-Vercoe, Moana, et Paul Zak. 2010. « Inductive modeling using causal studies in neuroeconomics: brains on drugs ». *Journal of Economic Methodology* 17 (2): 133–146.

-Veyne, Paul. 1970. *Comment on écrit l'histoire*. Paris: Seuil.

-Voon, Valerie., Mathias Pessiglione, Christina Brezing, Cecile Gallea, Hubert H. Fernandez, Raymond J. Dolan, et Mark Hallett. 2010. « Mechanisms underlying dopamine-mediated reward bias in compulsive behaviors ». *Neuron* 65 (1): 135–142.

-Vranas, Peter. 2000. « Gigerenzer's normative critique of Kahneman and Tversky ». *Cognition* 76 (3): 179–193.

-Weber, Bernd, Andreas Aholt, Carolin Neuhaus, Peter Trautner, Christian E. Elger, et Thorsten Teichert. 2007. « Neural evidence for reference-dependence in real-market-transactions ». *Neuroimage* 35 (1): 441–447.

-Wacongne, Catherine, Jean-Pierre Changeux, et Stanislas Dehaene. 2012. « A Neuronal Model of Predictive Coding Accounting for the Mismatch Negativity ». *The Journal of Neuroscience* 32 (11): 3665-3678.

-Weintraub, Roy E., et Evelyn L. Forget. 2007. «Introduction». *History of Political Economy* 39 (Suppl 1): 1-6.

-White, Hayden. 1973. *Metahistory: The Historical Imagination in Nineteenth-Century Europe*. Cambridge: Johns Hopkins University Press.

-Whitman, Glen. 2006. *Against the new paternalism: internalities and the economics of self-control*. *Working Paper.* Cato Institute.

-Willinger, Marc, et Nicolas Eber. 2005. *L'économie expérimentale*. Paris: Éditions La Découverte.

-Windmann, Sabine, Peter. Kirsch, Daniela Mier, Rudolf Stark, Bertram. Walter, Onur Güntürkün, et Dieter Vaitl. 2006. « On framing effects in decision making: linking lateral

versus medial orbitofrontal cortex activation to choice outcome processing ». *Journal of Cognitive Neuroscience* 18 (7): 1198–1211.

-Xue, Gui, Dara G. Ghahremani, et Russell A. Poldrack. 2008. « Neural Substrates for Reversing Stimulus–Outcome and Stimulus–Response Associations ». *The Journal of Neuroscience* 28 (44): 11196 -11204.

-Zak, Paul J., Angela A. Stanton, et Sheila Ahmadi. 2007. « Oxytocin increases generosity in humans ». *PLoS One* 2 (11): 1128.

-Zak, Paul J., Robert Kurzban, Sheila Ahmadi, Ronald S. Swerdloff, Jang Park, Levan Efremidze, Karen Redwine, Karla Morgan, et William Matzner. 2009. « Testosterone administration decreases generosity in the ultimatum game ». *PLoS One* 4 (12): 8330.

-Zak, Paul. J., Robert. Kurzban, et William. T. Matzner. 2004. « The neurobiology of trust ». *Annals of the New York Academy of Sciences* 1032 (1): 224–227.

-Zerki Semir M. 1974. « Functional organization of avisual area in the posterior bank of the superior temporal sulcus of the rhesus monkey ». *Journal of Physiology* 236: 549–73