Université de Montréal

Le néo-réductionnisme et le matérialisme éliminativiste de Paul M. Churchland

par
Mathieu Côté Charbonneau

Département de philosophie
Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de M.A.
en philosophie

Août 2008

Université de Montréal
Faculté des études supérieures

Ce mémoire intitulé :
Le néo-réductionnisme et le matérialisme éliminativiste de Paul M. Churchland

présenté par :
Mathieu Côté Charbonneau

a été évalué par un jury composé des personnes suivantes :

Jean-Pierre Marquis
président-rapporteur

Frédéric Bouchard
directeur de recherche

Daniel Laurier
membre du jury

.....
représentant du doyen de la FES (doctorat seulement)

Résumé

Le philosophe Paul M. Churchland a proposé un modèle de réduction interthéorique prenant en compte les théories scientifiques en cours de développement. À partir de ce modèle, il argumente pour le rejet de la psychologie cognitive de l'entreprise interdisciplinaire des sciences cognitives, position nommée 'matérialisme éliminativiste'. Ce rejet se fonde sur le développement des neurosciences cognitives (neurobiologie, neuropsychologie, etc.) et mise sur leur succès futur dans l'entreprise explicative des phénomènes mentaux et cognitifs.

Le présent mémoire a pour objectif de développer une analyse fine des thèses de Paul M. Churchland en proposant un cadre conceptuel permettant de structurer l'argumentation de l'auteur et de bien cerner la nature empirique (plutôt qu'analytique) de ses thèses. Parallèlement, il sera aussi question de montrer que les thèses de l'auteur s'inscrivent bien dans le cadre conceptuel général partagé par les scientifiques et philosophes sur la nature et le fonctionnement des explications en sciences cognitives.

Mots clefs

philosophie des sciences – philosophie de l'esprit – sciences cognitives - neurosciences – psychologie cognitive – Paul M. Churchland – réductionnisme – matérialisme éliminativiste

Abstract

Philosopher Paul M. Churchland designed a model of intertheoretic reductionism including the evolving character of theories. From this model, he argues for the elimination of cognitive psychology from the interdisciplinary project of cognitive science, a philosophical position named 'eliminative materialism'. This rejection is based on the recent development of cognitive neurosciences (neurobiology, neuropsychology, etc.) and the prediction of future success in the explanatory project of mental states and cognitive processes.

This thesis develops a fine-grained analysis of Paul M. Churchland's views by elaborating a conceptual framework able to structure his argumentation and by singling out the empirical character of the eliminativistic position (instead of a properly analytic one). Also, it has been shown that Churchland's theses are properly integrated in the common assumptions and theoretical perspectives shared by scientists and philosophers concerned by the nature and form of explanation in cognitive science.

Keywords

Philosophy of science – philosophy of mind – cognitive science – neuroscience – cognitive psychology – Paul M. Churchland – reductionism – eliminative materialism

Table des matières

Table des figures

À Charles Douzemille

Remerciements

Je tiens à remercier :

M. Frédéric Bouchard pour la structure offerte et beaucoup plus encore.
M. Jean-Pierre Marquis qui m'a convaincu de tenter la maîtrise.
M. François Duchesneau qui m'a fait voir que l'audace pouvait être une vertu.
Ces messieurs pour leur support tout au long de ma maîtrise.

Mon père et ma mère pour leur support moral et matériel.
Et mon frère Nicolas qui m'a toujours rappelé la percutante solidité du trottoir.

Introduction

Depuis les années 1950, le développement massif des sciences cognitives (Gardner, 1985) et de l'approche naturaliste en philosophie 'analytique' (Kim, 2003) a lentement transformé la discipline de la philosophie de l'esprit en philosophie des sciences de l'esprit. Un bon exemple de ce changement de perspective est la redéfinition du problème de la relation entre le corps et l'esprit. Selon John Bickle (1998, 2003), le problème philosophique de la relation entre le corps et l'esprit (*mind/body problem*) devra dorénavant être formulé dans une perspective épistémologique des sciences cognitives : comment les théories psychologiques (achevées) seront-t-elles misent en relation avec les théories neurobiologiques (achevées)? Le problème métaphysique de la relation entre esprit et corps prendra cette nouvelle forme épistémologique d'une relation interthéorique faisant le pont entre les entités et processus psychologiques postulés par les théories psychologiques (états mentaux, processus mentaux, contenu sémantique des représentations mentales) et l'ontologie des théories neurobiologiques (activations inter-neuronales, structures du système nerveux, réseaux de connexions neuronales, etc.). Dans la philosophie à venir, nous informe Bickle, la question de la relation du mental au physique passera par une théorie de la réduction interthéorique de la psychologie aux neurosciences.

Plusieurs camps se sont formés vis-à-vis de la question de la possibilité de former une telle réduction interthéorique. Une des positions les plus drastiques (et remarquées) insiste sur le fait que le cadre conceptuel utilisé par la psychologie cognitive n'est pas compatible avec celui des neurosciences et qu'une relation de réduction n'est pas possible *parce que* les entités et processus postulés par la psychologie cognitive n'existent tout simplement pas. Par conséquent, les théories psychocognitives ne seraient pas aptes à survivre au développement des sciences cognitives et devraient alors être remplacées par des théories neuropsychologiques plus justes, intégratives et prédictives. Cette position est nommée « matérialisme éliminativiste » et son porte-parole contemporain le plus reconnu est certainement Paul M. Churchland[1].

P.M. Churchland met en opposition les connaissances psychologiques que nous fournissent les neurosciences à celles tirées de la psychologie cognitive. Cette dernière

[1] Afin d'éviter toute ambiguïté (le simple nom 'Churchland' pouvant référer à deux auteurs), au risque d'alourdir le texte, lorsqu'il sera question de Paul M. Churchland, j'utiliserai soit le nom complet ou le diminutif P.M. Churchland, alors que pour Patricia S. Churchland, ce sera P.S. Churchland.

userait d'un réseau de concepts et de relations tirés de notre compréhension ordinaire des phénomènes psychologiques tels que nous les vivons au jour le jour. Ce réseau est nommé 'psychologie du sens commun' (*folk psychology*) et constituerait une des deux pierres fondatrices de toute l'entreprise des sciences cognitives (l'autre étant l'analogie de l'esprit comme processeur de traitement de l'information, soit la thèse computationnelle de l'esprit). Or, nous dit P.M. Churchland, ce réseau de concepts ne trouverait pas d'image neuropsychologique, i.e. qu'il ne trouverait aucun corrélat neurobiologique. De fait, la psychologie cognitive devra ou bien se renouveler en *neuro*psychologie cognitive ou disparaître (être éliminée du corpus scientifique). Inversement, l'élimination de la neurobiologie au profit d'une psychologie empirique ne fait pas sens.

Cette caractérisation de la relation de réduction souffre d'un problème important : le philosophe détermine *a priori* la possibilité ou l'impossibilité de la relation de réduction. Comme l'ont fait remarquer plusieurs auteurs (Oppenheim & Putnam (1958), Nagel (1961), P. M. Churchland (1979, 1981), P. S. Churchland (1980, 1986), Bickle (1996, 1998), Fodor (1974, 1975, 1987), etc.), la relation de réduction ne peut tenir qu'entre deux théories (ou ensemble de théories) suffisamment avancées (voir achevées) pour permettre de systématiser la relation qu'elles entretiendront. Cela signifie donc que la valeur de la thèse réductionniste (ou anti-réductionniste) devra être décidée par voies empiriques lorsque les deux entreprises seront suffisamment avancées pour permettre de décider de la faisabilité de la réduction interthéorique. Ironiquement, ces auteurs (excluant Nagel et Oppenheim) comme d'autres (ex : Davidson (1970, 1973)) ont cru bon de se prononcer sur le sort futur d'une telle relation. Le camp fonctionnaliste anti-réductionniste a rejeté la faisabilité d'une telle relation ou, du moins, la valeur de son apport explicatif et a plutôt poussé non pas vers un dualisme des substances (ce qui serait une thèse métaphysique) mais vers un dualisme explicatif (thèse épistémologique[2]) assurant une autonomie des sciences spéciales (psychologie et autres sciences humaines) des sciences naturelles dures (physique, chimie, biologie, etc.). Non seulement le vocabulaire utilisé dans les théories psychologiques lui serait propre mais, en plus, les sciences dures n'auraient rien à apprendre aux sciences spéciales. (Fodor, 1974) Inversement, des auteurs comme Feyerabend (1963), Rorty (1965) et P.M. Churchland (1979, 1981) ont critiqué vivement l'attitude chauvine de rejeter l'effort neuroscientifique dans l'entreprise des sciences cognitives avant même d'explorer

[2] Parfois aussi nommée « dualisme des propriétés » (*propriety dualism*).

quel apport elles pourraient fournir dans la production d'une théorie scientifique de l'esprit (*mind*). De plus, ils ont eux-mêmes dénigré le cadre conceptuel de la psychologie (cognitive et du sens commun) en l'accusant de ne pas être apte à rencontrer les exigences d'une bonne théorie matérialiste de l'esprit. Ironiquement (encore une fois), alors qu'il semble faire état de lieu commun dans la communauté philosophique concernant la nature empirique de la thèse réductionniste, chacun semble s'être adonné au prophétisme. Or jouer au prophète ne serait pas permissible dans une entreprise scientifique rigoureuse puisse que cela irait contre le projet explicatif des sciences (Popper (1959)) (pensons notamment aux prophéties de Lord Kelvin sur le devenir des sciences).

Si la relation de réduction est une thèse empirique concernant le statut des théories une fois leur structure suffisamment détaillée, le travail du philosophe concernant la validation ou le rejet de la relation de réduction devra patienter jusqu'à l'élaboration quasi-complète des théories cognitives, psycho- ou neuro-. Cela ne veut toutefois pas dire qu'il n'y a pas de travail philosophique (épistémologique) à faire d'ici-là : il semble déjà possible de projeter non pas les résultats finaux des investigations empiriques concernant la viabilité d'une relation de réduction mais plutôt d'investiguer par quels chemins un tel développement pourrait avoir lieu, ou pas, et quel appareillage conceptuel sera nécessaire pour y arriver. C'est en cela que consistera le présent mémoire. Partant d'une perspective épistémologique (il ne sera pas question de la vérité des théories ni des conséquences métaphysiques de leur acceptation mais bien de leur structure conceptuelle), il sera question d'éclairer par quelles voies conceptuelles devra passer celui qui cherchera à vérifier (ou infirmer) la thèse réductionniste.

Le premier pas consisterait à spécifier la nature de la relation de réduction interthéorique, ce qui sera le propos du premier chapitre. Deux familles de relation de réduction ont émergées au sein de la communauté scientifique depuis la naissance des sciences cognitives. La relation de réduction nagelienne, directement inspirée de la perspective des positivistes-logiques (Nagel, 1961; Oppenheim & Putnam, 1958; Carnap, 2003) puis celle du camp néo-réductionniste (ayant comme champions P.M. Churchland, J. Bickle, C.A. Hooker et, dans une moindre mesure, P.S. Churchland) inspirée plus directement par l'histoire des sciences et des réductions interthéoriques ayant déjà été menées auparavant (Bickle, 1996, 198, 2003; Hooker, 1981; P.M. Churchland, 1979, 1981; P.S. Churchland, 1986). La seconde famille de réductionnistes élargie la conception

classique (nagelienne) en permettant une gamme de relation de réduction intermédiaire distribuée sur un continuum continu entre les deux pôles radicaux de la réussite ou l'échec de celle-ci, soit la possibilité de révision interthéorique plus ou moins radicale et celle de l'élimination d'une théorie en faveur de l'autre.

C'est de cette seconde perspective dont il sera question dans ce mémoire et particulièrement des conséquences épistémologiques de la possibilité de l'éliminativisme. Depuis déjà plus de trente ans, il fait objet de consensus que la position nagelienne est trop rigide pour être valable et ce sera donc de la possibilité de mener une réduction telle qu'entendue par Paul M. Churchland qui rassemblera le propos de cette analyse. Il sera question ici d'élaborer à partir de la conception néo-réductionniste de P.M. Churchland un cadre conceptuel permettant de répondre aux exigences de son schème réductionniste et ainsi d'éclairer la situation actuelle vis-à-vis la relation de réduction entre les théories psycho- et neurocognitives.

En d'autres mots, l'objectif premier de ce travail consiste à élaborer un schème compréhensif de la stratégie de réduction proposée par Paul M. Churchland sans tomber dans les jeux spéculatifs concernant le futur de cette relation. Ce travail pourrait donc être considéré comme une tentative d'éclairer quelle situation épistémologique il faudra obtenir pour pouvoir juger de la faisabilité de l'un des résultats du schème proposé par P.M. Churchland. Étant donné la perspective épistémologique utilisée ici, il ne sera pas question des conséquences ontologiques du problème de la réduction en science mais bien du reflet épistémologique de celles-ci. Il ne sera pas question des entités réelles ou des propriétés de ceux-ci mais des objets et des prédicats postulés par les théories scientifiques. Il ne sera donc pas question de la nature du mental dans sa réalisation matérielle (matérialisme ontologique) mais plutôt de la réduction interthéorique de la psychologie aux neurosciences.

Pour y arriver, il m'aura fallu rendre plus explicite certains aspects de l'argumentation de P.M. Churchland qui sous-entend l'utilisation de certains outils épistémologiques développés par Paul M. Churchland lui-même (cadre néo-réductionniste, interprétation sémantique des réseaux connexionnistes, etc.) ainsi que par d'autres philosophes concernés par les sciences cognitives (niveaux explicatifs, explication fonctionnelle, relation de réalisation explicative, stratégies de recherche, énoncés ramséens, déshomoncularisation, etc.). L'analyse de ces outils et de leur intégration dans le schème néo-réductionniste de P.M. Churchland feront l'objet des chapitres 2 et 3.

En explorant, par le biais de ces outils, le programme réductionniste de Paul M. Churchland, nous verrons que son éliminativisme vis-à-vis la *psychologie du sens commun* n'est pas encore justifié : il peut donc être vrai ou faux, nous n'en savons rien. D'abord parce qu'il ne répond pas à une clause d'empiricité de la relation de réduction (défendue elle-même pas P.M. Churchland, voir chapitre 1), mais aussi parce que son programme ne fait que rejeter que le mode de computation propositionnelle (et la méthodologie de recherche employée par la psychologie cognitive à cet égard) et non pas le cadre conceptuel de la psychologie du sens commun *per se* (ce qui sera le propos du chapitre 4). Cette dernière conséquence détonnera à partir de l'élaboration du cadre conceptuel développé pour mieux comprendre la perspective de P.M. Churchland. Il est important de comprendre dès à présent que l'objectif premier de ce travail ne consiste pas à faire l'apologie ou la condamnation du matérialisme éliminativiste mais d'analyser plus à fond la conception néo-réductionniste de Paul M. Churchland et de la reporter dans son cadre empirique. Simplement, une telle analyse permet de distinguer certaines nuances qui sont floues dans le corpus philosophique de P.M. Churchland.

Cette approche sera limitée dans ses efforts par deux contraintes. D'abord, notre analyse étant limitée en espace, il ne sera pas question des méthodes de recherche utilisées par les neuroscientifiques (ex : localisation fonctionnelle par fMRI ou par les lésions cervicales, etc.) ou par les psychologues cognitivistes (ex : tâche de résolution de problème, méthodologie des recherches en intelligence artificielle, etc.). L'aspect méthodologique sera limité aux détails épistémologiques de la construction des théories explicatives des phénomènes mentaux (ex : explication fonctionnelle, vocabulaire théorique, etc.). Il ne sera donc pas question des méthodes appliquées mais de la méthodologie de la construction théorique des explications scientifiques.

Pour éviter des complications dépassant la portée du présent travail (mais des plus pertinentes pour une analyse plus ample), je considérerai que les neurosciences et la psychologie cognitive sont deux ensembles théoriques inscrits dans le même projet scientifique (sciences cognitives) et qu'ils ont pour même objectif l'explication des manifestations comportementales des créatures cognitives par une théorie des phénomènes mentaux. Cela permettra d'éviter une série de problèmes liés au débat réductionniste. Par exemple, Wimsatt (1976) propose une distinction entre la réduction entre les sciences (ex : la chimie se réduit à la physique) nommée '*inter-level reduction*' et la réduction d'une

théorie particulière à une autre au sein d'une même science (ex : réduction de la génétique des populations à la génétique moléculaire au sein de la biologie) nommée '*intra-level reduction*'. Encore, McCauley (1986, 1996[3], 2007) ajoute à ces distinctions une série de critères pour caractériser les théories et les sciences qui complexifient grandement le débat réductionniste[4]. Ces nuances, quoiqu'intéressantes et nourrissantes, ne participeront pas à la présente analyse (voir Sarkar (2005) pour un survol contemporain de ces questions).

La seconde contrainte est historique : nous n'en sommes pas encore au jour où les théories sont suffisamment avancées pour détailler la nature de la relation de réduction. Cela se fera particulièrement sentir lorsqu'il sera question de la relation de réalisation explicative qu'emploie P.M. Churchland pour assurer une réduction d'un niveau explicatif à un autre. Notre analyse ne pourra être que fragmentaire : le détail de la relation de réalisation explicative sera limitée aux connaissances neuroscientifiques actuelles explicitement utilisées par P.M. Churchland. Sur ces considérations négatives, passons à l'analyse positive du schème réductionniste de P.M. Churchland et à son intégration dans le projet explicatif des sciences cognitives.

---

[3] Dans McCauley (1996), la critique adressée à Paul et Patricia Churchland se trouve aux pages17-47 et est intitulée : « Explanatory Pluralism and the Co-Evolution of Theories in Science ».

[4] À la distinction du niveau d'analyse où se produit la réduction, il faudrait aussi prendre en compte le temps à laquelle la théorie a été émise déterminant la relation de succession des théories, les objectifs explicatifs, le contexte historique et socioculturel et les considérations normatives dirigeant les différentes entreprises scientifiques. (McCauley, 2007, pp.128-136)

Chapitre I – Le matérialisme éliminativiste

« Consensus holds that reductionism is dead. »
Bickle, 1998, p.1

La critique éliminativiste de Paul M. Churchland prend place au sein du problème réductionniste. L'histoire de se dernier s'étant développée bien avant l'incursion des idées de P.M. Churchland à son sujet, il est nécessaire de situer ce problème face au contexte intellectuel y ayant mené et d'explorer les premières analyses de la nature de ce problème, analyses auxquelles la position de P.M. Churchland réagit. Ce chapitre se concentrera donc d'abord à éclaircir la version classique de la relation de réduction interthéorique (section 1.1) pour ainsi permettre d'en exposer à la fois les premières réactions antiréductionnistes (section 1.2), puis d'analyser la position originale de P.M. Churchland (section 1.3). Il sera finalement possible d'énoncer déjà plus clairement en quoi le débat sur l'avenir de la relation de réduction interthéorique tend plutôt du côté prophétique que du côté analytique en montrant comment les tenants des deux partis (réductionnistes vs. antiréductionnistes) contredisent, par leurs prédictions, certaines des prémisses de leur conception même de la relation de réduction (section 1.4).

Le problème du réductionnisme émerge dans un contexte scientifique où l'unité des sciences est une valeur prédominante (Kitcher, 1981, 1989), aussi faut-il savoir ce que l'on entend par 'unité' puisqu'il est possible d'assigner plus d'un sens à ce terme[5]. Une première façon de la comprendre consiste à la considérer comme une thèse postulant que toutes les entités et phénomènes observables étudiés par les sciences peuvent ultimement être décrits par le langage des sciences physiques. Cette thèse est nommée 'physicalisme', ou '*token-physicalism*' (Fodor, 1974, p.100) et son acceptation semble faire objet de consensus en philosophie des sciences, du moins dans le cadre du débat entre les réductionnistes et les antiréductionnistes[6]. Une seconde formulation du principe d'unité, celui endossé par les réductionnistes, consiste à dire que touts les prédicats théoriques présents dans une théorie scientifique peuvent être identifiés à des prédicats théoriques

[5] Les types d'unité scientifique présentés ici sont tirés de Fodor (1974). Oppenheim & Putnam (1958) en dégagent au moins trois autres.

[6] Cette thèse a deux conditions : tous les phénomènes observables sont descriptibles et tous les phénomènes qui peuvent être décrits peuvent l'être par un même langage théorique, soit celui de la physique. Cette thèse est parallèle à un monisme matérialiste mais n'en découle pas nécessairement.

individuels de la physique ou à une composition de ceux-ci. Si les prédicats des deux théories réfèrent à des espèces naturelles[7] (*natural kind*), alors l'unité des sciences par réduction est obtenue. (Fodor, 1974, pp.100-104). Cette conception est aussi physicaliste, mais suppose que la physique a le privilège supplémentaire d'être la science fondamentale à laquelle toutes les autres se réduiront. C'est en ce sens que l'unité des sciences sera entendue ici et, pour en rendre compte plus à fond, une étude de la relation de réduction s'impose puisque ses défenseurs suggèrent qu'elle serait la ficelle qui unifierait les théories scientifiques entre elles.

Intuitivement, la relation de réduction possède plusieurs propriétés logiques. Premièrement, celle-ci devrait être asymétrique : une science se réduisant à une autre ne peut réduire cette dernière. Il n'est pas question de réduction circulaire et on peut donc déjà comprendre que cette relation impose un ordre dans l'organisation des sciences. La relation de réduction est aussi transitive : une science A réduit une science B, la science B réduit une science C, alors la science C est réduite à la science A. Ces deux propriétés de la réduction invitent à penser l'unité des sciences comme une hiérarchisation des sciences. Selon leur place dans l'organisation obtenue par une mise en relation de réduction complète de celles-ci, certaines sciences seront plus fondamentales (au sens où plusieurs sciences se réduisent à celles-ci) et d'autres sont plus spécifiques (elles ne réduisent pas ou peu d'autres sciences). Au sommet, on retrouverait les sciences qui n'en réduisent pas mais qui sont elles-mêmes réduites par d'autres. Les sciences réduites sont de plus haut niveau que leurs sœurs réductrices, assurant ainsi un amoncellement de paliers scientifiques. Ultimement, aux fondements de l'édifice, la physique *réduirait* toutes les autres sciences et ne serait réduite à aucune autre.

L'utilisation du conditionnel n'est pas fortuite : cette image de l'unité des sciences est une projection idéale. Cette conception des sciences n'est pas partagée par tous : les antiréductionnistes (Fodor, 1975) considèrent que cette image des sciences est erronée. La hiérarchisation des sciences serait possible pour les sciences naturelles 'dures' (ex : physique, chimie, etc.) mais ne pourrait pas intégrer les sciences spéciales (ex : psychologie, économie, etc.). Ces dernières seraient autonomes en ce sens où leur domaine explicatif ne gagnerait aucune information à être mis en relation avec les sciences de 'plus bas niveau'.

---

[7] Une espèce naturelle, au niveau épistémologique, est un prédicat x dont les objets pouvant satisfaire ce prédicats sont suffisamment similaires (naturellement similaires) pour permettre la projection d'un autre prédicat z par une loi 'Tous les x sont z'. (Quine, 1969, chapitre 5)

Mais avant de voir ce qui convainc les antiréductionnistes d'une image fragmentée de l'édifice des sciences et d'approfondir ce en quoi consiste l'idée d'autonomie des sciences, il est nécessaire de détailler plus à fond la relation de réduction telle qu'elle a été proposée originellement par les empiristes logiques. Pour ce faire, il sera d'abord question de l'aspect descriptif de la relation de réduction (son aspect formel). Puis il sera question des implications méthodologiques pour les sciences si elles devaient se réduire, ou pas.

1.1 Le réductionnisme classique – Nagel, Oppenheim et Putnam

Selon Oppenheim & Putnam (1958), la méthode privilégiée pour obtenir une unité entre les différentes branches scientifiques consiste à obtenir une relation particulière entre les différentes théories constituant ces branches, relation qu'ils nomment '*micro-reduction*'. Dans ce cadre conceptuel, deux théories (ou branches scientifiques, c'est-à-dire des ensembles de théories[8]) sont mises en relation de réduction si certaines conditions formelles et informelles sont remplies. Nagel (1961, chap. 11) en présente la formulation désormais classique, formulation autour de laquelle la première vague du débat s'est organisée. C'est de cette conceptualisation dont il sera question dans la présente section.

Une théorie scientifique est constituée d'un ensemble de propositions (lois, énoncés d'observations, définitions des entités et propriétés postulées) formalisées dans un langage logico-mathématique. (P.M. Churchland, 1979, chapitre 3) Une théorie qui est réduite par une autre est nommée 'théorie réduite' alors que la seconde est la 'théorie réductrice' ou 'théorie de base'.

Selon le schème conceptuel de Nagel, une théorie se réduit à une autre si certaines conditions formelles sont obtenues. Une théorie T_r (théorie réduite) est réduite à une théorie T_b (théorie de base) si l'ensemble des énoncés constituant T_r sont des conséquences logiques des énoncés constituant T_b. En d'autres mots, il faut pouvoir déduire le corpus théorique de T_r de celui de T_b.

---

[8] Dorénavant il ne sera question que de théorie, et pas de branches scientifiques. Une branche scientifique sera considérée comme un ensemble de théories ayant un certain air de famille (ex : l'ensemble des théories psychocognitives constituent la branche de la psychologie cognitive). La réduction d'une branche scientifique à une autre prendra place si l'ensemble des théories de la première se réduisent, d'une manière ou d'une autre, aux théories de la seconde. (Oppenheim & Putnam (1958, p.5); référant à Kemeny & Oppenheim (1956))

Pour permettre cette relation de réduction, il est nécessaire que tous les termes de T_r puissent trouver une expression dans T_b. Les sciences étant ce qu'elles sont, certains termes de T_r ne se retrouvent pas dans le vocabulaire de T_b et il faut alors faire appel à un troisième ensemble d'énoncés liant les termes particuliers de T_r à ceux de T_b. Cet ensemble est constitué de règles de correspondance ('*correspondence rules*' ou '*bridge-laws*)[9] permettant une traduction logique (déductive) entre deux ensembles d'énoncés faisant appel à des prédicats *prima facie* incommensurables (sans équivalences). Ces principes de correspondance préciseront de manière nomologique les relations d'identité entre les prédicats de T_b et ceux de T_r. (Nagel, 1961, chap.11; Fodor, 1974)

Une réduction requière aussi l'adhésion à certains critères informels. Parmi ceux-ci, on demande un dimorphisme dans le champ explicatif des deux théories. La théorie réductrice T_b doit avoir un champ explicatif plus large que l'autre (T_r) pour ainsi l'englober comme une sous-classe théorique (Oppenheim & Putnam, 1958, p.5). Le domaine explicatif de la théorie de base doit être plus général que celui de la théorie de plus haut niveau et doit le contenir. Ce 'dimorphisme descendant' suit l'idée d'une hiérarchie de généralité des sciences des plus englobantes (ex : physique) aux plus circonscrites (ex : sociologie). Les entités constituant la classe des phénomènes décrits dans les sciences plus spécifiques seraient plus complexes et décomposables en des entités faisant parti du vocabulaire théorique des sciences plus générales. (Oppenheim & Putnam, 1958, pp.9-11). Cette relation n'est pas réflexive mais la décomposition suivrait l'ordre spécifié par la propriété de transitivité de la réduction (Oppenheim & Putnam, 1958, p.7). Par exemple, une société est décomposable en individus humains puis en molécules, etc.; l'inverse est faux. D'ailleurs, Oppenheim et Putnam proposent une hiérarchisation par niveau des phénomènes naturels selon leur relation de décomposition et suggèrent donc une hiérarchie similaire des sciences liées à ces niveaux (dont la relation de transition serait celle de réduction) (Oppenheim & Putnam, 1958, p.9). Le dimorphisme est descendant puisque ce sont les sciences les plus fondamentales (dont les entités composent les organisations plus complexes des niveaux plus hauts) qui ont la plus grande généralité.

Cette perspective est clairement de tradition logico-positiviste puisqu'elle considère les théories scientifiques comme un ensemble langagier. La réduction serait la traduction logique (constituée de règles de correspondances rigides) d'un vocabulaire et d'un

[9] Nagel (1961) n'utilise pas l'appellation '*bridge-laws*'. Cette expression est apparue dans la littérature subséquente. (ex : Schaffner, 1967)

ensemble d'énoncés à un autre, assurant ainsi une systématicité entre les diverses branches scientifiques. L'unité des sciences serait alors à la fois une unité langagière, nomologique et pousserait vers une unité ontologique (un matérialisme physicaliste) (Oppenheim & Putnam, 1958, pp. 3-4).

Plusieurs conséquences épistémologiques résultent de la réussite d'une réduction. D'abord, l'entreprise scientifique gagne en cohérence. Une réduction par déduction assure que les théories à différents niveaux sont cohérentes les unes avec les autres. De plus, les deux théories se consolident mutuellement. La théorie réductrice gagne accès à toutes les réussites explicatives de la théorie réduite puisque, dans l'explication des énoncés d'observation de la théorie réduite, il est possible de substituer aux prédicats de la théorie réduite ceux de la théorie réductrice. Inversement, il arrive souvent que certains problèmes irrésolus dans la science réduite gagnent en clarté par l'apport de la science réductrice. Il est clair que la relation de réduction permet un échange d'information entre les deux théories et que la réduction subséquente de la théorie originellement réductrice à une autre théorie de plus bas niveau assure une systématisation des explications scientifiques. (Kitcher, 1989) Du point de vue du camp réductionniste, ces conséquences épistémologiques sont souhaitables et devraient donc être recherchées, ce qui fait de la relation de réduction un but dont la poursuite serait profitable (P.M. Churchland, 1979). Par conséquent, le réductionniste accorde à la réduction un caractère normatif : de deux théories au même niveau explicatif, une théorie qui trouverait réduction serait supérieure à une autre qui échouerait à cet égard.

1.2 L'antiréductionnisme classique en sciences cognitives

1.2.1 Problèmes de la conception classique

Dans le cadre du débat réductionniste en philosophie de l'esprit, la psychologie se réduira à neurobiologie s'il est possible de déduire les lois [10] et le vocabulaire psychologique en leurs alter-ego neurobiologiques par le biais des règles de correspondance. Cette position, nommée '*Type-Identity Theory*'[11], est défendue par Feigl (1958, 1967),

[10] Il ne sera pas question ici d'entrer dans la controverse du statut des lois en psychologie (ex : Cummins, 1983), ni même dans le détail de ce en quoi consisterait une loi *bona fide*. Le concept de loi adopté ici sera celui proposé par Hempel & Oppenheim (1948) repris dans Hempel (1965, pp.264-270).

[11] Et parfois même simplement la '*Identity-Theory*', voir Lycan (1999).

Smart (1959) et Place (1956). Un type d'état mental, par exemple la douleur (en général dira-t-on), est la classe englobant toutes les occurrences particulières de cet état mental (ex : toutes les douleurs particulières pouvant être ressenties). Selon la '*Type-Identity Theory*', par le biais des règles de correspondance (*bridge-laws*), cette classe d'états mentaux (la douleur en général) 'Px' est identifiable (identique) à une classe d'états neuronaux 'Nx' : 'Px ≡ Nx'.

La question de l'obtention de cette condition formelle pour les deux branches scientifiques a été très controversée et aujourd'hui il semble y avoir consensus quant à l'impossibilité d'obtenir une réduction telle que la propose la '*Type-Identity Theory*'. Deux arguments particulièrement ravageurs pour la conception classique de la réduction ont été avancés pour exposer l'impossibilité *a priori* de produire une réduction nagelienne de la psychologie aux neurosciences. Le premier argument, l'anomalie du mental, a été développé par Donald Davidson (1970, 1973) mais il ne sera pas considéré ici parce qu'il ne fourni pas d'apport significatif pour la question éliminativiste[12]. Le second argument, celui de la réalisation multiple des états mentaux, beaucoup plus importants dans le présent contexte, a été introduit par Hilary Putnam (1967a) et repris en force par Jerry Fodor (1974). L'argument de la réalisation multiple a été développé parallèlement à l'explication fonctionnaliste du mental. Il ne sera question ici que de l'argumentation générale soutenant cette conception antiréductionniste puisqu'au troisième chapitre seront illustrées plus à fond la thèse fonctionnaliste et son influence sur la méthodologie explicative en psychologie cognitive.

Putnam (1967a, 1975b) aurait été le premier à introduire l'argument connu sous le nom de 'réalisation multiple des états mentaux' (*multiple realizability*) pour indiquer qu'une réduction nagelienne des états mentaux aux états neuronaux ne serait pas possible. La thèse est la suivante : parce qu'il existe (hypothétiquement) un nombre indéfini de créatures qui peuvent partager les mêmes types d'états mentaux (ex : les humains et les androïdes), et que ces créatures peuvent être faites de matériaux différents (ex : organique pour les humains, différents alliages métalliques pour les androïdes), ou peuvent réaliser ces états mentaux par des mécanismes physiques différents (ex : neurones chez les humains ou microprocesseurs chez les androïdes), il n'existe pas un *type* de phénomènes physiques auxquels le réductionniste pourrait exclusivement identifier, par les règles de

[12] Voir toutefois Bickle (1992).

correspondance, un type de phénomène mental. Ainsi, les principes de correspondance ne pourront pas fournir des lois d'identité 'Px ≡ Nx' mais uniquement des principes de correspondance disjonctifs 'Px ≡ (M_1 v M_2 v M_3 v ...)'[13] et, comme le dit Jerry Fodor, ce ne semble pas être des lois acceptables en science. (Fodor, 1974, p.108-112, voir aussi Hempel & Oppenheim, 1948)

Une illustration de réalisation multiple classique va comme suit : un être humain, dont le mental est réalisé par un système nerveux organique, même s'il peut partager un même type d'état mental (ex : la douleur) avec un potentiel extra-terrestre dont le mental est réalisé par un système de valves en silicone (Lewis, 1980), ne pourra pas partager une description physique similaire pour ce même état mental puisque les deux systèmes sont constitués suffisamment différemment pour qu'on ne puisse déterminer une relation d'identité entre le phénomène mental de douleur et ses réalisations physiques multiples.

Il ne serait donc pas possible de réduire la psychologie aux neurosciences par le schème nagelien. Il ne serait pas possible de déduire les phénomènes psychologiques (même agrémentés de règles de correspondance) à partir des lois neurobiologiques puisque les phénomènes psychologiques ne se restreignent pas aux créatures organiques et dotées d'un système nerveux. Ainsi on ne pourrait déduire dans quel état mental se trouve un androïde asimovien à partir de la théorie neurobiologique, ce qui contredit la condition formelle de déduction de la théorie à réduire T_r à partir de la théorie de base T_b. Les phénomènes psychologiques des autres créatures non-neuronales et non-organiques ne seraient pas englobés par la théorie de base T_b (neurobiologique). Le champ explicatif de la psychologie dépasse donc celui de la neurobiologie, ce qui contredit la condition du dimorphisme descendant[14]. La relation de réduction ne pourra donc pas tenir. La psychologie serait une science autonome des autres puisqu'elle ne peut être systématiquement réduite à sa sœur de plus bas niveau (neurobiologie).

Même si les phénomènes psychologiques se limitaient aux créatures neurologiques, la thèse de la réalisation multiple tiendrait le coup. Au sein même des êtres organiques, on retrouve différentes réalisations neuronales pour des mêmes processus mentaux (Endicott, 1993), et plus encore :

---

[13] Px tient pour 'état psychologique type' et Nx tient pour 'état neurologique type'. M_1 (etc.) tiennent pour des états matériels particuliers et quelconques.

[14] Selon certains (Block, 1980), ne pas permettre à la psychologie de dépasser le domaine des états mentaux présents chez les êtres vivants terrestres actuels serait chauvin et (donc?) inacceptable.

> « The case most discussed is higher-level state-types being differently realized in different species or creature-kinds. But the same higher-level state-type might be multiply realizable in different members of the same species; or in a single individual at distinct moments in its own history; or even in a single individual at a single moment in its history. » (Horgan & Tienson, 1993, p.162)

Dans ce cas, ce ne serait pas la condition de dimorphisme descendant qui ne serait pas respecté mais l'identification systématique d'un type d'entité mentale en un seul type de phénomène neuronal.

1.2.2 Autonomie de la psychologie

De cet état de fait, les antiréductionnistes (principalement Fodor, 1974) en sont venus à concevoir la psychologie comme une science autonome des autres (ou, du moins, des sciences de plus bas niveau comme la biologie, la chimie et la physique) (Stich, 1978/1999, note 8). Cette autonomie peut être entendue soit ontologiquement (ce dont il ne sera pas question ici), soit épistémologiquement. (Kim, 1978) L'autonomie de l'entreprise psychologique (versant épistémologique) peut s'entendre de plusieurs façons complémentaires.

D'abord, l'autonomie de la psychologie vis-à-vis des autres sciences indique qu'elle ne peut être réduite par celles-ci. Elle aurait son propre domaine auquel aucune autre n'aurait accès (dans sa totalité). Cela ne va pas contredire la thèse physicaliste de l'unité des sciences telle qu'elle a été présentée précédemment. La physique pourrait très bien pouvoir décrire les objets et les phénomènes produisant les états mentaux (et il y a consensus en faveur de cette thèse[15]). C'est la seconde définition de l'unité des sciences qui est rejetée : il ne serait pas possible d'articuler *systématiquement* les régularités psychologiques dans un langage physique et de les identifier à des espèces naturelles de phénomènes physiques.

L'autonomie de la psychologie signifie aussi qu'il n'y a rien d'informatif à obtenir à partir des autres sciences puisque, la relation de réduction ne pouvant pas tenir entre la psychologie et les sciences plus fondamentales, la psychologie a donc un accès privilégié à une classe de phénomènes. Ces phénomènes ne peuvent pas être spécifiés dans un langage neurobiologique de manière intéressante et pertinente pour l'entreprise scientifique. La

[15] Fodor (1974) en est l'exemple par excellence.

psychologie aurait donc une licence complète à rendre compte des régularités et phénomènes mentaux et aurait un niveau d'abstraction suffisant (abstraction du substrat physique réalisant les états mentaux) pour permettre une psychologie universelle. Dans cette conception, les états mentaux seraient une classe de la réalité qui aurait ses propres lois et sa propre ontologie (ses propres espèces naturelles) et, pour en produire une image scientifique, ce sont les régularités qui tiennent entre les états mentaux qui seraient pertinentes, pas les mécanismes matériels qui les produisent[16].

Comme l'écrit Stich (1978/1999, p.261): « The autonomy principle serves a sort of regulative role in modern psychology, directing us to restrict the concepts we invoke in our explanatory theories in a very special way. ». Ce rôle normatif vis-à-vis de la méthodologie de recherche sera exploré au prochain chapitre lorsqu'il sera question de la stratégie explicative *top-down* (section 2.2.1). C'est en réaction à cet antiréductionnisme et au décret de l'autonomie de la psychologie cognitive vis-à-vis les neurosciences que Paul M. Churchland a élaboré sa position réductionniste originale.

1.3 Le programme néo-réductionniste de P.M. Churchland

La présente section sera divisée en trois sous-sections. Il sera d'abord question du cadre formel présenté par P.M. Churchland (1979) spécifiant les diverses relations de réduction que peuvent entretenir deux théories (section 1.3.1). Il sera par la suite question de déterminer le caractère normatif de ces diverses relations, particulièrement d'explorer le rôle normatif de la relation d'élimination (matérialisme éliminativiste) dans le cadre des sciences naturelles en général (section 1.3.2) pour ensuite spécifier sa nature et son rôle dans le débat du réductionnisme en sciences cognitives (section 1.3.3). Cela nous permettra à la fois de spécifier la nature empirique de la relation de réduction à la Churchland (section 1.4) puis de jeter les bases de pour une exploration systématique du statut actuel de la théorie réductionniste de P.M. Churchland (chapitres subséquents).

1.3.1 Cadre formel du néo-réductionnisme

Un quasi-consensus s'est établi dans les années 1970 quant à l'impossibilité de procéder à une réduction du type proposé par Nagel. Les relations d'identités entre les

[16] Voir Cartwright (1989) à propos de ce type d'abstraction.

entités postulées par les deux théories ainsi que la condition de déduction d'une théorie par l'autre sont trop rigides pour permettre la réduction de la psychologie aux neurosciences (Schaffner, 1967). Il aura fallu attendre jusqu'aux débuts des années 1980 pour qu'une alternative soit proposée. Intitulée '*New Wave Reductionism*' (Bickle, 1996) (traduit ici par néo-réductionnisme ou réductionnisme nouvelle-vague), cette alternative ne conçoit plus la réduction interthéorique comme une relation de déduction mais comme une relation d'analogie entre la théorie à réduire et une image de celle-ci formulée dans les termes de la théorie réductrice. Cette perspective n'est plus normative comme l'aurait été la position nagelienne, mais chercherait plutôt à intégrer les cas de réduction déjà obtenus par l'entreprise scientifique (Bickle, 1998, p.23).

Une première idée derrière le néo-réductionnisme consiste à expliquer les cas de réduction pour lesquels une théorie fausse (ex : théorie newtonienne du mouvement des corps) est réduite à une théorie considérée vraie (ex : théorie de la relativité einsteinienne). La réduction nagelienne ne peut permettre ce genre de cas puisqu'il est impossible de déduire du faux (mécanique classique) à partir d'énoncés vrais (mécanique relativiste) (Feyerabend, 1962; Hooker, 1979, 1980).

Paul M. Churchland (1979, section 11; 1989, chapitre 3, pp.47-52) a proposé un des premiers schèmes du néo-réductionnisme. La réduction ne serait pas une relation de déduction logique mais plutôt d'analogie par isomorphisme structurel des deux théories en potentielle relation de réduction.

Ce schème va comme suit : à partir de la théorie réductrice T_b, il faut construire une image I_b de la théorie à réduire T_r basée sur les prédicats et lois de base constituant le cœur de la théorie de plus bas niveau T_b. De plus, l'image I_b est construite de manière à refléter la structure des lois et des prédicats de T_r. C'est en ce sens que I_b est une version analogue de T_r dans T_b. L'image I_b analogue à T_r est construite pour couvrir le domaine explicatif de T_r. I_b est donc formulée dans le vocabulaire de T_b et il sera alors possible de produire des énoncés par I_b et vérifier leur validité dans T_b. De manière générale, cette image ne nécessitera pas l'appel à l'entièreté du corpus théorique de T_b puisque, selon le principe de dimorphisme descendant, T_b a un champ explicatif plus ample que T_r. S'il est possible de bâtir une telle image, T_r sera en relation de réduction à T_b.

Toujours selon P.M. Churchland (1979), si l'on tente une réduction de ce genre, quatre situations peuvent advenir. Une première possibilité consiste à obtenir une réduction

conforme au schème nagelien de la réduction. S'il est possible de construire une image I_b dont la structure serait parfaitement isomorphe avec la structure de la théorie T_r et que toutes les propositions dans I_b soient vraies dans T_b, alors nous obtenons une réduction parfaite des prédicats et lois de T_r aux prédicats et lois de T_b. L'isomorphisme permet d'établir une série de principes de correspondances (*bridge-laws*). Une situation de parfait isomorphisme répond aux critères de la réduction nagelienne puisque I_b est consistante à la fois à T_r et à T_b. Cette situation est toutefois rarement obtenue en sciences. (P.M. Churchland, 1979; Schaffner, 1967)

Deux autres situations peuvent survenir, dans lesquelles on obtiendrait une réduction quasi-parfaite mais requérant une certaine révision de T_r pour assurer une réduction souple de sa version corrigée. Dans les deux situations, T_r et T_b sont partiellement incommensurables. Dans un premier cas, l'image I_b peut refléter une large partie de T_r mais sans obtenir un isomorphisme parfait : une partie de l'image I_b ne trouverait pas d'équivalent dans T_r parce que la structure de la théorie réductrice T_b ne pourrait permettre cet isomorphisme. Dans un second cas, l'isomorphisme d'I_b à T_r pourrait être obtenu mais certains énoncés d'I_b seraient faux dans T_b. Dans un cas comme dans l'autre, il faudra produire une image T_r' de T_r permettant l'obtention de l'isomorphisme ou rectifiant la structure de T_r pour assurer la vérité des énoncés de I_b dans T_b. Ce ne sera donc pas T_r qui sera réduite mais une image (T_r') similaire à celle-ci. Nous pouvons donc considérer que la théorie T_r sera révisée de manière à se réduire conformément à T_b, ce que les néo-réductionnistes nomment 'réduction par révision'. (Bickle, 1996, p.66; P.M. Churchland, 1979, p.83-85) La théorie réduite subie des corrections à la lumière de la théorie réductrice.

Une quatrième et dernière situation consisterait en l'impossibilité de construire une image I_b isomorphe à T_r *vraie* dans T_b sans complètement dénaturer T_r par une image T_r'[17]. I_b serait constituée d'une forte majorité d'énoncés faux dans T_b. Dans ce cas, la réduction serait impossible et l'antiréductionnisme serait la position à adopter puisque T_b et T_r sont radicalement incommensurables. La réduction échouerait parce que T_r et T_b diffèrent trop pour permettre une relation d'analogie entre les deux.

---

[17] Il ne semble pas exister de critères formels mesurant si les corrections apportées à T_r par T_r' sont trop drastiques pour permettre le rejet de la relation de réduction ou pas. Toutefois Bickle (1998, p.199-203) propose trois critères informels permettant de distinguer la révision, le remplacement ou la rétention de la théorie T_r. Étant donné que ce qui nous intéresse ici est la position de P.M. Churchland, il ne sera pas question de ces critères.

Dans le schème néo-réductionniste, il est question de degré de réduction, contrairement à la conception nagelienne de la réduction où il y avait réduction par déduction ou pas. Les néo-réductionnistes illustrent ces différents degrés par un continuum du résultat de la réduction. À une extrémité nous retrouvons la réduction nagelienne, idéale et parfaitement logique (*smooth reduction* [18]). À l'autre nous trouvons une incommensurabilité radicale forçant l'antiréductionnisme (*bumpy reduction*[19]). Entre ces deux pôles, nous retrouvons une réduction plus ou moins aisée en fonction de la quantité de révision nécessaire pour permettre une relation d'isomorphisme entre les deux théories[20].

Le procédé de réduction doit donc suivre certaines étapes dans son élaboration. La première étape consiste à (1) déterminer le format de l'explication pour les deux théories (sont-ce des lois, des modèles, etc.?). Il faut ensuite (2) déterminer la structure de la théorie à réduire (T_r) pour ainsi obtenir un canevas à partir duquel construire l'image I_b. (3) La construction d'I_b débutera à partir de la théorie réductrice T_b dont il faudra préalablement déterminer la structure pour ainsi permettre une (4) élaboration de la traduction de T_r dans T_b par I_b. La dernière étape consistera à (5) analyser la validité des énoncés constituant I_b dans T_b et de déterminer si la structure d'I_b est parfaitement isomorphe avec T_r. Si cela n'est pas le cas, il faudra (6) réviser T_r par la construction d'une théorie similaire T_r' et revenir à l'étape précédente (5) jusqu'à ce que l'isomorphisme soit assuré ou que l'incommensurabilité radicale soit décrétée.

Cette 'recette' est ici capitale puisque c'est elle qui permettra, dans les chapitres suivants, de reconstruire le statut actuel de la relation de réduction telle que l'entend P.M. Churchland. Le chapitre II et le chapitre III (à l'exception de la section 3.3.3) toucheront particulièrement au point (1) alors que les sections suivantes (dont le chapitre IV) élaboreront les étapes (2) à (4). Les étapes suivantes sont affaire de recherches empiriques dépassant l'objectif du présent travail. Toutefois, il sera possible d'en esquisser quelques particularités à partir du travail effectué dans les pages qui suivent celle-ci (section 4.3).

---

[18] P.M. Churchland (1979, p.84)

[19] Ibid.

[20] Voir Bickle (1996) pour un schéma du continuum de la réduction nouvelle-vague et son équivalent prescriptif.

1.3.2 Le matérialisme éliminativiste de Paul M. Churchland

Malgré la prétention à une approche descriptive de la réduction telle que proposée par Hooker (1981) et Bickle (1998), une clause normative est implicite dans la perspective néo-réductionniste : c'est la théorie à réduire T_r qui devra être altérée pour permettre la relation de réduction. P.M. Churchland entrevoit la possibilité d'ajouter certaines clauses contextuelles ou contrefactuelles à T_b pour en former une image T_b' qui permettrait aux énoncés de I_b d'être vrais en T_b', mais jamais il n'est question de reformuler ou d'éliminer une partie de la théorie réductrice (P.M. Churchland, 1979, p.83-84).

Bien que la révision soit orientée, une approche purement descriptive n'a aucune force quant à l'attitude à prendre vis-à-vis l'une des situations résultantes possibles. Il n'est pas spécifié par quelle méthode la révision de T_r doit être entreprise, ni non plus quelles sont les limites de cette révision. Il n'est pas question non plus, jusqu'à présent, de ce qu'il faut faire si la réduction échoue. Il faudra donc un apport normatif pour permettre de répondre à ces problèmes : c'est à ce stade qu'entre en jeu l'éliminativisme.

Le caractère prescriptif du néo-réductionnisme de P.M. Churchland provient d'un ensemble de clauses normatives (valeurs métathéoriques ou *epistemic values*) dictant ce qui fait d'une théorie une bonne théorie scientifique et, par conséquent, indique aussi quelle attitude prendre vis-à-vis les théories dans une situation de réduction. P.M. Churchland ne fait pas une liste de ces valeurs, celles-ci sont évoquées de manière disparate au fil de ses analyses.

P.M. Churchland accorde une plus grande valeur à une théorie par l'amplitude de son intégration à l'ensemble de l'entreprise scientifique par sa cohérence avec les théories déjà bien assises dont les domaines explicatifs sont proximaux au sien.

> « [...] we must evaluate [a theory] with regard to its *coherence* and *continuity* with fertile and well-established theories in adjacent and overlapping domains because active coherence with the rest of what we presume to know is perhaps the final measure of any hypothesis. ». (P.M. Churchland, 1989, p.6; je souligne)[21]

Pour P.M. Churchland, l'unité en science n'est pas simplement un idéal épistémologique : c'est un critère normatif auquel devraient répondre toutes les théories

[21] L'extrait est tiré de l'article '*Eliminative Materialism and the Propositional Attitudes*' (P.M. Churchland, 1981) constituant le premier chapitre du recueil (pp.1-22) P.M. Churchland (1989). La page réfère au recueil.

scientifiques. Ce critère sera dorénavant invoqué sous le nom de 'critère d'intégration scientifique'.

P.M. Churchland invoque aussi un critère comparatif : de deux théories, la théorie à retenir devrait être celle qui permet la meilleure explication du domaine empirique qu'elles partagent. La qualité de la prédiction, sa précision et l'ampleur du champ explicatif d'une théorie joue en sa faveur vis-à-vis d'une autre qui ne saurait détenir ces qualités au même degré. « The choice is made rather on grounds of the relative "internal" virtues of the two alternative frameworks: on their inductive coherence, their explanatory unity, their informational richness, and suchlike (somehow understood) ». (P.M. Churchland, 1979, p.78) Il est à noter que P.M. Churchland ne donne pas de précision quant aux moyens de mesurer ces qualités. Ce critère sera dorénavant invoqué sous le nom de 'critère de qualité empirique'.

Avec ces deux valeurs métathéoriques en tête, il est dorénavant possible d'esquisser un portrait de l'attitude à prendre dans les cas de réduction par révision et d'incommensurabilité radicale. Favoriser une théorie vis-à-vis une autre consiste à mettre au rancart la moins performante des deux. C'est ce que l'histoire nous a appris. S'inspirant de la réduction de la mécanique classique à la mécanique relativiste, P.M. Churchland (1979, section 11) insiste que l'attitude à prendre (celle qui a été prise) consiste à rejeter la théorie moins performante ou ses éléments 'défectueux' pour intégrer la théorie plus performante. Ce rejet indique que la théorie moins fortunée (ou certains de ses éléments défaillants) ne sont plus adéquats pour faire partie de l'ensemble des théories acceptées. Dans T_r, ce qui correspondra aux énoncés faux de I_b dans T_b ou résistera à un isomorphisme entre I_b et T_r devra donc être rejeté comme étant simplement faux. Ce rejet consiste à éliminer du corpus scientifique accepté les éléments problématiques de la théorie réduite. Il faudra alors les remplacer par un ensemble de lois et/ou prédicats cohérents et en continuité avec T_b (réduction par révision).

Dans le cas d'une incommensurabilité radicale, c'est-à-dire dans les cas où I_b serait principalement constituée d'énoncés faux, il faudra rejeter T_r et léguer son domaine explicatif à T_b. La théorie qui répond le mieux aux critères susmentionnés doit être conservée alors que l'autre doit être éliminée, i.e. abandonnée comme explication scientifique pour un phénomène donné. C'est ce en quoi consiste la thèse de l'éliminativisme. Celle-ci stipule que dans le cas où une théorie échouerait à être réduite à

une autre théorie de plus bas niveau mieux intégrée et plus explicative, la théorie de plus haut niveau devrait être rejetée comme une théorie fausse et remplacée par la théorie de base.

P.M. Churchland suggère fortement que T_b répondrait toujours mieux aux deux ensembles de valeurs métathéoriques, d'où l'orientation de la révision et de l'élimination. Il ne semble pas exister (chez P.M. Churchland) d'analyse de l'attitude à prendre si la théorie de plus haut niveau (T_r) répondait mieux à ces valeurs que la théorie de plus bas niveau (T_b).

1.3.3 Le matérialisme éliminativiste en sciences cognitives

Dans la situation de la potentielle réduction de la psychologie aux neurosciences, la position de P.M. Churchland est claire et bien connue : la psychologie devra être éliminée au profit des neurosciences (P.M. Churchland, 1989, chapitre 1). P.M. Churchland partage donc la position antiréductionniste avec la communauté philosophique des années 1970, mais pour des raisons indépendantes. L'impossibilité de réduire la psychologie aux neurosciences ne serait pas causée par la réalisation multiple des états mentaux mais bien parce que la psychologie (cognitive) utilise un cadre conceptuel erroné, menant à des explications scientifiques radicalement fausses[22]. Les prédicats utilisés dans les théories psychocognitives ne seraient simplement pas des espèces naturelles et ne réfèreraient, en fait, à rien de réel.

Les prédicats utilisés par la théorie psychocognitive seraient tirés d'un cadre conceptuel issu du sens commun : la psychologie du sens commun (*folk psychology*). (P.M. Churchland, 1981) Cette dernière consiste en un vocabulaire incluant les états mentaux utilisés par tout un chacun pour s'expliquer ses propres comportements ainsi que ceux d'autrui tel que, par exemple, les états mentaux de peur, de douleur, de croyance, de désir, de joie, d'intention, etc. « "Folk psychology" denotes the prescientific, commonsense conceptual framework that all normally socialized humans deploy in order to comprehend, predict, explain, and manipulate the behavior of humans and the higher animals. ». (P.M. Churchland, 1998, p.3) Il sera question au quatrième chapitre de cette utilisation de la psychologie du sens commun comme cadre conceptuel de base pour les théories

[22] Il ne serait pas possible de produire à partir de T_r, même d'un T_r', une image I_b qui ne soit pas gravement infectée d'énoncés faux dans T_b, même si l'isomorphisme devait être atteint.

psychocognitives et, pour la suite, la critique que mène P.M. Churchland vis-à-vis la psychologie du sens commun sera assimilée à une critique de la psychologie cognitive dans son ensemble.

Malgré sa généralité, ce cadre conceptuel aurait plusieurs lacunes explicatives. Plusieurs phénomènes mentaux d'importance ne figureraient pas dans le domaine explicatif de la psychologie du sens commun. Il n'y aurait pas d'explication du sommeil (et des rêves), de plusieurs maladies mentales, de l'effet sur le mental de psychotropes et de lésions cervicales (ex : *split-brain*), d'une large gamme d'instincts et de curiosités perceptuelles (ex : *blind-sight*), de l'apprentissage, du développement psychologique, etc. (P.M. Churchland, 1989, pp.6-7) Ainsi, la psychologie cognitive ne répondrait pas bien au critère de qualité empirique, alors que les neurosciences promettraient un plus large domaine explicatif intégrant ce qui est ignoré par la psychologie cognitive. Pour ne donner que quelques exemples, l'effet des psychotropes trouve explication par leur effet sur les récepteurs dendritiques[23], champ particulier de la neuropharmacologie; les phénomènes de '*split-brain*' et de '*blind-sight*' trouveraient explication dans les contraintes du développement neuronal[24]; etc.

La psychologie cognitive souffrirait aussi d'une large inconsistance avec le reste de l'entreprise scientifique alors que les neurosciences seraient en continuité directe avec les autres sciences (ex : biologie). Cette accusation envers la psychologie cognitive en rapport avec le critère d'unité scientifique provient de deux critiques : la première concerne la formalisation des états mentaux dans les explications scientifiques en psychologie cognitive et la seconde porte sur l'argument de la réalisation multiple.

Les états mentaux postulés par la psychologie cognitive sont usuellement formalisés en attitudes propositionnelles (Russell, 1910; P.M. Churchland, 1970, 1979). Par exemple, une croyance (un état mental) est formalisée sous la forme « x croit que p » où *x* est l'agent observé et *p* une proposition telle que « la neige est blanche ». L'agent serait en relation avec la proposition et une théorie psychologique devrait définir cette relation. Le modèle computationnel classique utilisé en sciences cognitives postule que la cognition est un processus automatisé manipulant des chaînes de symboles représentant des propositions et que l'appareillage constituant la cognition ne procéderait plus qu'à une transformation

---

[23] (Bear et al., 2002, p.122-123)

[24] Arbib (1998, p.249,319); Bear et al. (2002, p.682-687); P.S.Churchland (1986, p.174-193, 224-228), Goldstein (2002)

logique des propositions présentes en elle. (Fodor, 1975; Newell, 1980, 1990) P.M. Churchland critique vivement cette conception et rejette la thèse que ce serait ce type d'opérations qui constituerait la cognition : il n'y a pas de symboles discrets (de propositions) dans le cerveau. (P.M. Churchland & P.S. Churchland, 1983) Une théorie computationnelle devra donc prendre en compte que ce sont des neurones et leurs interrelations causales d'activation et d'inhibition qui permettent la computation des informations provenant de l'environnement. Il est improbable que les créatures plus limitées au niveau cognitif (ex : oiseaux) computent des chaînes de symboles exprimés, par exemple, dans un langage de l'esprit. (P.M. Churchland, 1981, 1989, 2007 (chapitre 2); P.M. Churchland & P.S. Churchland, 1983) Toutefois, ce rejet ne concerne que la forme de computation effectuée par la cognition (manipulation de symboles selon des règles logiques) et pas la nature computationnelle des opérations mentales. (P.M. Churchland, 1989, chapitres 1, 5, 7, 9 et 10; P.M. Churchland & P.S. Churchland, 1983, p.8-13)

Cela permet d'ailleurs de comprendre l'éliminativisme de P.M. Churchland selon deux perspectives. D'abord, P.M. Churchland rejette le bagage conceptuel emprunté à la psychologie du sens commun, les concepts tels que ceux de désir et de croyance, peu importe leur structure computationnelle. Dans un second temps, P.M. Churchland rejette simplement la thèse de la computation symbolique voulant que les états mentaux soient adéquatement formalisés par les attitudes propositionnelles. Il sera plus amplement question de la formalisation computationnelle des états mentaux au prochain chapitre dans la discussion concernant les niveaux explicatifs; on y verra que le rejet de la computation de symboles discrets n'a que peu d'incidences directes sur le programme réductionniste de P.M. Churchland (la critique visant plutôt à administrer une réforme de la méthode en sciences cognitives et de l'attitude des psychologues). D'ici là, il faudra entendre l'éliminativisme de P.M. Churchland dans son premier sens, soit celui du rejet du cadre conceptuel classique (psychologie du sens commun) des états mentaux. C'est d'ailleurs contre celui-ci qu'est dirigé la seconde critique.

P.M. Churchland (2007, chapitre 2) mène une attaque directe contre l'autonomie de la psychologie en présentant une série de cas de réduction où il y a une situation de réalisation multiple de l'entité réduite. Par exemple, la température dans un gaz est réalisable dans de multiples substances (différents gaz) et pourtant cette espèce naturelle a été réduite (identifiée) à l'énergie cinétique moyenne des molécules constituant le gaz en

question. P.M. Churchland y offre aussi sept autres exemples de réduction interthéorique malgré la réalisation multiple des entités réduites. Si ces cas de réduction sont légitimes, alors il n'y a aucune raison de penser que la réalisation multiple des états mentaux joue un rôle antiréductionniste. On devrait plutôt s'attendre à obtenir une réduction et si celle-ci est impossible, c'est plutôt un signe que la théorie ne réfère pas à une espèce naturelle. (P.M. Churchland, 2007, pp.24-28)

> « On this alternative logical and historical pattern, legitimate molar-level theories that comprehend genuine natural kinds will thus be positively *expected* to find some such intertheoretic reduction. For if they eventually prove *not* to be thus reducible, we will have to reconsider the initial presumption that the molar-level theory really does embrace genuine high-level kinds governed by genuine high-level explanatory laws. » (P.M. Churchland, 2007, p.28; italiques dans l'original)

Dans chacun des cas présentés, il existe un principe à un niveau plus fondamental qui unifie les réalisations multiples. Ce principe ne fait pas simplement lier ensemble les différentes réalisations matérielles d'un phénomène : il en offre une explication informative. Dans le cas des processus mentaux, P.M. Churchland dégage deux possibilités (complémentaires) de principes permettant d'en unifier les réalisations multiples. (P.M. Churchland, 2007, pp. 27-33) Le premier principe est tiré de la nature thermodynamique de l'information et concerne alors la nature des processus computationnels. Pour assurer une meilleure continuité des idées, ce principe sera exposé au chapitre suivant. Le second principe, qui nous intéresse ici, consiste à postuler que toutes les entités dotées d'états mentaux sont en fait des créatures cognitives naturelles (*natural* (*wild*) *epistemic engines*). (P.M. Churchland & P.S. Churchland, 1983, p.7; conception reprise dans P.M. Churchland (2007, pp.28-33))

L'approche de P.M. Churchland est naturaliste : la nature de la connaissance n'est pas affaire d'épistémologie *a priori* mais devrait plutôt être déterminée par l'étude scientifique du fonctionnement de l'esprit humain (approche psychologique de la connaissance) (P.M. Churchland (1979); P.M. Churchland & P.S. Churchland (1983); Quine (1969, chapitre 3)) Ainsi la cognition ne devrait pas être étudiée dans son abstraction logique (il sera question de cette stratégie utilisée par la psychologie cognitive à la section 4.1) mais plutôt dans sa réalisation concrète. L'attitude naturaliste pousse donc à étudier les processus mentaux sous leur forme réalisée ici sur Terre et à intégrer à cette

étude les aspects évolutifs qui les ont forgés (évacuant ainsi la critique de chauvinisme évoquée précédemment).

Les capacités cognitives naturelles sont définies par leur fonction vis-à-vis de la sélection naturelle :

> « Call an epistemic engine any device that exploits a flow of environmental energy, and the information it already contains, to produce more information, and to guide movement. So far as natural (wild) epistemic engines are concerned, survival depends on a fit between the information contained and the world it inhabits. » (P.M. Churchland & P.S. Churchland, 1983, p.7)

Ce ne sont donc pas les types de fonctions cognitives exécutées par les états mentaux qui unifient le domaine de la psychologie mais plutôt la fonction plus globale de la cognition qui est de transformer l'information pertinente disponible dans l'environnement de la créature pour coordonner ses mouvements afin d'en assurer la survie et la reproduction (voir Godfrey-Smith (1991, 1996) pour une formalisation de cette fonction et de ses conditions adaptatives). Cette conception a son importance ici puisque nous verrons qu'elle supporte directement le format fonctionnel de l'explication psychologique (même dans le cadre conceptuel proposé par P.M. Churchland).

P.M. Churchland va plus loin encore en restreignant le domaine de la psychologie aux créatures terrestres :

> « If our alternative portrait of cognition is even roughly correct, the central job of cognitive psychology is to explore how it is that terrestrial brains are able to compute the extraordinary variety of functions displayed in diverse species of cognitive creatures. » (P.M. Churchland, 2007, p.35)[25]

Dans ce cas, imposer à toutes les créatures le cadre conceptuel de la psychologie du sens commun serait une activité anthropomorphe puisque les différentes cognitions animales seraient hétéroclites. « Clearly, such diverse creatures are not all computing the same functions, nor even remotly simlilar functions. » (P.M. Churchland, 2007, p.34) Ce qui unifierait le domaine cognitif des créatures terrestres serait exactement ce qui semblait disparate aux fonctionnalistes : les différentes cognitions sont toutes réalisées par des réseaux neuronaux.

Pour P.M. Churchland, la relation de réduction n'est pas simplement un critère esthétique qui fournirait aux scientifiques une image polie et bien structurée des théories. La réduction apporte avec elle une meilleure compréhension de la nature et du

[25] Voir aussi McCauley (1996, p.219).

fonctionnement des phénomènes à divers niveaux de réalité. La relation de réduction est explicative : l'unification de phénomènes en apparence différents permet une meilleure compréhension des phénomènes (mentaux) et surtout permettent d'améliorer la précision de la prédiction et la cohésion des théories scientifiques. (P.M. Churchland, 1979, section 11) Cette nature explicative de la relation de réduction permettra, au chapitre suivant, de déterminer la relation entre les divers niveaux explicatifs et ainsi de mieux cerner le contexte explicatif dans lequel devra s'inscrire le projet réductionniste de P.M. Churchland.

1.4 Le prophétisme de la réduction interthéorique

Dans son *Matter and Consciousness* (1988), P.M. Churchland compare les trois positions principales concernant la possibilité et la nature de la réduction interthéorique et relève un présupposé affectant autant la '*Type-Identity Theory*' que le fonctionnalisme antiréductionniste. Ces deux positions présupposent toutes deux la validité du cadre conceptuel de la psychologie du sens commun utilisé par la psychologie cognitive. Seule la position éliminativiste remettrait en question cette prémisse en la considérant comme une thèse en mal de confirmation empirique (P.M. Churchland, 1988, chapitre 2). Cette idée remet en perspective une caractéristique de la relation de réduction qui a été jusqu'ici implicite et dont l'analyse permettra de bien comprendre la pertinence de ce travail. Elle est constituée de deux facteurs. La réduction est une relation ne concernant que les théories achevées (critère temporel)[26] et ne peut donc être vérifiée (ou infirmée) que lorsque les deux théories concernées seront achevées (critère d'empiricité) (Fodor, 1974, p.100). P.M. Churchland (1981/1989, chapitre 1), appuyé par P.S. Churchland (1986, chapitre 9), est très clair à ce sujet : on ne peut rejeter la pertinence de l'investigation neuroscientifique par des arguments *a priori* car ceux-ci sont tous (jusqu'à présent) basés sur la validité du cadre conceptuel de la psychologie du sens commun. L'argument de la réalisation multiple suppose que les états mentaux pouvant être entretenus par divers substrats matériels ou produits par différents mécanismes sont des espèces naturelles. Or cette supposition nécessite une validation empirique que seule une hypothétique psychologie achevée pourra fournir. Dans ce cas, l'argument de la réalisation multiple n'est plus un argument *a priori*

[26] Dans la section « Formal Conditions for Reduction I.a. », Nagel dit explicitement : «In the highly developed science S... », (Nagel, 1961, p.346), puis élabore les critères formels de réduction. Quant à eux, Oppenheim et Putnam utilisent une variable temporelle pour établir la réduction d'une branche scientifique (e.g. chimie, physique) à une autre (Oppenheim & Putnam, 1958, p.5)

mais une thèse empirique *à vérifier* puisqu'elle est basée sur la validité du cadre conceptuel utilisé par la psychologie cognitive.

Si l'on accepte cette position, alors il en va de l'éliminativisme comme du fonctionnalisme : dans l'optique d'obtenir une théorie explicative des phénomènes mentaux (cognitifs), on ne peut, dès à présent, rejeter la pertinence de l'une ou l'autre des approches scientifiques à partir de la validité (ou de l'invalidité) de la thèse réductionniste. En ce sens, il faut donc éviter le prophétisme des fonctionnalistes qui cherche à assurer une autonomie de la psychologie tout comme il faut éviter de tomber dans le piège d'un éliminativisme assuré qui rejetterait tout bonnement les efforts scientifiques des psychologues[27]. La relation de réduction entre la psychologie cognitive et les neurosciences cognitives est une thèse empirique, i.e. dont il faudra non pas en prophétiser l'obtention ou l'échec mais plutôt en vérifier la teneure une fois les deux théories en compétition suffisamment complétées pour permettre d'apercevoir une image claire et nette de leurs affinités et divergences.

À l'époque où les divers camps se sont formés autour de la question réductionniste, tout comme aujourd'hui, les deux approches théoriques n'étaient pas (et ne sont toujours pas) suffisamment développées pour que l'on puisse aspirer dès à présent obtenir une analyse sérieuse de la possibilité de réduction interthéorique. Dans ce cas, être réductionniste ou antiréductionniste consisterait à jouer au prophète; il faudra encore attendre pour déterminer quel parti gagnera sa mise. Toutefois, il est possible d'entrevoir ce qu'il faudra obtenir pour mettre à l'épreuve la thèse réductionniste. Les six étapes présentées à la page 18 devront être complétées et il est déjà possible d'en débuter quelques unes.

Le cadre conceptuel du (néo)réductionnisme, duquel P.M. Churchland tire son éliminativisme, étant maintenant bien positionné vis-à-vis le cadre réductionniste classique ainsi que les critiques antiréductionnistes qui en découlèrent dans le champ des sciences cognitives, il est maintenant à propos de procéder à l'exploration de la nature et de la forme de l'explication en sciences cognitives. Le prochain chapitre s'enquerra de proposer un cadre conceptuel déterminant les relations entre l'explication proprement psychologique et l'explication neuroscientifique, soit la division tripartite des niveaux explicatifs. Le troisième chapitre (sections 3.1 et 3.2) raffinera le format de l'explication en sciences cognitives (l'explication fonctionnelle), permettant ainsi de débuter la construction d'un

[27] Clark (1996) est aussi de cet avis et critique l'attitude prophétique de Paul M. Churchland.

portrait schématique de l'état des faits concernant la relation de réduction telle que la conçoit P.M. Churchland. Cela permettra de délimiter l'horizon de la possibilité de réduction en spécifiant ce qui est déjà obtenu comme unité et quel travail il reste à faire pour en arriver à vérifier la validité de la thèse éliminativiste.

Chapitre II - Niveaux explicatifs en sciences cognitives

Le chapitre précédent a permis de déterminer le schème (néo)réductionniste global de P.M. Churchland et de le différencier de la conception nagelienne classique. Il a aussi été question de la marche à suivre pour déterminer empiriquement la validité de la thèse réductionniste entre la psychologie cognitive et les neurosciences. La marche à suivre donnée à la page 18 indique que la première étape logique à parcourir pour y parvenir consiste à déterminer le format de l'explication scientifique propre aux sciences cognitives. Toutefois, avant de s'y pencher, il est important de bien saisir le pluralisme explicatif immanent aux sciences cognitives contemporaines. Ce dernier se traduit par une approche sous trois perspectives explicatives qui entretiennent entre elles des relations précises. L'objectif de ce chapitre est de déterminer la nature de ces perspectives (section 2.1) ainsi que de déterminer les relations qui les unissent (section 2.1.3 et 2.1.4). Les chapitres suivants se construiront à partir du schème tripartite des niveaux explicatifs. De plus, ces trois approches ont des répercussions méthodologiques qui jouent un rôle important dans le cadre du débat de la réduction. Bien que celles-ci soient traitées à la fin du présent chapitre (section 2.2), leurs conséquences pour le débat seront particulièrement présentes lorsqu'il sera question de l'état actuel de la relation de réduction entre les deux théories scientifiques qui nous intéressent ici (section 4.1 et 4.3).

2.1 Les niveaux explicatifs en sciences cognitives

Nombre de philosophes, de psychologues et de théoriciens en intelligence artificielle s'entendent sur la nécessité d'approcher les systèmes cognitifs naturels et artificiels selon une perspective en trois niveaux d'explication[28]. Si Marr (1982) est généralement considéré comme le père de la première explicitation de la conception tripartite des niveaux explicatifs pour un phénomène mental (ou cognitif), conception

[28] Bechtel (1994); Bermudez (1995); P.S. Churchland & Sejnowski (1992); Dennett (1971); Dyer (1991); Horgan (1992); Horgan & Tienson (1993); Lycan (1987); Marr (1982); McCauley (1986); McClamrock (1991); Newell (1980, 1982, 1990); Peacocke (1986); Pylyshyn (1984); Van Eckardt (1993); Wimsatt (2006). Plusieurs philosophes ne distinguent pas trois niveaux mais bien deux : les niveaux personnels et sub-personnels (par exemple, Hurley (1998)). Ces auteurs intègrent les niveaux computationnel et neurobiologique dans le niveau sub-personnel. Le niveau personnel est à peu de choses près le niveau sémantique (voir explications plus bas).

relativement tacite dans la méthodologie explicative des sciences cognitives jusqu'à la publication de l'*opus magnum* de Marr, plusieurs variantes ont été proposées. La structure de base de cette organisation de l'explication semble faire objet de consensus malgré que de légères nuances (dont l'appellation donnée à chacun des trois niveaux) différencient la conception des uns et des autres selon leur perspective scientifique et/ou philosophique. L'approche utilisée ici est similaire à celle de Dennett (1971, 1987, 1991a, 1991b) puisque je ne traiterai pas des niveaux explicatifs comme des niveaux d'organisations ontologiques mais comme différentes perspectives théoriques concernées par un même problème scientifique. Cela aura une importance considérable quand il sera question de la relation de réalisation explicative (section 2.1.3).

Une étude des niveaux explicatifs s'impose ici parce que ceux-ci permettent de mieux cerner la structure de la réduction interthéorique entre les théories psychocognitives et neurocognitives. La possibilité de produire une réduction comme le propose Paul M. Churchland ne semble pas se faire simplement en produisant une image neurobiologique des processus psychologiques puisqu'il n'est pas évident, au premier regard, de déterminer ce qui, dans le cerveau, produit du mental. Ce sont les sciences cognitives, et les philosophes concernés par ce programme de recherche, qui ont permis de reformuler le problème ontologique de la relation du mental au corps sous une forme épistémologique. (Bickle, 1998)

Le niveau explicatif sémantique et celui de l'implémentation matérielle, comme nous le verrons, reflètent la problématique classique de la relation esprit/corps. La psychologie et le niveau sémantique ont une longue histoire et les études neurobiologiques (niveau d'implémentation matérielle) avaient connues un essor fulgurant au XIX^{ème} siècle : il était dorénavant sans conteste que le cerveau jouait le rôle de 'siège' de l'esprit et non plus d'un système de climatisation (conception aristotélicienne). Mais ce n'est qu'avec l'introduction de l'idée que le cerveau humain était une machine à traitement de l'information que le niveau computationnel a commencé à faire sens, c'est-à-dire que la nouveauté dans cette image provient de l'intégration du niveau computationnel comme intermédiaire entre les deux niveaux mentionnés précédemment. (Gardner, 1985) Cet intermédiaire est issue du développement des sciences cognitives et représente le paradigme central de ces 'jeunes' sciences : la cognition serait en fait un mécanisme complexe de transformation de l'information présente dans l'environnement. Le

développement de la théorie computationnelle en mathématique a fournie aux psychologues et aux neurobiologistes une méthode de formalisation des processus cognitifs, ce qui leur a permis de théoriser à propos des sous-procédures par lesquelles parvenir à une mécanisation de la cognition. Avant l'apparition du paradigme computationnel et de l'instauration du niveau explicatif approprié, la psychologie restait en mal d'un pont entre esprit et matière. (Hatfield, 2002)

Si cette image de la compréhension du fonctionnement du mental (*mind*) et de ses relations avec le système nerveux central est juste, il semble donc que pour permettre la réduction de la psychologie cognitive aux neurosciences il faille passer par un portrait computationnel des processus cognitifs. Les relations entre les phénomènes mentaux, leur forme computationnelle et leur substrat matériel sont rendues par la structure tripartite des niveaux explicatifs. Une analyse de ceux-ci et de leurs interrelations est de mise puisque c'est par elle que nous pourrons dresser un portrait de la structure que devra prendre l'image neurobiologique des processus cognitifs.

Plutôt que d'explorer la littérature concernant la caractérisation adéquate des niveaux explicatifs et leurs implications ontologiques, il sera plutôt question ici de montrer qu'il existe, entre ces trois niveaux, une place pour inclure la relation de réduction. Si l'analyse est adéquate, on pourra constater l'obtention d'un terrain d'entente minimal permettant une meilleure communication entre les théories psychocognitives et neurocognitives (toujours dans l'optique d'offrir un portrait de l'état des lieux pour réaliser la vérification de la thèse réductionniste de P.M. Churchland). Pour y arriver, il sera donc objet de la première partie de cerner les aspects propres aux différents niveaux puis d'analyser les relations qui les unifient ainsi que les contraintes qu'ils exercent les uns sur les autres. La seconde section de ce chapitre s'enquerra des impacts méthodologiques que cette répartition tripartite des niveaux explicatifs exerce sur l'entreprise des sciences cognitives.

2.1.1 Caractérisation de la nature explicative des trois niveaux

La logique derrière la répartition des niveaux explicatifs est simple : il y a trois façons de comprendre un même processus cognitif. Les trois niveaux sont organisés en fonction de la dichotomie esprit/matière *plus* la thèse computationnelle avancée par les sciences cognitives. Il y a le niveau sémantique auquel sont spécifiées la nature des états

mentaux, leurs interrelations causales globales ainsi que le contenu sémantique des représentations mentales. Le scientifique (psychologue) cherchera à déterminer en quoi consiste un état mental et quel est son rôle dans l'économie fonctionnelle de la cognition d'une créature. Dans le cas de la perception visuelle par exemple, le psychologue s'interrogera sur le fonctionnement ordinaire de la perception : que perçoivent les humains dans leur environnement visuel pour se former des représentations du monde qui les entoure et comment ces perceptions affectent-elles les autres états mentaux des individus? Inversement, au niveau de l'implémentation matérielle, on cherche à déterminer comment les mécanismes cognitifs sont réalisés dans la matière (dans le cerveau, un ordinateur, etc.) et quels sont les impacts de la structure matérielle du système cognitif sur ses opérations mentales. La question derrière ce niveau explicatif demande : « Comment des relations causales matérielles en viennent-ils à produire des processus cognitifs et quels sont les impacts des mécanismes matériels sur la cognition? ». Dans le cas où le mécanisme matériel est un cerveau, c'est le neurobiologiste qui se soucie de répondre aux questions posées à ce niveau. Dans le cas de la perception visuelle, le neurobiologiste s'intéressera au fonctionnement du nerf optique, de la décussation des capteurs rétiniens, etc. Finalement, le niveau computationnel est un niveau intermédiaire jouant le rôle de pont entre les deux autres niveaux. On y spécifie la logique des transactions informationnelles devant prendre place dans un système cognitif pour réaliser la fonction des différents processus mentaux. Le scientifique intéressé par la perception visuelle cherchera à comprendre comment l'information brute captée est organisée puis encodée en représentations mentales.

Chaque niveau spécifie un cadre conceptuel (accompagné d'un vocabulaire spécifiant l'ontologie des phénomènes étudiés) permettant de comprendre puis d'expliquer le fonctionnement d'un système cognitif. L'entreprise scientifique achevée devra avoir produit une explication à chaque niveau, cernant ainsi toutes les subtilités ontologiques (au sens quinien d'ontologie) et causales des phénomènes mentaux. Voyons plus en détail en quoi consistent ces niveaux explicatifs et le particularisme des questionnements qui y trouveront réponse.

Débutons par le niveau le plus intuitif mais aussi le plus abstrait : le niveau sémantique. C'est en celui-ci que s'organise l'entreprise psychologique à proprement parler. À ce niveau, le psychologue spécifie la nature des états mentaux, leurs relations

avec les autres états mentaux ainsi qu'avec les événements perçus et les actions ou comportements qui en résultent *normalement*. Le niveau sémantique spécifie les normes des comportements rationnels, des perceptions justes et du fonctionnement adéquat des processus mentaux. Cet aspect normatif du niveau sémantique divise le fonctionnement 'rationnel' d'une créature cognitive de ses troubles de fonctionnement. Dennett (1971) nomme la perspective donnée par ce niveau la '*intentional stance*' et considère que les troubles cognitifs doivent être expliqués aux autres niveaux (ex : les erreurs, les illusions d'optique, les maladies mentales, etc.)[29].

Le titre 'sémantique' indique que le système cognitif en question est conçu comme un système manipulant des représentations ayant un contenu intentionnel (étant à propos de quelque chose dans le monde). Par exemple, « Paul aime les poires » est un énoncé prenant place à ce niveau puisque l'amour de Paul réfère à un type d'objet, c'est-à-dire que son goût a pour objet (intention) les poires. Il n'est pas question de comment son cerveau en vient à 'aimer les poires' ni quelles opérations formelles sont requises pour en arriver à une telle affection. Il est question des relations qu'un certain état mental entretient avec son environnement (ou d'autres états mentaux) : comment une représentation obtient son statut de représentation de quelque chose en particulier. Le tenant de la psychologie du sens commun explique les comportements à ce niveau en fonction des croyances et désirs d'un sujet rationnel (voir Pylyshyn, 1984, p.34-38 à ce sujet). P.M. Churchland & P.S. Churchland (1983) spécifie aussi que c'est à ce niveau qu'il faut déterminer comment une créature cognitive est connectée (*hooked-up*) au monde qui l'entoure, c'est-à-dire comment il obtient des représentations du monde qui l'entoure. Newell (1990, p.51), quoiqu'il nomme ce niveau '*knowledge-level*', propose un schéma illustrant le système cognitif tel que décrit à ce niveau. Le système cognitif réagit à son environnement en percevant les événements qui le composent et, à partir des représentations qu'il détient, agit en conséquence (selon les normes de la rationalité). C'est ce schème qui devra trouver explication au niveau sémantique.

Le psychologue précise donc le contenu intentionnel des représentations du système et propose une logique de leurs interrelations, permettant ainsi de dessiner un portrait général des relations causales entre les différents états mentaux et la transformation de l'état global du système cognitif. On y spécifie donc la fonction spécifique des

[29] Toutefois, voir Bermudez (2001) pour un exemple d'une caractérisation au niveau sémantique de certains troubles mentaux.

différents types d'états mentaux, c'est-à-dire leur rôle vis-à-vis les autres états mentaux ou vis-à-vis la créature et ses actions. La compréhension d'un système cognitif *qua* système psychologique ne peut se faire qu'à ce niveau, c'est-à-dire que les théories psychologiques utilisent les ressources fournies par ce niveau explicatif. On peut donc aussi nommer ce niveau « niveau psychologique » et on y retrouvera, pour le défenseur de la psychologie cognitive, les concepts mentaux empruntés à la psychologie du sens commun, tandis que pour l'éliminativiste les concepts utilisés à ce niveau seront à déterminer par voies empiriques (neuropsychologie).

Le niveau intermédiaire est le niveau 'computationnel'. Pylyshyn (1984) nomme celui-ci 'niveau syntaxique' et Marr (1982) 'niveau algorithmique', mais j'ai préféré 'niveau computationnel' puisque les expressions utilisées par Marr et Pylyshyn connotent une manipulation automatisée de symboles discrets, ce qui n'est pas le cas pour tous les processus computationnels (ex : connexionnisme) pouvant y trouver explication. C'est à ce niveau que sera spécifié l'architecture computationnelle : l'économie des processus par lesquels les différents états mentaux procèdent à leur fonction (décrite au niveau sémantique) est présentée formellement (selon un modèle computationnel) et on y détermine les règles de transformation de l'information pour réaliser la fonction *input-output* (ce qui est perçu/action produite) donnée au niveau supérieur. (P.M. Churchland & P.S. Churchland (1983, pp.8-13); Stich (1978, 1983)) Le psychologue devra, à ce niveau, spécifier par quelles opérations formelles le système cognitif procède à la computation des informations entrantes (sensorielles) et déjà contenues dans le système. Le scientifique devra produire un portrait des opérations ne faisant plus référence à leur fonction mais uniquement à leurs transactions informationnelles.

À ce niveau, les représentations du niveau sémantique sont présentes en tant que réalités formelles. Le contenu de ces représentations, l'aspect proprement sémantique, ne s'y trouve pas explicitement. Les représentations peuvent y être considérées comme des objets sur lesquels opéreront les mécanismes computationnels, mais la nature du contenu n'y jouera aucun rôle *qua* contenu. Ce dernier pourrait être identifié à la forme particulière de la représentation une fois formalisée, mais il faudra pour cela déterminer un guide de traduction entre les représentations formalisées au niveau computationnel et les représentations imbues de contenu sémantique aû niveau supérieur. Ce niveau représente la thèse que les processus cognitifs sont, en fait, des processus purement syntaxiques, c'est-à-

dire qu'ils manipulent des représentations en vertu de leur structure uniquement, aveugles à leur contenu sémantique. (P.M. Churchland & P.S. Churchland (1983); Stich (1978, 1983)) Le niveau computationnel ne pourra pas faire *sens du monde* puisque la computation n'illustre que le fonctionnement formel des états mentaux en faisant abstraction des réalités auxquelles elles réfèrent. Ce n'est pas à ce niveau non plus que les relations causales produisant les comportements et les changements dans les états mentaux seront spécifiés mais plutôt au niveau suivant, soit au niveau de l'implémentation matérielle.

Le niveau de l'implémentation matérielle[30] est le niveau le plus concret puisqu'il consiste à déterminer par quels mécanismes physiques (matériels) la computation sera réalisée. C'est à ce niveau que l'on spécifie quel substrat physique réalisera la computation - qui elle-même réalisera la fonction des états mentaux – et que le scientifique produira les explications concernant les interactions causales des mécanismes matériels réalisant les processus cognitifs. La description des opérations électroniques d'un microprocesseur se fait à ce niveau, tout comme celle du système nerveux central.

C'est ce niveau que les fonctionnalistes négligent pour l'explication psychologique complète parce qu'il y aurait une panoplie de moyens matériels de réaliser le substrat d'un système cognitif. Pour le théoricien intéressé à faire une description complète d'un système cognitif particulier, une caractérisation à ce niveau est nécessaire et impliquera les lois physiques pertinentes au phénomène matériel réalisant le système cognitif. Dans le cas où le système cognitif serait celui d'une créature naturelle terrestre, ce sera la neurobiologie (neurophysiologie, neurochimie, biologie cellulaire, etc.) qui offrira l'explication du système cognitif à ce niveau. Si l'on se restreint alors à une science cognitive concernant ces créatures (comme le voudrait le naturalisme de P.M. Churchland), on pourrait nommer ce niveau le niveau 'neurobiologique', mais cette appellation pourrait être malheureuse s'il fallait que l'argument de la réalisation multiple des états mentaux soit valide (la neurobiologie ne pourrait référer aux intelligences artificielles dont la théorie explicative à ce niveau serait plutôt l'électronique). Le choix de la théorie apte à prendre place au niveau de l'implémentation matérielle est largement dépendant de la stratégie explicative utilisée par le scientifique, ce dont il sera question à la section 2.2.

---

[30] Aussi connu sous le nom de 'niveau de réalisation matérielle', mais j'ai préféré le terme 'implémentation' puisque la relation de réalisation explicative prend place aussi entre les niveaux computationnel et sémantique (section 2.1.3).

2.1.2 Mise en application des trois niveaux explicatifs

Pour mieux illustrer la méthodologie explicative usant de la typologie tripartite susmentionnée, prenons un cas exemplaire d'un système très simple, soit le réseau de neurone calculant une fonction de conjonction logique :

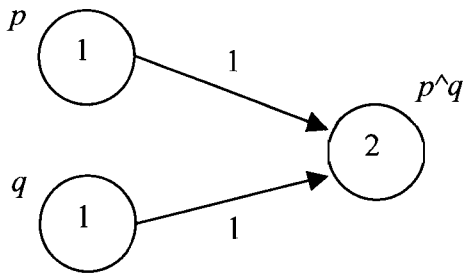

Table de vérité

p^q	q	$\neg q$
p	V	F
$\neg p$	F	F

Figure 1 - Réseau connexionniste de 'conjonction logique'

Au niveau sémantique, ce réseau est une instance d'un détecteur qui a pour fonction de déterminer si une paire de proposition (p,q), en fonction de leur valeur de vérité, sont en conjonction logique, i.e. que la proposition moléculaire (p^q) est vraie. Pour qu'une fonction soit une conjonction logique, elle doit répondre à la table de vérité qui est associée à ce connecteur logique (la conjonction est vraie si et seulement si les deux propositions qui la constituent sont eux-mêmes vrais). C'est la table de vérité qui détermine la sémantique des propositions car elle représente tous les mondes possibles dans lesquels existent p et q. Cette fonction est celle connue sous l'étiquette vernaculaire du 'ET'.

Au niveau computationnel, il faut utiliser la méthode PDP (*parallel-distributed processing* (McClelland & al., 1986)) de computation pour rendre compte du réseau de neurones[31]. Deux neurones d'*input* ont chacun une valeur d'activation dont le seuil est calibré à 1. Le neurone de sortie est calibré à 2. Chaque neurone d'entrée est connecté au neurone de sortie par une connexion synaptique dont la valeur d'excitation est de 1. Parce que ce réseau a été conçu pour instancier une fonction de conjonction, on sait que l'activation de chacun des neurones d'entrée représente une des deux propositions (p,q) et que l'activation du neurone de sortie représente la conjonction de ces deux propositions. Dans le cas où on l'ignorerait, ce serait la tâche du psychologue de déterminer (au niveau

[31] Pour une explication plus détaillée du fonctionnement des réseaux de neurones artificiels, voir l'analyse de la théorie neurocomputationnelle de P.M. Churchland au chapitre III section 3.3.2 ainsi qu'au chapitre IV section 4.2.

sémantique) quel est le contenu sémantique d'une activation des neurones d'entrée en spécifiant la fonction réalisée par le réseau et en l'identifiant à la définition fonctionnelle de la conjonction. Si la proposition *p* est vraie, le neurone correspondant s'activera; de même pour *q*. Si les deux neurones d'entrées s'activent (parce que les deux propositions sont vraies) alors la force d'excitation administrée au neurone de sortie sera suffisamment forte pour l'activer. Dans tous les autres cas, le neurone de sortie restera coi.

Au niveau d'implémentation, qui n'existe pas parmi ces pages puisque le réseau présenté en exemple n'a aucune force causale, on peut imaginer plusieurs substrats physiques exécutant le réseau de neurone. Un système de valves, un réseau de neurones biologiques, un réseau virtuel sur ordinateur, etc. Une fois le choix de réalisation physique fait, il restera à analyser quelles lois (ex : hydrologiques, biologiques ou électroniques) entreront en jeu pour prédire (et construire) le réseau de la conjonction.

2.1.3 Les niveaux explicatifs et le schème de la réduction

L'appellation de 'niveaux' pourrait sembler, jusqu'ici, maladroite : il a plutôt été question ici de perspectives différentes (*stances*) que de niveaux superposés. Les niveaux explicatifs ne forment pas une hiérarchie dont la base serait plus fondamentale et le sommet plus complexe et dépendant de ses piliers. Il y a une certaine structure dans les relations mais, d'un point de vue explicatif, celle-ci n'est pas nécessairement hiérarchique. Les philosophes et les psychologues invoquent souvent la nécessité d'approcher un phénomène cognitif par les trois niveaux ou, du moins, qu'une explication complète d'un phénomène cognitif requière une explication de celui-ci à chaque niveau. De ce point de vue, les trois niveaux sont liés par un critère pragmatique de l'explication et la prédominance de l'un sur l'autre dépend de la question posée par le scientifique. (Dennett, 1971) Par exemple, il serait possible de donner une description sémantique d'une illusion d'optique (ex : Muller-Lyer) en spécifiant que le segment de droite ressemblant à une double flèche semble plus court que l'autre mais l'explication du mécanisme cognitif responsable de l'illusion d'optique devra être spécifié à un autre niveau.

Les trois niveaux concernent différents aspects d'une même réalité et utilisent pour y arriver un différent vocabulaire théorique. Toutefois, ces trois perspectives, si elles sont concernées par une même classe de phénomènes, doivent être systématiquement reliées pour permettre d'unifier la description et l'explication des différentes facettes d'un même

processus cognitif. D'ailleurs, le fait même de parler d'un *même* phénomène cognitif malgré la divergence théorique des trois approches présuppose un principe d'unification. Il faut donc déterminer une relation entre ces trois perspectives. Un premier pas dans cette direction est fourni par la relation de réalisation. Étant donné la perspective épistémologique de notre analyse, il ne sera pas question des conditions ni des implications ontologiques de la relation de réalisation (*material realization*) entre divers niveaux d'organisations matérielles mais bien de son apport explicatif (*explicative realization*). (Wilson & Craver, 2007)

Il a été question de cette relation dans le premier chapitre. L'argument antiréductionniste de la réalisation multiple utilise explicitement cette relation. Toutefois, cet argument porte sur la réalisation matérielle des phénomènes mentaux : quels sont les substrats matériels qui sont impliqués dans l'histoire causale des états mentaux? Ce qui nous intéresse ici ce n'est pas simplement quelles structures physiques (neurobiologiques) sont impliquées dans le fonctionnement des états mentaux mais bien comment les processus neurobiologiques produits par ces structures (mécanismes) permettent de comprendre (expliquer) les processus mentaux. La relation de réduction se veut explicative et non pas purement descriptive. Comme Wilson & Craver (2007) le soulignent bien :

> « When attention shifts from describing to explaining, however, entities and their material realizers are no longer the primary focus. Instead, attention shifts to properties and activities as realized kinds and, correlatively, to component parts, their properties, their activities and their organization. Explanatory forms of realization are not just exhaustive lists of material constituents, but selective descriptions of the relevant parts for some explanatory purpose. » (Wilson & Craver, 2007, p. 88)[32]

C'est de ce sens dont nous nous enquerrons. Dans la perspective de P.M. Churchland, celle qui nous intéresse ici, la relation de réalisation explicative tiendrait tout de même entre le niveau sémantique et celui de l'implémentation matérielle. Plusieurs mécanismes physiques pourraient réaliser un même type d'état mental et l'exploration scientifique du fonctionnement de ces mécanismes nous aidera à comprendre les propriétés des états mentaux (ex : plasticité du cerveau vis-à-vis de l'apprentissage). Toutefois, il semble aussi exister une telle relation entre le niveau sémantique et le niveau computationnel. (Horgan, 1992; Horgan & Tienson, 1993; Marr, 1982) Il existe plusieurs

[32] Wilson & Craver (2007) dégagent plusieurs types de réalisations explicatives. Il sera question de celles pertinentes pour le projet de P.M. Churchland au chapitre III section 3.3.3 et au chapitre IV sections 4.2 et 4.3.

formats computationnels et plusieurs fonctions algorithmiques pouvant réaliser les mêmes fonctions *input/output*[33] entretenues par les états mentaux spécifiés au niveau sémantique. Cela indique donc que le choix d'une technique computationnelle pour formaliser les relations internes et externes des processus mentaux est relativement indépendant des entités et prédicats de la théorie psychologique au niveau sémantique. Il existe une panoplie de possibilités de réalisations des états mentaux sous forme computationnelle, ce qui n'impose pas *a priori* un schème computationnel pour une théorie au niveau sémantique (P.M. Churchland dans McCauley (1996, p.219-225)). Marr nous dit :

> « The choice of an algorithm is influenced for example, by what it has to do and by the hardware in which it must run. But there is a wide choice available at each level, and the explication of each level involves issues that are rather independent of the other two. » (Marr, 1982, p.23)

Toutefois, il ne semble pas qu'il y ait une réalisation multiple aussi libérale entre le niveau computationnel et le niveau d'implémentation physique. Un ordinateur de bureau peut être décrit selon une approche computationnelle de traitement sériel de symboles mais pas selon la perspective connexionniste. Inversement, un cerveau procéderait par la transformation de matrices d'activations neuronales[34] et, selon P.M. Churchland (1979) du moins, ne procéderait pas à un traitement sériel de symboles (voir Clark (1996); Dennett (1991); Densmore & Dennett (1999) pour un point de vue opposé). Il est important de ne pas confondre l'idée qu'il n'y a pas, dans certains cas donnés, une aussi grande liberté de manière de réaliser un complexe computationnel au niveau de l'implémentation physique avec l'idée qu'il serait impossible d'établir une équivalence entre deux modèles computationnels utilisant un mode de formalisation différent. Par exemple, un programme d'ordinateur pourrait simuler des réseaux virtuels de neurones artificiels (ce qui est en fait la technique utilisée pour étudier les propriétés de ces réseaux (McClelland et al., 1986). Toutefois, la réalisation physique de ce réseau virtuel est en fait une série d'états de l'ordinateur sériel qui en produit la simulation.

Le principe d'unification et la structure de la hiérarchie des niveaux explicatifs émergent lorsque l'on approche ceux-ci avec la relation de réalisation explicative et le schème réductionniste. Si la relation de réduction est une relation explicative entre les différents niveaux, elle pourra passer par une systématisation de la relation de réalisation

---

[33] Concernant les explications fonctionnelles, voir le prochain chapitre section 3.2.
[34] Voir la section 3.3.2 du prochain chapitre à ce sujet.

explicative entre les niveaux. C'est cette approche qui sera utilisée ici et elle consistera à organiser les niveaux explicatifs selon la hiérarchisation des théories scientifiques telle que conçue ordinairement dans le cadre réductionniste. Ces perspectives joueront alors le vrai rôle d'un niveau : les éléments au niveau de l'implémentation matérielle devront être spécifiés pour en dériver une explication des phénomènes au niveau computationnel et puis pour finalement produire une explication au niveau sémantique. C'est en définissant la relation de réalisation explicative que le réductionniste pourra déterminer quelles parties mécaniques et quelles propriétés neurobiologiques seront pertinentes pour offrir un portrait computationnel du système nerveux central. « The explanation neglects some properties of the material realizers and accentuates others. For explanatory realization (as opposed to material realization), it frequently does matter which parts one attends to : arbitrary parts will not be explanatorily relavant » (Wilson & Craver, 2007, pp. 88-89) De même lorsqu'une relation de réalisation sera recherchée entre le niveau computationnel et sémantique. Il sera d'autant plus évident aux chapitres suivants que P.M. Churchland fait de même lorsqu'il propose de réduire les théories psychocognitives aux théories neurocognitives et je tenterai de montrer que le modèle néo-réductionniste de P.M. Churchland cherche à systématiser les explications produites aux différents niveaux explicatifs et que les relations de réduction et de réalisation explicative se confondent. Si l'on peut spécifier une relation de réalisation explicative entre un mécanisme neurobiologique et un modèle computationnel, c'est qu'il existe une transformation de l'explication neurobiologique en explication neurocomputationnelle. Cela ne tient pas entre les relations de réalisation matérielle et de réduction, comme le montre l'argument de la réalisation (matérielle) multiple, mais cela ne concerne pas le projet de P.M. Churchland. Nous examinerons plus en détail comment P.M. Churchland défini les principes de réalisation explicative entre les différents niveaux au chapitre aux sections 3.3.3, 4.2 et 4.3.

Le niveau sémantique et le niveau de l'implémentation matérielle semblent donc être le lieu explicatif où prendront place, respectivement, la théorie psychologique et la théorie neurobiologique. (Newell, 1990, chapitre 2) Le cadre conceptuel emprunté à la psychologie du sens commun ne spécifie ni la manière dont les états mentaux sont réalisés par le cerveau ni quelles sont les transactions informationnelles permettant d'organiser les fonctions mentales. Par exemple, le concept de douleur n'indique en aucun cas comment les neurones doivent parvenir à détecter les agressions physiques faites au corps, ni par quel

processus informationnel un système cognitif en vient à détecter et différencier les différentes douleurs puis d'y réagir. Quant à elle, la neurobiologie est un ensemble théorique qui reste aveugle aux processus informationnels mais ignore aussi les différents états mentaux constituant la cognition de la créature étudiée. Le fonctionnement des neurones et leur typologie, la propagation des influx nerveux, le mécanisme de production des neurotransmetteurs, sont tous des exemples de théories constituant le corps de la neurobiologie. On pourrait agrémenter la théorie neurobiologique d'une théorie neurocomputationnelle, celle des PDP par exemple, mais cette dernière prendra place au niveau computationnel et pas au niveau de l'implémentation matérielle. De même, la neuropsychologie serait une théorie au niveau psychologique (sémantique) et, encore une fois, pas au niveau de l'implémentation matérielle.

On peut donc constater que le niveau sémantique est le siège explicatif de la théorie psychologique alors que le niveau d'implémentation physique est utilisé pour spécifier le fonctionnement du matériel réalisant la cognition. Dans l'optique du réductionnisme, il semble donc que la relation de réduction doive tenir entre ces deux niveaux : comment traduira-t-on le niveau sémantique (les prédicats et les régularités constituant la théorie psychologique) en son image au niveau de l'implémentation matérielle (comment le système nerveux central en vient à produire les états mentaux spécifiés au niveau sémantique)? La réponse que donne P.M. Churchland à cette question consiste à utiliser le niveau computationnel comme pont entre les deux théories. Les réseaux de neurones, par leurs capacités computationnelles, instancieraient des théories. P.M. Churchland (1989, chapitre 5) propose un schème permettant de réduire les phénomènes du niveau sémantique au niveau computationnel, puis au niveau de l'implémentation matérielle (neurobiologie). Il en sera question dans les chapitres suivants.

Pour l'instant, il sera question des contraintes que les niveaux exercent les uns sur les autres et on verra que les sciences cognitives sont fondées sur la thèse que les phénomènes cognitifs peuvent être décrits et expliqués *qua* systèmes computationnels et que certaines machines matérielles peuvent réaliser des systèmes computationnels (ex : machine de von Neumann, système nerveux central, etc.). L'idée consiste à concevoir l'élaboration du niveau computationnel comme une traduction méthodique de l'un des deux autres niveaux et, en fonction de la méthodologie explicative choisie, de déterminer quel niveau devra être traduit et les propriétés de cette relation de traduction (ce qui, pour P.M.

Churchland, est la porte d'entrée de la relation de réduction). (P.M. Churchland, 1979, section 11; 1989; P.M. Churchland & P.S. Churchland, 1983)

2.1.4 Relations entre les niveaux

Le niveau computationnel a été décrit comme une perspective explicative basée sur la notion de manipulation et de transformation de l'information. Les sciences cognitives ont introduites l'analogie du cerveau/ordinateur : le cerveau serait comme un ordinateur programmé pour effectuer les opérations cognitives constitutives des processus mentaux. Ces procédés font appel aux ressources fournies par deux théories : la théorie de l'information (originellement développée par Shannon (1948)) et la théorie de la computation (originellement développée par Church (1932) et Turing (1936)). L'apparition de la théorie de la computation (machine de Turing universelle) a permis de produire cette analogie et d'en obtenir une formalisation plus systématique. (voir Gardner, 1985) Toutefois, il est important de noter que la stratégie la plus utilisée dans le domaine de la psychologie cognitive et de l'intelligence artificielle est celle du traitement syntaxique de symboles (Fodor, 1975, 1987; Newell 1980, 1990; Newell & Simon, 1972; Newell et al., 1989; Rey, 1997).

Toutefois, dans l'optique d'une réduction des processus mentaux aux processus neuronaux, ce type d'approche ne semble pas indiquée : le cerveau n'est pas un ordinateur sériel mais parallèle et il ne semble pas y avoir de symboles discrets dans le cerveau (revoir, à ce sujet, la section 1.3.3 du présent mémoire ainsi que P.M. Churchland (1989, chapitre 1 et 4; 2007, chapitre 2); P.M. Churchland & P.S. Churchland, 1983)). Selon P.M. Churchland (1989, chapitres 5, 6), l'approche la plus pertinente dans cette optique serait l'approche dynamiste des réseaux de neurones artificiels (McClelland & al., 1986; Horgan & Tienson, 1992a; van Gelder, 1998; van Leeuwen, 2005). Il en sera question plus en détail aux chapitres suivants.

La réduction de la psychologie cognitive aux neurosciences devra donc passer par une explication computationnelle. Pour mieux cerner comment intégrer cette répartition tripartite au tableau du schème de la réduction, il est en ordre d'examiner les contraintes que les différents niveaux explicatifs jouent les uns sur les autres.

Le choix de la méthode de computation sera sujet à plusieurs contraintes imposées par le niveau sémantique ainsi que par le niveau d'implémentation matérielle. Le temps de

réaction d'un sujet lors d'une tâche, réalité du niveau psychologique, impose une contrainte temporelle au traitement de l'information par le système computationnel, aussi faudra-t-il choisir une forme de computation suffisamment performante pour répondre à cette contrainte (Rey, 1997). Il est important de noter que l'aspect proprement psychologique du système a disparu à ce niveau. On ne parle plus de croyance et de désirs mais de symboles ou de fonctions mathématiques de transformations. (Horgan & Tienson, 1993; Marr, 1982) Le vocabulaire du niveau computationnel est aussi bon pour décrire le programme d'un logiciel que les opérations formelles des états mentaux. Il n'y a rien d'intrinsèquement psychologique à ce niveau : l'aspect proprement mentaliste disparaît pour laisser place à une construction mathématico-logique. Dans cette construction, il n'y a pas de restrictions sur la forme utilisée (réalisation multiple des états mentaux au niveau computationnel) : une approche algorithmique basée sur la manipulation syntaxique de symboles (Newell, 1980) ou une approche connexionniste PDP transformant des vecteurs d'activations (P.M. Churchland, 1989a; McClelland et al., 1986) n'ont pas, à cet égard, aucune priorité intrinsèque, bien que la communauté des sciences cognitives ait favorisée la première option.

Une autre contrainte du niveau sémantique sur le niveau computationnel est reflétée par le débat de la nature du contenu sémantique des représentations mentales. Putnam (1975a, 1988) insiste sur le fait que le contenu sémantique des représentations ne peut être donné purement en vertu d'une théorie psychologique mais nécessite aussi une histoire causale transcendant le simple système cognitif étudié. C'est ce qui est entendu par la notion de '*wide content*' (Burge (1979, 1983)). Fodor (1975) a développé la thèse du langage de l'esprit pour restreindre l'aspect sémantique des représentations mentales à la description du système cognitif. Cette thèse stipule que le contenu des représentations est purement affaire de l'état interne du système cognitif (*narrow content*) et requière donc un solipsisme méthodologique (un système cognitif pourrait être complètement expliqué sans faire référence à l'environnement dans lequel il se trouve). (Fodor, 1987) Cette problématique ressurgie dans la conception de P.M. Churchland vis-à-vis le contenu des représentations et sera donc traitée plus à fond à la section 4.2.1.

Le niveau inférieur impose lui-aussi des contraintes sur le niveau computationnel. La thèse du langage de l'esprit (Fodor, 1975, 1987) est une théorie concernant la nature des processus computationnels et, selon P.M. Churchland (1981/1989) et P.M. Churchland &

P.S. Churchland (1983), la possibilité d'un langage interne semble peu probable étant donné l'absence flagrante de langage chez la majorité des espèces non-humaines. De plus, l'idée d'une computation utilisant des symboles discrets impliquerait des symboles physiques, or le cerveau n'étant qu'un large réseau de neurones (niveau d'implémentation), une formalisation des réseaux de neurones (PDP, donc au niveau computationnel) sera plus réaliste et il n'y a pas de possibilité de retrouver les symboles et propositions postulées par les attitudes propositionnelles. (P.M. Churchland & P.S. Churchland, 1983)

Beaucoup de contraintes seront tirées de clauses normatives imposées par la communauté scientifique (Newell, 1990). Une clause de réalisme peut contraindre le niveau computationnel sur divers plans. Prenons, par exemple, la contrainte temporelle du niveau sémantique. Celle-ci n'a d'importance que si le scientifique cherche à respecter une clause de réalisme dans la performance du système. Imaginons qu'un être humain moyen prend deux secondes à reconnaître une forme géométrique qui lui est présentée. Formellement, un système qui réaliserait cette fonction cognitive en trois semaines mais atteindrait les mêmes résultats qu'une autre qui le ferait en deux secondes seraient computationnellement équivalents : ils exécutent, chacun à leur manière, la même fonction; mais le premier système est moins intéressant parce que moins réaliste. Cette contrainte de 'réalisme' peut provenir des deux autres niveaux. D'abord, à la lumière de la théorie de l'évolution, les computations doivent fournir des résultats au moins dans le temps de vie de la créature, voir même dans un temps suffisamment restreint pour assurer la survie de la créature (on voit mal un système perceptuel prendre deux jours pour reconnaître un prédateur menaçant bondissant sur nous). Une seconde provenance de la clause de temporalité peut parvenir du niveau sémantique : le processus computationnel doit parvenir à computer la fonction cognitive tout en permettant de prédire le temps réel de calcul d'un sujet observé. (Newell, 1990)

Toutefois, ces contraintes normatives ne sont pas suffisantes pour déterminer quelle fonction abstraite pourra réaliser un état mental particulier. Leur rôle consiste à discriminer entre des modèles computationnels parvenant à computer la même fonction. Il sera à nouveau question de ces clauses normatives lors de la présentation, au chapitre IV, de l'approche neurocognitive de P.M. Churchland.

Ce portrait semble organiser les contraintes autour du niveau computationnel sans que ce dernier n'affecte réellement les deux autres niveaux. Ce n'est pas un accident : ces

relations générales de contraintes (ex : réalisme) font état de consensus dans la communauté des sciences cognitives (Reed, 1999), bien que dans le détail chacun ne soit pas tout à fait en accord avec l'importance à accorder à celles-ci. Toutefois, il existe d'autres relations de contraintes, certaines provenant du niveau computationnel. Celles-ci sont méthodologiques, liées au procédé de construction (dans le temps) d'explications à différents niveaux des phénomènes cognitifs. Ces contraintes méthodologiques ne font pas objet de consensus et dépendent de la stratégie explicative utilisée par le scientifique, ce dont il sera question à présent.

2.2 Stratégies explicatives

Il a été question, jusqu'à présent, exclusivement du type d'explication à chacun des niveaux et des relations qui les unifient (relation de réalisation explicative ainsi que des contraintes qu'ils exercent les uns sur les autres). Ce portrait s'est dessiné de manière suffisamment générale pour faire abstraction du statut actuel des sciences cognitives : celles-ci sont incomplètes et sont toujours en processus d'avancement des connaissances. Ces niveaux explicatifs jouent autant un rôle dans le procédé d'explication que dans celui de l'élaboration et la découverte des théories explicatives des phénomènes cognitifs. Toutefois, ils ne peuvent pas, en eux-mêmes, fournir les principes dirigeant l'organisation du projet scientifique des sciences cognitives ni non plus de méthodologie pour assurer la découverte des théories aux différents niveaux. Une stratégie explicative doit être fournie, c'est-à-dire qu'il faut déterminer la méthode (dont l'exécution prendra place dans le temps) pour permettre une mise en relation systématique des trois niveaux explicatifs. Le débat entre les tenants du cadre conceptuel fourni par la psychologie du sens commun et les éliminativistes illustrent bien ce en quoi consiste une stratégie explicative. Pour le fonctionnaliste, la psychologie du sens commun pourvoit un cadre conceptuel au niveau sémantique et c'est à partir de ce cadre conceptuel que la formalisation computationnelle devra prendre place. Quant à l'éliminativiste, il lui faudra construire le détail du niveau computationnel puis celui du niveau sémantique en fonction des découvertes sur la structure du système nerveux chez une créature donnée (niveau de l'implémentation matérielle).

Une stratégie explicative attribue donc un rôle méthodologique différent à chacun des niveaux explicatifs. Une panoplie de stratégies explicatives a été proposée au travers la littérature mais celles-ci peuvent être divisées en quelques groupes types. Les deux stratégies explicatives qui nous intéresseront ici sont celles proposées par Paul M. Churchland (la méthode *bottom-up*) et celle prescrite par les fonctionnalistes antiréductionnistes (la méthode *top-down*).

2.2.1 La Stratégie Top-Down

Cette stratégie est nommée *top-down* parce qu'elle procède du niveau le plus abstrait, le niveau sémantique (ou psychologique), pour élaborer ensuite les explications au niveau computationnel. En définissant d'abord les fonctions mentales par leurs interrelations causales et les contraintes de manifestation (ex : temps de réponse) et ce à partir du cadre conceptuel fourni par la théorie psychologique du sens commun, le psychologue fonctionnaliste (aidé par les chercheurs des autres disciplines des sciences cognitives) produira par la suite un portrait computationnel des fonctions mentales déjà définies. Le niveau de l'implémentation matérielle ne prend qu'une place très périphérique : les fonctionnalistes adeptes de l'argument de la réalisation multiple relèguent la question de l'implémentation matérielle et des problèmes techniques qui y sont liés à d'autres disciplines (ex : psychiatrie). Une fois le modèle computationnel des fonctions cognitives prédéfinies (issues du cadre conceptuel de la psychologie du sens commun), le psychologue cherchera à déterminer le particularisme psychologique de ses sujets (par la psychologie humaine, l'éthologie cognitive, robopsychologie, etc.).

Il faut distinguer deux gammes de stratégies *top-down*. La première est celle du 'fonctionnalisme *a priori*' (Armstrong, 1968; Jackson & Pettit, 1993; Lewis, 1972; McGinn, 1991) pour lequel les concepts mentaux découverts *a priori* (soit les concepts classiques empruntés à la psychologie du sens commun) sont confirmés par leur origine introspective immédiate. Cette stratégie est imperturbablement *top-down* puisque toute l'investigation scientifique qui suivra dépendra du cadre conceptuel élaboré au niveau sémantique (psychologique). Une alternative nommée 'psychofonctionnalisme' (Block, 1980, p.272) permet un certain révisionnisme des concepts d'états mentaux et des relations fonctionnelles qu'ils entretiennent entre eux à la lumière des découvertes empiriques faites en psychologie. Cette stratégie débute elle aussi son investigation avec les concepts donnés

par la psychologie du sens commun mais, explorant empiriquement leur fonctionnement et dérivant des manifestations comportementales certaines propriétés de ces processus cognitifs, elle permet une altération de nos préjugés à leur propos, altération à l'écoute des découvertes empiriques faites par le psychologue de laboratoire. Ce psychofonctionnalisme défend la thèse selon laquelle l'analyse fonctionnelle adéquate d'une fonction mentale sera celle faite par la psychologie achevée. (Fodor, 1968) Toutefois, ces altérations au cadre conceptuel de base se feront de manière méliorative : les concepts visent déjà dans la bonne direction, il n'est question que d'ajuster plus précisément le tir.

Si l'on retourne à notre réseau de neurones représentant une conjonction, le psychologue débuterait son investigation en déterminant les critères spécifiques à la conjonction logique (en spécifiant sa table de vérité, les *inputs* propositionnels). En second lieu, le psychologue invoquera le modèle computationnel des réseaux de neurones (selon l'exemple, mais une machine de Turing pourrait tout aussi bien être choisie) et cherchera à formaliser la conjonction à l'aide de celui-ci. En spécifiant les conditions de la conjonction au niveau sémantique, le psychologue sait que l'entrée sensorielle doit être minimalement binaire (la valeur de vérité de *p*, la valeur de vérité de *q*), et la sortie unique (la valeur de vérité de $p \wedge q$). Il organisera les liens entre neurones de manière à ce que les conditions spécifiées par la table de vérité soient toujours respectées. Finalement, le psychologue (ou son technicien) déterminera par quel moyen physique il devrait implémenter ce réseau.

Cette stratégie explicative sera étudiée plus en détail au prochain chapitre et nous en dégagerons un type d'explication qui permettra d'organiser les explications aux différents niveaux explicatifs dans une même structure. Ce type d'explication est l'explication fonctionnelle (section 3.2).

2.2.2 La Stratégie Bottom-Up

La critique de Paul M. Churchland défendue dans son *Scientific Realism and the Plasticity of Mind* (particulièrement au chapitre 4) et dans P.M. Churchland (1981/1989) vise directement la stratégie *top-down* en indiquant qu'elle présuppose une valeur aux concepts empruntés à la psychologie du sens commun (la *p-theory,* P.M. Churchland (1979)) et que toute théorie devrait d'abord trouver une confirmation par les sciences empiriques plutôt que dans des préjugés introspectifs *a priori*. La stratégie *top-down* ne peut être valable en science puisqu'elle prend son point de départ dans une théorie en mal

de confirmation empirique. La voie que propose P.M. Churchland pour déterminer quels types d'états mentaux seront présent dans une psychologie achevée consiste à débuter les recherches au niveau d'implémentation physique, stratégie typiquement *bottom-up*.

La stratégie *bottom-up* débute par l'étude de la réalité matérielle du mental et rejette donc implicitement la validité d'une science psychologique autonome, déconnectée des sciences physiques et biologiques. Cela mène le scientifique à fonder ses recherches sur les sciences biologiques, particulièrement les neurosciences (qui sont *empiriquement construites*, on pense notamment aux découvertes de Ramon y Cajal, Gogli, Broca, etc.). L'étude du cerveau révélera les structures par lesquelles il y a computation (actuellement représentée le plus fortement par les théories PDP). Finalement, une interprétation sémantique (psychologique) de ces processus permettra de fonder une psychologie au cadre conceptuel empirique.

Encore une fois, tout comme la stratégie *top-down* rigide a une alternative moins radicale, il existe une stratégie *bottom-up* plus flexible, défendue par Patricia S. Churchland (1986, p.362-376). Cette stratégie cherche à explorer le cerveau et à localiser les fonctions psychologiques pour ainsi plus aisément interpréter la machinerie neuronale. Le neuroscientifique utilise les concepts de la psychologie du sens commun et cherche à déterminer la localisation de ces fonctions dans le système nerveux et, à partir de cela, débute son analyse interprétative de la structure de la zone fonctionnelle pour déterminer comment l'organisation neuronale à cet endroit permet l'actualisation de la fonction mentale. (P.S. Churchland & Sejnowski, 1982) Le révisionnisme des concepts mentaux est permis puisque la structure peut (et doit, selon Patricia S. Churchland) nous informer sur les nuances computationnelles de la fonction étudiée. Éventuellement, la psychologie devra se fondre dans une neuropsychologie achevée. Cette stratégie est fortement controversée, entre autre parce qu'elle cherche à concilier les deux stratégies explicatives présentées ici. (voir (P.S. Churchland, 1989; P.S. Churchland & Sejnowski, 1992; Kosslyn, 1997; von Eckardt Klein, 1978)).

L'étude cervicale n'est pas le seul point de départ de la stratégie *bottom-up*. La théorie de l'évolution par sélection naturelle impose une série de contraintes orientant celle-ci. Dans les recherches neurologiques, l'utilisation des primates et, dans une moindre mesure, des autres créatures dotées d'un système neuronal, est basée sur l'idée d'une certaine continuité phylogénétique dans l'anatomie et la physiologie des créatures

apparentées, thèse néodarwiniste. (voir Hardcastle, 2007, pp. 299-303) D'autres contraintes proviennent de la théorie de l'évolution. Par exemple, il serait pertinent de rejeter une thèse voulant qu'un module mental spécialisé dans la lecture et l'écriture existe et soit à découvrir puisque la thèse est phylogénétiquement improbable (ce serait un module trop récent et peu pertinent pour affecter la survie naturelle). Cela pourrait pousser les neuroscientifiques à postuler qu'il y a là, en quelques sortes, une exaptation (ou récupération d'une structure déjà présente) et que la connaissance de la fonction d'origine peut être informative sur sa nouvelle utilisation. Plusieurs philosophes (P.M. Churchland & P.S. Churchland (1983); Dennett (1983, 1987, 1998a)) ont critiqué l'absence flagrante de ces contraintes (adaptatives) de bas niveau en psychologie cognitive alors que d'autres en ont critiqué l'abus (voir Buller (2005) par exemple, et indirectement Gould & Lewontin (1979)).

Le cadre conceptuel de base pour permettre une mise en relation des théories psychocognitives et neurocognitives a été suffisamment élaboré pour permettre une analyse directe de chacun des niveaux dans une perspective réductionniste. Nous verrons dans la section 4.1 que la stratégie *top-down* des fonctionnalistes permettra de formaliser les phénomènes cognitifs constituant la faune des théories psychocognitives en fonctions computationnelles et que cela permettra un rapprochement entre celles-ci et les fonctions computationnelles découvertes dans le système nerveux humain (ou animal). Le prochain chapitre dégagera des théories psychocognitives (niveau sémantique) un type d'explication - l'explication fonctionnelle - et il sera montré que ce format explicatif peut être utilisé aux trois niveaux explicatifs, ce qui permettra d'assurer un premier pas vers l'isomorphisme des théories psychocognitives aux théories neurocognitives. Suivant la stratégie explicative *bottom-up* de P.M. Churchland, il sera question à la fin du chapitre III (section 3.3) de l'isomorphisme explicatif entre les niveaux neurobiologique et computationnel. Au chapitre IV, les niveaux computationnel et sémantique trouveront aussi une telle mise-en-relation et nous serons en mesure de déterminer l'état actuel des sciences cognitives ainsi que les lacunes concomitantes dont souffre la perspective réductionniste de P.M. Churchland.

Chapitre III – Isomorphisme fonctionnel

Il a été vu au chapitre II que la structure des niveaux explicatifs permet de mieux situer le rôle de la réduction interthéorique proposée par P.M. Churchland (section 1.3) : celle-ci devra permettre une liaison systématique des processus et entités postulés à un niveau et les lier à d'autres du niveau inférieur. Tout ce qui figurera au niveau sémantique devra pouvoir être rendu au niveau computationnel, puis au niveau neurobiologique[35]. Il est désormais possible d'amorcer la construction d'une image de la psychocognitive dans le vocabulaire des neurosciences. Cela devra se faire en assurant un isomorphisme dans la structure explicative utilisée par chacune de ces entreprises scientifiques (voir section 1.3.1., plus particulièrement la page 20). Un outil est disponible pour faciliter cette tâche : il existe un format de l'explication générale en sciences cognitives, l'explication fonctionnelle, qui semble à propos pour permettre de structurer l'isomorphisme entre les explications typiquement psychologiques (théorie réduite ou T_r) et leur potentiel reflet (image de la théorie réduite dans le vocabulaire de la théorie réductrice ou I_b) dans un schème d'explications neuroscientifiques (théorie réductrice ou T_b).

> « In order for a theory to count as functionalist, it is necessary only that it claims that token psychological states are to be identified with token physical states under *some* scheme for identification, so long as that scheme identifies the psychological states in a vocabulary by making essential reference to their role in the psychology of their bearers, where that role is characterized by reference to their relations to other such states, to processes operating over them, and to functionally characterized inputs and outputs. » (Garfield, 1988, p.37)[36]

L'utilisation de l'explication fonctionnelle semble adéquate à la fois au niveau sémantique, computationnel et de l'implémentation matérielle. Aux trois niveaux il est des plus pertinents d'expliquer le déroulement des processus qui y sont observés par un schème fonctionnel spécifiant les entrées sensorielles et l'état des processus cognitifs pour déterminer quelles réactions comportementales prendront place.

---

[35] Il ne sera pas question des processus mentaux (niveau sémantique) qui ne seraient peut-être pas proprement computationnels (ex : émotions). Toutefois, ceux-ci devront tout au moins trouver réalisation au niveau neurobiologique et nécessiteront alors une relation de réalisation liant le niveau sémantique au niveau neurobiologique.

[36] Bien que la citation précédente suggère le nom « fonctionnaliste » pour une théorie répondant à ce format explicatif, pour ne pas brouiller le propos le terme sera employé uniquement pour faire référence au camp de penseurs en faveur de l'autonomie de la psychologie. Donc lire : « For a theory to count as [a functional explanation]... ».

C'est ce dont il sera question dans ce chapitre. Il sera tout d'abord question de déterminer la nature de l'explication fonctionnelle en spécifiant d'abord son origine (point de vue fonctionnaliste en psychologie, section 3.1), puis de préciser son format explicatif (section 3.2). Il sera aussi montré comment ce type d'explication peut être inséré dans le schème des niveaux explicatifs introduit au chapitre précédent et ainsi de former une structure explicative permettant de répondre à la demande d'un isomorphisme des théories en relation de réduction interthéorique telle que la propose P.M. Churchland (section 3.2). Une fois ce format explicatif bien inséré dans le cadre conceptuel développé dans les chapitres précédents, ce chapitre se clora sur une première mise en application de la relation de réduction interthéorique en utilisation la relation de réalisation explicative (introduite à la section 2.1.3) pour lier les explications fonctionnelles du niveau computationnel au niveau de l'implémentation matérielle (voir section 3.3.).

Le choix d'analyser ces deux niveaux avant de poursuivre vers le niveau sémantique est justifié par deux raisons. Premièrement, il est question d'être cohérent avec la stratégie explicative prescrite par P.M. Churchland : son schème méthodologique étant *bottom-up*, l'enquête réductionniste devra débuter en déterminant les relations entre le niveau neurobiologique (implémentation matérielle) et les processus neurocomputationnels qui en résultent. En second lieu, la relation de réalisation explicative entre les niveaux sémantique et computationnel nécessitera une double analyse devant à la fois permettre de déterminer (a) comment construire le niveau sémantique (neuropsychologie) à partir des deux autres niveaux inférieurs et (b) comment évaluer si la structure de la psychologie cognitive peut trouver confirmation, révision ou élimination à partir du schème neuropsychologique. Étant donné que P.M. Churchland s'est particulièrement penché sur la relation de réalisation explicative entre les niveaux computationnels et neurobiologiques, cela nous donnera les outils pour mieux spécifier la nature de cette relation vis-à-vis le niveau sémantique.

3.1 L'approche fonctionnaliste de la psychologie cognitive

Avant d'entrer dans une analyse du format fonctionnel de l'explication, il est important de différencier la doctrine fonctionnaliste comme point de vue en philosophie des

sciences cognitives de l'explication fonctionnelle *per se*. La théorie fonctionnaliste de l'explication des phénomènes mentaux prend racine dans une critique adressée au programme béhavioriste développé au milieu du XX^{ème} siècle. Les béhavioristes rejetaient de leur ontologie les états mentaux. Les créatures pouvant réagir aux stimulations fournies par l'environnement étaient dotées de régularités dispositionnelles : lorsque l'environnement fournit au système psychologique certaines stimulations sensorielles, la créature était disposée à réagir (par comportements observables) selon certaines régularités. Ces régularités devaient être relevées par le théoricien et faire état de lois psychologiques dans la théorie béhavioriste. L'explication béhavioriste consistait en un schème nomologique liant les stimulations à certains comportements types (en fonctions de clauses *ceteris paribus*). (Malcolm, 1968; Ryle, 1949; Skinner, 1957; Watson, 1913, 1930) Fortement influencé par la pensée réductionniste de l'unification des sciences du Cercle de Vienne, les béhavioristes se proposaient de construire une science des comportements faisant fi des concepts provenant d'un vocabulaire mentaliste (Carnap, 1933; Chisholm, 1957; Hempel, 1949; Ryle, 1949). Ce vocabulaire devait être remplacé par une liste de propensions comportementales stéréotypées (quelles classes de comportements résultaient, selon une relation nomologique, de certaines classes de stimulations)et ainsi former une science objective (le mental ne pouvant être observé *per se*). (Zuriff, 1985)

Alliées au développement des sciences cognitives, plusieurs critiques philosophiques ont ravagées le programme béhavioriste, ce qui a mené à son remplacement par l'entreprise fonctionnaliste[37]. (Gardner, 1985) Relevant le fait que les clauses *ceteris paribus* utilisées par les chercheurs béhavioristes présupposent toujours des états mentaux (utilisant, par exemple, l'idée qu'un acteur pouvait témoigner de comportements sans recevoir les stimuli appropriés (colère actée) ou celle d'un hypothétique 'super-spartiate' qui savait les refouler malgré la présence des stimuli correspondants - un bon spartiate doit sembler insensible à la douleur - et bien d'autres exemples encore), les critiques ont mis à mort l'option béhavioriste comme science psychologique. (Putnam, 1963, 1967b; Chomsky, 1959)

L'approche fonctionnaliste des sciences cognitives s'est ainsi développée en réintégrant dans l'explication psychologique un vocabulaire mentaliste. Les états mentaux

[37] Il existe plusieurs familles de fonctionnalisme (certaines seront évoquées plus loin), mais leurs principaux aspects sont recoupés par l'analyse de ce chapitre et, à moins d'être explicitement évoquées, le terme fonctionnaliste doit être entendu comme tout fonctionnalisme. Voir Rey (1997, chapitres 6 et 7 à ce sujet).

devaient dorénavant jouer le rôle d'intermédiaires entre les entrées sensorielles et les sorties comportementales, c'est-à-dire que l'explication psychologique devait faire référence à une 'boîte noire'[38] transformant les *inputs* sensoriels en *outputs* comportementaux, boîte noire constituée d'une interaction de divers états mentaux. Il n'était plus question de se limiter aux régularités observables mais de produire des théories sur les mécanismes cognitifs sous-jacents à ces régularités. (Putnam, 1967b; Fodor, 1968) Les sciences cognitives ont droit à un type d'objet d'étude fort particulier : des systèmes complexes dont les mécanismes internes sont cachés à l'œil de l'observateur extérieur. En ce sens, le scientifique qui étudie un aspect des êtres vivants procède alors comme un ingénieur qui cherche à comprendre la nature du fonctionnement et l'utilité (la fonction[39]) d'un artefact ingénieux qu'on lui présente. (Dennett, 1994; 1995, pp.212-220) Cette stratégie est nommée '*reverse engineering*' : plutôt que de procéder à la construction d'une solution optimale pour un problème déterminé par une série de contraintes (ce qui serait le '*forward engineering*'), l'ingénieur devra approcher l'objet de son étude comme la solution à un problème d'ingénierie et devra alors déterminer les raisons du 'choix' (ou de la rétention) d'un *design* plutôt que d'un autre. Se basant d'abord sur le problème pour lequel l'objet a été créé et la tâche apparente à accomplir par l'artefact (naturel), le chercheur tentera lui-même de trouver une solution optimale en fonction des contraintes ayant délimitées les possibilités de solutions et cherchera à analyser l'objet à la lumière de celles-ci. (Dennett, 1990, 1994) On peut donc reformuler l'objectif scientifique des sciences cognitives comme une tentative de résoudre les problèmes de '*reverse engineering*' spécifiquement liés à l'appareillage cognitif. Ainsi, adopter la stratégie du *reverse engineering* en psychologie cognitive consiste à donner une importance primordiale à deux questions scientifiques. Le psychologue cherche à déterminer quels sont les effets des mécanismes cognitifs, spécifiant ainsi la classe des comportements intéressants dont il faudra expliquer la production. Puis le psychologue fonctionnaliste cherchera à déterminer ce qui se déroule dans la 'boîte noire' pour en dégager les états mentaux jouant un rôle causal pour la production des comportements. (Lycan, 1987)

---

[38] L'origine du terme proviendrait du Hixon Symposium (Jeffress, 1954), acte de naissance des sciences cognitives (Gardner, 1985).

[39] Ici utilisé au sens plus téléologique, le terme fonction, par ailleurs (à moins d'indications contraires), est entendu dans sa forme logique : une fonction est une mise en relation systématique d'éléments d'un domaine vers une image.

Les processus mentaux prenant place dans cette boîte noire seraient affectés par les entrées sensorielles mais aussi par les autres processus mentaux s'y déroulant, pour éventuellement produire un comportement externe et observable. De ce point de vue, un état mental doit être défini selon la fonction qu'il réalise, soit par la description complète des effets que causent sur lui les entrées sensorielles, des relations causales qu'il entretient avec les autres états mentaux et finalement comment toute cette interaction causale permet la manifestation de certains comportements. (Putnam, 1963, 1967a, 1967b; Fodor, 1975; mais aussi Dennett, 1975; Lycan, 1981, 1987) Pour le fonctionnaliste, un état ou processus mental est une fonction, c'est-à-dire qu'il est défini par le rôle qu'il joue (les mises en relations causales qu'il permet) au sein du système cognitif. Un état mental ne peut pas être défini sans référence au système qui l'intègre et en exploite les produits. Le fonctionnaliste rejette donc l'approche nomologique simpliste des béhavioristes pour rendre compte du fonctionnement du mental et favorise plutôt une approche analytique, décomposant le système cognitif en différentes fonctions constitutives.

Le fonctionnaliste a donc trois classes de phénomènes à déterminer : la nature des entrées sensorielles, le fonctionnement des états mentaux et la nature des comportements pouvant être produits. L'explication psychologique, une fois la science achevée, devra contenir dans son *explanans* une spécification des entrées sensorielles à un moment *t* ainsi qu'une spécification de l'état global des états mentaux au même moment pour en dériver l'*explanandum*, soit le comportement produit après *t*. L'objet de l'explication psychologique (pour une psychologie achevée) n'est donc pas le fonctionnement du mental *per se*, mais plutôt les occurrences des comportements produits par un système cognitif relativement aux stimulations qui l'affectent. Toutefois, la recherche en sciences cognitives doit, pour parvenir à élaborer des explications psychologiques, fournir une image du mécanisme des fonctions cognitives permettant de lier stimulations sensorielles et comportements. Comprendre les mécanismes sous-jacents aux régularités comportementales est l'objectif de *recherche* des sciences cognitives. (Gardner, 1985) L'analyse fonctionnelle d'un système cognitif, particulièrement celle proposée par Cummins (1983), sera traitée plus à fond lors de la discussion de la théorie psychocognitive au prochain chapitre.

3.2 L'explication fonctionnelle

Il est possible de soustraire de la position fonctionnaliste un format explicatif qui sera identifié ici à l'expression 'explication fonctionnelle'. Le format est très simple et intuitif : une explication fonctionnelle est une mise en relation systématique (une fonction) d'un domaine spécifiant une classe d'entrées (*input*) affectant un système à une image constituée d'une classe de sorties (*output*) produites en réaction à ces entrées. (Putnam, 1967a; Fodor, 1975; Lycan, 1981) Le format est le suivant : <<entrée, état initial>, <sortie, état final>>[40]. Tout mécanisme reliant systématiquement une classe donnée d'*inputs* à une classe donnée d'*outputs* aura une fonction du même type. Ce format d'explication est suffisamment vague pour ne pas être exclusif aux sciences cognitives. Un des exemples classiques utilisés pour expliquer cette notion est celui du piège à souris dont l'*input* est une souris vivante, l'*output* une souris morte, et la fonction serait celle de 'piège à souris'. La biologie utilise aussi ce type d'explication. (Sober, 1985) Le propre de l'explication fonctionnelle en sciences cognitives sera défini par le détail de la classe générale de tous les entrées possibles et de celle de toutes les sorties possibles. Les entrées devront être sensorielles et les sorties devront être comportementales.

Illustrée par une analogie désormais classique, cette thèse stipule que le mental est une boîte noire (un mécanisme caché dans la tête dont on n'aurait pas accès directement) qui transformerait les informations entrantes (stimulations sensorielles) en des réactions corporelles (comportements ou actions).

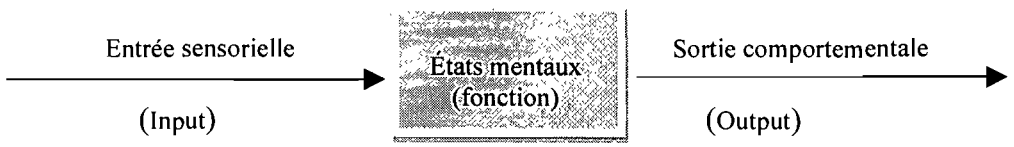

Figure 2 – Schéma de l'explication fonctionnelle

Pour procéder à une explication fonctionnelle, le scientifique devra (1) caractériser la classe des entrées sensorielles possibles ainsi que (2) la classe des sorties comportementales possibles, puis (3) spécifier comment les différents processus mentaux en viennent à coordonner l'information entrante pour produire les comportements qui y

[40] Adaptation de l'explication fonctionnelle dans P.M. Churchland (2007, p.19)

sont associés. (Rey, 1997, chapitre 6) Il est à noter que les étiquettes utilisées pour nommer les trois éléments constitutifs de l'explication fonctionnelle peuvent porter à confusion s'il n'y a pas de distinctions marquées entre l'explication fonctionnelle globale d'un système cognitif et l'explication fonctionnelle locale d'un processus cognitif. (Fodor, 1983) Pour simplifier l'explication de la distinction, limitons-nous à la perspective du niveau sémantique.

Une explication fonctionnelle locale consiste à déterminer, pour un processus cognitif donné (ex : prise de décision), le fonctionnement de ce mécanisme, la classe des *inputs* pouvant en affecter le mécanisme ainsi que les résultats de l'interaction entre ces apports externes et le processus même. Les entrées ne sont pas nécessairement sensorielles dans une explication fonctionnelle locale, tout comme les sorties ne sont pas nécessairement comportementales. Par exemple, dans le processus de prise de décision, les entrées pourraient être décrites comme des croyances sur l'état du monde (et de l'agent) et des désirs spécifiant les états du monde (et de l'agent) à obtenir. Ces entités auront été construites par d'autre processus cognitifs et ne sont pas à proprement parler des stimulations sensorielles. Les résultats de la délibération seraient des décisions, ce qui devra être transposé par d'autres mécanismes cognitifs en comportements en vue d'atteindre l'objectif escompté. Mais pour un processus cognitif donné, une explication fonctionnelle locale pourrait spécifier une entrée sensorielle (ex : perception) ou une sortie comportementale (ex : processus moteurs). (Fodor, 1983) Afin de ne pas créer de confusion, les entrées et sorties d'un processus cognitifs, relevées par une explication fonctionnelle, seront nommés respectivement *inputs* et *outputs*. Cette nomenclature sera avantageuse lorsqu'il sera question de la variation typologique de ces classes d'entités aux différents niveaux explicatifs.

L'explication fonctionnelle globale consiste à déterminer les relations entres toutes les formes d'entrées sensorielles pouvant affecter le système cognitif et les formes de sorties comportementales pouvant être produites par ce même système cognitif. Ces relations sont spécifiées en déterminant les mécanismes internes (processus cognitifs) permettant de produire les comportements adéquats aux stimulations de l'environnement. Ainsi considérée, l'explication fonctionnelle concerne le système cognitif dans sa globalité puisque la description des états mentaux consiste en fait à rabouter tous les processus cognitifs décrits localement par une explication fonctionnelle. (Fodor, 1983; Newell, 1990;

von Eckardt, 1993) Ce portrait fera nécessairement appel à des stimulations sensorielles pour *input*s et à des sorties comportementales pour *outputs* mais, comme nous le verrons dans l'analyse fonctionnelle des différents niveaux explicatifs, même au niveau global la typologie de ces classes d'entités varie grandement.

Si l'on met en relation ce type d'explication avec la répartition tripartite des niveaux explicatifs (niveaux sémantique, computationnel et d'implémentation matérielle) et les stratégies explicatives (*top-down* versus *bottom-up*) présentées au chapitre précédent, nous obtenons un schème explicatif beaucoup plus complexe mais surtout beaucoup plus riche par les contraintes inter-niveaux qu'il prescrira[41]. L'explication fonctionnelle peut prendre place à chacun des trois niveaux (selon chacune des trois perspectives). Cela signifie qu'à chaque niveau serait spécifié un vocabulaire explicatif en déterminant les prédicats utilisés pour caractériser les *inputs*, les *outputs* ainsi que les mécanismes des processus cognitifs (mécanismes dans la 'boîte noire'). Le niveau sémantique devra spécifier ces trois éléments selon un cadre conceptuel psychologique alors que le niveau computationnel devra y procéder en formalisant ceux-ci en structures informationnelles ou en processus de transformation de l'information. Le niveau de l'implémentation physique devra y parvenir en déterminant par quels phénomènes matériels les trois classes d'entités sont réalisées. Tout ceci découle directement de ce qui a été dit précédemment dans cette section ainsi qu'au chapitre II (section 2.1).

La hiérarchie des trois niveaux explicatifs a été construite en fonction de la relation de réalisation explicative (section 2.1.3). Il semble donc approprié que la spécification de cette relation doive prendre en compte les trois classes d'entités de l'explication fonctionnelle. Les *inputs* au niveau sémantique ne devront pas être réalisés par les *outputs* au niveau computationnel ni par les processus computationnels. Il semble qu'il faille réserver leur réalisation à la classe des *inputs* informationnels. De même pour les *inputs* du niveau neurobiologique. Cela pousse à penser qu'il faille lier (par la relation de réalisation explicative) les entités postulées à un niveau par les entités équivalentes à un autre niveau et, ce, à partir de leur rôle dans l'explication fonctionnelle (*input*, *output* ou mécanisme

[41] Il est à noter qu'à partir de ce point jusqu'à la fin de la présente section, la mise en relation des niveaux explicatifs au format fonctionnel de l'explication en sciences cognitives n'existe pas tel quel dans la littérature. C'est en fait une construction conceptuelle originale. Toutefois, cette 'construction' de ma part (ainsi que la figure 3 (et son titre) qui en offre une présentation visuelle) découlent du cadre conceptuel utilisé jusqu'à présent, comme les explications à venir permettront d'en convenir, et ils permettent une meilleure compréhension de la relation de réduction proposée par P.M. Churchland.

interne à la 'boîte noire'). Ce type de relation sera exploré dans la dernière section de ce chapitre (section 3.3) et on y trouvera une mise en relation suffisamment aisée pour illustrer la force du cadre conceptuel utilisé ici pour analyser le débat réductionniste.

La direction de la spécification théorique des explications fonctionnelles dépend de la stratégie explicative utilisée. L'approche *bottom-up* demande au scientifique de spécifier quelle famille de machines matérielles elle étudiera (processeurs électroniques, systèmes nerveux, etc.) pour ainsi déterminer la caractérisation adéquate des prédicats utilisés pour l'explication fonctionnelle à ce niveau. (Dennett, 1994) Par exemple, dans le cas du neurobiologiste, il sera question de spécifier la nature et le fonctionnement des neurones, les stimulations matérielles affectant les organes perceptuels de la créature étudiée, etc. De plus, étant donné la stratégie *bottom-up*, les types d'entrées sensorielles au niveau de l'implémentation matérielle restreindront les types d'entrées sensorielles au niveau computationnel puis au niveau sémantique. Cette distinction est particulièrement importante lorsque le scientifique cherche à déterminer l'image fonctionnelle globale d'un niveau : on voit mal comment spécifier une relation de réalisation entre fonctions mentales (mécanismes dans la 'boîte noire') à partir d'éléments constitutifs de la description des classes d'entrées sensorielles (*input*). Il sera question, à la section 4.1.3, d'une infection 'mentaliste' des classes d'entrées sensorielles et des sorties comportementales dans la caractérisation du niveau sémantique par la psychologie cognitive, ce qui s'avérera problématique pour la réduction puisqu'il ne semble pas y avoir une telle superposition de ces éléments fonctionnels aux deux autres niveaux. (Pylyshyn, 1984)

Au contraire, une approche *top-down* débutera avec la caractérisation fonctionnelle au niveau sémantique des états mentaux, des *inputs* et des *outputs* (ce que la psychologie cognitive fonctionnaliste fait depuis près de soixante ans) pour ainsi spécifier leurs équivalents computationnels. La transition d'un niveau à l'autre sera alors déterminée par les normes de traduction constituant la relation de réalisation explicative adéquate pour relier les niveaux concernés. La relation de réalisation explicative permet donc de systématiquement traduire un phénomène prenant place à un niveau donné (ex : computationnel) à un autre (ex : implémentation matérielle). Cette bidirectionnalité de la relation permet la distinction réalisation/implémentation. Une approche *bottom-up* pousse à déterminer ce que réalise un certain agencement matériel au niveau computationnel (*reverse engineering*), alors que l'implémentation d'un processus computationnel consiste à

déterminer quelles conditions devront être produites au niveau matériel pour construire un système computant la fonction computationnelle de ce processus (*foward engineering*). (Dennett, 1994; P.S. Churchland, 1986, chapitre 9) Ces considérations nous permettent d'illustrer par la figure 3 la structure des niveaux explicatifs selon le schème fonctionnel (global).

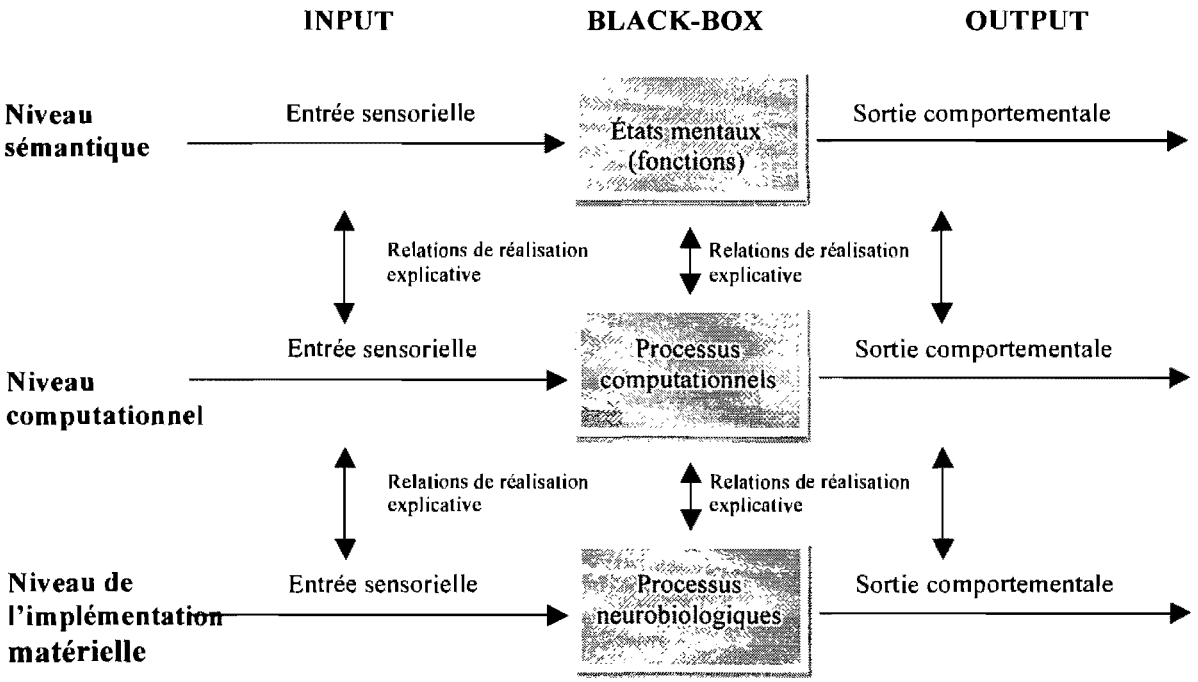

Figure 3 – The Great Computational Sandwich

Comme il en a été question aux sections 2.1.3 et 2.1.4, la relation de réalisation explicative impose des contraintes d'un niveau à l'autre : tous les éléments du niveau computationnel devront trouver une réalisation neurobiologique sans quoi le scientifique devra faire face à des propriétés mystérieuses dont la réalité matérielle ne saurait en rendre compte la nature (physicalisme). (Fodor, 1974) De même pour le niveau sémantique vis-à-vis le computationnel. (P.M. Churchland (1989, chapitre 5; 2006; 2007, chapitre 2)) Cette direction des contraintes, forcée par la stratégie *bottom-up* de P.M. Churchland, permet alors de spécifier quels éléments du vocabulaire théorique de chaque niveau doivent être invoqués pour permettre une image systématique d'un niveau par un autre. Si une entité théorique spécifiée à un niveau ne trouve pas de relation de réalisation explicative avec le

niveau inférieur, il faudra alors penser à sa révision (ou à son élimination) puisqu'il n'y a pas d'image théorique (I_b) cohérente avec la théorie de base du niveau inférieur (T_b).

Le format de l'explication fonctionnelle ainsi que la répartition tripartite des niveaux explicatifs sont des outils épistémologiques largement reconnus dans la communauté philosophique concernée par les sciences cognitives (par exemple Dennett, 1987, 1994; Fodor, 1975, Lycan, 1981; Putnam, 1967a, 1967b; Marr, 1982; Newell, 1980, 1990, etc.). L'utilisation en parallèle de ces outils dans le cadre du débat réductionniste semble, au premier œil, ne pas porter à controverse. Pourtant, comme il en sera question au chapitre suivant, l'unification de ces outils épistémologiques a des répercussions drastiques sur les possibilités de réduction des théories psychocognitives aux théories neurocognitives. Certains de ces impacts ont une aura positive : il semble que cela permette au débat réductionniste de trouver une formulation plus systématique et plus riche. La section suivante illustrera ce côté harmonieux introduit par l'unification des deux outils épistémologiques en analysant les relations des éléments fonctionnels au niveau neurobiologique et neurocomputationnel telles que l'entend P.M. Churchland. Toutefois, certaines conséquences sont plus troubles puisqu'il semble que de nouveaux problèmes émergent à la fois pour le réductionniste et l'antiréductionniste : il devient clair que les entrées sensorielles et les sorties comportementales caractérisées au niveau sémantique résistent plus agressivement à une évaluation de la thèse réductionniste que les processus mentaux caractérisés à ce niveau. Il en sera question au prochain chapitre aux sections 4.2.2 et 4.2.3. Pour l'instant, il est en ordre de procéder à une analyse des deux premiers niveaux par lesquels la stratégie explicative *bottom-up* de P.M. Churchland prescrit de débuter les recherches en vue d'obtenir une évaluation de la thèse réductionniste, soit la relation de réalisation entre le niveau de l'implémentation matérielle (neurobiologique) et le niveau computationnel (neurocomputationnel).

3.3 Réduction du niveau neurocomputationnel au niveau neurobiologique

3.3.1 Explication fonctionnelle au niveau neurobiologique

Les explications données au niveau de l'implémentation matérielle consistent à déterminer quels éléments et processus neurobiologiques permettent de rendre compte de l'activité cognitive de la créature étudiée (P.S. Churchland, 1986 1ère partie). Étant donnée

l'approche *bottom-up* prescrite par le naturalisme de P.M. Churchland, le vocabulaire théorique utilisé dans cette perspective est celui des sciences neurobiologiques (neurophysiologie, neuroanatomie, neurochimie, etc.).

L'unité fonctionnelle de base dans la théorie neurobiologique est le neurone. Ce dernier est une structure complexe dont les éléments constitutifs et ses processus internes jouent un rôle fondamental dans son fonctionnement et ont des impacts directs sur les processus computationnels produits au niveau supérieur. Il ne sera pas question ici d'entrer dans les détails ni du fonctionnement du neurone ni dans celui des réseaux de neurones naturels formant les diverses structures du système nerveux central (voir Bear et al. (2002) à ce sujet), mais une brève explication s'impose.

Les neurones sont connectés les uns aux autres par la proximité de leur projections synaptiques et dendritiques. Si un neurone est suffisamment excité, son axone propagera un influx nerveux (potentiel d'action) et les synapses en son extrémité relâcheront des neurotransmetteurs qui, s'ils sont captés par les dendrites d'un autre neurone, déclencheront un influx électrochimique qui exciteront cet autre neurone. Ce sont les neurotransmetteurs qui permettent l'activation d'un neurone par un autre. De cette explication schématique, il est aisé de voir que le fonctionnement du neurone peut être rendu par une explication fonctionnelle : les *inputs* sont les neurotransmetteurs dégagés par les neurones afférents, les *outputs* des neurotransmetteurs dégagés vers d'autres neurones et le fonctionnement interne est déterminé par l'équilibre biochimique des cellules neuronales, de l'état électrique de la membrane de l'axone et des différentes organelles constituant le soma du neurone. (Bear et al. ,2003, chapitres 2 à 6)

Les neurones peuvent se connecter les uns aux autres formant ainsi une diversité de réseaux neuronaux. Ces réseaux sont en fait constitués par un raboutage des neurones. P.S. Churchland & Sejnowski (1992, p.11) proposent une hiérarchie des niveaux d'organisation neurobiologique[42] (neurone, réseau, carte neuronale, système particulier, système nerveux central). Ceux-ci répondent tous au format fonctionnel. Les *inputs* et les *outputs* seront déterminés en fonction de l'organisation des connections entre les neurones, suivant la direction dendrites-soma/axone-synapses. Il est à noter que les *inputs* des réseaux neuronaux (et donc aussi des neurones individuels) peuvent provenir de neurotransmetteurs transportés dans le sang produits par des glandes éloignées. On peut aussi concevoir la

[42] À différencier des niveaux explicatifs dont il a été question jusqu'ici.

fréquence de la décharge des neurotransmetteurs (*spiking frequency*), alors calculable par l'état d'innervation des axones.

Du point de vue de l'organisation globale du système nerveux (central et périphérique), les *inputs* et *outputs* seront considérés d'une autre manière, quoique les réseaux de neurones décrits précédemment continuent de jouer le rôle de mécanismes transformant les *inputs* en *outputs*. (Bear et al., 1992, chapitres 8 à 12; P.S. Churchland, 1986, chapitres 3 à 5) Certains neurones ne reçoivent pas de neurotransmetteurs mais servent en fait de capteurs sensoriels pour la créature étudiée. Les cônes et les bâtonnets constituant la rétine en sont un exemple bien connu. Ces capteurs sont excités par des stimulations externes provenant de l'environnement et y réagissent en activant les neurones connectés à leurs synapses (phénomène de la transduction). Les neuroscientifiques devront alors spécifier quels phénomènes physiques externes activent ces neurones sensoriels (ex : mouvement de la membrane basilaire frottant les cellules ciliées). Ceux-ci seront nommés 'stimulations'. Du côté des *outputs*, les phénomènes physiques pouvant être produits par les interactions des réseaux neuronaux naturels sont multiples : activation des muscles, changements physiologiques (par exemple les réactions 'émotives' ou la sécrétion d'hormones), etc.

Comme il a été vu à la section 1.3.3 et 2.1.4, P.M. Churchland & P.S. Churchland (1983) prescrivent une approche individualiste (solipsisme méthodologique) des systèmes cognitifs. (voir Burge, 1979, 1986; ainsi que Fodor, 1980) Cela implique alors que les *outputs* au niveau neurobiologiques ne peuvent faire référence à des éléments environnementaux au système cognitif (ex : prendre un objet) et donc que les notions classiques d'intentionnalité n'ont aucune prise à ce niveau (Brentano, 1973; Quine, 1960, chapitre 6). Les *outputs* comportementaux tout comme les *inputs* sensoriels devront être qualifiés uniquement à partir de leur rôle au niveau de la cognition et pas en fonction de leur provenance ou de leurs effets sur l'environnement. Le neurobiologiste pourrait faire référence aux muscles et autres dispositifs biologiques dont le fonctionnement peut être altéré par les processus cognitifs mais, comme il a été expliqué précédemment, dans le cadre d'un projet réductionniste il faudra parvenir à une caractérisation des *inputs* et *outputs* permettant de systématiquement lier les trois niveaux.

Ainsi, pour déterminer la classe des *outputs* neurobiologiques, le neurobiologiste pourrait y inclure des réactions motrices, mais celles-ci ne sont pas des comportements à

proprement parler. Millikan (1993, chapitre 7) propose une nuance entre les réactions motrices et les comportements à proprement parler : les comportements auraient une charge référentielle à l'environnement car ceux-ci sont des réactions motrices servant une fonction biologique (évolutionniste). Ainsi, la description complète des réactions motrices aux suites d'une série d'entrées sensorielles pour un système cognitif ne permettrait pas de qualifier ces réactions motrices de comportements : il faut déterminer ce qui fait qu'une série de mouvements produits par un système cognitif peut être qualifiée de comportement ayant une fonction et entretenant donc une relation avec l'environnement. Cette nuance serait réservée au niveau sémantique. Maintenant que le niveau neurobiologique (implémentation matérielle) a été cerné, il est temps de passer au niveau supérieur suivant, soit le niveau computationnel pour ensuite déterminer (section 3.3.3) quelles relations de réalisation explicatives pourront être proposées pour lier les deux niveaux fonctionnels.

3.3.2 Explication fonctionnelle au niveau computationnel

Le format computationnel prescrit par P.M. Churchland est le *Parallel-Distributed Processing* (PDP), sous-famille des réseaux connexionnistes. (McClelland et al., 1986). Ce format computationnel s'inspire directement des réseaux de neurones naturels quoiqu'un grand nombre de propriétés neuronales ne sont pas transférées dans l'image neurocomputationnelle des PDP, dont plusieurs propriétés ayant un effet neurocomputationnel (ex : rôle des cellules gliales, rôle fonctionnel des différents neurotransmetteurs, différentes architectures des neurones naturels, etc.).

Les réseaux de neurones artificiels suivent une logique très simple et permettent de réaliser des fonctions très complexes. Deux éléments théoriques constituent les matériaux de construction de la théorie : les unités d'activations (ou neurones[43], à différencier des neurones au niveau de l'implémentation matérielle) et les connexions inter-unités. Les unités sont interconnectées à d'autres unités, à des senseurs et à des effecteurs[44]. Les connexions reliant les unités sont généralement directionnelles (le signal part d'une unité et

[43] Les neurones dans la perspective computationnelle ne sont pas identiques aux neurones au niveau matériel. Toutefois, l'un est réalisé par l'autre, ce dont il sera question dans la section suivante.

[44] Un effecteur sera ici défini comme un dispositif produisant un effet en fonction de l'information reçue par le réseau de neurone. Dans un cadre strictement neurobiologique, ces effecteurs sont en fait les neurones connectés aux muscles, aux glandes (tout organe activé par une stimulation neuronale, mais dans un cadre d'intelligence artificielle plus généralisé, les effecteurs pourraient très bien être connectés à une imprimante, des hauts-parleurs ou même par de l'information envoyée dans un logiciel.

va en activer une autre). Une unité peut donc émettre un signal par une connexion efférente (par son synapse vers une dendrite) ou en recevoir une par une connexion afférente (neurotransmetteurs reçus par sa dendrite). La toile des connexions constitue le réseau de neurone. (P.M. Churchland, 1989, chapitre 5; P.S. Churchland & Sejnowski, 1992; Bechtel & Abrahamsen, 2002)

On peut produire une explication fonctionnelle des unités d'activation et de leurs connexions. Chaque unité émet des signaux par ses connexions efférentes si elle est suffisamment activée. Chacune des unités comporte donc un seuil d'activation minimal spécifiant la force de l'activation nécessaire devant être fournie à l'arrivée synchrone de signaux transportés par les connexions afférentes (*input*). Chaque connexion est déterminée par son unité d'activation de départ, la force d'activation de la connexion et l'unité d'arrivée qui sera affectée du signal. La force totale des signaux afférents est comparée avec la valeur du seuil d'activation de l'unité (fonction). Si le seuil est atteint ou dépassé par la force d'activation à un certain moment, l'unité s'active à son tour et émet un signal par toutes ses connexions efférentes (*fonction*). Le signal ira stimuler d'autres unités. Une unité peut donc être dans deux états d'activation : activée ou pas[45]. Quant à elles, les connexions peuvent être de deux natures : ou bien elles sont dites 'excitatrices', dans ce cas elles envoient un signal dont la force d'activation est positive; ou bien elles sont 'inhibitrices', alors leur force est négative, c'est-à-dire qu'elles diminuent l'excitation envoyée au neurone récepteur.

Un réseau d'unités d'activation et des connexions qui les lient est aussi candidat à l'explication fonctionnelle. Celui-ci est constitué par le raboutage des unités d'activations et cela lui octroie trois parties constitutives (dont une est facultative). Les unités ayant une connexion afférente provenant de senseurs constituent la couche d'entrée (ou '*input layer*'). Celle-ci reçoit des activations des senseurs selon la fonction de transduction (spécifiant comment les effets physiques sur les senseurs (niveau de l'implémentation matérielle) donnent lieu à certains *patterns* d'activation (niveau computationnel). Un *input* au niveau computationnel consiste donc en un signal afférent à la couche d'entrée. Les unités connectées à des sorties effectrices constituent la couche de sortie (ou '*output layer*') et les effecteurs recevant les *patterns* d'activation de la couche de sortie produiront des réactions

[45] À fin de simplification, j'escamote les détails du 'calcul' des signaux d'activations. Le lecteur intéressé à pousser plus à fond sa compréhension des réseaux de neurones artificiels pourra se référer à Bechtel & Abrahamsen (2002), McClelland et al. (1986) ainsi qu'à Anderson (1995) et Arbib (1998).

mécaniques selon la fonction de transduction spécifié par leur structure (spécifiée au niveau de l'implémentation matérielle). Dans un réseau de neurone artificiel, le *output* est un *pattern* d'activation efférent de la couche de sortie. (McClelland et al., 1986)

Il doit y avoir au moins deux couches pour former un réseau et les unités d'activation de l'une ne peuvent pas (toutes) être présentes dans l'autre. Cela oblige donc un réseau constitué d'au moins deux couches d'unités différentes. Une troisième couche d'unités (facultative) peut relier la couche d'entrée et la couche de sortie. Cette dernière sera nommée couche intermédiaire (ou '*hidden layer*') et c'est celle-ci qui procédera à la manipulation de l'information fournie à la couche d'entrée par les senseurs (fonction). (McClelland et al., 1986; P.M. Churchland, 1989, chapitre 9) Il n'y a pas de limites au nombre de couches intermédiaires (sinon celle du système nerveux étudié dans les cas de modèles théoriques de fonctions neurobiologiques). Dans le cas où le réseau ne serait constitué que de deux couches, c'est la structure des connexions inter-unités qui déterminera les opérations de transformation de l'information. Il est important de comprendre que cette description des réseaux artificiels cherche à refléter la constitution (partielle) des réseaux de neurones naturels. (McClelland et al., 1986, volume 2) Au niveau computationnel, cette description trouve sa réalisation dans une formalisation mathématique.

C'est par une représentation vectorielle que les réseaux de neurones artificiels trouvent leur description formelle. (McClelland et al., 1986, chapitre 2) Les valeurs d'activations envoyées à une unité sont indexées (dans un ordre arbitraire), permettant ainsi de représenter celles-ci par un vecteur. Supposons, pour simplifier, que toutes les unités constituant la couche d'entrée reçoivent pour *input* le même vecteur d'activation. Ce vecteur représente donc la structure du *pattern* de l'activation de la couche d'entrée et un autre représentera la valeur d'activation produite par les unités de la couche de sortie (*output*). La fonction de transformation de l'information par un réseau consiste en une multiplication matricielle du vecteur *input* à une matrice représentant la force d'activation de chacune des connexions afférentes aux différentes unités de la couche d'entrée. Chaque unités a un seuil d'activation et enverra, par ses connexions efférentes, une valeur d'activation vers d'autres unités (ou effecteurs). La fonction de transformation des vecteurs d'activations afférents (*input*) en vecteurs d'activation efférents (*output*) sera donnée par une matrice spécifiant le seuil d'activation (ou poids) de chacun des neurones constituant la

couche observée. Le vecteur efférent (*output*) sera le résultat du produit du vecteur afférent (*input*) et de la matrice spécifiant le poids de chacune des connexions alimentant la couche. Le vecteur efférent (*output*) sera envoyé comme vecteur *input* à une autre couche de neurones (elle-même représentée par une autre matrice). Les réseaux de neurones artificiels ont donc pour mode de computation la multiplication matricielle de vecteurs. (McClelland et al., 1986, chapitre 2)

3.3.3 Relations entre le niveau neurobiologique et le niveau computationnel

À partir de cette explication succincte des réseaux de neurones artificiels et naturels, Paul M. Churchland (et les chercheurs connexionnistes) proposent une relation de réalisation très simple entre les différentes entités décrites aux deux niveaux. Chaque unité d'activation tiendrait pour un neurone, chaque connexion entre les unités d'activation serait réalisée par des connexions synaptiques. Les *patterns* d'activation dans les réseaux de neurones artificiels sont représentés par la fréquence d'activation afférente des neurones naturels. P.M. Churchland (1989, chapitre 5, pp.98-102) illustre plus en détail cette relation de réalisation en proposant un modèle computationnel des cellules de Purkinje :

> « The output frequency of spike emissions for each cell is determined by the simple *frequency* of input stimulations it receives from all incoming synaptic connections and by the *weight* or *strength* of each synaptic connection, which is determined by the placement of the synapses and by their cross-sectional areas. These strength values are individually represented by the coefficients of [a] matrix [...]. *The neural interconnectivity thus implements the matrix.* » (P.M. Churchland, 1989, pp.99-100, italiques dans l'original)

Bien sûr, il n'est pas spécifié en détail par quels processus les cellules de Purkinje parviennent à effectuer cette fonction de multiplication matricielle (ex : comment les neurones parviennent à effectuer la respiration cellulaire, etc.). Ce type d'explication ne serait pas pertinent vis-à-vis le détail des processus neurocomputationnels pour permettre de produire une relation de réalisation explicative entre les deux niveaux. Toutefois, ce type d'explication neurobiologique pourrait jouer un rôle primordial s'il fallait expliquer pourquoi un réseau de neurones naturels ne parviendrait pas à implémenter les fonctions computationnelles (ex : la carence en apport de sodium par le sang au neurone empêcherait la production d'un potentiel d'action).

La relation de réalisation spécifie donc les relations pertinentes entre l'ontologie et les processus au niveau computationnel à ceux du niveau neurobiologique. Elle explique aussi pourquoi un ordinateur sériel multipliant des vecteurs et des matrices ne serait pas réduit au niveau neurobiologique : l'ordinateur sériel ne parviendrait pas à remplir les conditions neurobiologiques spécifiées par la relation de réalisation explicative. Il n'y aurait pas de manière systématique de produire un isomorphisme entre la description physique des processus électroniques d'un ordinateur et les opérations computationnelles de multiplication matricielle réalisées par celui-ci répondant aux conditions de la relation de réalisation explicative spécifiée par la stratégie *bottom-up* de P.M. Churchland. (Ramsey et al., 1991 ; Von Eckardt, 2005)

Le manque de réalisme (de correspondance) entre les réseaux de neurones artificiels vis-à-vis les réseaux de neurones naturels est bien connu[46]. Cela n'est pas un problème pour la stratégie réductrice de P.M. Churchland puisqu'elle ne pourra prendre place qu'une fois les neurosciences achevées. Ainsi il est encore temps pour les chercheurs de déterminer plus amplement et plus précisément les conditions de la relation de réalisation explicative entre ces deux niveaux et de construire un portrait neurocomputationnel plus réaliste vis-à-vis les réseaux de neurones naturels (ex : en intégrant les fréquences d'activation des neurones, le caractère continu des activations, etc.). Cela constitue d'ailleurs une partie du champ de recherche des neurosciences actuelles. (McClelland et al., 1986, volume 2)

Nous avons vu dans ce chapitre (section 3.1 et 3.2) que le format de l'explication fonctionnelle, emprunté à la psychologie cognitive et aux philosophes fonctionnalistes, pouvait être utilisé pour rendre compte des processus cognitifs aux différents niveaux explicatifs (section 2.1). À fin de démontrer cette thèse, il a été question dans la section 3.3 d'appliquer ce schème conceptuel pour d'abord caractériser les niveaux neurobiologique et neurocomputationnel, puis de spécifier certaines propriétés de la relation de réalisation explicative, assurant ainsi un début d'isomorphisme dans l'explication scientifique à ces

[46] McClelland et al. (1986) offrent, dans le deuxième volume, six chapitres dédiés à l'évaluation et à la similarité des réseaux neuronaux artificiels aux réseaux neuronaux naturels. Arbib (1998) offre aussi plusieurs articles à ce sujet. Les principaux problèmes consistent à déterminer quelles sont les caractéristiques des neurones naturels qui sont pertinents pour la computation et à trouver manière à les modeler avec le plus grand réalisme dans les modèles artificiels. De plus, grand nombre de recherches connexionnistes ne sont pas intéressées à produire des modèles réalistes mais simplement à implémenter des algorithmes (Bechtel & Abrahamsen, 2002).

deux niveaux. Le pas suivant consistera donc, selon la stratégie *bottom-up* de P.M. Churchland (section 2.2.2), à mettre en relation les processus computationnels avec les processus cognitifs au niveau sémantique (section 4.2), complétant ainsi le schème de l'explication fonctionnelle et à trois niveaux explicatifs. Pour y arriver, et ainsi déterminer l'état des lieux vis-à-vis le néo-réductionnisme et la possibilité de réduire la psychologie cognitive aux neurosciences à partir d'une relation de réalisation explicative (section 2.1.3), il faudra tout d'abord rendre compte de la structure de la théorie psychocognitive et du vocabulaire théorique dont ses théoriciens font usage (section 4.1).

Chapitre IV : Image neuropsychologique des théories psychocognitives

Il a été vu au dernier chapitre comment les réseaux PDP (*parallel-distributed processing*) pouvaient être réalisés par les réseaux neuronaux naturels (section 3.3.3). Une telle thèse ne porte pas ou peu à controverse. Il semble que ce mode de computation soit le plus apte à produire un modèle des opérations cognitives menées dans les systèmes cognitifs naturels. (P.M Churchland, 1989; P.S. Churchland, 1986; McClelland et al., 1986) Le débat réductionniste prend de l'ampleur lorsqu'il est question d'expliquer les états mentaux postulés au niveau sémantique. Selon le modèle de l'explication fonctionnelle aux trois niveaux explicatifs proposé au deuxième chapitre, le réductionnisme de P.M. Churchland, pour être validé, devra tout d'abord produire un portrait théorique global de la psychologie achevée (théorie à réduire T_r). De ce portrait, à la lumière des niveaux neurobiologique et neurocomputationnel (théorie réductrice T_b) expliqués au chapitre précédent (section 3.3), il faudra produire l'image (image I_b de la théorie réduite à partir de la théorie réductrice) de la psychologie cognitive dans le vocabulaire neurocognitif (neuropsychologie). La méthode pour produire cette traduction consistera à établir une relation de réalisation explicative (voir section 2.1.3) pour assurer une liaison entre les processus automatisés aux niveaux décrits précédemment (neurobiologiques et neurocomputationnels) et les aspects proprement sémantiques (le contenu des représentations et les opérations cognitives sensibles à ce contenu). Toutefois, la construction du détail nécessaire pour la production d'une image neuropsychologique nécessitera de clarifier une nuance importante au sein de la structure même de la méthodologie de la psychologie cognitive. La distinction consiste à bien comprendre le statut et le rôle de la psychologie du sens commun au sein de la méthodologie de recherche de la psychologie cognitive et de distinguer le cadre conceptuel de base des modèles des sous-processus permettant d'expliquer la nature et le fonctionnement du mental. Les modèles psychocognitifs tendent à être limités à certaines facultés mentales, offrant ainsi une explication locale des processus cognitifs, alors que la psychologie du sens commun offre un portrait global des interactions causales menant aux sorties comportementales. Cette nuance a de l'importance car la spécification des entités théoriques diffère selon qu'elle provient d'un modèle psychocognitif ou du portrait global de la psychologie du sens

commun. En vue d'enrichir cette distinction, il faudra procéder à une analyse de la structure des théories psychocognitives, analyse qui sera menée dans la section 4.1.

Toutefois, pour parler de réduction réussie, il ne faudra pas simplement construire une image neuropsychologique (I_b) mais aussi montrer s'il est possible de rendre compte des théories psychocognitives à partir de cette image, c'est-à-dire comment les fonctions mentales (au niveau sémantique) postulées par les théories psychocognitives peuvent être représentées dans le vocabulaire neuropsychologique. Dans le cas de révisions ou même d'élimination, la simple production d'une image neuropsychologique différente de la théorie psychocognitive ne suffit pas : il faut montrer aussi *pourquoi* il n'y a pas de ponts possibles (ou, du moins, montrer pourquoi certains ponts échouent à se faire) et par quelle méthode réviser les éléments défectueux de la théorie en cours de réduction. Il faut donc déterminer quelles relations entretiennent les théories psychocognitives à leur alter-égo neuropsychologiques. La première section de ce chapitre (section 4.1) s'enquerra donc d'analyser le niveau sémantique tel que les tenants de la psychologie cognitive le conçoivent et de préciser le rôle de la psychologie du sens commun (*folk psychology*) dans la méthodologie de la psychologie cognitive. À la lumière de l'analyse faite à la section 3.3.3, la section suivante (section 4.2) dressera un portrait de la théorie neuropsychologique, soit des relations de réalisations explicatives entre les éléments du niveau computationnel et ceux du niveau sémantique (*input/output*/processus de la 'boîte noire'). Des conclusions tirées à partir de l'analyse de la relation de réalisation explicative, il sera possible de produire un premier portrait de l'état des lieux vis-à-vis le réductionnisme de P.M. Churchland et d'examiner la possibilité de son matérialisme éliminativiste Il sera alors question des parallèles et des divergences entre les deux conceptions du niveau sémantique (théorie psychocognitive T_r *versus* neurocognitive I_b) et explorerons brièvement quelques voies dont l'exploration future permettra de mieux concevoir la situation de la vérification empirique de la thèse réductionniste (section 4.3).

4.1 Structure des théories psychocognitives

4.1.1 Cadre conceptuel de la psychologie cognitive

> « If that sounds like grandmother psychology, that's because much of grandmother's psychology is sound, not in its explanation of underlying mechanisms but in the *kinds* of regularities it addresses. We begin with these categories because that's where we at least have a clue that there is

> something systematic about human behaviour. We do not end there, however; an explanation requires much more. Yet a theory that *ignores* truisms has started off on the wrong foot. » (Pylyshyn, 1984, p.258)

Cette citation nous permettra de ne pas entrer dans le débat concernant les origines et le statut ontologique de « l'objet » 'psychologie du sens commun'[47]. Celle-ci sera considérée comme un cadre théorique emprunté par la psychologie cognitive pour amorcer ses recherches sur les mécanismes sous-jacents aux facultés qui y sont postulées. Une telle caractérisation montre clairement quel est l'objectif scientifique de la psychologie cognitive : le niveau sémantique (ou psychologique) est déjà suffisamment complet par l'apport du cadre conceptuel de la psychologie du sens commun; il ne reste plus au scientifique qu'à développer une théorie des processus computationnels qui réalisent les états mentaux et leurs interrelations causales. Le niveau sémantique ne sera pratiquement pas altéré par les découvertes empiriques au niveau de son cadre conceptuel. Le détail des relations causales qu'entretiennent les états mentaux entre eux sera raffiné par les expériences empiriques et les mécanismes computationnels qui les produisent devront être reconstruits au travers une caractérisation plus fine des processus de traitement de l'information les réalisant. Il faut donc d'abord connaître le détail de la psychologie du sens commun.

David Lewis (1972/1980) a proposé une stratégie de formalisation des états mentaux permettant de déterminer quelles conditions devront remplir certains états de fait pour être considérés comme des instances d'un type d'état mental. Cette caractérisation se borne à rendre compte des états mentaux tirés de la psychologie du sens commun uniquement en vertu de leur rôle causal dans l'économie du mental. La stratégie va comme suit. Le psychologue (ou le philosophe) doit retrouver toutes les régularités causales reliées à un état mental donné, soit ses interactions causales avec les entrées sensorielles, les autres états mentaux et finalement les comportements. Formalisant ces régularités en propositions logiques (le type de relation causale étant formalisé par un prédicat), il constituera alors un énoncé disjonctif assemblant ces relations les unes avec les autres. Une fois cette liste constituée, par une science psychologique achevée (Lycan, 1987; Block, 1980, p.272) ou par une analyse *a priori* des 'platitudes du sens commun' (Lewis, 1972/1980, p.212), il en formera un énoncé existentiel où tous les termes théoriques (les concepts d'états mentaux

[47] Concernant ce débat, voir P.M. Churchland (1981, 1989), P.M. Churchland & P.S. Churchland (1996), Clark (1996), Fletcher (1995), Greenwood (1991), Horgan (1992), Horgan & Woodwart (1985), Stich (1983).

ainsi que les caractérisations des *inputs* et *outputs*) seront remplacés par des variables sans noms. Cela produit alors la définition d'un état mental type et si cet énoncé disjonctif est transformé en énoncé conditionnel, alors notre psychologue détient un schème pour identifier un phénomène comme une occurrence particulière de l'état mental type formalisé par cette méthode. Cette dernière, ici présentée informellement, est connue sous le nom de 'ramséfication'[48] et permet de rendre compte de la fonction d'un processus mental en spécifiant ce qu'il fait (mais pas comment il le fait). L'énoncé ramséen ne fait pas référence à la manière dont le système cognitif devrait être réalisé matériellement ou computationnellement; il ne fait que spécifier les relations causales abstraites de la transformation des entrées sensorielles par les états mentaux en comportements (Rey, 1997). Cela est parfaitement en accord avec la perspective fonctionnaliste[49].

Dans la perspective épistémologique utilisée ici, cela signifie que les énoncés ramséens permettent d'identifier un phénomène à un type d'état mental donné s'il répond à un énoncé ramséen pour ce type d'état mental. Cela a son importance car, comme nous le verrons à la section 4.3, le rejet du mode de computation classique (computation de propositions par des règles) n'implique pas un rejet des définitions ramséennes des états mentaux caractérisant la psychologie du sens commun. Si l'on s'en tient à cette caractérisation de la psychologie du sens commun, le psychologue débutera ses recherches en proposant un modèle computationnel répondant à la définition ramséenne de ce processus mental.

Le psychologue a donc déjà entre les mains les grandes lignes de sa théorie finale. La perception transforme les entrées sensorielles en représentations mentales qui sont conservées dans la mémoire, le processus attentionnel filtre les informations de manière à permettre un traitement plus en profondeur des informations pertinentes, la faculté de langage permet de traduire les états mentaux en sons significatifs, etc. (Reed, 1999) Les manuels de psychologie cognitive sont construits selon un canevas reflétant les grandes

---

[48] Le terme anglais est '*ramsification*'. Voir Rey (1997, p.172-175) pour un exposé simple et schématique de la technique ainsi que de la formalisation en logique. Pour le développement original par David Lewis, voir Lewis (1966, 1970, 1972). Il existe d'autres formes d'explications fonctionnalistes, dont notamment la '*machine functionalism*', introduite par Putnam (1960, 1967) selon laquelle une description en termes de machine de Turing probabiliste complète la description d'un système cognitif, où un état mental est en fait identifié à un état de la machine de Turing universelle. Certains philosophes (notamment Rey, 1997) considèrent que ce type d'explication est compatible avec une description par les énoncés ramséens.

[49] Ce procédé permet d'introduire de nouveaux termes théoriques à partir d'un vocabulaire déjà développé uniquement en vertu du rôle causal qu'ils jouent. Ceci permet de définir les termes mentaux dans la perspective fonctionnaliste sans même approuver cette doctrine. Comme le souligne Rey (1997, p.176) ainsi que Block (1980), ce procédé est tout à fait applicable dans les autres sciences naturelles.

lignes de son cadre conceptuel et ce canevas ne varie généralement que dans la présentation des détails. Les thématiques abordées sont souvent exactement les mêmes et ce n'est que rarement que l'ordre de la présentation de celles-ci varie.

Le système cognitif est conçu comme une machine transformant des entrées sensorielles en comportements, mais puisque le mécanisme de ce système est complexe, aucune approche nomologique simple ne pourra rendre compte des relations entre les entrées sensorielles et les sorties comportementales. Cela ne signifie pas qu'il n'existe pas de lois en psychologie, au contraire (ex: la loi de Fitts), mais ces lois dénotent des régularités dont le ou les mécanismes sous-jacents responsables restent encore à déterminer. (Newell, 1990) Le psychologue devra spécifier dans quel état est le système lorsqu'il reçoit certaines stimulations sensorielles (quelles sont les interrelations entre les différents états mentaux à un moment *t* pour un système cognitif particulier). Cela appelle donc vers une explication plus mécaniciste que nomologique[50]. Les énoncés ramséens définissent alors la nature des états mentaux d'un point de vue local, mais leur interconnections définitionnelles permet de reconstruire le portrait global du système cognitif.

Une fois le schème conceptuel du niveau sémantique établi, le psychologue se tournera vers la réalisation computationnelle du processus mental étudié. En accord avec la stratégie explicative *top-down* (section 2.2.1), le psychologue passera au niveau computationnel pour déterminer le mécanisme de traitement de l'information qui expliquera le processus mental du niveau sémantique. Le pas suivant consiste donc à spécifier par quels sous-processus (expliqués par le modèle computationnel) les grandes facultés mentales et intellectuelles procèdent pour effectuer leurs fonctions. Cette tâche consiste à traduire le niveau sémantique en un vocabulaire computationnel : quel type de modélisation de processus computationnels permet d'expliquer les relations causales observées au niveau psychologique? Il faudra donc déterminer ce en quoi consiste la relation de réalisation explicative dans la stratégie explicative de la psychologie cognitive.

4.1.2 Relation de réalisation explicative dans la stratégie *top-down*

La relation de réalisation explicative la plus utilisée par les fonctionnalistes et les chercheurs en intelligence artificielle (excluant ceux qui s'intéressent au connexionnisme et

[50] Voir Bechtel & Abrahamsen (2005), Craver (2001), Wright & Bechtel (2007) pour une analyse plus approfondie de ce genre d'explication.

aux modèles dynamicistes[51]) lie les phénomènes cognitifs du niveau sémantique à des programmes donnés par des modèles de computation classique (traitement sériel de symboles discrets par des règles logiques, ou cognition par inférence). (Newell, 1980) Cette stratégie analytique de traduction du niveau sémantique au niveau computationnel (*top-down*) est parfois nommée déshomoncularisation. (Lycan, 1987) C'est de cette perspective dont il sera question ici et elle sera reprise plus loin lorsqu'il sera question de la relation de réalisation explicative proposée par P.M. Churchland pour lier les deux niveaux (section 4.2).

Robert Cummins (1983) offre l'analyse standard de ce qu'il nomme '*functional analysis*' (déshomoncularisation), soit un portrait détaillé de la méthode appropriée pour produire une explication psychologique, c'est-à-dire pour analyser les capacités (ou dispositions) d'un système cognitif. Les capacités sont, en fait, identiques aux états mentaux formalisés dans les énoncés ramséens. (Rey, 1997) Refusant l'approche nomologique utilisée entre autre par le béhaviorisme scientifique, Cummins défend une approche d'analyse interprétative qui fournit, pour un système cognitif donné, une description des capacités dont il est l'instanciation (soit la ramséfication). Le fonctionnement d'un système n'est pas explicable par des lois scientifiques reliant par une loi de covariation deux variables (*input/output*) parce que l'état interne du système affecte la fonction *input/output*. Ces états peuvent, pour un même *input*, transformer ce dernier en différents *outputs* et vice-versa, contrevenant donc à la notion de fonction mathématique simple utilisée dans les explications nomologiques traditionnelles (mais pas au concept de fonction utilisé ici, voir section 3.1 et 3.2). De plus, même si un rendu nomologique était donné, celui-ci ne serait pas un compte rendu des liens causaux qui entrent en jeu dans le système, c'est-à-dire qu'il y aurait un *mapping* entre les *inputs* et les *outputs*, mais cette projection ne tiendrait pas lieu d'explication causale (Cummins (1983), Pylyshyn (1984, 1989). L'analyse interprétative (ou analyse fonctionnelle) de Cummins procure au psychologue une stratégie pour formuler des descriptions explicatives d'un système donné sans toutefois en fournir une explication causale.

> « My topic, however, is psychological explanation. [...] It seems to me that most psychological explanation makes no sense when construed as causal subsumption but makes a great deal of sense when construed as analysis.

[51] Van Gelder (1998)

> Hence, an understanding of the analytic strategy is essential to an understanding of psychological explanation. » (Cummins, 1983, p.27)

L'analyse fonctionnelle se borne donc à atomiser une fonction psychologique en en dégageant des sous-fonctions pouvant la produire par leur interaction. Cette décomposition d'une disposition (d'un état mental ayant un énoncé ramséen), en organisant les sous-dispositions dans une '*flow-chart*' par exemple, doit permettre de rendre compte des manifestations observables de la disposition, c'est-à-dire de ses relations causales avec les *inputs*, les *outputs* et les autres états mentaux. Les capacités permettant de réaliser la disposition analysée seront donc moins complexes et fourniront un apport de connaissance considérable sur le système.

> « The explanatory interest of functional analysis is roughly proportional to (i) the extent to which the analyzing capacities are less sophisticated than the analyzed capacities, (ii) the extent to which the analyzing capacities are different in kind from the analyzed capacities, and (iii) the relative sophistication of the program appealed to - i.e., the relative complexity of the organization of component parts/processes that is attributed to the system. » (Cummins, 1983, p.30)

Cette stratégie est connue sous le nom de 'déshomoncularisation' : le psychologue étudiant une fonction cognitive cherchera à la caractériser d'abord dans toute sa complexité. Il divisera alors ce processus en d'autres (sous-processus) qui effectueront des tâches plus simples quoique tout de même complexes, à l'image de petits hommes intelligents qui vivraient dans nos têtes (homoncules) et qui se partageaient les tâches nécessaires pour effectuer la fonction cognitive de base (celle définie par un énoncé ramséen). (Lycan, 1981, 1987) Récursivement, ces processus plus simples (homonculaires) seront eux-mêmes divisés en d'autres sous-processus, et ainsi de suite jusqu'à ce qu'il n'y ait plus place pour des homoncules mais uniquement à des processus complètement idiots et relativement simples (Cummins, 1983; Dennett, 1975, Lycan, 1981, 1987). Il n'y a pas de régression à l'infinie d'homoncules qui expliquent d'autres homoncules (thèse anti-ryléenne) puisqu'un premier homoncule fonctionnel est divisé en homoncules effectuant des tâches plus simples. Le psychologue s'attend alors à heurter un plancher de simplicité, terminant ainsi le processus de déshomoncularisation et lui procurant une fonction aisément formalisable au niveau computationnel. (Lycan, 1987)

Cette stratégie explicative permet de rendre compte, sans faire appel à la structure physique du système, de son organisation et des activités que celle-ci lui procure,

organisation sans laquelle les parties matérielles qui la composent ne pourraient, par simple agrégat, produire les mêmes effets. Pour utiliser un exemple classique de cette stratégie, une analyse fonctionnelle d'une automobile porterait le 'voiturologue' (généralement représenté dans la littérature par un extra-terrestre curieux) à observer la machine en marche, à remarquer sa capacité à se mouvoir en faisant tourner ses roues. Il décomposerait alors cette disposition en sous-dispositions (e.g. pivoter les roues, dégager une énergie latente pour en faire de l'énergie motrice, un système pilote, etc.) et aurait ainsi une analyse fonctionnelle s'il organise ces sous-dispositions selon une organisation appropriée.

Une psychologie achevée serait une psychologie où tous les états mentaux du niveau sémantique seront représentés par des sous-processus caractérisés dans un vocabulaire de forme computationnelle. La méthode est explicitement *top-down* puisque l'on ne peut déterminer les processus de traitement de l'information qu'à la lumière des tâches à effectuer, tâches qui ne sont spécifiées que dans un vocabulaire psychologique, soit au niveau sémantique. Cette méthode est contraire à l'approche neurocomputationnelle *bottom-up* pour laquelle un réseau de neurone instancie déjà une série de processus computationnels et où le neuropsychologue doit, à partir de cette fonction, déterminer quel état mental il accompli (i.e. quelle signification psychologique donner au réseau de neurone). La stratégie de déshomoncularisation utilisée par la psychologie cognitive cherche usuellement à déterminer par quels processus propositionnels les opérations de transformation de l'information prendront place. (Newell, 1980, 1990; Fodor & Pylyshyn, 1988) Il ne sera pas question ici d'analyser les modèles faisant appel à la computation symbolique utilisés pour comprendre les processus cognitifs postulés par la psychologie cognitive puisque ceux-ci ne prendront pas de place dans le schème réductionniste de P.M. Churchland. (P.M. Churchland, 1989; P.M Churchland & P.S. Churchland, 1983) Toutefois, comme il en sera question à la section 4.2, puisqu'il y a réalisation multiple des processus cognitifs au niveau sémantique par des systèmes computationnels au niveau inférieur, il ne semble pas y avoir de justifications pour refuser une déshomoncularisation faisant appel aux modèles PDP. Cela aura son importance puisque c'est justement par cette possibilité que les états mentaux formalisés par les énoncés ramséens pourraient trouver réduction au niveau (neuro)computationnel.

La stratégie de construction de la relation de réalisation explicative proposée par la psychologie cognitive pour lier les états mentaux du niveau sémantique à des processus de traitement de l'information au niveau computationnel a maintenant été définie : c'est un procédé de déshomoncularisation des dispositions définies par des énoncés ramséens. Toutefois, pour compléter l'analyse, il faudra déterminer la nature des *inputs* et *outputs* au niveau sémantique pour compléter le détail de l'explication fonctionnelle à ce niveau et permettre ainsi de débuter notre analyse de la relation de réalisation explicative entre les niveaux sémantique et computationnelle telle que P.M. Churchland la conçoit.

4.1.3 Analyse des *inputs* et *outputs* au niveau sémantique

Le cœur de l'explication fonctionnelle est constitué des états mentaux postulés par la psychologie du sens commun intégrés au corpus théorique de la psychologie cognitive. L'énoncé ramséen définit ainsi les relations causales entre les états mentaux et les *inputs* et *outputs* au niveau sémantique. Du point de vue de l'explication fonctionnelle globale, les *inputs* ne sont pas des stimulations simples (ex : vibration dans l'air) à partir desquelles construire des perceptions qui seront subséquemment cogitées mais plutôt des perceptions déjà construites à partir d'une ontologie ordinaire (du sens commun). Pylyshyn (1984) en caractérise la nature dans ses grandes lignes :

> « What serves as the functional stimulus depends on how a person interprets the situation (for example, the stimulus in the pedestrian-automobile example is accident; but, of course, if that person is told it is a rehearsal for a television show, the stimulus is no longer accident but rehearsal and engages the habits appropriate for that category). » (Pylyshyn, 1984, p.9)

Bien que Pylyshyn utilise le terme 'stimulus', il est clair qu'il entend ce terme dans une autre perspective que neurobiologique. Il n'est pas question non plus de la structure informationnelle complète du stimulus physique. Une telle description est suffisante au niveau computationnel qui ne s'enquière pas de la signification des représentations sur lesquelles procède le système cognitif, mais ultimement il faudra savoir déterminer quels *patterns* d'information seront intéressants pour l'explication complète des phénomènes mentaux au niveau sémantique. Une description complète des processus de computation basée sur le format mathématico-logique de la stimulation n'est pas suffisante : comme l'illustre bien la citation précédente, ce qui importe pour l'explication psychologique c'est

surtout le *sens* (voir même la référence) qui sera donné à cette stimulation, i.e. « qu'est-ce que le sujet perçoit? ».

Pylyshyn (1984, p.2-3) donne pour exemple un individu qui assiste à un accident de voiture et courre à une cabine téléphonique pour alerter des secours. Les entrées sensorielles intéressantes dans cette situation ne sont pas cernées par la description des informations sensorielles captées par les divers sens mais la perception générale que le système cognitif a de sa situation. Pour comprendre pourquoi l'individu est accouru à la cabine téléphonique plutôt que de danser ou de ne rien faire, il faudra comprendre ce que l'individu a perçut *qua* situation. Dans l'exemple, ce qui importe pour le psychologue (dans une perspective sémantique), c'est que l'observateur a perçu un accident. Il existe une vaste variété d'entrées sensorielles qui auraient pu donner exactement la même explication causale au niveau psychologique. Un sourd aurait été tout autant témoin de l'accident qu'un individu à l'oreille fine et il aurait pu agir de manière quasi-identique. Ce qui importe selon Pylyshyn (ainsi que Fodor, 1975, 1980; et Dennett, 1978, 1987, 1991b), c'est le *pattern* de l'information distinguée par le sujet *dans la stimulation sensorielle*. Ce *pattern* joue un rôle causal : c'est la perception de l'accident qui a poussé l'individu à courir à la cabine téléphonique, pas l'ensemble des stimulations l'affectant au même moment. Spécifier la nature structurelle du stimulus physique ne donne pas une explication intéressante au niveau psychologique parce que le sens de ce qui est perçu ne serait pas disponible dans cette seule structure. De nombreuses variations dans la stimulation physique pourraient être provoquées et le sujet y percevrait tout de même un même *pattern* (le même accident). Il y a, en quelques sortes, une réalisation multiple des stimuli qui peuvent donner lieu à une perception (Dennett, 1987, p.25-26), et cela rend le point de vue purement informationnel peu intéressant au niveau psychologique :

> « This is not to deny that some causal chain connects stimulation and perception. The point is simply that there exist regularities stateable over properties of the stimulation itself. This, in turn, is true because it appears that virtually no physical properties (including arbitrarily complex combinations and abstractions over such properties) are necessary and sufficient for the occurrence of certain perceptions; yet it is these perceptions that determine psychological regularities in behavior. » (Pylyshyn, 1984, p.15)

L'*input* 'perception' est donc spécifié en faisant abstraction du détail des processus de perception. Ce qui importe au niveau psychologique ce n'est pas par quels processus le

signal physique est interprété mais le résultat de ceux-ci : le sens qui lui sera donné. Le psychologue inclus donc un aspect normatif dans la description des *inputs* : le système cognitif fonctionnant adéquatement percevra un accident si certaines conditions environnementales (et ontogénétiques et autres clauses *ceteris paribus*) sont respectées. La perception est donc une entrée sensorielle déjà porteuse d'un sens. Ce ne sont pas les contractions stomacales qui me poussent à manger, c'est la faim que je perçois en moi.

Puisqu'il y a réalisation multiple des stimulations pour un même type de perception (mon estomac peut se contracter de différentes manières mais cela m'indiquera toujours une seule chose : que j'ai faim), il faudra établir une relation permettant de relier systématiquement les stimulations différentes, dénuées de sens, aux perceptions du niveau sémantique. Le psychologue cognitif, usant de la méthode *top-down*, débute alors avec le cadre conceptuel de la psychologie du sens commun pour ensuite déterminer quelles structures informationnelles doivent être présentes (sont nécessaires) pour donner lieu à un type de perception donné. (Pylyshyn, 1984) Toutefois, ce ne sont pas uniquement les éléments pertinents à la reconnaissance de la situation perçue qui devront être indiqués dans la structure informationnelle du stimulus, mais aussi une large part des attentes du système cognitif. Les psychologues ont pris conscience qu'il y avait un *feedback* important de la part du système cognitif dans l'interprétation perceptuelle des signaux physiques. Ce phénomène est connu sous le nom d'apport '*top-down*'[52] : l'état du système cognitif détermine en partie la manière dont un signal sera interprété. Un système cognitif dans un état donné interprétera une stimulation d'une manière différente d'un même système mais dans un autre état. Le fameux exemple de l'image du lapin-canard est un exemple typique de ce phénomène. (Reed, 1999) L'idée de connaissances tacites (*tacit knowledge*) en est un autre. (Pylyshyn, 1981) De plus, certaines 'perceptions' font appel à des dimensions très abstraites. Dennett (1971) suggère qu'un individu ne fait pas simplement qu'entendre des émissions sonores lorsqu'une autre personne lui parle mais perçoit un message. De même, dans le cadre des expériences en psychologie cognitive, les chercheurs présentent au patient une tâche à accomplir (un problème à résoudre) (voir Newell & Simon (1972) par exemple). Le psychologue prend alors pour acquis que le patient a 'compris' le problème (la tâche à

[52] À différencier du *top-down* méthodologique, l'apport *top-down* au niveau cognitif consiste en l'impact qu'ont des processus cognitifs de haut niveau (ex : imagerie mentale, attentes, etc.) sur des processus cognitifs plus proximaux aux sens (ex : traitement visuel, auditif, etc.). Un bon exemple est le processus d'attention qui sélectionne dans les entrées sensorielles ce qui sera traité et surveillé et rejette les informations périphériques qui ne seront pas ou peu analysées. (Reed, 1999)

effectuer) ou le message (ex : indications du chercheur) et cherchera alors à le résoudre (l'expérience de Kosslyn (1980, 1997) en est un exemple très simple)[53]. Ainsi, pour la psychologie cognitive, les *inputs* ne sont pas simplement des structures perçues dans l'environnement mais peuvent aussi comporter des aspects plus cognitifs, voir même interpersonnels (ex: l'autorité du chercheur dans l'expérience de Milgram (1974)) ou sociaux (ex: manifestation agressive ou passive).

La caractérisation des perceptions est donc imbue de la théorie psychologique du sens commun. La raison est simple : au niveau sémantique, l'*input* est défini comme le produit des processus perceptif et des attentes (elles-mêmes spécifiées par une théorie du mental), certains mêmes de processus cognitifs de haut niveau. Le psychologue invoque ainsi une clause normative déterminant qu'est-ce que le système cognitif devrait avoir perçu s'il fonctionne adéquatement.

> « Similarly, the only inputs of interest among the possible stimulations patterns that might impinge on the organism are those that the organism is designed to use. Not only must they not break or jam the organism's system; they must also be described not arbitrarily but in accordance with their own functional forms, those forms that accord with the properly functioning organism's way of using its inputs. » (Millikan, 1993, p.152)

Si l'on présente un livre à un patient et qu'il nous répond que c'est un lion, on inférera qu'il n'a pas bien perçu l'objet. Si l'on demande à un patient de déterminer si un élément se trouve sur une carte (expérience de Kosslyn (1980, 1997)) et qu'il répond « jaune », on insinuera que le patient n'a pas compris la situation qui lui est présentée. Pour déterminer ce qui est une perception adéquate, il faut déjà savoir (du moins dans les grandes lignes) comment un individu *normal* répondra à celle-ci. De plus, lorsque le psychologue cherche à déterminer par quels processus computationnels les perceptions sont construites, il se base sur cette caractérisation normative (ex : tout bon système perceptuel se trompera là où il y a des illusions d'optiques)[54]. La 'déshomoncularisation' des processus perceptifs se fera à la lumière de la notion *a priori*[55] d'une bonne perception et

---

[53] Voir von Eckardt (1993) pp.32-45 pour un aperçu succinct de l'expérience de Kosslyn ainsi que Pylyshyn (1981).

[54] von Eckardt (1993) va plus loin en spécifiant que les sciences cognitives sont concernées spécifiquement par « the study of the human adult's normal, typical cognition » (p.6), ce qu'elle nomme la clause ANTCOG.

[55] *A priori* parce que la perception normale n'est pas déterminée par les recherches psychologiques ou neuropsychologiques à proprement parler. Les clauses normatives déterminant ce qui fait d'une perception une perception adéquate ne sont pas recherchées par voies empiriques mais bien déterminées par le sens commun. Cela ne veut pas dire que ces clauses normatives peuvent être raffinées ou agrémentées par les recherches empiriques (ex : illusions d'optiques).

les expériences menées en laboratoire pour déterminer ces processus se fera toujours avec un ensemble prédéterminé de bonnes réponses à un stimulus donné. (von Eckardt, 1993)

Du côté des *outputs*, il est question au niveau sémantique non plus de réactions motrices ou de signaux moteurs mais de comportements et d'actions. Millikan (1993, chapitre 7 et 8) distingue les comportements des réactions motrices en spécifiant qu'un comportement a une fonction (au sens biologique du terme)[56]. La fonction biologique d'un mécanisme est définie par les effets pour lesquels le mécanisme a été sélectionné puis reproduit. Seules les réactions motrices ayant une fonction biologique en ce sens seraient intéressantes pour le psychologue. Ainsi, un comportement qui aurait une fonction biologique devra donc être avoir été sélectionné (naturellement). Cette thèse pousse à penser que les comportements et les processus cognitifs et moteurs qui les produisent se développent normalement chez les individus d'une même espèce. Le mécanisme cognitif pour produire ces comportements serait relativement similaire parmi les individus d'une même espèce.

Il n'est pas à propos ici d'explorer plus à fond la nature des fonctions biologiques ni non plus les conséquences que Millikan en infère, mais ce qui est important pour notre analyse c'est de remarquer que ce ne sont pas toutes les réactions motrices, ni non plus tous les détails de ces activités, qui sont pertinents pour une caractérisation des comportements. Fodor (1975, chapitre 2) va encore plus loin en proposant une classe distincte de comportements : les actions. Celles-ci seraient expliquées quand les croyances, les désirs et l'intention même d'agir d'une manière déterminée sont (adéquatement) causalement liés, i.e. qu'un comportement est une action si elle a été causalement initiée par l'intention de la faire. Cette classe de comportements nécessiterait donc une théorie du mental pour pouvoir l'individuer. Davidson (2001) l'illustre bien par ce passage : « Whenever someone does something for a reason, therefore, he can be characterized as (a) having some sort of *pro attitude* toward actions of a certain kind, and (b) believing (*knowing, perceiving, noticing, remembering*) that his action is of that kind. » (Davidson, 2001, pp.3-4, l'emphase est la

[56] Le concept étiologique de fonction proposé par Millikan (1984a) (voir aussi Wright (1973)) ne doit pas être confondu avec la notion de fonction utilisée ici, soit celle défendue par Cummins (1975, 1983). La conception étiologique de fonction détermine 'à quoi sert' un système produisant certains effets selon que le système a été construit et reproduit parce qu'il produisait ces effets. Le concept de fonction de Cummins consiste plutôt à déterminer 'à quoi sert' une partie d'un système, quel est la contribution causale d'un élément dans l'économie du système auquel il participe.

computationnels d'une certaine profondeur (là où la perception fournie des représentations pouvant jouer un rôle au sein des théories psychocognitives jusqu'au 'moment' où les intentions pour l'action sont utilisées par les processus moteurs).

La théorie psychocognitive du niveau sémantique a maintenant été caractérisée sous sa forme fonctionnelle : les *input*s sont des situations déjà analysées, les *outputs* des actions déjà pleines d'intentions et les états mentaux sont définis par leur rôle causal dans l'économie du mental, rôle formalisé par les énoncés ramséens dont les mécanismes sous-jacents sont décortiqués par le procédé de déshomoncularisation. Cela représente donc un portrait schématique de la théorie psychocognitive (T_r) qui devra entrer en relation de réduction avec son image neuropsychologique (I_b). La prochaine section aura pour objectif d'expliquer comment P.M. Churchland conçoit la construction d'une image neuropsychologique (I_b) à laquelle la théorie psychocognitive devra se comparer pour pouvoir déterminer de son sort (voir les sections 1.3.1 et 1.3.2). On y procédera d'abord en spécifiant comment un état neurocomputationnel peut contenir de l'information *à propos* de l'environnement d'une créature et comment ceux-ci sont liés à l'organisation des neurones dans le cerveau (section 4.2.1). Cela donnera donc les outils nécessaires pour examiner la relation de réalisation explicative entre les *inputs* et *outputs* au niveau neurocomputationnel et neuropsychologique (section 4.2.2). Par après il sera question de la relation de réalisation explicative entre les processus neuropsychologiques per se et leur nature neurocomputationnelle.

4.2 Réduction du niveau sémantique au niveau computationnel

Il a été question à la section 2.1.1 de l'une des problématiques fondamentales du niveau sémantique: une bonne théorie psychologique devra fournir une explication du lien qui unit la *structure du monde* dans lequel vivent les créatures cognitives et les *représentations* utilisées par celles-ci pour y circuler adéquatement. Cette problématique demande donc à ce que le psychologue fournisse une explication du contenu sémantique des représentations utilisées par la cognition d'une créature et comment ces représentations reflètent la structure réelle de son environnement. Dans l'optique des niveaux explicatifs, cela signifie qu'il faille déterminer comment une représentation purement formelle donnée au niveau computationnel (ici des vecteurs d'activation, voir section 3.3.2) contient de

l'information à propos d'une chose et comment cette information joue le rôle de contenu sémantique des représentations au niveau psychologique.

P.M. Churchland & P.S. Churchland (1983), ainsi que P.M. Churchland (2007, chapitre 2, pp.28-33), avancent la thèse que le contenu des représentations (leur valeur informative) provient de l'utilisation par les êtres vivants de la structure disponible dans le flot de l'énergie environnante[59]. En fonction de la définition donnée à l'expression 'créature cognitive naturelle'[60], P.M. Churchland nous indique que l'information utilisée par les systèmes cognitifs est en fait une structure de flot d'énergie dans l'environnement :

> « Specifically, a creature that learns about the world is a creature that exploits the low-entropy internal structure or information that it already possesses, in such a fashion that, if the creature is placed in an environment with an information-rich energy flow, it comes to embody additional information-bearing structure about its environment, typically in the form of a progressively rewired brain. » (P.M. Churchland, 2007, p.30)[61]

Il est important de noter que l'information ainsi captée est redistribuée en structure neurologique (soit en restructuration des réseaux de neurones), permettant ainsi de représenter l'information fournie par l'environnement par le détail de la structure du réseau. L'information ici n'est pas tant dans le vecteur d'activation mais plutôt dans la configuration des connexions du réseau de neurone qui l'interprétera. Ainsi l'information n'est pas décryptée par les réseaux neuronaux : ce sont les vecteurs d'activation fournis par les entrées sensorielles qui activent dans les réseaux de neurones l'information incarnée dans leur structure (*embodiement*). (P.M. Churchland, 2007, chapitre 8) Il est en lieu d'analyser plus en détail comment, selon P.M. Churchland, les réseaux de neurones représentent, par leur structure d'activation, l'information fournie par le monde et d'éclairer les relations entre l'environnement et les représentations cognitives qu'une créature a de celui-ci. Bien sûr, il ne peut être question ici que de l'interprétation de Paul M. Churchland vis-à-vis les réseaux de neurones (toute critique de cette approche ne pouvant pas être explorée ici, mais voir par exemple Laakso & Cottrell (2006)).

---

[59] Simplement, cela signifie qu'il existe de l'organisation et des régularités dans l'environnement et que les créatures exploitent celles-ci pour s'organiser et survivre.

[60] Voir la citation de P.M. Churchland & P.S. Churchland (1983) donnée à la page 17 de ce mémoire.

[61] Le même phénomène serait constitutif des systèmes biologiques dans leur généralité : leur structure physique et procédurale serait elle-même issue de cette exploitation de la structure de l'énergie dans des systèmes à basse entropie. Ce principe définirait d'ailleurs la nature des être vivants. (P.M. Churchland, 2007, p.29)

4.2.1 Les prototypes comme système de représentation[62]

Un réseau de neurone artificiel et naturel (avec l'utilisation adéquate de la relation de réalisation explicative) peuvent être représentés dans tous leurs états computationnels par un espace à *n*-dimension (où *n* est le nombre de neurones ou d'unités d'activation présents dans le réseau[63]) et où chaque dimension est divisée par les différents états d'activation possible pour le neurone en question. Chaque point dans cet espace est défini par une valeur d'activation que prendra chacun des neurones[64]. Un vecteur d'activation à la couche d'entrée (*input layer* recevant un *input*) est donc représenté par un point dans l'espace définissant l'état d'activation de cette couche. On peut donc dire que ce point dans l'espace représente l'information issue de la transduction effectuée après la réception de stimulations externes par les senseurs. Dans la couche intermédiaire, le vecteur d'activation aura été transformé en un autre vecteur d'activation, lui-même représentable dans un espace reflétant le niveau d'activation des neurones de cette couche. (P.M. Churchland (2007, chapitre 8, pp.144-147))

En entraînant des réseaux de neurones artificiels, les chercheurs en PDP (*parallel-distributed processing*) se sont aperçus qu'un réseau pouvait discriminer différents *inputs* (produisant une réponse particulière pour une classe d'*input* et une autre réponse pour une autre classe d'*input*) en partitionnant son espace d'activation neuronale au niveau de la couche intermédiaire en sous-espaces dans lesquels tombaient respectivement les activations de chacune des classes de vecteurs afférents. Les neurones effecteurs répondraient différentiellement selon la classe d'*input*. Bien sûr, les réseaux étaient entraînés à discriminer les classes d'*inputs*; la question n'étant pas ici comment entraîner les réseaux pour y arriver mais plutôt de déterminer comment les réseaux s'organisent pour représenter de l'information. Dans les sous-espaces d'activation discriminant des classes de vecteurs, il existe un autre sous-espace pour lequel tous les vecteurs tombant dans ce

---

[62] Cette section est une analyse directe de P.M. Churchland (1989, chapitres 5 et 9) ainsi que de P.M. Churchland (2007, chapitre 8).

[63] Fodor & Lepore (1996) ont critiqué cette approche en croyant que les espaces d'activations étaient en fait des espaces sémantiques dont les dimensions représenteraient en fait des concepts simples (primitifs) à partir desquels les concepts complexes (prototypes) étaient constitués. Cette confusion a amorcé un débat entre F&L et P.M. Churchland (repris dans McCauley (1996)) dont le fond peut être mis de côté ici puisque P.M. Churchland (2007, chapitre 8) permet d'éviter celui-ci en spécifiant clairement que les dimensions sont les états d'activation des unités (neurones). C'est d'ailleurs cette conception plus récente qui sera explorée dans les pages qui suivent.

[64] Voir P.M. Churchland (1989, p.104) pour une représentation graphique d'un espace à trois dimensions (à trois neurones).

dernier produisent une activation maximale de la réponse discriminatrice des neurones constituant la couche de sortie : le vecteur d'activation répondrait très fortement à la structure typique des vecteurs constituant la classe d'*input*. Cet espace est nommé prototype (ou attracteur). Pour un même réseau de neurone (entraîné), il existerait plusieurs prototypes (un prototype par classe d'*input* à discriminer). (P.M. Churchland, 1989; McClelland et al., 1986)

C'est à partir de ce point que se joue le saut au niveau sémantique : les relations spatiales (dans l'espace à *n*-dimension de la couche observée) déterminent la structure informationnelle captée par les réseaux de neurones, c'est-à-dire qu'un ensemble de prototypes joueraient le rôle de concepts. Chaque prototype détecte un élément dans la structure du vecteur d'activation dans la couche d'entrée et ce sont les relations spatiales entre les différents prototypes dans l'espace à *n*-dimension qui déterminent la structure informationnelle dont le réseau de neurone est imbu. Le passage du vecteur d'activation dans ce réseau permettrait d'en dégager les informations auxquelles le réseau est sensible. (P.M. Churchland, 2007, p.150) Il est à noter ici que cette conception ne fait absolument pas appel à la provenance de la stimulation sensorielle ayant été transformée en vecteur d'activation par la transduction, ni non plus au résultat moteur qui en découlera. (P.M. Churchland (2007, p.) Un prototype a la charge informationnelle qu'il a selon sa position relative aux autres prototypes dans un même réseau de neurone.

Cela reflète la nature purement formelle (syntaxique) des opérations computationnelles, ce qui n'a pas toujours été la position de P.M. Churchland (voir McCauley (1996, pp.275-276)). Le prototype n'a pas de contenu sémantique en vertu d'une relation au monde mais plutôt selon sa position dans l'espace à n-dimensions du réseau dans lequel il figure. P.M. Churchland (2007, chapitre 8) en arrive à conclure à plusieurs conséquences épistémologiques. Cette caractérisation, aidée de quelques outils (mesure de la similarité des espaces prototypiques (Laakso & Cottrell, 2000) permet d'inférer des mesures de similarité conceptuelle (selon les distances des prototypes dans l'espace à n-dimension d'un réseau donné), de contredire la thèse lockéenne de primitifs tirés de la perception (tous les prototypes sont complexes et indécomposables (voir les textes de P.M. Churchland dans McCauley (1996), particulièrement dans les réponses aux critique de Fodor et Lepore[65] ainsi que de réfuter la thèse propositionnelle de la

[65] Voir la note 60 pour les références à cet échange.

computation (P.M. Churchland & P.S. Churchland, 1983). D'ailleurs il est intéressant de noter que la méthode pour mesurer la similarité des concepts entre deux réseaux de neurones permet d'unir (au moins d'une manière) la réalisation matérielle multiple de différents réseaux neuronaux computant la même fonction (Laakso & Cottrell, 2000).

La relation entre l'environnement et les représentations mentales (prototypes) consisterait alors non pas en une histoire causale de la formation de la représentation (Putnam, 1975) ni non plus à partir d'un langage de l'esprit (Fodor (1975)), mais bien en une relation d'isomorphisme des éléments constitutifs de l'information contenue dans l'environnement (dans les stimulations sensorielles affectant adéquatement l'appareillage cognitif d'une créature) et la structure des éléments détectés par les réseaux de neurones (et son complexe de prototypes). (P.M. Churchland, 2007, chapitre 8) P.M. Churchland conçoit l'attribution sémantique d'un réseau de neurone comme une fonction de *mapping* de ces deux ensembles structurels. Contre Fodor (1975, chapitre 1) qui prétend que tout les systèmes de représentations requièrent un symbolisme et donc un langage pour exprimer celui-ci, la notion de *mapping* de P.M. Churchland ne fait appel qu'à une équivalence structurelles (de relations entre différents éléments) entre les prototypes et les éléments physiques réels des entités perçues. Cela ne requière pas ni un langage de l'esprit, ni un format propositionnel (symbolique) de la computation. De plus, les structures en question ne sont pas spécifiquement liées à des vecteurs provenant de stimulations sensorielles. Un réseau local pourrait même procéder à l'interprétation de cartes structurelles représentant des éléments abstraits (ex : système politique ou conception morale). (McCauley, 1996, chapitres 16-17) Cela signifie de plus qu'un même réseau de neurones pourrait, par une même valeur d'activation globale, représenter deux choses différentes. Le réseau n'est sensible qu'à la structure du vecteur d'activation, pas à sa provenance, et le neuropsychologue ne pourra déterminer le contenu de la représentation qu'une fois le système en action dans une situation donnée.

> « It is the map's internal structure that makes it the specific portrayal that it is. And it is the existence of an abstract, relation-preserving projective mapping, from some external domain to that map, that makes it a good or an accurate portrayal of that external domain. Causal connections enter the picture only if, and only when, the map is finally put to some use or other. »
> (P.M. Churchland, 2007, p.156)

Le '*mapping*' des prototypes à des éléments structurels de la réalité va bien plus loin que la simple reconnaissance d'objets ou l'apprentissage de concepts. P.M. Churchland

prétend même que ce serait le même type d'explication qui permettrait de rendre compte des fonctions motrices (*output*) ainsi que de toutes les autres fonctions intermédiaires. (P.M. Churchland, 2007, p.158) Les fonctions cognitives seraient en fait une interaction de telles cartes structurelles et chaque étape dans la cognition serait en fait la mise en relation de ces différentes cartes structurelles (partition d'espaces d'activation) à la lumière de l'organisation des connexions et de la valeur d'activation des neurones. (voir P.M. Churchland, 1989, pp.82-96 pour un exemple de ce type de *mapping*).

Il est clair à ce point qu'il manque à la théorie de P.M. Churchland un lourd support empirique que seuls les résultats des recherches neuroscientifiques permettront de fournir pour préciser la nature de ces interactions. P.M. Churchland n'en est pas inconscient : il ne prétend que faire une esquisse des processus neurocomputationnels de base pour permettre d'entrevoir un portrait plus global des éventuels développements neuroscientifiques et de leurs conséquences philosophiques, épistémologiques et scientifiques. Mais cette esquisse est faite à partir du principe de base de l'interprétation sémantique des réseaux neuronaux : soit, selon la perspective de P.M. Churchland, par les prototypes. Cela nous permet d'en dégager quelques conséquences pour le projet réductionniste, particulièrement vis-à-vis la réduction des *inputs* et *outputs* du niveau sémantique tels que les théories psychocognitives en font usage.

4.2.2 Réduction des *inputs* et *outputs* du niveau psychologique au niveau computationnel

Avec ce qui a été dit dans les sections précédentes, il est possible d'explorer la relation de réalisation explicative qui lie les *inputs* et *outputs* caractérisés à partir du niveau computationnel à ceux du niveau sémantique pour ainsi en produire une image (I_b) neuropsychologique. Cette section est particulièrement technique, mais elle a pour objectif de mieux cerner le détail de la relation de réalisation entre les *inputs/outputs* au niveau neurocomputationnel et sémantique. Elle permettra, par ailleurs, d'éclairer l'état actuel de la relation de réduction entre les théories psychocognitives et neurobiologiques (ce qui sera exposé dans la section suivante).

Pour en arriver à traduire les *inputs* sensoriels caractérisés au niveau neurobiologique et neurocomputationnel (voir section 3.3.3), le neuroscientifique devra

reconstruire la structure informative des éléments constitutifs de l'environnement extérieur aux systèmes cognitifs (une sorte d'ontologie structurelle d'éléments pertinents à la reconnaissance) et ainsi permettre de produire une relation de *mapping* entre ceux-ci et les éléments structuraux distingués par les réseaux de neurones. Une large part de ce travail consistera à déterminer comment les différents organes sensitifs (capteurs sensoriels) en viennent à produire des vecteurs d'activations porteurs d'information pertinente pour la computation, i.e. qu'il faudra déterminer la *fonction de transduction* des différents organes sensitifs. (Neisser, 1967, 1976)

La fonction première des sens est de capter l'information concernant l'environnement du système cognitif et de la transformer sous une forme de représentation pouvant être traitée par celui-ci. Cette fonction est nommée 'transduction' : un phénomène physique affecte le capteur sensoriel et celui-ci, en vertu de sa structure (ou mécanisme), traduit le signal pour qu'il puisse être manipulé par les processus mentaux subséquents Pour un même type de signaux physiques, il est possible d'utiliser une grande gamme de capteurs sensoriels. Pour caractériser l'*input* au niveau computationnel, le psychologue doit pouvoir déterminer la structure de l'information qui pénètre dans le système cognitif. Ultimement, même si la fonction de transduction dépend de la structure du capteur sensoriel, il doit exister une relation constante assurant que l'information pertinente contenue dans les signaux émis par l'environnement sera captée puis utilisée par le système cognitif. C'est ce que Akins (1996) nomme la clause de la 'préservation de la structure du stimulus' (PSS) : « The relevant structure of the external events or properties must also be preserved by the sensory signals : the representational relations among the sensory signals must mirror the relevant relations in the sensed domain. » (Akins, 1996, p.343) Sans cette condition, on voit mal en quoi nos sens nous informeraient sur la réalité externe et comment l'évolution des espèces 'cognitives' aurait bien pu faire sens.

Cette clause spécifie ce qu'il nous faut entendre par 'stimulation' au niveau neurobiologique : le signal affectant le capteur, peu importe la nature de celui-ci, est traduit en signal (décrit au niveau computationnel) utilisable par le système cognitif. L'*input* au niveau computationnel (neuro- ou pas) est donc la structure informationnelle du signal physique fournissant au système cognitif les informations pertinentes aux comportements conséquents. C'est d'ailleurs ce qui permet de déterminer la nature des sens et de les différencier. (Neisser, 1967, 1976) La clause PSS indique aussi que toute la structure du

médium physique, porteuse d'information, n'est pas pertinente. La chaleur dégagée par l'absorption des rais lumineux ne concerne pas le signal visuel, le médium dans lequel l'onde de vibration affecte les capteurs auditifs non plus (dans les limites où le capteur reste fonctionnel, mais cela n'est pertinent qu'au niveau physique). Ce sont ces *patterns* dans la structure des stimulations qui permettent de distinguer un signal informatif du 'bruit' environnant et il en incombe au psychologue (de la perception) de déterminer quelles opérations d'analyse des stimulations permettra de soutirer des *inputs* bruts les informations pertinentes pour la construction de percepts adéquats. Ce sont ces *patterns* et les événements extérieurs auxquels ils sont liés qui spécifieront la nature sémantique des processus de perception (voir Marr, 1982). La spécification du *input* sous forme de stimulation est importante car ce n'est que par elle que le psychologue pourra comprendre ce sur quoi le processus de perception s'applique dans son travail.

P.M. Churchland insiste que les éléments structurels ne sont pas limités à ceux que la perception détecte. Il y aurait des éléments structurels abstraits dont les processus cognitifs de haut niveau seraient particulièrement aptes à détecter et à y réagir pour assurer une réponse adéquate en temps et lieu.

> « [...] there is no reason for the brain to show any such preference in what it constructs maps of. Abstract state spaces are just as mappable as concrete physical ones, and the brain surely has no advance knowledge of which is which. We should expect it, rather, to evolve maps of what is functionally significant, and that will frequently be an abstract state space. » (P.M. Churchland, 1989, p.96)

Cela pousse à croire qu'il n'y a pas de contradiction *a priori* motivant le rejet des *inputs* pour lesquels il y a une utilisation implicite du portrait théorique qu'offre la psychologie cognitive du fonctionnement global des systèmes cognitifs. L'apport sémantique *top-down* des processus de haut niveau sur les perceptions pourrait être déterminé par une relation de réalisation explicative suffisamment avancée. Les perceptions riches en contenu 'mentaliste' seraient des représentations prenant place à un certain niveau de profondeur dans les processus cognitifs (ce qui est une thèse très peu controversée). Il n'est toutefois pas clair comment les dimensions structurelles des apports *top-down* permettraient de rendre compte de la notion de 'problème' ou d'autres 'perceptions' plus abstraites. Nous sommes en droit de supposer que les recherches en sciences cognitives offriront éventuellement un portrait cognitif des apports *top-down* plus

'abstraits' ou en viendront peut-être à les rejeter (élimination)[66], la question ici étant seulement décidable empiriquement.

Les différentes réactions motrices seraient computées de la même manière que les entrées sensorielles. P.M. Churchland (1989, pp.107-109) illustre comment incorporer l'aspect sensori-moteur dans une analyse par espace d'activation (*motor state-space*). Ce sont les prototypes moteurs qui définiraient les *patterns* de comportements récurrents (ex : locomotion) par leur structure relationnelle dans l'espace d'activation. Ceux-ci auraient bien sûr systématiquement un aspect temporel intégré (un comportement se déroulant toujours dans le temps). Cela ne cause pas de problème, indique P.M. Churchland. L'aspect temporel est identifié à un trajet dans l'espace d'activation : le prototype temporel correspond alors à ce trajet.

> « The graceful step cycle of the galloping cat will be very economically represented by a closed loop in that joint-angle state space. If the relevant loop is specified or "marked" in some way, then the awesome task of coordinated locomotion reduces to a simple tracking problem: make your motor state-space position follow the path of the loop. » (P.M. Churchland, 1989, p.107)

La thèse des prototypes moteurs semble appuyée par la définition que donne Millikan (1993) des comportements et de leur fonction. Si cette dernière est déterminée par la rétention d'une génération à une autre de ces dispositifs comportementaux en vertu des effets de ces comportements, il semble adéquat de penser que ceux-ci sont représentés dans l'économie d'un système cognitif. Toutefois, il n'est pas clair comment identifier les actions à des prototypes. Plusieurs actions sont uniques, tout comme plusieurs intentions sont uniques, et il ne semble pas que les prototypes puissent répondre à de telles actions puisqu'elles n'ont pas été apprises. P.M. Churchland ne spécifie pas non plus ce qui fait le propre des actions en vue d'un but à atteindre (ex : résoudre un problème). Si celles-ci font appel à des croyances et des désirs (et à d'autres concepts de processus mentaux), il n'est pas clair comment spécifier l'ontologie des actions à partir d'une relation de réalisation explicative au niveau computationnel. Cela requiert une théorie avancée des processus cognitifs de haut niveau et, encore une fois, elle ne pourra être fournie que par voies

---

[66] La possibilité de rejeter (éliminer) de l'ontologie des sciences cognitives certaines classes d'*input* me semble particulièrement intéressante puisque l'éventualité d'une telle élimination pourrait remettre en question la simple valeur théorique de ces entités 'mentalistes' extérieure aux systèmes cognitifs. Ces répercussions devraient se faire ressentir plus amplement dans les sciences humaines. Mais cet impact de l'éliminativisme n'est pas dans la lunette de visée de la présente analyse.

empiriques. Seulement alors pourrons-nous comparer la structure des théories du mental rendant compte des actions uniques. Cela n'est toutefois pas une objection au projet de P.M. Churchland; cela ne fait qu'illustrer l'immaturité des théories neurocognitives. De plus, il ne semble pas, à ce point, y avoir de problèmes conceptuels qui indiqueraient l'impossibilité de réduire les théories psychocognitives aux théories neurobiologiques. Une incompatibilité pourrait émerger dans la tentative de réduire une théorie psychocognitive particulière au corpus neuroscientifique, mais cela est du ressort des sciences empiriques (et donc des recherches futures). Il est clair que les recherches neurocognitives ont encore du chemin à faire et que la réussite dans la réduction des actions aux *outputs* au niveau computationnel n'est pas encore à portée de main.

4.3 Réduction des états mentaux au niveau sémantique

Il est maintenant temps de procéder au point chaud du débat concernant le réductionnisme. Les chapitres I, II et III ont proposés un cadre conceptuel pour évaluer les critères de réussite de la réduction et la forme que devrait prendre le schème de l'explication réductrice. Dans la section précédente, il a été question de la réduction des *inputs* et *outputs* du niveau sémantique au niveau neurocomputationnel. Il ne reste plus qu'à déterminer si les états mentaux du niveau sémantique, postulés par la psychologie cognitive à l'aide du cadre conceptuel de la psychologie du sens commun, pourront ou non trouver place à réduction. Il faut toutefois éclaircir un peu la question en spécifiant ce qui cherche à être réduit et ce qui n'a pas à l'être.

Le premier chapitre (section 1.3.3) a spécifié quels aspects de l'entreprise psychocognitive seront le centre d'intérêt de notre analyse de la relation de réduction, soit les états mentaux postulés par la psychologie du sens commun. Nous avons vu à la section 4.1.1 que son cadre conceptuel est défini par un processus de ramséfication qui déterminera le portrait des relations causales qu'entretient un état mental type avec les *inputs*, *outputs* et autres états mentaux, tous spécifiés dans un vocabulaire approprié au niveau sémantique. Cette limitation avait pour but d'éviter d'entrer dans les détails de la compatibilité des approches symboliques avec un modèle de computation PDP. Toutefois, l'approche symbolique doit tout de même trouver sa place (ou son rejet) dans l'approche réductionniste de P.M. Churchland puisqu'elle constitue le cœur de la construction de modèles théoriques en psychologie cognitive. (Newell, 1980, 1990) Le procédé de déshomoncularisation fait

appel à ce format computationnel lorsqu'il est question de déterminer par quelles opérations simples (suffisamment idiotes pour ne plus faire appel à des homoncules) il sera possible de reproduire la fonction complexe du processus cognitif. Cela nous permet donc de poser deux questions vis-à-vis le statut de la relation de réduction. Premièrement, comment se comporte le schème conceptuel de la psychologie du sens commun donné par les définitions ramséennes face aux niveaux neurocomputationnel et neurobiologique? Secondement, comment intégrer les modèles computationnels spécifiés par une stratégie de déshomoncularisation dans l'optique neurocomputationnelle?

Pour répondre à la première question, il faudra spécifier les relations de réalisation explicative permettant de produire une image neurocomputationnelle des énoncés ramséens. Suivant la stratégie de Lewis (1972/1980), les énoncés ramséens ne posent pas de contraintes fortes vis-à-vis le choix d'un type de réalisation matérielle. « Using ramsification, one can describe a causal organization without committing oneself to specific features of the realization of that organization, i.e. all the other, often physical properties that aren't mentioned in the Ramsey sentence. » (Rey, 1997, p.177) Cela offre donc une possibilité de concilier ceux-ci au modèle neurocomputationnel de P.M. Churchland: il faudra découvrir, dans le portrait du système cognitif au niveau neurocompuationnel, un schème pouvant répondre au portrait causal de l'énoncé ramséen affecté à chacun des types d'états mentaux du niveau sémantique. Le produit de cette recherche devra être évalué à la lumière des valeurs métathéoriques invoquées à la section 1.3.2, soit le critère d'intégration scientifique et le critère de qualité empirique et, selon les résultats de l'évaluation, il faudra prendre l'une des quatre voies possibles de la réduction (section 1.3.1). Si un assemblage de neurones (niveau neurobiologique) réalise un schème d'opérations neurocomputationnelles dont les relations causales reflètent ceux de l'énoncé ramséen d'un état mental type (niveau sémantique), alors il y aura réduction. Si une large partie de l'énoncé ramséen peut être identifié mais qu'une partie ne trouve pas de corrélat neurocognitif, il faudra alors rejeter cette partie et redéfinir l'état mental au niveau sémantique par un énoncé ramséen qui ne contiendrait pas dans sa forme disjonctive la portion incohérente. Toutefois, les énoncés ramséens ont ce caractère holistique puisqu'ils spécifient les interrelations causales entre les divers états mentaux. Ainsi, l'énoncé ramséen définissant le processus attentionnel pourrait contenir une référence au processus de la conscience, et vice-versa. La révision ou l'élimination d'un concept d'état mental

aura donc pour conséquence de changer non pas seulement le concept révisé mais aussi tous les autres y faisant référence, parfois de manière superficielle, parfois de manière drastique, altérant ainsi le portrait global de la psychologie du sens commun. Comme il en a été question à la section 1.3.2, il n'est pas clair par quels principes ajouter ou altérer une telle portion de l'énoncé ramséen ni non plus comment déterminer les impacts que cette révision aura sur les autres éléments du réseau de concepts. Il reste encore aux néo-réductionnistes à spécifier la méthodologie de révision en sciences cognitives.

De cette image de la réduction, il ne semble pas découler dès à présent que les énoncés ramséens définissant les concepts tirés de la psychologie du sens commun seront radicalement altérés pour produire une image neurocomputationnelle adéquate de ceux-ci. On pourrait même proposer que ces énoncés sont suffisamment généraux et abstraits (ils ne spécifient pas le détail matériel convenant à leur implémentation) pour ne pas radicalement s'heurter aux découvertes neurobiologiques. Même si cela reste une question empirique dont la réponse ne sera fournie que par des théories achevées, il semble qu'une thèse éliminativiste vis-à-vis la psychologie du sens commun soit précipitée et peu plausible pour l'instant.

La situation des modèles explicatifs de la psychologie cognitive n'en est pas tout à fait aussi facile. Une théorie neurocognitive aurait toujours certaines vertus qui lui permettraient de mieux réussir que les modèles psychocognitifs basés sur la computation propositionnelle. Comme il en a été question plus tôt, le modèle computationnel PDP et la relation de réalisation explicative entre les niveaux neurocomputationnel et sémantique proposés par P.M. Churchland obligeraient un rejet des modèles à computation symbolique et sérielle puisque le système nerveux serait en fait un processeur à traitement parallèle et aux représentations distribuées. (P.M. Churchland, 1989; P.S. Churchland, 1986) Cette thèse est encore controversée étant donné la possibilité théorique d'une machine virtuelle, celle-ci sérielle et symbolique, qui serait réalisée par le fonctionnement des réseaux de neurones (voir à ce propos Dennett (2006) pour une vue très conciliatrice). Dans un cas comme dans l'autre, il semble tout de même que les modèles computationnels propositionnels doivent être rejetés, non pas par leur manque de réalisme, mais par leur potentielle incapacité à répondre aussi adéquatement que les modèles neurocomputationnels aux deux valeurs métathéoriques invoquées par P.M. Churchland (section 1.3.3). Les réseaux de neurones artificiels semblent aisément être plus cohésifs

avec le restant du corpus scientifique grâce à leur nature analogue aux neurones naturels (malgré les quelques problèmes de réalisme dont il a été question), ce qui leur assure une plus grand réussite face au critère d'intégration scientifique. De plus, les réseaux de neurones artificiels offrent des capacités computationnelles permettant un plus grand réalisme dans les résultats que la computation symbolique, ce qui les favorise vis-à-vis le critère de qualité empirique. (Arbib, 1998;Bechtel, 1991; McClelland et al., 1986)

Il pourrait être possible de rejeter l'aspect propositionnel des théories psychocognitives même si elles offraient des modèles aussi fonctionnels et prédictifs que leur compétiteur neuropsychologique. Les théories neurocognitives répondraient mieux au critère de qualité empirique simplement parce qu'elles engloberaient un plus grand nombre de phénomènes que les théories psychocognitives. Une théorie neurocognitive prend place sur trois niveaux explicatifs et peut donc couvrir à la fois les aspects propres à ces trois niveaux (ex : défectuosités matérielles, normativité des états mentaux, processus de transduction, plasticité du processus, etc.) alors que la théorie psychocognitive pour un domaine équivalent au niveau sémantique ne pourrait pas répondre aux phénomènes matériaux. De plus, comme il en a été question plus haut, au niveau sémantique, le portrait neurocognitif des phénomènes mentaux englobe le schème fonctionnel global alors que la psychologie cognitive doit reléguer aux neurosciences le détail des processus perceptuels et moteurs. À qualité de prédiction égale, l'approche neuroscientifique promet une compréhension plus large et intégrative des phénomènes cognitifs que celle de la psychologie cognitive. Bien sûr, il faudrait en arriver à un point où nous aurions deux théories compétitrices aussi prédictives, ce qui est une question empirique dont seul l'avenir en décidera le sort.

L'avantage va encore à l'approche neuroscientifique si l'on considère le critère d'intégration scientifique. Le fait même qu'une théorie neurocognitive prenne place sur trois niveaux permet d'assurer des relations étroites avec l'aspect physique des systèmes cognitifs. Mais plus encore, le niveau de l'implémentation matériel est proprement biologique. Tous les systèmes cognitifs connus sont d'origine biologique et cela permet donc de déterminer les influences des entités présentes dans les autres sciences (ex : chimie, biologie moléculaire, etc.) sur l'aspect neuropsychologique (ex : Bickle (2003)).

Rejeter le format propositionnel de la cognition ne signifie pas qu'il faille rejeter d'emblée la stratégie de déshomoncularisation classique proposée par les tenants de la

psychologie cognitive (et donc toute la méthodologie de celle-ci). Au contraire, il semble même que, pour pouvoir en venir à déterminer si la réduction des états mentaux postulés par la psychologie du sens commun est possible ou pas, il faille justement procéder par une telle stratégie car ce n'est qu'ainsi que nous pourrons construire une image neuroscientifique de la psychologie cognitive au niveau sémantique. La déshomoncularisation devrait donc se faire en fonction en vue de produire une image neurocomputationnelle des états mentaux au niveau sémantique et ce, en fonction des contraintes qu'exerce le niveau neurocomputationnel sur le niveau sémantique et vice-versa. L'exemple de réseau de neurone de la conjonction présenté à la section 2.1.2 en est un exemple simpliste. En spécifiant la table de vérité de l'opérateur de conjonction, on en vient à déterminer qu'il faudra deux neurones d'entrée (pour les 2 propositions connectées), ainsi que les conditions à respecter pour réaliser cette fonction. L'histoire du connexionnisme a même montré comment les modèles primitifs (ex: le perceptron (Rosenblatt (1958)) ne pouvaient parvenir à computer certaines fonctions définies au niveau sémantique (ex: disjonction exclusive (Minksy & Papert (1969))) sans intégrer un niveau intermédiaire de neurones (*hidden-layer*). Ce genre d'analyse prend son envol au niveau sémantique et cherche à construire, en divisant le processus en sous-processus, un modèle connexionniste (plutôt que symbolique) permettant de produire la fonction désirée. Il ne resterait plus qu'à trouver confirmation de l'existence d'un tel type de réseau en localisant ces processus dans le système nerveux animal.

Conclusion

Too much thought and not enough experiment.
Patricia Churchland (1986, p.363)

La réduction interthéorique, comme l'indiquent bien Oppenheim & Putnam (1958) et Kitcher (1981, 1989), est une méthode d'unification des sciences. Gaardner (1985) et von Eckart (1993) (pour n'en nommer que deux) déplorent justement un manque d'unité en sciences cognitives. Si on se fie au schéma en couverture du "Report of the State of the Art Committee to the Advisors of the Alfred P. Sloan Foundation" (repris dans von Eckardt (1993, p.2)), les sciences cognitives seraient divisées en six pôles de recherche, dont un neuroscientifique et un autre psychologique; et cet éclatement limiterait les avancées dans le domaine s'il n'y avait pas de communication et de coopération entre ces perspectives. L'interdisciplinarité dans ce domaine est depuis longtemps une vertu célébrée (Jeffress, 1954), mais il semble que la formalisation d'outils explicatifs pouvant être employés autant dans une perspective qu'une autre puisse unifier encore plus les différentes entreprises constitutives des sciences cognitives. P.M. Churchland & P.S. Churchland (1998, chapitre 15) valorisent l'unification des sciences puisque cela permet à la fois d'obtenir un appareillage conceptuel plus efficace et cohérent (voir aussi P.M. Churchland (1979, chapitre 3)), mais aussi de stimuler les recherches empiriques et la construction d'un cadre théorique pour les mener. P.M. Churchland[67], dans McCauley (1996, pp.219-221) revient sur sa position éliminativiste radicale et prêche une version modérée, celle de la coévolution interthéorique proposée par P.S. Churchland (1986, pp.362-376), selon laquelle la psychologie cognitive et les neurosciences devraient plutôt coopérer et s'influencer mutuellement pour éventuellement se fondrnte dans un même ensemble théorique où les deux théories seraient confondues en une seule, soit une neuropsychologie achevée :

> « Neuroscience and psychology need each other. Crudely, neuroscience needs psychology because it needs to know what the system does; that is, it needs high-level specifications of the input-output properties of the system. Psychology needs neuroscience for the same reason: it needs to know what the system does. That is, it needs to know whether lower-level specifications bear out the initial input-output theory, where and how to revise input-output theory, and how to characterize processes at levels below the top. » (P.S. Churchland, 1986, p.373)

[67] Ainsi que Patricia S. Churchland.

Le schème explicatif du néo-réductionnisme présenté dans les chapitres précédents (sections 2.1 et 3.2) cherche à produire un terrain d'entente minimale entre les deux perspectives cognitivistes pour d'abord assurer, dans un futur plus ou moins proche, la possibilité d'évaluer la relation de réduction interthéorique et le détail de sa faisabilité, mais aussi, du même coup, de pouvoir à la fois localiser la nature et l'impact des efforts scientifiques produits par les différentes approches en sciences cognitive et d'organiser les résultats empiriques dans un schème unificateur.

La position éliminativiste de Paul M. Churchland repose sur deux thèses. D'abord, P.M Churchland (1979) propose une nouvelle manière d'entendre la relation de réduction interthéorique. Le premier chapitre de ce mémoire avait pour objectif d'expliquer cette nouvelle conception (néo-réductionniste) tout en la situant dans son contexte philosophique. Cette nouvelle position est originale d'abord parce qu'elle permet d'introduire une échelle de degrés de réussite de la réduction d'une théorie scientifique à une autre (1.3.1), mais aussi parce qu'elle propose certaines normes permettant à la communauté scientifique de déterminer le sort des théories en jeu (1.3.2). Cela a permis de cerner la nature de la position anti-réductionniste de P.M. Churchland (1979, 1981) et ainsi de déterminer les conditions de validation de celle-ci.

La seconde thèse tenue par P.M. Churchland, nécessaire pour endosser la possibilité éliminativiste dans son schème réductionniste, n'est pas tant philosophique que prophétique. Il déjà clair que les neurosciences vont proposer, dans un futur plus ou moins lointain, une théorie (beaucoup) plus complète et précise des processus cognitifs que les théories psychocognitives compétitrices ne le pourront. (P.M. Churchland, 1981) Certains marqueurs illustrent bien la possibilité d'un tel remplacement, comme le peu de plausibilité de la thèse de la computation propositionnelle (section 1.3.3), la force des découvertes et des explications neuroscientifiques ainsi que des prouesses computationnelles des modèles de réseaux de neurones (PDP) (section 4.2.1). À la lumière de la démarche proposée par le néo-réductionnisme churchlandien, le *modus ponens* permettant de décider du sort de la psychologie cognitive est aisé à accomplir : il faut éliminer la psychologie du sens commun de l'entreprise des sciences cognitives et, avec elle, la méthodologie de la psychologie cognitive (section 1.3.3).

Selon Paul M. Churchland, la psychologie du sens commun *est* fondamentalement erronée et les neurosciences vont construire *avec succès* une théorie neuropsychologique

alternative, soit une théorie (au niveau sémantique) qui remplacera, lorsqu'elle sera suffisamment mature, la psychologie "fantasque" tirée de nos croyances traditionnelles. Pour faire sens de ceci à la lumière de l'analyse conduite dans les pages précédentes, une telle lecture est ou bien adéquate, et alors P.M. Churchland tombe dans des spéculations plus prophétiques qu'analytiques (section 1.4) étant donné qu'il ne semble y avoir aucune raison disponible dans son cadre conceptuel de soutenir sa seconde thèse, soit de rejeter dès aujourd'hui la psychologie du sens commun (voir section 4.3). C'est ce que l'on pourrait nommer la lecture 'négative' de l'éliminativisme de P.M. Churchland. Au contraire, l'affirmation prédictive selon laquelle P.M. Churchland endosserait le rejet nécessaire et inévitable de la psychologie du sens commun pourrait être erronée et dans ce cas il faudrait alors comprendre son éliminativisme comme une prescription méthodologique, voir même comme une certaine 'propagande' pour amorcer (ou encourager) un nouveau programme de recherche[68]. Il me semble que cette seconde lecture soit la bonne (voir McCauley (1996, pp.219-221) et le retour qu'y fait Paul M. Churchland sur son 'éliminativisme'), que l'on pourrait nommer la lecture 'positive' de son éliminativisme.

Accepter comme véridique ou incorrect un cadre conceptuel permettant de catégoriser et d'expliquer des phénomènes naturels en vue de mener des recherches scientifiques requiert un appui empirique. La psychologie du sens commun est un tel cadre conceptuel. (P.M. Churchland, 1981; Feyerabend, 1963; Rorty, 1965) Or, rejeter ou accepter *a priori* la valeur de la psychologie du sens commun contredit ce précepte scientifique. La possibilité de l'élimination d'un cadre conceptuel du vocabulaire théorique des sciences (ici cognitives) implique que celui-ci est en fait spéculatif et ne doit être admis que par sa valeur scientifique (et non pas par l'apparence intuitive de sa véracité).

> « The claim that psychology comprehends a distinct level of phenomena comprehended by a distincts set of laws uniquely appropriate to that level *is not an assumption that our opposition can have for free.* It is part of what is at issue – *empirically* at issue – in this broad debate » (P.M. Churchland dans McCauley (1996, p.224); italiques dans l'original)

[68] Tout à fait à l'image de l'éliminativisme de Feyerabend (1962, 1963) mais surtout Feyerabend (1979). Il est d'ailleurs assez remarquable de constater à quel point la stratégie propagandiste de P.M. Churchland répond aux prescriptions de Feyerabend. L'œuvre de P.M. Churchland semble avoir été très influencée par celui-ci (voir Churchland (1998, chapitre 15) par exemple), mais c'est là le sujet d'une autre étude plutôt philologique que philosophique. Toutefois, le caractère prophétique de P.M. Churchland ressemble étrangement au prophétisme anti-matérialiste décrié par Feyerabend (1963).

La dernière section du chapitre précédent (section 4.3) souligne qu'il y a probablement une confusion dans l'éliminativisme de P.M. Churchland : l'élimination de la psychologie du sens commun ne devrait peut être pas être identifiée au rejet de la formalisation propositionnelle au niveau computationnel. Notre hiérarchie des niveaux explicatifs (section 2.1, plus particulièrement la section 2.1.3) montre que cela n'est pas une relation nécessaire (selon la nature de la relation de réalisation explicative) et que le rejet de l'approche symbolique de la computation n'a en fait pas les mêmes conséquences que le rejet des entités théoriques au niveau sémantique ordinairement modélisées de cette manière au niveau computationnel. P.M. Churchland rejetterait donc en bloc le cadre conceptuel du niveau sémantique et les processus computationnels qui y sont liés par la relation de réalisation explicative de déshomoncularisation classique (section 4.1.2).

Toutefois, si l'on contraint aux énoncés ramséens la description du niveau sémantique utilisée par la psychologie cognitive, il ne semble pas y avoir de raisons valables dès aujourd'hui pour rejeter la psychologie du sens commun comme cadre conceptuel *potentiellement* véridique par rapport aux états mentaux. D'abord, une telle formalisation de ce cadre conceptuel permet des conditions de vérifications assez larges pour que nous ne puissions trouver, dès à présent, la moindre contradiction empirique. Patricia Churchland (1986, pp.368-373) fait référence à la fragmentation du concept d'apprentissage à partir des découvertes neuroscientifiques. Toutefois, bien que cette fragmentation soit bien réelle, les divers types d'apprentissage continuent tout de même à répondre au concept plus général d'apprentissage, soit l'intégration pour un rappel futur d'expériences passées. Dans ce cas, il y aurait révision ou élimination si et seulement si le concept tiré de la psychologie du sens commun ne pourrait cohabiter avec les détails empiriques concernant les divers types d'apprentissage. Bien que cela reste une question empirique, les nouvelles catégories de conditionnement, d'apprentissage procédural ou déclaratif continuent de répondre au concept général d'apprentissage (tiré de la psychologie du sens commun) qui ne fait que spécifier une relation générale entre certains *inputs* (expériences passées) et *outputs* (rappels déclaratifs (ex : connaissances) ou performatifs (ex : comportements). Cette généralité du cadre conceptuel emprunté par la psychologie cognitive pourrait assurer sa survie.

De plus, le cadre conceptuel utilisé au niveau sémantique pourrait très bien être éliminé ou retenu (avec ou sans modifications) selon le type de relation de réalisation

explicative qui est choisie pour lier les niveaux computationnel sémantique (voir même les niveaux de l'implémentation matérielle et computationnel). Il a été question à la section 4.3 de l'incohérence entre la déshomoncularisation classique et sa formalisation en système de computation sériel et symbolique face aux *corpus* des neurosciences, ce qui jouait en la faveur de l'éliminativisme puisque cette relation ne respecte pas suffisamment le critère de l'intégration aux sciences (section 1.3.2). Rejeter ce type de relation de réalisation ne semble pas contrevenir à la possibilité d'établir une relation de réalisation explicative à la fois satisfaisante pour la communauté scientifique et soutenant le cadre conceptuel de la psychologie du sens commun. Il restera toutefois à spécifier par quelle méthodologie déterminer la conception appropriée des entités et prédicats utilisés pour former les énoncés ramséens au niveau sémantique. Si la proposition de les réduire est correcte, deux méthodes pourraient s'avérer correctes selon la stratégie explicative utilisée par le chercheur. Pour le tenant de la stratégie explicative *top-down* (section 2.2.1), il faudra déterminer une méthode de ramséfication ne faisant pas référence explicitement à un mode de computation (ce qui ne veut pas dire que les définitions données doivent répondre au modèle neurocomputationnel en vigueur mais simplement ne pas le contredire explicitement) (voir section 4.3). Pour celui qui défendrait une approche *bottom-up* (section 2.2.2), ce sera par la définition d'une relation de réalisation explicative basée sur les modèles neurocomputationnels (stratégie empruntée par P.M. Churchland (2007)) et/ou sur la biologie évolutive. (Godfrey-Smith, 1991; Millikan 1984b, 1993) P.M. Churchland en est même venu à encourager les deux stratégies :

> « We therefore regard it as completely unproblematic both that hierarchy of independently worthy sciences should exist and that major illumination should often flow from the lower-level sciences upwards. Also, we have argued that illumination often flows from the high-level sciences *down*wards. » (P.M. Churchland & P.S. Churchland dans McCauley (1996, pp.220-221); italiques dans l'original.)

Un éliminativisme cohérent avec le cadre conceptuel du néo-réductionnisme churchlandien vise donc bien plus le rejet de la relation de réalisation explicative telle que la psychologie cognitive l'utilise par la formalisation symbolique et l'appel à des règles d'inférence déductive que le rejet du cadre conceptuel au niveau sémantique utilisé par la psychologie cognitive. Cette méthode de formalisation est une thèse concernant la nature de la relation de réalisation explicative entre les niveaux sémantique et computationnel et son rejet n'affecte que les contraintes que les niveaux exercent les uns sur les autres. La

psychologie du sens commun pourrait donc survivre à une reformalisation de son équivalent computationnel et ne serait pas éliminable pour ces raisons. Cela ne veut pas dire que la psychologie du sens commun survivra à l'évolution des théories neuroscientifiques; et cela ne veut pas dire non plus qu'elle soit complètement erronée, ou du moins significativement tarée pour être éliminée sans lendemains. Cela est encore une question ouverte et j'espère que cette analyse montre bien que la dernière chose qui fera avancer le débat du réductionnisme en sciences cognitives est l'apparition de nouveaux prophètes qui dicteraient à l'histoire elle-même ce qu'elle devra être.

P.M. Churchland semble voir juste (selon l'analyse conduite aux sections 1.3 et 4.3) lorsqu'il rejette la stratégie explicative *top-down* de déshomoncularisation. La méthode *top-down* et la déshomoncularisation en processus sériels de traitement symbolique de l'information au niveau computationnel trouve sa proéminence dans l'optique d'une psychologie cognitive autonome usant de la stratégie de recherche *top-down* (section 1.2.2). L'antiréductionnisme de P.M. Churchland, de ce point de vue, devrait plutôt être conçu comme une critique méthodologique : affirmer *a priori* la validité scientifique du cadre conceptuel de la psychologie du sens commun ainsi que de notre compréhension intuitive des rouages fonctionnels qui constitue la mécanique des processus mentaux qu'il postule contrevient à la nature empirique des sciences naturelles (section 1.4). D'une manière ou d'une autre, il semble alors que les recherches procédant par une stratégie purement *top-down* couvrent le risque d'élaborer des théories à partir d'un cadre conceptuel erroné, totalement ou partiellement. Cette lecture (positive) de P.M. Churchland évite de tomber dans le prophétisme et suggère une version méthodologique de l'éliminativisme : toute théorie, en vertu même du fait qu'elle soit une théorie, est spéculative et nécessite donc un appui empirique, et cet appui empirique consiste en l'intégration par réduction de cette théorie au *corpus* scientifique déjà admis. (P.M. Churchland, 1979, 1981, 1989; voir aussi Feyerabend, 1963)

Bibliographie

Akins, Kathleen (1996), "Of sensory systems and the "aboutness" of mental states", *Journal of Philosophy* 91:337-372.

Anderson, James A. (1995), *An Introduction to Neural Networks*. Cambridge, Massachusetts: MIT Press.

Arbib, Michael A., ed. (1998), *The Handbook of Brain Theory and Neural Networks., A Bradford Book*. Cambridge, Massachussetts: The MIT Press.

Armstrong, D. M. (1968), *A Materialist Theory of the Mind*. London: Routledge & Keagan Paul.

Bear, Mark F., Barry W. Connors, & Michael A. Paradiso (2002), *Neurosciences : À la découverte du cerveau*. Traduit par André Nieoullon. Second Edition ed. Rueil-Malmaison: Groupe Liaisons S.A. Original edition, 2001.

Bechtel, William (1991), "Connectionism and the Philosophy of Mind: An Overview", in Terence Horgan and John Tienson (eds.), *Connectionism and the Philosopy of Mind*: Kluwer Academic Publishers, 30-59.

——— (1994), "Levels of Description and Explanation in Cognitive Science", *Minds and Machines* 4:1-25.

Bechtel, William, & Adele Abrahamsen (2002), *Connectionism and the mind: A introduction to parallel processing in networks*. Oxford: Blackwell Publishing.

——— (2005), "Explanation: A Mechanist Alternative", *Studies in History and Philosophy of Biological and Biomedical Sciences* 36:421-441.

Bermudez, Jose Luis (1995), "Syntax, Semantics and Level of Explanation", *The Philosophical Quarterly* 45:361-367.

——— (2001), "Normativity and Rationality in Delusional Psychiatric Disorders", *Mind and Language* 16.

Bickle, John (1992), "Mental Anomaly and the New Mind-Brain Reductionism", *Philosophy of Science* 59 (2):217-230.

——— (1996), "New Wave Psychophysical Reductionism and the Methodological Caveats", *Philosophy and Phenomonological Research* 56 (1):57-78.

——— (1998), *Psychoneural Reduction*. Cambridge, Massachussetts: MIT Press.

——— (2003), *Philosophy and Neuroscience: A Ruthlessly Reductive Account*. Dordrecht: Kluwer Academic Publishers.

Block, Ned (1980), "Troubles with Functionalism", in Ned Block (ed.), *Readings in Philosophy of Psychology*, Cambridge, Mas.: Harvard University Press, 268-305.

Brentano, Franz (1973), *Psychology from an Empirical Standpoint*. London: Routledge and Kegan Paul.

Buller, David J. (2005), *Adapting Minds : evolutionary psychology and the persistent quest for human nature*. Cambridge: The MIT Press.

Burge, Tyler (1979), "Individualism and the Mental", *Midwest Studies in Philosophy* 4:73-121.

——— (1986), "Individualism and Psychology", *The Philosophical Review* 95:3-46.

Carnap, Rudolf (1933), "Psychology in physical language", *Erkentnis* 4:107-142.

——— (2003), *The Logical Structure of the World and Pseudoproblems in Philosophy*. Chicago: Open Court Publishing.

Cartwright, Nancy (1989), "Capacities and Abstractions", in Philip Kitcher & Wesley C. Salmon (eds.), *Scientific Explanation*, Minneapolis: University of Minnesota Press, 349-356.

Chisholm, Roderick (1957), *Perceiving*. Ithaca: Cornell University Press.

Chomsky, Noam (1959), "A Review of B. F. Skinner's *Verbal Behavior*", *Language* 35 (1):26-58.

Christensen, Scott M., & Dale R. Turner, eds. (1993), *Folk Psychology and the Philosophy of Mind*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Church, Alonzo (1932), "A set of Postulates for the Foundation of Logic", *Annals of Mathematics* 33:346-366.

Churchland, Paul M. (1970), "The logical character of action-explanations", *Philosophical Review* 2 (79):214-236.

——— (1979), *Scientific realism and the plasticity of mind*. Cambridge, England: Cambridge University Press.

——— (1981), "Eliminative Materialism and the Propositional Attitudes", *The Journal of Philosophy* 78 (2):67-90.

——— (1988), *Matter and Consciousness: A Contemporary Introduction to the Philosophy of Mind*. revised edition ed, *A Bradford Book*. Cambridge, Massachusetts: MIT Press. Original edition, 1984.

——— (1989), *A Neurocomputational Perspective - The Nature of Mind and the Structure of Science, Bradford Book*. Cambridge, Massachusetts: MIT Press.

——— (2007), *Neurophilosophy at Work*. New York: Cambridge University Press.

Churchland, Paul M., & Patricia S. Churchland (1983), "Stalking the Wild Epistemic Engine", *Noûs* 17 (1):5-18.

——— (1998), *On the Contrary*, Cambridge, Massachusetts: MIT Press.

——— (1996), "The Future of Psychology, Folk and Scientific", in Robert N. McCauley (ed.), *The Churchlands and their Critics*, Cambridge, Mass.: Blackwell Publishers, 219-255.

Churchland, Patricia Smith (1986), *Neurophilosophy : Toward a Unified Science of the Mind-Brain*. Cambridge: The MIT Press.

Churchland, Patricia S., & Terrence J. Sejnowski (1992), *The Computational Brain*. Edited by Terrence J. Sejnowski & Tomaso A. Poggio. 1 vols, *Computational Neuroscience*. Cambridge, Massachusetts: The MIT Press.

Clark, Andy (1996), "Dealing in Futures: Folk Psychology and the Role of Representations in Cognitive Science", in Robert N. McCauley (ed.), *The Churchland and their Critics*, Cambridge, Massachussets: Blackwell, 86-103.

Craver, Carl F. (2001), "Role, Mechanisms, and Hierarchy", *Philosophy of Science* 68:53-74.

Cummins, Robert (1975), "Functional Analysis", *Journal of Philosophy* (72):741-765.

——— (1983), *The Nature of Psychological Explanation*. Cambridge, Mass.: The MIT Press.

Davidson, Donald (1963), "Actions, Reasons, and Causes", *The Journal of Philosophy* 60 (23):685-700.

——— (1970), "Mental Events", in Lawrence Foster & J. W. Swanson (eds.), *Experience and Theory*, Amherst, Mass: University of Massachusetts Press, 79-101.

——— (1973), "The Material Mind", in Patrick Suppes (ed.), *Logic, Methodology, and the Philosophy of Science*: North-Holland, 709-722.

——— (2001), *Essays on Actions and Events*. Oxford: Oxford University Press.

Dennett, Daniel C. (1971), "Intentional Systems", *Journal of Philosophy* 68 (4):87-106.
——— (1975), "Why the Law of Effect Will Not Go Away", *Journal of the Theory of Social Behaviour*:169-176.
——— (1978), *Brainstorms - Philosophical Essays on Mind and Psychology*. Edited by Margaret A. Boden, *Harverster Studies in Philosophy*. Hassocks, Sussex: The Harvester Press Limited.
——— (1983), "Intentional Systems in Cognitive Ethology: the 'Panglossian Paradigm' Defended", *Behavioral and Brain Sciences* 6:343-390.
——— (1987), *The Intentional Stance*. Cambridge, Mass.: The MIT Press.
——— (1990), "*The Interpredation of Texts, People, and Other Artifacts*", Philosophy and Phenomenological Research 50:177-194.
——— (1991a), *Consciousness Explained*. Boston: Back Bay Books.
——— (1991b), "Real Patterns", *The Journal of Philosophy* 88:27-51.
——— (1994), "Cognitive Science as Reverse Engineering: Several Meanings of "Top-Down" and "Bottom-Up"", in D. Prawitz, B. Skyrms & D. Westerstahl (eds.), *Logic, Methedology, and Philosophy of Science IX*, Amsterdam: Elsevier Science ,BV, 679-689.
——— (1995), *Darwin's Dangerous Idea*. New York, Touchstone.
——— (1998a), *Brainchildren - Essays on Designing Minds*. Cambridge, Mass.: The MIT Press.
——— (1998b), "Artificial Life as Philosophy", in, *Brainchildren - Essays on Designing Minds*. Cambridge, Mass.: The MIT Press, 261-264.
——— (1998c), "Cognitive Science as Reverse Engineering: Several Meanings of "Top-Down" and "Bottom-Up"", in, *Brainchildren - Essays on Designing Minds*. Cambridge, Mass.: The MIT Press, 249-260.
——— (2006), "Two Steps Closer on Consciousness", in Brian L. Keeley (ed.), *Paul Churchland*, Cambridge, Massachusetts: Cambridge University Press, 193-203.
Densmore, S, and Daniel Dennett (1999), "The Virtues of Virtual Machines", *Philosophy and Phenomenological Research* 59 (3):747-761.
Dyer, Michael G. (1991), "Connectionism versus Symbolism in High-Level Cognition", in Terence Horgan & John Tienson (eds.), *Connectionism and the Philosophy of Mind*, Dordrecht: Kluwer Academic Publishers, 382-416.
Eccles, John C. (1989), *Evolution of the Brain: Creation of the Self*. London: Routledge.
Endicott, Ronald P. (1993), "Species-Specific Properties and more Narrow Reductive Strategies", *Erkenntnis* 38:303-321.
Feigl, Herbert (1958), "The 'Mental' and the 'Physical'", in Herbert Feigl, M Scriven & G. Maxwell (eds.), *Concepts, Theories, and the Mind-Body Problem*, Minneapolis: University of Minnesota Press.
——— (1967), *The "Mental" and the "Physical", The Essay and a Post-Script*. Minneapolis: University of Minnesota Press.
Feyerabend, Paul (1962), "Explanation, Reduction, and Empiricism", in H. Feigl & G. Maxwell (eds.), *Scientific explanation, space and time*, Minneapolis: University of Minnesota Press, 28-97.
——— (1963), "Materialism and the Mind-Body Problem", *The Review of Metaphysics* 17:49-66.
——— (1979), *Contre la méthode: Esquisse d'une théorie anarchiste de la connaissance*. Traduit par Baudouin Jurdant et Agnès Schlumberger. Paris: Seuil.

Fletcher, Garth (1995), *The Scientific Credibility of Folk Psychology*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Fodor, Jerry A. (1968), *Psychological Explanation*. New York: Random House.

——— (1974), "Special Sciences", *Synthese* 28:97-115.

——— (1975), *The Language of Thought*, Cambridge, Massachusetts: Harvard University Press.

——— (1980), "Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology", *Behavioral and Brain Sciences* 3 (1):63-73.

——— (1983), *The Modularity of the Mind*. Cambridge, Massachusetts: MIT Press.

——— (1987), *Psychosemantics*. Edited by Margaret A. Boden, *Explorations in Cognitive Sciences*. Cambridge, Mas.: The MIT Press.

Fodor, Jerry A., & Zenon W. Pylyshyn (1988), "Connectionism and cognitive architecture: A critical analysis", *Cognition* 28:2-71.

Gardner, Howard (1985), *The Mind's New Science*. New York: Basic Books.

Garfield, Jay L. (1988), *Belief in Psychology*. Cambridge, Mass: The MIT Press.

Godfrey-Smith, Peter (1991), "Signal, Decision, Action", *Journal of Philosophy* 88:709-722.

——— (1996), *Complexity and the Function of Mind in Nature*. Cambridge, Massachussetts: Cambridge University Press.

Goldstein, E. Bruce (2002), *Sensation and Perception*. Sixth Edition ed. Pacific Grove: Wadsworth Group.

Gould, Stephen Jay, & Richard C. Lewontin (1979), "The spandrels of San Marco and the Panglossian paradigm", *Proceedings of the Royal Society of London* B. 205:581-598.

Greenwood, John D., ed. (1991), *The future of folk psychology - Intentionality and cognitive science*: Cambridge University Press.

Hardcastle, Valerie Gray (1996), *How to Build a Theory in Cognitive Science*. Albany, New York: State University of New York Press.

——— (2007), "The Theoretical and Methodological Foundations of Cognitive Neuroscience", in Paul Thagard (ed.), *Philosophy of Psychology and Cognitive Science*, Amsterdam: Elsevier, 295-311.

Hatfield, Gary (2002), "Psychology, Philosophy, and Cognitive Science: Reflections on the History and Philosophy of Experimental Psychology", *Mind and Language* 17 (3):207-232.

Hempel, Carl G. (1949), "The logical analysis of psychology", in Herbert Feigl & Wilfrid Sellars (eds.), *Readings in Philosophical Analysis*: Appleton Century Crofts.

——— (1965), *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. New York: The Free Press.

Hempel, Carl G., & Paul Oppenheim (1948), "Studies in the Logic of Explanation", *Philosophy of Science* 15:135-175.

Hooker, C. A. (1979), "Critical Notice: R. M. Yoshida's *Reduction in the Physical Sciences*", *Dialogue* 18:81-99.

——— (1981), "Towards a General Theory of Reduction. Part I: Historical and Scientific Settings. Part II: Identity in Reduction. Part III: Cross-Categorial Reduction", *Dialogue* 20:28-52, 201-236, 496-529.

Horgan, Terence (1992), "From cognitive science to folk psychology: computation, mental representation, and belief", *Philosophy and Phenomonological Research* 52:449-484.

Horgan, Terence, & John Tienson (1992a), "Cognitive systems as dynamical systems", *Topoi* 11:27-43.

——— (1993), "Levels of Description in Nonclassical Cognitive Science", in Christopher Hookway & Donald Peterson (eds.), *Philosophy and Cognitive Science*, New York: Cambridge University Press.

Horgan, Terence, & James Woodward (1985), "Folk Psychology is Here to Stay", *The Philosophical Review* 94 (2):197-226.

Jackson, Frank, & P. Pettit (1993), "Folk Belief and Commonplace Belief", *Mind and Language* 8:298-305.

Jeffress, Lloyd A. (1954), "*Cerebral Mechanisms in Behaviour. The Hixon Symposium*", New Yor, Hafner.

Keeley, Brian L., ed. (2006), *Paul Churchland, Contemporary Philosophy in Focus*. New York, New York: Cambridge Uniersity Press.

Kim, Jaegwon (1978), "Supervenience and nomological incommensurables", *American Philosophical Quarterly* 15 (2):149-156.

——— (2003), "The American Origins of Philosophical Naturalism", *Journal of Philosophical Research* APA Centennial Volume:83-98.

Kemeny, J. G., & Paul Oppenheim (1956), "On Reduction", *Philosophical Studies* 7:6-19.

Kitcher, Philip (1981), "Explanatory Unification", *Philosophy of Science* 48:507-531.

——— (1989), "Explanatory Unification and the Causal Structure of the World", in Philip Kitcher and Wesley C. Salmon (eds.), *Scientific Explanation*, Minneapolis: University of Minnesota Press, 410-505.

Kosslyn, Stephen M. (1980), *Image and Mind*. Cambridge, Massachussetts: Harvard University Press.

——— (1997), "Mental Imagery", in Michael S. Gazzaniga (ed.), *Conversations in the Cognitive Neuroscience*, Cambridge, Mass.: The MIT Press.

Laakso, Aarre, & Garrison W. Cottrell (2000), "Content and Cluster Analysis: Assessing Representational Similarity in Neural Systems", *Philosophical Psychology* 13:47-76.

Lewis, David (1966), "An Argument for the Identity Theory", *Journal of Philosophy* 63 (1):17-25.

——— (1970), "How to Define Theoretical Terms", *Journal of Philosophy* 67 (13):427-446.

——— (1972/1980), "Psychophysical and Theoretical Identifications", *Australasian Journal of Philosophy* 50:249-258. (citations dans Ned Block (ed.), (1980), *Readings in Philosophy of Psychology*, Cambridge, Mass.: Harvard University Press, 207-215).

Lycan, Willam G. (1981), "Toward a Homuncular Theory of Believing", *Cognition and Brain Theory* 4:139-159.

——— (1987), *Consciousness*. Cambridge, Mass: The MIT Press.

Lycan, William G., ed. (1999), *Mind and Cognition*. 2nd edition ed. Oxford: Blackwell. Original edition, 1990.

Malcolm, Norman (1968), "The Conceivability of Mechanism", *The Philosophical Review* 77 (1):45-72.

Marr, David (1982), *Vision*. San Francisco: Freeman & Co.

McCauley, Robert N. (1986), "Intertheoretic Relations and the Future of Psychology", *Philosophy of Science* 53:179-198.

———, ed. (1996), *The Churchlands and their critics*. Edited by Ernest Lepore, *Philosophers and their critics*. Cambridge, Massachussets: Blackwell.
——— (2007), "Reduction: Models of Cross-Scientific Relations and their Implications for the Psychology-Neuroscience Interface", in Paul Thagard (ed.), *Philosophy of Psychology and Cognitive Science,* Amsterdam: North-Holland Elsevier, 105-158.
McClelland, James L., David E. Rumelhart, & G.E. Hinton, eds. (1986), *Parallel distributed processing: Explorations in the microstructure of cognition*. 2 vols. Cambridge, Massachussetts: The MIT Press.
McClamrock, Ron (1991), "Marr's Three Level: A Re-evaluation", *Minds and Machines.*
McGinn, Colin (1991), *The Problem of Consciousness*. Oxford: Blackwell.
Milgram, Stanley (1974), *Obedience to authority : an experimental view*. New York: Harper & Row.
Millikan, Ruth Garrett (1984), "Proper Functions", in, *Language, Thought, and Other Biological Categories*, Cambridge, Massachusetts: The MIT Press.
——— (1993), *White Queen Psychology and Other Essays for Alice*. Cambridge, Massachussetts: MIT Press.
Minsky, Marvin, & Seymour Papert (1969), *Perceptrons: An introduction to Computational Geometry*. Cambridge, Massachussetts: MIT Press.
Nagel, Ernest (1961), *The Structure of Science*. New York: Harcourt, Brace, and World.
Neisser, Ulric (1967), *Cognitive Psychology*. New York: Meredith Publishing Company.
——— (1976), *Cognition and Reality: Principles and Implications of Cognitive Psychology*. San Francisco: W. H. Freeman and Company.
Newell, Alan (1980), "Physical Symbol Systems", *Cognitive Science* 4 (2):135-183.
——— (1982), "The Knowledge Level", *Artificial Intelligence* 18 (1):87-127.
——— (1990), *Unified Theories of Cognition*. Cambridge, Massachussetts: Harvard University Press.
Newell, Alan, & Herbert A. Simon (1972), *Human Problem Solving*. Englewood Cliffs, New Jersey: Prentic-Hall Inc.
Newell, Alan, Paul S. Rosenbloom, & John E. Laird (1989), "Symbolic Architectures for Cognition", in Michael I. Posner (ed.), *Foundations in Cognitive Science*, Cambridge, Mass.: The MIT Press, 93-132.
Oppenheim, Paul, & Hilary Putnam (1958), "Unity of Science as a Working Hypothesis", *Minnesota Studies in the Philosophy of Science* II:3-36.
Peacocke, P. (1986), "Explanation in Computational Psychology: Language, Perception and Level 1.5", *Mind and Language* 1:101-123.
Place, Ullin T. (1956), "Is Consciousness a brain process?", *British Journal of Psychology*:44-50.
Popper, Karl (1959), *The Logic of Scientific Discovery*. London: Hutchinson.
Putnam, Hilary (1960), "Minds and Machines", in S. Hook (ed.), *Dimensions of Mind*, New York: New York University Press.
——— (1963), "Brains and Behavior", in R. J. Butler (ed.), *Analytical Philosphy*, Oxford: Blackwell, 211-235.
——— (1967a), "Psychological Predicates", in W. H. Capitan & D. D. Merrill (eds.), *Art, Mind, and Religion*, Pittsburgh: University of Pittsburgh Press.
——— (1967b), "The Mental Life of Some Machines", in Hector-Neri Castaneda (ed.), *Intentionality, Minds and Perception*, Detroit: Wayne State University Press, 177-200.

——— (1975a), "The Meaning of 'Meaning'", in, *Mind, Language, and Reality*, Cambridge, Massachussetts: Cambridge University Press, 215-271.

——— (1975b), "The Nature of Mental States", in, *Mind, Language, and Reality: Philosophical Papers*, Cambridge, Massachusetts: Cambridge University Press, 429-440.

——— (1988), *Representation and Reality*. Cambridge, Massachussetts: The MIT Press.

Pylyshyn, Zenon (1981), "The Imagery Debate: Analogue Media vs. Tacit Knowledge", *Psychological Review* 88:16-45.

——— (1984), *Computation and Cognition: Toward a Foundation for Cognitive Science*. 2nd ed. Cambridge, Mass.: The MIT Press.

——— (1989), "Computing in Cognitive Science", in Michael I. Posner (ed.), *Foundations in Cognitive Science*, Cambridge, Mass.: The MIT Press, 49-92.

Quine, Willard Van Orman (1960), *Word and Object*. Cambridge, Massachusetts: MIT Press.

——— (1969), *Ontological Relativity and Other Essays*. New York: Columbia University Press.

Ramsey, William, Stephen P. Stich, & Joseph Garson (1991), "Connectionism, Eliminativism, and the Future Of Folk Psychology", in William Ramsey, Stephen P. Stich & David E. Rumelhart (eds.), *Philosophy and Connectionist Theory*, Hillsdale, New Jersey: Lawrence Erlbaum Associates, 199-228.

Reed, Stefen K. (1999), *Cognition : Théories et applications*. Bruxelles: DeBoeck Université.

Rey, Georges (1997), *Contemporary Philosophy of Mind*. Cambridge, Mass.: Blackwell Publishers.

Rorty, Richard (1965), "Mind-Body Identity, Privacy, and Categories", *Review of Metaphysics* 19:24-54.

Rosenblatt, Frank (1958), "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain", *Psychological Review* 65 (6):386-408.

Ryle, George (1949), *The Concept of Mind*. London: Hutcheson.

Russell, Bertrand (1910), "Knowledge by acquaintance and knowledge by description", *Proceedings of the Aristotelian Society* 11:108-128.

Sarkar, Sahotra (2005), "Reduction: A Philosophical Analysis", in, *Molecular Models of Life*, Cambridge, Massachusetts: MIT Press, 105-114.

Schaffner, Kenneth F. (1967), "Approaches to Reduction", *Philosophy of Science* 34 (2):137-147.

Shannon, C. E. (1948), *A mathematical theory of communication.*

Skinner, B. F. (1957), *Verbal Behavior*. New York: Appleton-Century-Crofts.

Smart, J. J. C. (1959), "Sensations and brain processes", *Philosophical Review* 68:141-156.

Sober, Elliot (1985), "Panglossian Functionalism and the Philosophy of Mind", *Synthese* 64 (2):165-193.

Stich, Stephen P. (1978/1999), "Autonomous Psychology and the Belief-Desire Thesis", in William G. Lycan (ed.), *Mind and Cognition*, Oxford: Blackwell, 259-270.

——— (1983), *From Folk Psychology to Cognitive Science*, *Bradford Books*. Cambridge, Massachusetts: The MIT Press.

Turing, Alan (1936), "On Computable Numbers with an Application to the Entscheidungsproblem", *Proceeding of the London Mathematical Society* 2 (42):230-265.

van Gelder, Tim (1998), "The Dynamical Hypothesis in Cognitive Science", *Behavioral and Brain Sciences* 21:1-14.

van Leeuwen, Marco (2005), "Questions For The Dynamicist: The Use of Dynamical Systems Theory in the Philosophy of Cognition", *Minds and Machines* 15:271-333.

von Eckardt, Barbara (1993), *What Is Cognitive Science?* Cambridge, Massachussetts: MIT Press.

——— (2005), "Connectionism and the Propositional Attitudes", in C. Erneling & D. Johnson (eds.), *The Mind as a Scientific Object: Between Brain and Culture*, New York: Oxford University Press.

von Eckardt Klein, Barbara (1978), "Inferring Functional Localization from Neurological Evidence", in Edward Walker (ed.), *Explorations in the Biology of Language*, Cambridge, Mass.: The MIT Press, 27-66.

van Gelder, Tim (1998), "The Dynamical Hypothesis in Cognitive Science", *Behavioral and Brain Sciences* 21:1-14.

Watson, John B. (1913), "Psychology as the Behaviorist Views It", *Psychological Review* 20:158-177.

——— (1930), *Behaviorism*. Chicago: University of Chicago Press.

Wilson, Robert A., & Carl F. Craver (2007), "Realization: Metaphysical and Scientific Perspectives", in Paul Thagard (ed.), *Philosophy of Psychology and Cognitive Science,* Amsterdam: North-Holland Elsevier, 81-104.

Wimsatt, William C. (2006), "Reductionism and its heuristics: Making methodological reductionism honest", *Synthese* 151:445-475.

Wright, Larry (1973), "Functions", *Philosophical Review* (82):139-168.

Wright, Cory, & William Bechtel (2007), "Mechanisms and Psychological Explanation", in Paul Thagard (ed.), *Philosophy of Psychology and Cognitive Science,* Amsterdam: North-Holland Elsevier, 31-79.

Zuriff, G. E. (1985), *Behaviorism: A Conceptual Reconstruction*. New York: Columbia University Press.