Ni élimination, ni réduction, ni intégration alors quoi? Sciences de l'esprit, neurosciences et modèle instrumental

Denis Forest

Intellectica, 2018/1-2, **69**, pp. 167-185

Ni élimination, ni réduction, ni intégration, alors quoi ? Sciences de l'esprit, neurosciences et modèle instrumental

Denis FOREST*

RÉSUMÉ. Depuis la publication de *Neurophilosophie* en 1986, plusieurs théories ont visé à rendre compte en philosophie de ce que les relations entre psychologie et neurosciences devraient être – en particulier, l'éliminativisme, le réductionnisme et le modèle intégratif plus récent qui conçoit l'explication psychologique comme une simple esquisse par rapport à l'explication neurocognitive de type mécaniste. Cet article présente une discussion critique de ces modèles. En particulier, faire de la psychologie un composant à l'intérieur du programme multidisciplinaire plus large des neurosciences peut conduire à négliger les préoccupations des chercheurs dans les diverses sciences de l'esprit et à méconnaître les raisons pour lesquelles ils attachent de l'importance aux neurosciences lorsqu'ils cherchent à atteindre les buts qui sont constitutifs de leurs disciplines. En conséquence, un modèle alternatif est proposé pour penser l'utilité que peuvent avoir les neurosciences pour la psychologie.

Mots-clés : Psychologie, neurosciences, éliminativisme, coévolution, réductionnisme, intégration, analyse fonctionnelle, pluralisme.

ABSTRACT. No Elimination, no Reduction, no Integration: Then What? Sciences of the Mind, Neuroscience and the Instrumental Model. Since the publication of *Neurophilosophy* in 1986, several accounts have been offered in philosophy of what the relations between psychology and neuroscience should be -namely, eliminativism, reductionism, and the more recent integrative model which construes psychological explanation as a mere sketch of a mechanistic, neurocognitive explanation. In this paper a critical discussion of these various attempts is offered. In particular, presenting psychology as a component of the wider, multidisciplinary program of neuroscience comes with the risk of neglecting what the researchers involved in the different sciences of the mind worry about, and the very reason why they value neuroscience as an essential tool to reach the goals constitutive of their disciplines. Accordingly, I suggest an alternative model to understand the usefulness of neuroscience to psychology.

Keywords: Psychology, neuroscience, eliminativism, coevolution, reductionism, integration, functional analysis, pluralism.

INTRODUCTION

Quand les philosophes des sciences s'intéressent aux sciences de l'esprit, le problème qu'ils soulèvent est souvent de savoir ce que fait (ou peut faire) aux sciences de l'esprit le développement de certaines branches des sciences de la nature, en particulier des neurosciences. Il est remarquable en soi que la question posée au sujet de sciences de l'esprit soit celle de leur relation à autre chose qu'elles, alors qu'on peut très bien se poser des questions philosophiques qui s'adressent à la psychologie elle-même, comme celle de la nature des

* Université Paris 1 Panthéon-Sorbonne et IHPST, Paris – UFR de Philosophie – 75005 Paris.
denis.forest@univ-paris1.fr

explications en psychologie, ce qu'a fait par exemple Robert Cummins[1]. Tout en analysant au préalable quelques représentations communes de la solution qu'on peut donner à ce problème des relations entre sciences de l'esprit et neurosciences, je proposerai une autre manière de l'envisager.

Avant d'aborder une revue des solutions déjà proposées et une alternative à celles-ci, sans doute faut-il justifier l'expression « sciences de l'esprit » qui est une manière de ne pas dire « psychologie ». Il y a deux raisons à ce choix. La première est une question de filiation historique des débats contemporains. L'expression « sciences de l'esprit » (*Geisteswissenchaften*) est celle qui a été retenue pour rendre « sciences morales » (*moral sciences*) dans la première traduction allemande de la *Logique* de Stuart Mill[2]. Dans le Livre VI de la *Logique*, Mill se posait en philosophe des sciences le problème des relations entre lois de l'esprit (*laws of mind*) et lois physiologiques, en particulier entre association mentale et association neurale (comme on le verra théorisé plus tard par Donald Hebb). Constatant une asymétrie entre ce que *connaît déjà* la psychologie (les lois de l'association des idées) et ce que *postule seulement* la physiologie cérébrale (des mécanismes rendant compte d'une telle association) il défendait à la fois l'unité méthodologique de la science et l'autonomie *de fait* de la psychologie entendue comme connaissance des lois fondamentales de l'esprit humain, actuellement non dérivables de lois physiologiques établies. Le débat actuel me semble beaucoup plus tributaire du débat de Mill avec la psychophysiologie naissante qu'on ne reconnaît généralement. En effet, Mill voyait dans la psychologie non pas seulement un ensemble de lois (actuellement non réduites à d'autres lois) mais comme un domaine d'objets pour la connaissance scientifique et comme un accès spécifique à ces objets (d'où sa critique de la critique par Comte de l'observation intérieure)[3].

La seconde raison est que mettre « psychologie » au singulier c'est risquer de prendre la partie (la psychologie cognitive) pour le tout. On trouvera en psychologie des différences méthodologiques et des différences d'objet qui interdisent de faire de la psychologie cognitive, la branche des sciences de l'esprit qui a le moins d'aversion vis-à-vis des neurosciences, une partie sans spécificité de celles-ci, et qui interdisent tout autant de gommer a priori la différence entre psychologie expérimentale et psychologie clinique, ou de minorer la différence entre le programme de la psychologie évolutionniste et celui, par exemple, de la psychiatrie culturelle (la seule manière de minorer l'importance de ces différences serait d'entrer dans des considérations normatives qui consisteraient à juger non pas seulement de la valeur des travaux individuels, mais de la valeur des types de recherche eux-mêmes). À la question : les diverses branches des sciences de l'esprit possèdent-elles une forme d'unité propre qui permettrait d'envisager l'unification de « la » psychologie avec les neurosciences, la réponse doit être aujourd'hui négative si elle veut tenir compte de la variété effective des recherches. Il

[1] Cummins, 1983. Sur philosophie et psychologie, voir Engel, 1996.
[2] Mill, 1843. La traduction allemande de 1846 était due à J. Schiel. Cf. Hatfield, 1993, p. 544, note 56.
[3] Mill, 1865.

semble beaucoup plus réaliste de se donner d'emblée deux pluriels désignant deux familles de disciplines, de chercher à quoi peuvent ressembler des points de contact entre membres de ces deux familles et à quoi pourrait ressembler la multiplication de ceux-ci. Cette approche prudente n'est pas nécessairement une approche défaitiste. Par exemple, la neuropsychologie se présente à la fois comme une tradition de recherches relativement autonome au sein des neurosciences et comme une discipline dont les liens avec la psychiatrie font l'objet d'interrogations récurrentes. Une relation plus étroite entre neuropsychologie et psychiatrie[4], à supposer qu'elle soit possible, serait locale, et non globale, mais il y aurait certainement des leçons à en tirer d'une portée plus vaste.

Je procèderai en trois temps. Dans les deux premières parties, partant de l'éliminativisme, je présenterai deux autres types d'analyse récentes en philosophie des relations entre neurosciences et sciences de l'esprit, qui pensent ces relations en termes de réduction et en termes d'intégration. Ces analyses ne sont sans doute pas très attrayantes pour les chercheurs en sciences de l'esprit : en effet, à l'intérieur de chacune de ces approches, les relations entre sciences de l'esprit et neurosciences sont représentées par les philosophes des sciences sur le mode d'une unification où les sciences de l'esprit jouent le rôle peu exaltant d'un actionnaire minoritaire de l'entreprise scientifique, d'un détenteur de vérités seulement provisoires, partielles et superficielles. Dans la troisième partie, je présenterai deux objections au modèle de l'intégration, tout en explicitant les raisons pour lesquelles ce modèle a pour les philosophes l'intérêt d'une forme de « position par défaut » qu'on peut être tenté d'adopter par provision. Enfin, dans une dernière partie, je présenterai une analyse alternative, le modèle des neurosciences comme instrument. Je mettrai en particulier ce modèle à l'épreuve d'un cas, celui des relations entre mémoire et simulation du futur, avant de conclure sur la contribution des neurosciences à une pluralité de projets distincts.

I – LES TROIS SCÉNARIOS DE PATRICIA CHURCHLAND

Ce qu'on retient souvent de Patricia Churchland est ce qu'on appelle l'éliminativisme, que l'on présente comme une position parmi les options disponibles aujourd'hui en philosophie de l'esprit. Dans son livre *Neurophilosophie*[5], l'éliminativisme concerne bien le problème des relations corps-esprit (voir le titre du chapitre 7, « Réduction et le problème corps-esprit ») , mais le propos de Patricia Churchland est un propos qui s'inscrit pour une large part dans le cadre de la philosophie des sciences et non de la philosophie de l'esprit. La question de la réduction possible du mental au physique est en effet présentée par Patricia Churchland comme devant être posée à partir de la question de la réduction interthéorique telle qu'elle a été pensée par un philosophe des sciences comme Ernst Nagel (une théorie est réductible à une autre théorie si et seulement si les propositions de la théorie qu'on se propose de réduire sont déductibles des propositions de la théorie

[4] Murphy, 2006, propose de voir la psychiatrie comme la partie clinique des neurosciences cognitives.
[5] Churchland, 1986.

réductrice)[6]. La question est donc moins directement celle de l'esprit et du corps (ou du cerveau) que celle des relations entre deux types de théorie au lexique hétérogène. Cette conception de la réduction a été depuis très discutée, mais ce n'est pas ce qui importe pour le présent propos. Ce qui importe est plutôt de relever les points suivants.

Premièrement l'éliminativisme n'est pas présenté par Patricia Churchland comme une doctrine qui aurait sa valeur en soi (comme l'épiphénoménisme ou le fonctionnalisme) mais comme une manière de lever un obstacle sur la voie de la réduction de la connaissance de l'esprit à la connaissance neuroscientifique : en s'inspirant de certains épisodes de l'histoire des sciences, ce que veut souligner Patricia Churchland est que toute réduction n'est pas nécessairement « rétentive » ; en clair, il peut se trouver des cas où la théorie ancienne au lieu d'être réduite par une nouvelle théorie plus puissante, est plutôt remplacée par elle. En clair, la chimie n'a pas réduit la théorie du phlogistique en la conservant, pas plus que la physique n'a réduit la théorie du calorique ; ces théories ont été simplement abandonnées ou éliminées. Il importe alors peu qu'en matière de « théorie psychologique » Patricia Churchland s'intéresse à ce qu'elle nomme « psychologie populaire » (le parler ordinaire au sujet des états mentaux) plutôt qu'à telle théorie psychologique plus sophistiquée : la position de Patricia Churchland présente dans tous les cas la connaissance actuelle de l'esprit (que ce soit sous forme scientifique, ou infrascientifique) comme ce à quoi nous devrions tout simplement devoir renoncer.

La seconde chose à souligner, c'est que Patricia Chuchland a recours à la métaphore du spectre, pour exprimer l'idée selon laquelle il y a *a priori* tout une variété de formes et de résultats possibles de la réduction. La réduction non-rétentive (qui procède par élimination) est située à une extrémité de ce spectre, dans le cas où l'ancienne théorie disparaît : comme le phlogsitique, la théorie psychologique désuète rejoint alors le cimetière des idées. Mais le point qu'il faut souligner est que la métaphore du spectre implique que si la réduction n'est pas nécessairement rétentive (sans cela, l'éliminativisme ne serait pas une option), la réduction n'est pas nécessairement non-rétentive non plus : les relations entre psychologie et neurosciences pourront prendre diverses formes, certaines plus proches de l'élimination, et d'autres de la rétention pure et simple à l'autre extrémité du spectre. En supposant que la conception chomskyenne du langage soit vraie, par exemple, la réalisation physique de l'opération « merge » étant connue, on aurait affaire à une identification et donc à une réduction rétentive ou conservatrice[7]. Patricia Churchland est donc loin de promettre l'élimination des théorie psychologiques connues comme un avenir inévitable pour elles. Elle dit au contraire, on ne peut plus clairement, que le point qui importe est que la réduction *peut* être non-rétentive ; et comme la question est une question empirique à laquelle la réponse ne peut venir que de changements scientifiques largement imprévisibles, cela n'aurait tout simplement pas de sens d'affirmer

[6] Nagel, 1979.

[7] Berwick & Chomsky, 2016, chapitre 4.

dans le cadre de sa philosophie qu'une bonne réduction *doit être* non-rétentive. Autant dire que Patricia Churchland ne propose pas une *doctrine* éliminativiste, mais une *stratégie* éliminativiste à laquelle la recherche peut avoir ou non recours.

La troisième chose à retenir, c'est que non seulement ce qui importe à Patricia Churchland est de penser les relations entre théorie psychologique et théorie neuroscientifique en les inscrivant dans le temps de leur dynamique (c'est la fameuse idée de la « coévolution des théories »), mais en outre, elle n'a pas une représentation homogène d'une telle coévolution. Un de ses lecteurs, Robert N. McCauley[8], a ainsi pu distinguer ce qu'il appelle coévolution$_S$, coévolution$_M$ et coévolution$_P$. La coévolution$_S$ est celle qui aboutit à une élimination (S désignant une révolution scientifique, avec abandon de l'ancienne théorie qui n'a pu être sauvée). La coévolution$_M$ est celle qui aboutit à une réduction rétentive (M désignant une microréduction) où la théorie réduite demeure correcte bien qu'elle soit désormais déductible de la théorie réductrice. Mais à quoi peut bien correspondre la coévolution$_P$? McCauley s'appuie pour la concevoir sur les passages où Patricia Churchland reconnaît que pendant la phase de coévolution, chaque théorie (en clair la théorie psychologique *aussi bien que* la théorie neuroscientifique) peut *informer* et *corriger* l'autre[9]. La théorie à réduire n'est donc pas conçue par Patricia Churchland uniquement comme une « proie » passive qui se borne à attendre la réduction qui la fera accéder à une forme de scientificité nouvelle tout en lui faisant perdre son autonomie. Un exemple que prend Patricia Churchland est la psychologie de la mémoire. Au chapitre neuf, elle souligne en particulier l'importance qu'a eu le cas HM, ce patient épileptique sur lequel on a pratiqué en 1953 une résection bilatérale du lobe temporal interne[10]. Ce patient ne formait plus de nouveaux souvenirs alors qu'il pouvait apprendre de nouvelles formes de coordination visuo-motrice, ce qui a invité à distinguer entre « mémoire descriptive » et « mémoire procédurale », mais aussi à postuler des systèmes de mémoire indépendants et à supposer que la mémoire procédurale est indépendante des lobes temporaux. La neuropsychologie clinique (ici, considérée comme une branche de la psychologie) contribue donc à façonner les *explananda* des neurosciences de la mémoire, mais aussi leur agenda lorsqu'il s'agit de postuler, de rechercher et d'identifier des mécanismes spécifiques. La science « réductible » exerce donc des contraintes sur la science réductive en régime de coévolution$_P$. Pour paraphraser Kant, sans neurosciences, la psychologie est vide, mais sans psychologie, les neurosciences sont aveugles.

Bien entendu, on pourrait soutenir que McCauley a tort de distinguer la coévolution$_P$ (p pour pluraliste) de la coévolution$_M$, en tant que la coévolution$_P$, où il y a enrichissement mutuel sans réduction en vue, est simplement présentée par Patricia Churchland comme une étape sur la voie de la coévolution$_M$. Mais à nouveau, il en va de la coévolution des théories comme

[8] Mc Cauley, 1996.
[8] Churchland, *op. cit.*, p. 284.
[9] Scoville et Milner, 1957 ; Corkin, 2013.

du spectre des options en matière de réduction : Patricia Churchland admet explicitement qu'il est *possible* que la réduction ne se produise finalement pas (réduction de la psychologie de la mémoire aux neurosciences de la mémoire, ou de la génétique de la transmission des caractères à la génétique moléculaire). Ce qui paraissait être la préparation d'une forme de réduction pourrait s'avérer être à terme un processus co-évolutif sans fin. Le propre de la position de Patricia Churchland est de penser qu'on aurait tort de prédire avec assurance que la réduction ne se produira pas, il n'est pas de l'exclure a priori, puisqu'une fois encore ce qu'elle décrit est une gamme d'éventualités et que le renoncement à la philosophie première coïncide pour elle avec le renoncement à trancher ce type de question a priori. L'idée de coévolution$_P$ est donc celle de disciplines qui se développent au contact l'une de l'autre ; elle est donc celle d'une alternative à l'élimination comme à la réduction qui n'est pourtant pas un état de stagnation de la recherche. « L'éliminativisme de Patricia Churchland » pouvant être « réduit » à une sorte d'épouvantail naturaliste et scientiste, il semble utile de présenter la philosophie des relations entre psychologie et neurosciences qu'elle a proposé comme une revue systématique des scénarios alternatifs pour un avenir ouvert, et ce tout autant qu'une défense argumentée du réductionnisme.

II – DE LA RÉDUCTION À L'INTÉGRATION

La perspective sur les neurosciences que propose Carl Craver dans une série de travaux[11] est différente de celle de Patricia Churchland, et cela au moins pour deux raisons. La première est qu'il propose de remplacer le modèle déductif-nomologique (DN), que Patricia Churchland accepte, par un modèle différent de l'explication scientifique, appelé mécaniste[12]. La seconde, qui nous concerne plus directement ici, est que l'unification n'est plus conçue par lui en termes de réduction des théories. En effet, le modèle alternatif de l'unification dont il a proposé deux variantes dérive en partie des propositions de Nancy Maull et Lindley Darden[13] sur l'unification non-réductive. Les travaux de Maull et Darden sont importants dans la mesure où ils rompent avec le type de position du problème (dérivation ou non-dérivation des lois psychologiques) que partageaient Stuart Mill et Patricia Churchland, quelle que soit la différence entre le scepticisme du premier et l'optimisme gnoséologique de la seconde. La description qui suit du modèle de Maull et Darden, puis des deux propositions de Craver, permet de présenter une des conceptions les plus discutées de l'unification dans la littérature philosophique contemporaine, et de présenter aussi, de ce fait, une conception de la manière dont les sciences de l'esprit pourraient aujourd'hui être appelées à être intégrées aux neurosciences, en accord avec une conception renouvelée de ce qu'est l'unité de la science.

---

[11] Craver, 2007 , Piccinini & Craver, 2011.

[12] Machamer, Darden & Craver, 2000. Soulignons cependant que Churchland était prête à reconnaître dans *Neurophilosophy* qu'un changement de modèle de l'explication ne modifierait pas en profondeur son analyse. Elle évoquait alors le travail de Cummins dont toute la philosophie des mécanismes s'est inspirée.

[13] Maull, 1977 ; Darden & Maull, 1977.

Avec leur modèle des « théories inter-champs », Darden et Maull ont proposé trois inflexions par rapport au modèle de la réduction inter-théorique. Pour penser l'unification des sciences en visant à une forme d'adéquation descriptive, il faut tout d'abord selon elles non pas penser l'unification en termes de relation entre des *théories*, mais de relation entre des *champs de recherche*. Un champ est une « aire » de la science (*area of science*) qui peut être identifié à partir de plusieurs éléments qui le constituent : un problème central, un domaine constitué d'items reliés à ce problème, des contraintes sur la solution possible à celui-ci, des techniques et des méthodes appropriées, et aussi, éventuellement, des concepts des lois et des théories reliées au problème. Par exemple, la génétique et la cytologie étaient deux champs au début du XX^{e} siècle. Le problème central de la cytologie était l'analyse de la cellule en composants et le problème central de la génétique était l'explication de l'héritabilité des traits. En second lieu, Darden et Maull ont proposé de changer de modèle de l'unification : une théorie inter-champ est une théorie qui explique les relations entre deux champs sans les réduire l'un à l'autre. Une telle théorie répond aux problèmes qui se posaient dans ces champs sans pouvoir être résolus avec les ressources internes à un seul de ces champs. Ainsi, pour reprendre le même exemple, les gènes étaient des entités hypothétiques avec des fonctions connues ; et les chromosomes étaient des entités visibles au microscope dont la fonction était inconnue. L'apport de Walter Sutton et de Theodor Boveri a été de proposer une théorie interchamp dans laquelle les gènes étaient situés sur les chromosomes. Le changement de modèle de l'unification est ainsi aisé à percevoir : en tant que la théorie chromosomique de l'hérédité mendélienne procure un pont entre génétique et cytologie, elle est l'instrument d'une unification qui ne passe pas par la réduction d'une « théorie » à une autre, ni d'un champ à un autre, mais par la contribution de chaque champ à la résolution du problème central du champ scientifique connexe. Dès lors, et en troisième lieu, c'est la conception de l'unité de la science elle-même qui est modifiée : celle-ci ne consisterait pas en une « succession hiérarchique de réductions entre théories, mais plutôt en l'établissement de ponts entre des champs au moyen de théories interchamps »[14]. *In fine*, le réseau complexe de telles connexions serait ce qui réalise l'unité de la science.

Dans son livre de 2007, Carl Craver adresse deux critiques à Darden et Maull, tout en reprenant leur idée d'unification non-réductive comme alternative correcte au modèle de la réduction inter-théorique[15]. La première est que selon lui, la structure des théories inter-champs n'est pas assez explicitée par ces auteurs. La seconde, qui est la plus importante, consiste à dire que, en ce qui concerne les neurosciences en tout cas, la charge de la connexion entre champs, et de l'explication qui spécifie la nature de cette connexion, n'est pas assurée par une « théorie », mais par la description d'un mécanisme. Ce mécanisme réunit plusieurs champs en tant qu'il est constitué de plusieurs niveaux articulés entre eux, et que chaque champ de recherche a vocation à préciser la description correcte d'un de ces niveaux ou la manière dont deux

[14] Darden & Maull, 1977.
[15] Craver, 2007, p. 255.

d'entre eux peuvent être ajustés l'un à l'autre. C'est ce qui justifie l'usage de l'expression « modèle de l'unité en mosaïque des neurosciences » : « les découvertes dans différents champs des neurosciences sont utilisées, comme les tesselles d'une mosaïque pour élaborer[l'explication]à partir [d'un] mécanisme abstrait et pour façonner l'espace des mécanismes possibles »[16].

Deux notions clés sont mobilisées par Craver pour expliciter son modèle de l'unité en mosaïque, celle de niveau, et celle de contrainte. Concernant les niveaux d'un mécanisme, Craver insiste sur le fait que « l'intégration inter-champ » se produit à la fois entre les niveaux et à l'intérieur de chaque niveau : les conceptions traditionnelles de l'unité par la réduction ont négligé la question de l'intégration interchamp à un même niveau. Celle-ci se produit lorsque des chercheurs avec une formation différente, un lexique différent, des intérêts et des techniques d'investigation différents, s'intéressent au même processus à un niveau donné et complètent sa description. Mais à l'évidence, l'intégration inter-niveau demeure très importante. Dans l'article « Penser les mécanismes », l'idée était que les grands changements qui affectent l'explication ne sont pas ordinairement le fruit d'un intérêt exclusif pour des entités et des activités situées à des niveaux plus fondamentaux, mais qu'ils consistent en une articulation plus explicite et plus pertinente entre des niveaux hiérarchisés de mécanisme interconnectés[17]. Dans *Expliquer le cerveau*[18], l'exemple de la mémoire spatiale, déjà abordé par Bechtel et Richardson, montre comment l'intégration entre des niveaux de mécanismes est aussi une intégration entre des champs de connaissance différents[19]. Ces champs sont, dans ce cas, au nombre de quatre : l'étude du comportement (avec, par exemple, les travaux utilisant le labyrinthe de Morris pour tester les capacités d'orientation des rongeurs[20]), la neurobiologie (qui identifie la responsabilité causale des neurones de l'hippocampe dans la production des cartes de l'environnement), l'étude des liaisons synaptiques (ou étude du « niveau cellulaire-électrophysiologique ») et enfin, au fondement de cette hiérarchie de niveaux, l'étude des mécanismes moléculaires (activation des récepteurs NMDA) qui rendent possible, aux niveaux supérieurs, les liaisons synaptiques, la formation des cartes, et les altérations du comportement. La relation entre multiplicité des champs et unité en mosaïque est donc une relation entre multiplicité des *ressources* de l'explication, et unité de l'explication qui les intègre dans une hiérarchie de niveaux.

Quant à la notion de contrainte, il faut entendre par là « une découverte qui soit façonne les limites de l'espace des mécanismes possibles soit change la

[16] *Ibid*., p. 228.
[17] Machamer, Darden & Craver, 2000.
[18] Bechtel & Richardson, 2000/2010.
[19] Craver, 2007, p. 166.
[20] Le labyrinthe de Morris est une plateforme immergée dans une piscine et par là rendue invisible, dont l'emplacement peut être déterminé indirectement par rapport à des repères visuels.
[21] *Ibid*., p. 247.
[22] *Ibid*., p. 259.

distribution des probabilités à l'intérieur de cet espace »[21]. L'idée est que les découvertes venant de chaque champ excluent certaines options à l'intérieur des champs connexes et augmentent ou diminuent la probabilité selon laquelle certaines options sont correctes. Elle est aussi que la régulation de la recherche se fait selon une norme de la connaissance qui est la cohérence de chaque contribution avec les autres, une cohérence qui n'est pas le produit de la simple compatibilité ou non-contradiction logique entre propositions mais qui impose que chaque composant du mécanisme s'insère dans un modèle donné de la production du phénomène étudié.

À l'intérieur de son modèle de l'unité en mosaïque, Craver retrouve la thématique qui était celle de Patricia Churchland lorsqu'elle prenait en compte, avant la phase d'unification réductrice, les relations réciproques entre des recherches menées dans des domaines hétérogènes (la coévolution$_P$ de McCauley). Ainsi les chercheurs impliqués dans des « champs où on caractérise des phénomènes de haut niveau influencent ceux qui travaillent sur des phénomènes de plus bas niveau »[22]. L'exemple en est la distinction par Tolman entre mémoire des cartes et mémoire des routes, qui prédisait une ségrégation entre des mécanismes neuraux distincts. Les sciences de la cognition et du comportement vont donc non seulement fournir des phénomènes à expliquer, mais de ce fait orienter la recherche dans telle ou telle direction, en filtrant les phénomènes de bas niveau et en déterminant lesquels méritent d'être étudiés. Des recherches et des controverses sur le comportement animal ont pu motiver des travaux aboutissant à des découvertes importantes en neuroscience[23]. À ces efforts pour apparier le niveau inférieur au niveau supérieur répondent les efforts opposés. Craver reprend l'exemple de HM, en en caractérisant la portée d'une manière différente de celle de Patricia Churchland que nous avons mentionnée plus haut. Alors que Patricia Churchland voyait dans les dissociations que met en évidence la neuropsychologie une motivation pour rechercher des mécanismes neuraux dissociés, Craver présente le même cas comme une raison de postuler des formes de mémoires distinctes. La neuropsychologie apparait comme une discipline qui suscite aussi bien la recherche d'une description adéquate des mécanismes neuraux épargnés ou atteints par les lésions (ce qui retient l'attention de Patricia Chuchland) que celle des systèmes cognitifs correspondants (ce qui retient celle de Craver).

Dans leur article de 2011, Piccinini et Craver[24] proposent ce qu'ils appellent un cadre pour la construction d'une science unifiée de la cognition. À la différence du livre de 2007, où les phénomènes cognitifs constituaient un niveau parmi d'autres, certes le plus élevé, et où les sciences de l'esprit n'étaient pas une cible privilégiée de l'analyse générale des neurosciences, cette fois les relations entre psychologie et neurosciences sont bien l'objet unique de l'analyse. En outre, en 2007, les sciences de la cognition et du

[23] O'Keefe & Nadel, 1978.
[24] Piccinin & Craver, 2011.

comportement étaient seulement la source d'une *caractérisation* des phénomènes que venaient expliquer les apports conjoints de disciplines d'une autre nature ; cette fois la psychologie est bien reconnue comme source d'explications en bonne et due forme, dont le type est l'analyse fonctionnelle de Cummins[25]. Piccinini et Craver se prononcent cependant contre l'idée selon laquelle l'autonomie de la psychologie serait de droit, et non simplement de fait. Pour eux les analyses fonctionnelles de la psychologie doivent être comprises comme des « esquisses de mécanismes ». Le concept d'esquisse de mécanisme avait été introduit dans l'article « Penser les mécanismes ». Machamer, Darden et Craver écrivaient alors : « Une esquisse est une abstraction pour laquelle les entités et les activités *constituantes* ne peuvent pas (encore) être précisées ou dans laquelle certaines étapes restent des cases vides. Dans la continuité productive qui relie deux étapes qui se suivent, il y a des parties manquantes, des boîtes noires que nous ne savons pas encore remplir »[26]. Parler d'esquisse de mécanisme c'est donc poser que les explications de la psychologie ont vocation à être complétées par les neurosciences pour atteindre leur *propre* but explicatif. L'explication en psychologie sera complète lorsque les détails de la réalisation neurale seront fournis. Comme la coévolution des théories aboutissait pour Patricia Churchland à la possibilité d'une unification réductive, la convergence de l'analyse fonctionnelle et de la connaissance de la réalisation neurale correspondante doit selon Piccinini et Craver permettre de parvenir à une explication intégrée des phénomènes de perception, de reconnaissance ou de mémoire. Là où Marr estimait que la distinction entre niveaux était utile et que l'étude des processus cognitifs gagnait à s'abstraire de la connaissance de leur implémentation[27], Piccinini et Craver complètent l'analyse *d'Expliquer le cerveau* en faisant de l'explication psychologique l'esquisse de l'explication neurocognitive.

III – LIMITES ET ATTRAIT DU MODÈLE DE L'INTÉGRATION

Il y a au moins deux raisons de demeurer sceptique vis-à-vis du modèle de l'intégration des sciences de la cognition aux neurosciences tel qu'il peut être défendu, soit à partir du modèle de « l'unité en mosaïque des neurosciences », soit dans la perspective de Piccinini et Craver selon laquelle les explications psychologiques sont de simples « esquisses de mécanisme ».

La première raison a été formulée par Jacqueline Sullivan[28]. Elle touche à la question des conditions sous lesquelles deux domaines scientifiques peuvent être reliés. Selon Sullivan, il y a une contrainte traditionnelle sur l'unité réductive qui demeure dans l'intégration explicative (cette contrainte pèse sur toute unification, quel qu'en soit le modèle), c'est ce qu'elle appelle la « connectabilité ». De même que dans le modèle de la réduction il faut qu'il y

[25] Cummins, 1975. L'idée de Cummins est celle de l'explication d'une capacité par un ensemble de sous-capacités, comme la circulation sanguine dépend des parties du système circulatoire et de leurs activités.
[26] Machamer, Darden & Craver, 2000.
[27] Marr, 1980.
[28] Sullivan, 2016.

ait coréférence entre le terme qu'utilise la théorie réductrice (chaleur) et celui qu'utilise la théorie réduite (énergie cinétique moyenne), de même pour intégrer les sciences de l'esprit et les sciences du cerveau il faut s'assurer que les concepts sont stabilisés, c'est-à-dire que la capacité cognitive qui est étudiée par le psychologue et celle dont les neurosciences entendent donner une explication mécaniste coïncident. Or il n'est pas certain que le laboratoire de psychologie et le laboratoire de neurosciences s'accordent dans la définition des termes, ni que les laboratoires de psychologie eux-mêmes s'accordent entre eux. À cela s'ajoute que les neurosciences ne peuvent pas résoudre tous les problèmes des sciences psychologiques à leur place. Dans une représentation optimiste de la recherche expérimentale, un paradigme expérimental (comme le labyrinthe de Morris) permet de « prendre sur le fait » l'exercice de la capacité A ; puis une recherche connexe peut permettre d'identifier les conditions neurales de A. Mais l'analyse de la littérature menée par Sullivan[29] tend à montrer qu'il est difficile de trancher la question de savoir quel phénomène cognitif le comportement des rongeurs dans le contexte du labyrinthe de Morris permet d'objectiver, les chercheurs oscillant dans leurs publications entre plusieurs descriptions alternatives de cette capacité (mémoire des lieux, cognition spatiale, apprentissage des lieux, etc.). « Morris a développé un moyen de détecter un ensemble de changements comportementaux mais il n'a pas identifié la fonction psychologique ou l'ensemble des changements représentationnels que ces changements comportementaux indiquent »[30]. Dès lors il est difficile de prétendre que l'on va donner grâce aux neurosciences des explications mécanistes des phénomènes cognitifs puisque la pré-condition du succès de l'entreprise explicative (la définition du phénomène à expliquer) n'est pas remplie. Le *motto* des philosophes mécanistes « il n'y a pas de mécanismes *simpliciter*, seulement des mécanismes *pour* des phénomènes »[31] se heurte aux difficultés de la spécification des dits phénomènes. Les neurosciences ne peuvent pas remédier à l'absence d'*explananda* bien définis.

La seconde difficulté est liée à la valeur accordée à l'unification dans le cadre du modèle de l'unité en mosaïque. Craver soutient que l'unité en mosaïque de la science ne fournit pas un critère de démarcation entre science et non-science (sans doute on peut imaginer beaucoup de liens entre deux pseudo-sciences qui mimeraient l'unité en mosaïque) mais qu'elle a la vertu épistémique de produire des explications robustes, et de permettre par conséquent une démarcation entre bonne et mauvaise (neuro)science. En pratique, des champs peuvent agréger des découvertes et parvenir à une forme d'unité par convergence tout à fait plausibles (et tenue pour telle dans des articles de revue à comité de lecture ayant passé toutes les barrières de l'évaluation) sans produire pour autant des explications robustes. La théorie du miroir brisé qui est une explication neuroscientifique de l'autisme est fondée 1. sur les propriétés connues des neurones miroirs et les fonctions qui leur sont généralement attribuées, 2. sur les difficultés de l'imitation dans l'autisme dont

[29] Sullivan, 2010.
[30] *Ibid*., p. 270.
[31] Craver & Bechtel, 2006.

font état certaines études, enfin 3. sur l'observation d'activation atypiques de l'activité des neurones miroirs chez les autistes. La théorie du miroir brisé est rendue attrayante par le fait que la gamme des domaines cognitifs dans lesquels les neurones miroirs sont impliqués coïncide assez bien avec la gamme des domaines dans lesquels des anomalies et des déficits sont constatés dans l'autisme. Si la théorie du miroir brisé n'est pas pour autant correcte[32], il semble que l'unité en mosaïque admette plusieurs formes, dont certaines ne coïncident pas avec le succès épistémique.

Il sera bien sûr possible de répondre que l'unité en mosaïque permet de rendre compte des raisons pour lesquelles on peut rejeter le modèle du miroir brisé (certaines découvertes diminuent la probabilité selon laquelle la théorie du miroir brisé serait correcte). Plus généralement, on pourrait réaffirmer que le point de vue philosophique sur la science est un point de vue normatif et que les difficultés d'un ordre ou d'un autre qui peuvent être rencontrées (instabilité persistante des concepts, formes dévoyées d'unité) sont seulement des embarras passagers qui ne doivent pas la concerner. Cependant, on peut souhaiter que la philosophie, sans abandonner son point de vue normatif, s'interroge sur les causes de la rivalité persistante entre des explications alternatives. Qu'elle s'interroge sur le fait que des formes d'unité seulement possibles ou plausibles peuvent passer durablement pour des formes d'unité associées à des formes d'explication plus robustes. Qu'elle réfléchisse sur ceci que l'addition des contraintes est un phénomène rendu délicat par la disparité des protocoles, des champs scientifiques et des intérêts de recherche, à tel point que la leçon à tirer des données disponibles et de la littérature sur un sujet donné peut faire débat entre experts du même champ.

Si le modèle de l'intégration de la psychologie aux neurosciences (qui succède aux formes non rétentive et rétentive de réduction qui ont été évoquées plus haut) demeure attirant malgré les problèmes qu'on vient de rappeler, c'est sans doute pour un ensemble de raisons assez hétérogènes qu'il n'est pas inutile d'expliciter.

La première raison qu'on peut invoquer est un présupposé d'homogénéité. L'analyse fonctionnelle sera conçue comme une esquisse d'explication mécaniste pour deux raisons : la première est que la stratégie explicative est la même et la seconde est que ce qu'il faut expliquer doit coïncider quel que soit le type de l'explication choisi, par analyse psychologique ou par description d'un mécanisme neural. On pourrait s'étonner que Piccinini et Craver ne prennent pas la peine d'examiner les explications psychologiques qui ne procèdent pas par analyse fonctionnelle (il y en a certainement : voir les explications narratives, ou motivationnelles par exemple). Mais on verra (avec la question de la profondeur explicative, ci-dessous) que leur parti-pris est en fait normatif. Ils estiment donc, implicitement, qu'il *faut* préférer des explications psychologiques procédant par analyse fonctionnelle à d'autres ne procédant pas ainsi. Quant à la question de ce qu'il s'agit d'expliquer, il est frappant que Piccinini et Craver considèrent comme allant de soi que les chercheurs en psychologie et en neuroscience ont des intérêts de recherche convergents, que ce qu'ils cherchent doit naturellement coïncider. On admet

[32] Southgate & Hamilton, 2008.

qu'on peut faire ou bien une décomposition psychologique, ou bien une décomposition neurocognitive de la mémoire, mais c'est une hypothèse de travail assez forte que d'admettre d'emblée que les fins de la recherche seront les mêmes dans les deux cas.

La seconde raison concerne l'idée qu'on se fait d'une bonne explication, ce qu'on appelle parfois la question de la *profondeur explicative*. Piccinini et Craver le disent explicitement : « les explications qui intègrent [l]es détails sur les mécanismes [neuraux]sont plus profondes que celles qui ne les intègrent pas ». Ce serait une justification indirecte du privilège accordé à l'analyse fonctionnelle qui procède en deux temps : 1. les explications qui détaillent davantage de niveaux du mécanisme sont préférables aux explications qui en détaillent moins parce qu'elles obligent à admettre ou *constater simplement* moins de choses ; elles sont donc plus profondes et pas simplement plus détaillées. 2. parmi toutes les explications qui ne décrivent pas des niveaux de mécanisme articulés entre eux, les meilleures explications seront celles qui sont au moins *compatibles* avec ces spécifications supplémentaires ; autant dire que les meilleures explications psychologiques seront celles qui sont le mieux susceptibles d'intégrer les connaissances issues des neurosciences. Il semble qu'une autre considération, l'utilité, détermine également la nature de ce que sera une explication plus profonde, ou préférable : une explication préférable est une explication qui permet de meilleures prédictions quant aux réponses du système étudié à des conditions de fonctionnement diverses, or l'idée est qu'une explication qui intègre davantage de « détails structuraux » est en meilleure position pour faire de telles prédictions. On peut remarquer que des chercheurs dans les sciences de l'homme et de la société pourront mettre en doute que la qualité de l'explication et la précision de la prédiction soient fonction de la connaissance de tels détails structuraux (voir les recherches en sociologie de la dépression[33]). En second lieu, on peut également remarquer que Craver et Piccinini comparent deux explications idéalement abouties et correctes, donc de qualité égale, l'une par analyse psychologique-fonctionnelle, l'autre par description d'un mécanisme. Il est bien évident pourtant que la supériorité de l'explications mécaniste est fonction de sa pertinence, et que si une telle explication n'atteint qu'au plausible, ou au possible, alors sa rivale, l'explication psychologique par analyse fonctionnelle (Mill parlait de lois, mais c'était à peu près son point) restera préférable à l'intérieur du cadre défini. Le cimetière de l'histoire est sans doute rempli d'explications mécanistes qui prétendaient à une profondeur avec laquelle la psychologie ne pouvait rivaliser, mais qui se sont avérées être dénuées de valeur. Ajouter des détails neuroscientifiques hasardeux à une analyse psychologique des processus émotionnels ou mémoriels ne la rend pas meilleure, elle l'affaiblit plutôt.

La troisième raison tient sans doute à la présentation des neurosciences définies comme étant un « programme de recherche multi-champs » dans lequel on retrouve, au milieu d'autres champs, la psychologie expérimentale,

[33] Brown & Harris, 1978.
[33] Craver, 2007, p. 228-29.

l'éthologie et la psychiatrie »[34]. Sans doute on peut remonter à Francis Otto Schmitt et aux premières occurrences significatives du terme « neuroscience » dans les années 1960 (le fameux *Neuroscience research program*) pour trouver les sources de cette caractérisation[35]. Il est parfaitement légitime de présenter les neurosciences ainsi, mais il faut certainement souligner aussi que la participation de certains chercheurs d'un champ donné à un programme de recherche pluridisciplinaire n'implique nullement que l'agenda de la recherche dans le champ en question soit entièrement déterminé par le dit programme. C'est ce point que je développerai dans la quatrième partie. À partir du moment où on assimile les sciences de l'esprit à des champs mobilisés par le programme des neurosciences, il n'est plus guère étonnant qu'elles soient perçues comme en partageant largement les visées.

Enfin, quatrième raison, le modèle de l'intégration paraît une voie moyenne (et par là raisonnable) entre d'une part les variantes de l'option réductionniste et d'autre part des formes de dualisme méthodologique, dont certaines (mais pas toutes[36]) aboutissent à la revendication d'une forme d'incommensurabilité.

IV – LE MODÈLE INSTRUMENTAL

Dans cette dernière section, je présenterai une alternative aux modèles de l'élimination, de la réduction et de l'intégration que nous venons de passer en revue, que j'appellerai le modèle instrumental. Dans le modèle instrumental, les sciences de l'esprit puisent dans les ressources des neurosciences pour atteindre des buts qui leur sont propres. Les neurosciences ne détiennent pas une part de vérité sur les phénomènes mentaux en tant qu'elles en étudient la réalisation matérielle, mais en tant qu'elles contribuent, à travers la connaissance de cette réalisation, à en parfaire l'analyse psychologique. Avec ce modèle, les neurosciences ne sont pas à l'horizon d'une réduction ou d'une intégration qui abolirait toute forme d'autonomie des sciences de l'esprit, et ceci parce qu'obtenir des explications neurales n'est pas ce qui motive les sciences de l'esprit à avoir recours aux neurosciences. L'idée n'est pas de présenter un modèle qui rendrait compte de tous les usages des neurosciences, mais de rendre compte de certains de ces usages en accord avec ce que la recherche se propose de faire dans de nombreux cas.

Un obstacle qui empêche de prendre le modèle instrumental en considération est l'idée selon laquelle, comme nous l'avons déjà vu, les neurosciences sont un programme de recherche multi-champs auquel certaines sciences de l'esprit participent. Le projet explicatif de « la » science en ce cas, se confond naturellement avec le projet explicatif des neurosciences. Mais rien n'empêche de considérer que les sciences de l'esprit sont elles-mêmes, pour paraphraser Darden et Maull, *un programme de recherche multi-champs* auquel diverses branches des neurosciences peuvent contribuer, et contribuent de fait. On peut alors soutenir que dans cette perspective, la contribution des neurosciences aux sciences de l'esprit n'équivaut pas au programme d'une

[34] Craver, 2007, p. 228-29.

[35] Sur Schmitt, cf. Rose & Abi-Rached, 2013, chapitre 1.

[36] On relira avec profit von Wright, 1971 sur la compatibilité entre explication causale et explication intentionnelle-historique.

intégration progressive des sciences de l'esprit aux neurosciences. Comme point de comparaison, prenons la contribution de l'intelligence artificielle à la psychologie. Dans un article qui a fait date[37], Jeffrey Elman a montré que le fait de disposer d'une mémoire de travail limitée permettait à un réseau connexionniste de traiter en priorité des structures syntaxiques simples et que l'augmentation progressive de cette même mémoire de travail permettait ensuite à ce même système de faire fond sur cette aptitude précoce pour traiter des structures plus complexes. On voit que le but que cherchait à atteindre Elman était celui d'obtenir une compréhension de l'apprentissage linguistique en termes de contraintes « chronotopiques », les « débuts modestes » et l'augmentation progressive des ressources mémorielles du système étant ce qui canalise l'évolution ultérieure des capacités de ce même système. On voit aussi que le *Design* du modèle est entièrement guidé par une préoccupation de psychologue (comment l'enfant qui commence à se familiariser avec une langue humaine fait-il pour ne pas s'égarer dans le labyrinthe de la complexité syntaxique). Avec Elman, il ne s'agit pas de *remplacer* l'étude psychologique de l'apprentissage par l'étude des réseaux connexionnistes, ni de *réduire* les phénomènes développementaux chez les humains à des changements dans de tels réseaux, ni de traiter l'étude de l'apprentissage linguistique comme « l'esquisse » de ce qui est « complété » par la description du mécanisme qui régit l'évolution de tels réseaux, et ce pour parvenir à une parfaite *intégration* des deux. Il s'agit seulement de *rendre plausible* un apprentissage au sens fort, c'est-à-dire sans bagage cognitif inné, en utilisant ce qu'Elman appelle des « arguments indirects », pour montrer comment des limitations initiales de ressources cognitives comme les ressources mémorielles, loin d'empêcher l'acquisition, peuvent au contraire canaliser le développement et rendre possibles des acquis ultérieurs. L'intelligence artificielle, en cas, est au service des sciences de l'esprit ; elle pèse dans un débat sans abolir l'autonomie du questionnement dans un champ donné.

Qu'en est-il des neurosciences elles-mêmes ? On peut se demander si l'usage d'outils de recherche comme l'imagerie fonctionnelle ou plus récemment l'optogénétique ont pour vocation de permettre l'intégration des sciences de l'esprit aux neurosciences. Quand un psychologue cherche à déterminer quels sont les *corrélats neuraux de*, que cherche-t-il exactement ? On prendra l'exemple d'une recherche sur le substrat neural de l'imagination du futur qui a eu recours à l'imagerie fonctionnelle[38]. Trois tâches sont proposées aux sujets de l'expérience : 1. Penser à des événements dans son futur personnel. 2. Penser à des événements dans son passé personnel. 3. Penser à des épisodes de la vie d'un tiers (une célébrité). L'imagerie fonctionnelle permet d'établir que a) certaines régions (groupe A) présentent la *même* activation durant des tâches relatives au passé et au futur personnel (cingulum postérieur bilatéral, région parahippocampique, cortex occipital gauche). b) d'autres régions (groupe B) présentent des activations plus importantes dans les tâches qui concernent la vision du futur que dans celles impliquant le souvenir personnel (cortex prémoteur gauche latéral, précunéus

[37] Elman, 1993.

[38] Szpunar, Watson & McDermott, 2007.

gauche, etc.). c) Les régions appartenant à ces deux groupes sont plus activées que lors de la construction d'événements concernant un personnage public.

Chercher quelles régions corticales sont impliquées dans la pensée épisodique tournée vers le futur, pensée que l'on contraste avec des pensées d'un autre type, n'est pas chercher un explication mécaniste pour un phénomène mental. Le psychologue n'est pas motivé à effectuer sa recherche par le souci de produire une telle explication. Son problème demeure un problème de psychologue : chercher une caractérisation plus précise des processus mentaux impliqués dans la pensée épisodique du futur ; effectuer une décomposition mentale. Certaines régions du groupe A concernent le traitement contextuel (*contextual processing*) (comme la région parahippocampale dans le lobe temporal médian). Ce qui suggère qu'en nous pensant dans le futur nous nous imaginons dans des contextes déjà rencontrés. Un ingrédient de la pensée épisodique ordinaire serait d'être une pensée située dans des lieux familiers. De leur côté, certaines régions du groupe B (cortex prémoteur latéral, cortex postérieur pariétal médian, cervelet postérieur) qui sont donc plus activées dans la pensée du futur, sont fréquemment impliquées dans des tâches faisant appel à l'imagerie motrice. L'asymétrie du passé et du futur serait liée au fait de *se voir faire des choses* dans le futur tandis que nous ne pouvons modifier le passé. Ce serait notre puissance d'agir qui distingue le futur du passé.

La question importante est alors : est-ce que Szpunar et ses collègues, en faisant appel à l'imagerie fonctionnelle, en identifiant des régions, veulent « compléter » avec des détails structurels une « esquisse de mécanisme » ? Il s'agit plutôt d'apprendre de l'imagerie fonctionnelle en quoi consiste psychologiquement une capacité comme l'imagination du futur et quelle relation elle a avec d'autres capacités. La caractérisation fonctionnelle ou psychologique d'une région impliquée dans une tâche n'est pas une sorte d'à côté inessentiel de ce qui compterait vraiment, à savoir l'identification du réseau neural proprement dit. Ce n'est pas le cortex prémoteur latéral qui intéresse le psychologue, mais ce qu'il fait, et ce que nous apprend son activation sur le contexte de celle-ci. Le modèle instrumental est bien un modèle au sens où il ne se contente pas de constater que le psychologue peut faire appel aux données des neurosciences : il pose que ces données *doivent* non pas seulement permettre d'identifier le mécanisme neural correspondant de manière correcte, mais aussi rendre capable de compléter par ce moyen de manière non-triviale l'analyse psychologique.

Les sciences du cerveau sont prises en compte par les sciences de l'esprit (psychologie, ou psychiatrie[39]) en tant qu'elles sont un moyen d'identification des processus mentaux et des relations qu'ils entretiennent. C'est la raison pour laquelle il y a un débat sur la qualité de la contribution des sciences du cerveau à une telle identification. On peut très bien en effet avoir un *consensus* sur « la destruction de la région X compromet l'exécution des tâches de type Y » et un *débat* sur la description psychologique adéquate de X. D'une part, aller des

[39] Voir les conclusions qu'on peut tirer au sujet du délire de la découverte d'une activation aberrante du réseau du défaut (*Default network*) dans la schizophrénie : in Gerrans, 2013.

activations aux rôles fonctionnels (inférence inverse[40]) ne va pas de soi et il faut sans doute croiser le maximum d'informations pour espérer une identification correcte de la contribution de X [41]. Ensuite, si on admet des rôles multiples pour les mêmes composants[42], cela peut compliquer considérablement l'analyse.

La position dans la littérature dont la proposition du modèle instrumental est la plus proche est à ma connaissance celle de Roth et Cummins[43]. Ces auteurs ont en effet récemment proposé une distinction entre expliquer un « effet », c'est-à-dire un phénomène psychologique robuste, au moyen d'une analyse fonctionnelle de nature psychologique (explication horizontale) et expliquer comment une analyse fonctionnelle est implémentée (explication verticale). Penser que l'explication horizontale est une explication verticale incomplète serait méconnaître qu'elles sont en fait complémentaires et qu'elles poursuivent des buts différents, Il est possible que, dans bien des cas, cette complémentarité ne soit pas inerte, et que spécifier un mécanisme neural (explication verticale) amène à choisir une analyse fonctionnelle plutôt qu'une autre. Mais cela ne revient pas à proprement parler à *compléter* telle analyse fonctionnelle, mais seulement à la *confirmer* (ou à en infirmer une autre). La confirmation, et Roth et Cummins reprennent sur ce point un passage célèbre de Fodor, est « isotrope » et tout est bon pour confirmer une analyse fonctionnelle[44], en incluant des données issues des neurosciences. En ce sens, il n'y aura pas d'autonomie de la psychologie (et il n'y a pas à en avoir). Mais l'absence d'autonomie de la confirmation n'est pas l'absence d'explications et, on peut l'ajouter, d'intérêts de recherche *sui generis* dans les sciences de l'esprit. Que la recherche dans les sciences de l'esprit intègre des données venues de champs connexes n'implique pas qu'elles n'aient pas des raisons bien à elles de faire appel à ces données[45].

CONCLUSION

Chercher « le » bon modèle des relations entre sciences de l'esprit et neurosciences serait faire comme s'il n'y avait qu'un seul type de progrès en la matière, un seul type d'issue heureuse aux débats, un seul ensemble de

---

[40] Poldrack, 2006.

[41] C'est l'idée de « congruence qualifiée » in Forest, 2014.

[42] Anderson, 2014.

[43] Roth & Cummins, 2017.

[44] Fodor, 1983, p. 137 : « lorsque je dis que la confirmation est isotrope, je veux dire que toute vérité empirique (ou bien sûr déductive) préalablement établie peut être pertinente, pour la confirmation d'une hypothèse scientifique ».

[45] Avec le modèle instrumental, l'idée est aussi d'offrir une alternative à la représentation commune d'un impérialisme des « disciplines en neuro », qui viendraient prétendre se substituer aux sciences humaines et sociales (Vidal & Ortega, 2017). Quelles que soient les relations de la neuro-esthétique à l'esthétique, ou de la neuro-économie à l'économie, c'est un fait robuste qu'une demande de neurosciences vient, par exemple, des psychologues de la mémoire eux-mêmes dans l'étude des phénomènes mémoriels, dans un esprit qui est plus celui de la collaboration scientifique que de la concurrence. C'est cette appropriation des neurosciences par les sciences de l'esprit qui doit être comprise et il ne me semble pas qu'on puisse parler de dénaturation des objets de recherche ni de voie sans issue *a priori*, à la place des chercheurs eux-mêmes et de l'examen de l'ensemble de leurs résultats.

préoccupations légitimes des chercheurs. Pourtant, si on apprend que telle région est activée dans tel contexte, ou que sa détérioration a tel effet, cela peut être « significatif »[46] à plus d'un titre, et cela peut intéresser le chercheur en neurosciences qui cherche à établir une « biographie fonctionnelle » de la région en question, le psychologue qui cherche à décomposer tel type d'activité mentale et à comprendre en quoi elle consiste, comme le neuropsychologue qui cherche à éclairer un tableau clinique : chacun d'eux va puiser dans les mêmes ressources, et tous trois vont éventuellement collaborer temporairement, mais chacun conservera en propre un agenda différent, sinon des convictions à l'abri de la révision.

La dynamique de la recherche a certainement offert dans l'histoire des exemples d'élimination (les facultés de Gall) comme des cas d'intégration (les niveaux de la mémoire), et en cela, l'idée d'un éventail d'issues possibles à la coévolution des connaissances, proposée par Patricia Churchland, reste intéressante. Mais cette dynamique passe aussi par la contribution des neurosciences à l'analyse des phénomènes de l'esprit et c'est pour cette raison que le modèle instrumental proposé ci-dessus entend compléter d'autres modèles, rendre plus intelligible une partie de la recherche, et situer les problèmes qui se posent à l'intérieur de celle-ci. On a beaucoup écrit sur l'explication neuroscientifique du mental, il reste à penser la place des neurosciences *dans* les sciences de l'esprit.

RÉFÉRENCES

Anderson, M. L. (2014). *After phrenology. Neural Reuse and the Interactive Brain*. Cambridge, MA, MIT Press.

Bechtel, W. & Richardson, R. [2000]. *Discovering Complexity*. Cambridge, MA, MIT Press, 2010.

Berwick, R.C. & Chomsky, N. (2016). *Why Only Us. Language and Evolution*, Cambridge, MA, MIT Press.

Brown, G.W. & Harris, T. (1978). *Social Origins of Depression. A Study of Psychiatric Disorders of Women*. New York, The free Press.

Churchland, P. (1986). *Neurophilosophy. Towards a Unified Science of the Mind/ Brain*. Cambridge, MA, MIT Press.

Corkin, S. (2013). *Permanent Present Tense. The Man with no Memory, and What he Taught the World*. Londres, Allen Lane.

Craver, C. (2007). *Explaining the brain*, Oxford University Press.

Craver, C. & Bechtel, W. (2006). "Mechanisms", in J. Pfeifer & S. Sarkar (éds.), The Philosophy of Science: An Encyclopedia. Psychology Press, pp. 469-478.

Cummins, R. (1975) Functional Analysis, *Journal of Philosophy*, 72, pp. 741-64.

Cummins, R. (1983). *The Nature of Psychological Explanation*, Cambridge, MA, MIT Press.

Darden, L. & Maull, N. (1977). Interfield theories, *Philosophy of Science*, *44*, 43-64.

Elman, J. (1993).Learning and development in neural networks: the importance of starting small. *Cognition*, *48*, 71-99.

Engel, P. (1996) *Philosophie et psychologie*, Paris, Gallimard, Folio Essais.

Fodor, J. (1983). *La modularité de l'esprit*, traduction Abel Gerschenfeld, Paris, Éditions de Minuit.

Forest, D. (2014). *Neuroscepticisme*. Paris, Ithaque.

[46] Kitcher, 2001, chapitre VI.

Gerrans, P. (2013). Delusional attitudes and default thinking. *Mind & Language*, *28*, *1*, 83-102.
Hatfield, G. (1993). Helmholtz and Classicism: the science of aesthetics and the aesthetics of science, in D. Cahan (éd.), *Hermann von Helmholtz, and the Foundations of 19th Century Science*, University of California Press, pp. 522-558.
Kitcher, P. (2001). *Science vérité et démocratie*, traduction Stéphanie Ruphy, Paris, Presses Universitaires de France.
Machamer, P., Darden, L & Craver, C. (2000), Thinking about mechanisms, *Philosophy of science*, *57*, 1-25. Traduction de Marion Le Bidan et Denis Forest à paraître in *Philosophie de la biologie*, Textes clés, direction Gayon (Jean) et Pradeu (Thomas), Paris, Vrin.
Marr, D. (1982). *Vision*, New York, Freeman.
McCauley, R. (1996). Explanatory pluralism and the Co-evolution of Theories in Science in R. McCauley (éd), *The Churchlands and their Critics*. Oxford, Blackwell Books, pp. 17-47.
Maull, N. (1977). Unifying science without reduction*, Studies in the History and Philosophy of Science*, *8*, 143-162.
Mill, J.S. [1843]. *The Logic of the Moral Sciences* (*Logic*, Livre VI), réédition Chicago, Open Court, 1999.
Mill, J.S. [1865]. *Auguste Comte and Positivism*, in *Collected Works of John Stuart Mill*, Volume X. Toronto, University of Toronto Press, London: Routledge & Kegan Paul, 1985, pp. 261-368.
Murphy, D. (2006). *Psychiatry in the Scientific Image*, Cambridge, MA, MIT Press.
Nagel, E. (1979). *The Structure of Science*, Indianapolis & Cambridge, Hackett Publishing.
O'Keefe, J. & Nadel, L. (1978). *The Hippocampus as a Cognitive Map*, Oxford University Press.
Poldrack, R. (2006). Can Cognitive Processes Be Inferred from Functional Imaging Data? *Trends in Cognitive Science*, *10(2)*, 59-63.
Piccinini, G. & Craver, C. (2011). Integrating Psychology and Neuroscience: Functional Analysis as Mechanism Sketches, *Synthese 183(3)*, 283-311.
Rose, N. & Abi-Rached, J.M. (2013). *Neuro. The New Brain Sciences and the Management of the Mind.* Princeton University Press.
Roth, M. & Cummins, R. (2017). Neuroscience, Psychology, Reduction, and Functional Analysis. In D.M. Kaplan (éd.) *Explanation and Integration in Mind and Brain Science*. Oxford, Oxford University Press, pp. 29-43.
Scoville, W.B. & Milner, B. (1957). Loss of recent Memory after Bilateral Hippocampal Lesions, *Journal of Neurology, Neurosurgery and Psychiatry*, *20*, 11-20.
Southgate, V. & De Hamilton, A.F. (2008) Unbroken mirrors; challenging a theory of autism, *Trends in Cognitive Science*, *12(6), 225-229*.
Sullivan, J. (2010). Reconsidering 'spatial memory' and the Morris Water Maze, *Synthese*, *177*, 261-83.
Sullivan, J. (2016). Construct stabilization and the Unity of Mind-Brain Sciences, *Philosophy of Science*, *83(5)*, 662-673.
Szpunar, K., Watson, J.M. & McDermott, K.B. (2007). Neural substrates of envisioning the future, *Proceedings of the National Academy of Sciences*, pp. 642-647.
Vidal, F. & Ortega, F. (2017). *Being Brains. Making the Cerebral Subject*. New York, Fordham University Press.
Wright, G.H. von (1971). *Expliquer et comprendre*. Traduction Olivier Fontaine, Ithaque.