Université de Montréal

Analyse du débat entre le cognitivisme classique et le connexionnisme sur la question des concepts

par

Éric D. Racine

Département de philosophie

Faculté des sciences et des arts

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

octobre, 2000

Université de Montréal
Faculté des études supérieures

Ce mémoire intitulé :

Analyse du débat entre le cognitivisme classique et le connexionnisme sur la question des concepts.

présenté par :

Eric D. Racine

a été évalué par un jury composé des personnes suivantes :

Mémoire accepté le : 16 janvier 2001

ÉPIGRAPHE

Nommons-nous le savoir, le concept, l'essence ou le vrai en revanche l'étant ou l'objet le concept, et entendons-nous en revanche l'étant ou l'objet, l'examen consiste alors pour nous à aller voir si l'objet correspond à son concept. L'on voit bien que les deux sont la même chose; mais l'essentiel pour toute l'investigation est de tenir fermement que ces deux moments, concept et objet, être pour un autre et être en soi-même, tombent dans le savoir que nous explorons, et que du coup nous n'avons pas besoin d'apporter avec nous des unités de mesure, et d'appliquer nos lubies et pensées lors de cette investigation; de ce que nous laissons tomber celles-ci, nous atteignons au fait de considérer la Chose comme elle est en et pour soi-même.

G.W.F. Hegel,
La phénoménologie de l'esprit (1807)

SOMMAIRE

Les concepts sont au coeur de nombreuses interrogations philosophiques traditionnellement retrouvées en épistémologie et en logique. Plus récemment, la philosophie de l'esprit a connu un développement important qui l'a amené a abordé de front cette question. Son objectif est de comprendre le rôle des concepts en tant que représentations mentales et en cela, elle se place en situation de collaboration avec les sciences cognitives contemporaines. Conséquemment, une littérature se situant au carrefour de ces deux disciplines se développe présentement autour de certains débats fondamentaux, notamment au sujet des concepts.

En fait, deux « paradigmes » s'affrontent aujourd'hui dans les sciences cognitives sur la question des concepts. D'un côté, le cognitivisme classique présente les concepts comme des symboles sur lesquels les processus mentaux viendraient opérer. D'un autre côté, le connexionnisme, puisant dans les réseaux de neurones formels, propose une définition des concepts en termes de région de classification dans des espaces d'activation. Des propriétés différentes sont attribuées aux concepts dans ces deux perspectives. D'une part, les concepts intègrent les exigences de compositionalité, de productivité et de systématicité. Ils ont une signification fixe qui leur permet d'être des éléments constitutifs des représentations mentales dans le cadre d'une syntaxe et d'une sémantique combinatoires. L'accent est donc mis sur les aspects sémantiques des concepts mais au détriment d'une certaine fragilité et d'une non-plausibilité neurobiologique. D'autre part, les concepts sont des ensembles dynamiques et statistiques de sous-caractéristiques, souples et robustes qui rendent compte de la subsomption conceptuelle, de l'application de concepts (généralisation spontanée), de l'apprentissage et des assises empiriques de la conceptualité mais aux dépens des aspects sémantiques.

Le but du présent mémoire est de présenter l'opposition fondamentale entre ces deux théories des concepts et de la clarifier. Pour ce faire, nous élaborons une typologie constituée de quatre dimensions d'explication : la dimension fonctionnelle-causale (F-C), la dimension relationnelle-causale (R-C), la dimension fonctionnelle-descriptive (F-D) et la dimension relationnelle-descriptive (R-D). La première décrit le fonctionnement causal d'un sous-système, la deuxième la relation causale entre un sous-système et un super-système, la troisième fournit une description abstraite du fonctionnement d'un sous-système et la quatrième une description abstraite des relations entre un sous-système et un super-système.

Notre analyse conclut que le cognitivisme se cantonne dans la dimension R-D de la cognition et le connexionnisme dans la dimension F-D. Par conséquent, nous pouvons constater deux choses. Premièrement, pour rendre compte pleinement de toutes les facettes de la cognition, nous devons faire appel à toutes les dimensions d'explication disponibles étant donné le caractère partiel des théories courantes. Deuxièmement, nous ne pouvons nous contenter d'opter pour une seule théorie afin de rendre compte de la conceptualité mais nous devons plutôt tenter de dégager une explication plus globale intégrant les quatre dimensions d'une explication de la conceptualité.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

REMERCIEMENTS

J'aimerais remercier mon directeur de mémoire, Daniel Laurier du Département de philosophie de l'Université de Montréal. Ses commentaires précis, ses conseils judicieux, son respect pour mes idées parfois très embryonnaires, sa disponibilité ainsi que son soutien m'ont été d'un grand secours dans toutes les étapes de ma recherche. En outre, je dois le remercier pour l'appui financier qu'il m'a accordé sous la forme d'une bourse de recherche et d'un contrat de traduction. J'aimerais aussi remercier le Département de philosophie de l'Université de Montréal pour une bourse d'excellence qu'il m'a offerte lors de mon arrivée au Département. Cette bourse m'a permis, entre autres, de travailler pour trois professeurs, Jean Roy, Pierre Poirier et Christian Nadeau que j'aimerais remercier. Je tiens à remercier le groupe d'introduction au connexionnisme de l'UQÀM pour un cadre stimulant ainsi que Michel Seymour et le groupe de recherche sur le langage, l'innéité et l'interprétation. Je salue amicalement Stéphane Potvin et Jean Frigault avec lesquels j'ai pu discuter de mon projet. Enfin, je voudrais souligner l'appui de ma conjointe, Nathalie Prud'homme, ainsi que celui de ma famille, notamment de ma mère, au cours de mes études universitaires.

INTRODUCTION

Qu'est-ce qu'un concept? Voilà une question qui a animé la réflexion philosophique depuis ses débuts. Car mieux comprendre la conceptualité a été un enjeu fondamental lié intrinsèquement à de nombreuses questions épistémologiques et métaphysiques. La théorie des Idées de Platon et l'interrogation d'Aristote sur les catégories de l'être en sont des témoignages révélateurs. Pour ces deux philosophes, la clarification de la nature du concept était la clef de voute d'une étude de la connaissance et en retour une source de clarification pour certains enjeux philosophiques. Elle était l'enquête dernière de l'esprit à la fois dans la perspective de la compréhension du monde et dans la perspective de sa propre compréhension. En fait, une prémisse centrale que nous pouvons extirper des débuts de la pensée occidentale est que l'esprit humain est conceptuel ou, selon une tournure plus pragmatique, l'esprit est quelque chose ayant des capacités conceptuelles. Sans l'ombre d'un doute, cette prémisse a marqué profondément la conception occidentale de l'esprit et, d'une façon plus globale, de l'être humain. La question des concepts a donc pris rapidement une importance fondamentale. Elle surplombe encore nos réflexions sur l'esprit.

Loin de demeurer confinée à la philosophie, l'étude de la conceptualité est maintenant au coeur des préoccupations d'une discipline scientifique interdisciplinaire relativement nouvelle mais dont l'importance se fait grandissante, les sciences cognitives. Les philosophes de l'esprit contemporains[1], voyant l'intérêt et la complexité philosophiques de la question, participent aux échanges théoriques fondamentaux. Car étudier la conceptualité humaine n'est pas une mince tâche et n'est surtout pas dépourvu d'ambiguïtés

[1] Il faut comprendre ici la philosophie de l'esprit au sens de *philosophy of mind* dont l'objet n'est pas une philosophie spiritualiste mais plutôt une enquête critique sur la nature des phénomènes mentaux. (Voir Engel, 1994 à ce sujet)

philosophiques. D'ailleurs, cette question est d'autant plus délicate qu'elle fait l'objet d'un débat entre deux approches en sciences cognitives, soit le cognitivisme classique et le connexionnisme.

Jusqu'au milieu des années quatre-vingt, les sciences cognitives étaient dominées théoriquement et pratiquement par le cognitivisme classique. Dans ce cadre, l'esprit est le *software* reposant sur un support physique, le *hardware* (le cerveau). La tâche qui incombe aux sciences cognitives est de découvrir les processus mentaux dans le *software*. Ces processus sont des algorithmes mentaux dans lesquels s'effectue une manipulation de symboles. Les concepts ce sont des symboles, des éléments constitutifs de la pensée.

Par contre, depuis une quinzaine d'années, un nouveau « paradigme », le connexionisme ou l'approche PDP[2] (*parallel distributed processing*) gagne en importance. Il diffère du cognitivisme classique sur de nombreux points. Selon le connexionnisme, les processus mentaux ont lieu à l'intérieur de réseaux regroupant un grand nombre d'unités (ou neurones). Brièvement, chaque unité se trouve, à tout moment, dans un certain état d'activation. L'état d'activation d'un neurone est déterminé à un temps donné par une règle d'activation. Celle-ci permet de calculer l'apport des entrées du neurone en fonction d'une matrice de connectivité qui établit l'apport de chaque entrée pour un neurone donné. Une fonction de sortie détermine ensuite si l'état d'activation du neurone est suffisant pour déclencher une sortie. Enfin, une règle d'apprentissage permet de modifier la connectivité entre les neurones de façon à modifier leurs interactions. Les connexions entre les neurones déterminent ce que le réseau « représente » ou « connaît »[3] car, pour une

---

[2] Ou encore neurocomputationnalisme ou perspective neurocomputationnelle, chaque dénomination mettant l'accent sur une dimension particulière (par ex. : le traitement en parallèle, le stockage des connaissances dans les matrices de connectivité ou l'allure neuronale des modèles). Nous optons pour l'expression « connexionnisme » qui est la plus largement répandue.

[3] Nous n'entrons pas pour l'instant dans les questions complexes soulevées par l'interprétation appropriée des propriétés des réseaux de neurones car chaque propriété des réseaux fait l'objet de controverses substantielles

computation donnée, c'est l'état des connexions du réseau qui déterminera de quelle façon le stimulus sera traité. Mais qu'est-ce qu'un concept dans cette perspective?

Une interprétation répandue est que le concept est contenu dans le schéma d'activation du réseau. Plusieurs techniques permettent d'ébaucher les linéaments d'une théorie prototypique des concepts à partir des réseaux de neurones. Le plus souvent, il s'agit de représenter l'activation du réseau à l'aide d'un espace d'activation à n-dimensions. Les concepts ce sont les régions de classification à l'intérieur de cet espace. Cependant, une telle position semble entrer en conflit avec la conception classique des concepts. Qui a le mot juste dans cette histoire? Les concepts sont-ils des symboles ou des connexions?

En ce qui nous concerne, la question est de déterminer quelle est l'interprétation juste de la capacité de ces deux modèles à rendre compte de la conceptualité.[4] Il s'agit donc de déterminer quel est le rapport entre le cognitivisme et le connexionnisme sur la question des concepts. Y a-t-il un modèle fondamentalement juste et un autre erroné? ou bien y a-t-il une complémentarité possible entre ces deux approches? Notre hypothèse générale est que le cognitivisme et le connexionnisme sont complémentaires sur la question des concepts. Cependant, un travail de présentation et de clarification doit être effectué afin de rendre cette affirmation plausible.

L'enjeu du présent mémoire est donc de clarifier le rapport entre le cognitivisme et le connexionnisme sur la question des concepts. Ceci dit, nous sommes conduits à déterminer localement si les deux approches sont compatibles ou incompatibles. Pour ce faire, nous suivons la présentation suivante. Premièrement, nous présentons les concepts dans le

---

et chaque choix terminologique peut refléter une distinction dans l'interprétation. Nous tentons d'être neutre pour l'instant.

[4] Ce sont des modèles dans la mesure où ils tentent de décrire à un niveau abstrait et schématique le fonctionnement de l'esprit. Ils sont computationnels dans la mesure où ils accordent à l'esprit des capacités de calcul (inférences, résolution de problèmes, etc.).

cognitivisme classique selon lequel les concepts sont des symboles. L'exposé se fait essentiellement à partir des écrits de Fodor mais aussi, dans une moindre mesure, de Pylyshyn et de Johnson-Laird.[5] Deuxièmement, nous présentons les concepts dans le connexionnisme selon lequel les concepts sont des régions de classification dans des espaces d'activation, ou plus grossièrement « des connexions ». À ce niveau, c'est l'interprétation de Paul M. Churchland qui est exclusivement présentée.[6] Enfin, dans un troisième temps, nous faisons un effort pour clarifier le débat sur les concepts. Nous faisons intervenir des éléments tirés des écrits de Clark, Smolensky et Rey à cet effet. Cette démarche nous conduit à la systématisation de trois éléments de clarification (le niveau d'analyse, le type d'analyse et le type d'entreprise) sous la forme de quatre dimensions d'explication (fonctionnelle-causale / relationnelle-causale / fonctionnelle-descriptive / relationnelle-descriptive) s'appliquant à différents niveaux d'analyse. La stratégie d'analyse développée avance que les deux types de modélisation ne sont pas incompatibles puisqu'ils s'appliquent à deux dimensions différentes de la conceptualité. Cependant, la compatibilité est fonction inverse des prétentions accordées aux deux approches étant donné que les interprétations de Fodor et de Churchland sont en tant que telles incompatibles. Néanmoins, avec des clarifications visant à dissiper certaines ambiguïtés d'interprétation, il est possible d'attribuer des prétentions plus modestes aux deux approches. Les concepts revêtent alors des caractéristiques relevant à la fois du cognitivisme classique et du connexionisme.

---

[5] L'appel à différents auteurs pose moins problème dans le cadre du cognitivisme classique étant donné l'homogénéité relative des positions exprimées par ces auteurs.

[6] Cette précaution est de mise étant donné la variété d'interprétations des capacités conceptuelles des réseaux de neurones.

Chapitre 1 : Les concepts dans le cognitivisme classique

Ce premier chapitre présente les concepts selon le cognitivisme classique.[7] La première section contient une brève mise en contexte explicitant les racines historiques et théoriques du cognitivisme classique. Vient ensuite une présentation générale du cognitivisme classique, c'est-à-dire de sa conception des symboles et des processus mentaux. Une deuxième section est consacrée à l'interprétation de Jerry Fodor. Fodor soutient que seulement les symboles du cognitivisme classique peuvent rendre compte de certaines caractéristiques essentielles des concepts dont, entre autres, la compositionalité, la productivité et la systématicité. Enfin, dans une troisième section nous considérons les critiques connexionnistes adressées aux thèses de Fodor et au traitement des concepts dans le cognitivisme classique. Ces reproches visent la rigidité et la fragilité excessives des concepts, leur insensibilité à l'environnement externe, leur implausibilité neurobiologique, leur lien avec la question de l'autonomie de la psychologie, leur rôle dans la relation entre le langage et la pensée, le lien exclusif entre le principe de compositionalité et le cognitivisme classique ainsi que l'hypothèse du langage de la pensée.

1.1. Introduction générale au cognitivisme classique

Cette première section est une introduction générale aux thèses du cognitivisme classique. Une brève mise en contexte précède une présentation générale des symboles et des processus mentaux.

---

[7] Nous désignons de façon équivoque le cognitivisme classique comme la perspective classique ou symbolique.

1.1.1. Mise en contexte

Le cognitivisme classique est l'approche dominante en sciences cognitives. Il puise ses racines dans la logique, la théorie de la computation et l'intelligence artificielle. (Bechtel et Abrahamsen, 1991, 9-11) Il est né dans les années cinquante et soixante en réaction au béhaviorisme. Car sa thèse principale est que la cognition est un élément explicatif central de l'être humain, contrairement au béhaviorisme pour lequel la considération des comportements observables (stimuli et réponses) suffisait. Trois figures importantes ont pavé la voie au cognitivisme classique, soit Alan Turing, Allen Newell et Herbert A. Simon.[8]

Le logicien et mathématicien Alan M. Turing a formalisé la notion intuitive de computation.[9] (Turing, 1936) Il a fondé sa démonstration sur l'hypothèse d'une machine capable d'effectuer des calculs, une machine de Turing. (Kim, 1998, 80; Boolos et Jeffrey, 1996, 19-33) Sa conclusion était que si une machine de Turing pouvait calculer une opération, alors cette dernière était effectivement computable. Ce résultat constituait une formalisation de la notion intuitive de computabilité. L'idée a été ensuite de définir la capacité de penser en termes d'opérations sur des symboles internes effectuées à la manière des machines de Turing. (Turing, 1936 et 1950) Ces hypothèses sont à la base du cognitivisme pour lequel les processus mentaux sont des computations effectuées sur des symboles. L'intuition de traiter l'esprit comme une machine manipulant des symboles vient donc en partie des écrits de Turing. D'ailleurs, il est régulièrement mentionné comme

---

[8] Il ne s'agit pas ici d'une histoire complète du cognitivisme mais de quelques éléments historiques nous permettant d'introduire certaines thèses du cognitivisme. Bien évidemment, il y a les contributions importantes de Chomsky avec sa grammaire générative, de Church au niveau de la théorie de la computation, de Von Neumann pour l'architecture de l'ordinateur tel que nous le connaissons, de Shannon, Wiener et McCarthy pour la théorie de l'information, etc.

l'un des principaux responsables de la diffusion de cette idée. (Fodor, 1998; Andler, 1992; Clark, 1989; Jonhson-Laird, 1988; Pylyshyn, 1985)

Deuxièmement, Simon et Newell se sont inspirés de Turing pour formuler « l'hypothèse du système symbolique physique » (*physical symbol system hypothesis*). Un système symbolique physique (SSP) répond à aux moins trois conditions. (Clark, 1989, 11; Allen, 1982 et 1980) (1) Il doit contenir des symboles qui sont des schémas physiques (*physical patterns*) qui peuvent être structurés. (2) Il doit contenir des structures de symboles et un ensemble de processus (eux-mêmes codés sous forme de structures symboliques) opérant sur ces structures. (3) Il est situé dans un monde plus large et peut lui être relié par désignation ou interprétation. En fait, un SSP est tout système capable de manipuler des *tokens* de symboles[10] en vertu de leur contenu sémantique. L'hypothèse du SSP peut selon Newell et Simon servir de critère pour attribuer l'intelligence et la pensée à des systèmes car, selon eux, tout système intelligent ou pensant est un SSP. La réflexion de Newell et Simon a eu une importance centrale pour le cognitivisme qui définit l'intelligence, la pensée et la conceptualité en termes de manipulation de symboles.

1.1.2. Le cognitivisme classique

Les contributions de Turing, Newell et Simon amènent un point de vue où la cognition est entendue en un sens modérément restreint, soit la manipulation d'information codée sous forme de symboles.[11] (Lemaire, 1999, 24; Johnson-Laird, 1988, 34) Selon cette conception de la cognition le but des sciences cognitives est de découvrir les processus cognitifs

---

[9] Turing n'est pas le seul a avoir travaillé fructueusement sur le problème de la computation. Emil Post et Alonzo Church ont développé des formalismes leur permettant d'arriver à des conclusions similaires. (Pylyshyn, 1985, 50)

[10] C'est-à-dire des symboles (*types*) instanciés physiquement.

impliqués dans les tâches cognitives entendues comme manipulation de symboles. Étant donné le rôle central des symboles, il est donc primordial de clarifier ce que sont les symboles et en quoi consiste leur manipulation.

1.1.2.1. les symboles

Qu'est-ce qu'un symbole? À première vue, les symboles sont des entités représentatives qui renvoient à des éléments du monde. Ils correspondent généralement au langage ordinaire que nous utilisons dans la vie quotidienne tels les mots « chien », « table », etc. On les retrouve dans des systèmes présentant trois composantes : (1) un ensemble de symboles primitifs et un ensemble de principes permettant de contruire des symboles complexes à partir des plus simples; (2) un ensemble d'entités constituant le domaine représenté (*symbolized*) et (3) une méthode pour relier les symboles aux entités qu'ils représentent. (Johnson-Laird, 1988, 28-9 et Fodor et Pylyshyn, 1988, 7-9) Mais pourquoi faire appel aux symboles?

Le recours aux symboles est fait en vertu de la valeur des explications formulées en termes de leur manipulation.[12] La plupart des ouvrages du cognitivisme classique aborde cette question en distinguant différents niveaux d'analyse. Le cadre de référence est le plus souvent inspiré de la distinction de David Marr (*Vision*, 1982)[13] entre les niveaux computationnel, algorithmique et implémentationnel. (Voir tableau 1.2.1.1.a.) En gros, le niveau computationnel est le niveau auquel on décrit quel est *le but* de la computation d'un système donné. Au niveau algorithmique, on détermine *comment* la théorie

---

[11] Il faut prendre note, sans en dire plus, qu'il y a des débats sur l'interprétation des notions de « computation » et de « symbole » à l'intérieur même du cognitivisme classique. Voir Pylyshyn (1985) à ce sujet. L'exposé qui suit tente de demeurer général en évitant les controverses.

[12] Rappelons que les processus mentaux, objets d'étude du cognitivisme classique, sont des manipulations de symboles.

[13] Voir, entre autres, Rumelhart et McClelland (1986b et 1985) et Clark (1989, 18-9) sur cette question.

computationnelle peut être implémentée, c'est-à-dire quel algorithme sera responsable de la transformation des entrées et des sorties. Enfin, le niveau implémentationnel est le niveau où l'on détermine comment le système sera *physiquement réalisé*. De façon générale, les cognitivistes insistent pour restreindre l'explication cognitive au niveau computationnel et, dans une moindre mesure, au niveau algorithmique, étant donné que la manipulation de symboles se fait selon eux à ce niveau. (Johnson-Laird, 1988; Fodor et Pylyshyn, 1988; Broadbent, 1985; Pylyshyn, 1984) Mais comment concilier le traitement de symboles avec une ontologie physicaliste des sciences naturelles dont, après tout, les *sciences* cognitives sont censées faire partie?

Théorie computationnelle	**Représentation et algorithme**	**Implémentation dans le *hardware***
Quel est le but de la computation, pourquoi est-il approprié, et quelle est la logique de la stratégie par laquelle il peut être mené à terme?	Comment cette théorie computationnelle peut-elle être implémentée? En particulier, quelle est la représentation pour l'entrée et de la sortie, et quel est l'algorithme pour la transformation?	Comment la représentation et l'algorithme peuvent-ils être physiquement réalisés?

Tableau 1.1.2.1.a. : Les trois niveaux selon Marr (1982, 25)[14]

Celui qui parle de manipulation de symboles s'engage à expliquer comment son entreprise cadre avec l'ontologie des autres sciences. Les cognitivistes justifient l'usage des symboles par le fait qu'ils permettent d'accéder à un niveau d'analyse plus abstrait que les simples niveaux algorithmique ou implémentationnel. Les symboles sont les réalisations physiques des états représentationnels (sémantiques ou ayant une désignation/interprétation) d'un système et ils permettent de capturer un niveau de généralisation qui n'est pas accessible

[14] Traduction libre.

aux deux niveaux inférieurs.[15] (Pylyshyn, 1985, 1985, 26-8; Johnson-Laird, 1988, 28-9 et Fodor, 1998, 28-9) Prenons le cas de deux systèmes tels qu'un être humain et un automate partageant le même but au niveau computationnel, soit de vouloir consommer de l'énergie. L'humain peut atteindre ce but à l'aide d'un algorithme lui permettant de consommer des aliments « naturels ». L'automate, quant à lui, pourrait remplir cette exigence en consommant de l'électricité. Par conséquent, nous pouvons décrire ces systèmes identiquement au niveau computationnel car tous deux consomment de l'énergie mais ils diffèrent au niveau algorithmique puisque la série d'opérations (ou l'algorithme) qu'ils déploient pour parvenir à consommer de l'énergie n'est évidemment pas la même. Les descriptions diffèrent encore plus au niveau physique car la façon dont l'humain et l'automate consomment de l'énergie est instanciée physiquement de manière encore plus divergente, l'humain étant fait de composés organiques et l'automate de composés inorganiques. On en vient alors à une certaine conception des relations entre les différents niveaux d'analyse où ceux-ci sont relativement autonomes et liés approximativement (*loosely*). (Marr, 1982, 25)

Une façon répandue d'articuler les rapports entre niveaux d'analyse est de soutenir qu'un niveau plus abstrait est toujours dans une situation de réalisation multiple (*multiple-realization*) par rapport à un niveau d'analyse inférieur. (Pylyshyn, 1985, 33 et 38) (Voir la figure 1.1.2.1.a. qui illustre la relation entre les différents niveaux dans le cognitivisme classique.) Par exemple, le niveau computationnel peut être réalisé de différentes façons au niveau algorithmique et un algorithme peut lui-même être réalisé de plusieurs façons au niveau physique. Par conséquent, ce qui unit deux choses à un niveau computationnel n'est

[15] Un peu à la manière de Dennett (1987 et 1981), Pylyshyn soutient que le niveau d'explication computationel (intentionnelle pour Dennett) est utilisé par défaut lorsque les deux autres niveaux (conception et physique pour Dennett) ne sont pas disponibles. (Pylyshyn, 1988, 26)

peut-être pas apparent au niveau algorithmique et encore moins au niveau physique. [16] C'est donc le niveau supérieur qui permet de trouver la similitude. Dans ce schéma, les niveaux d'analyse inférieurs vont venir baliser l'implémentation des niveaux supérieurs. Cependant, il découle d'un tel cadre d'analyse que les trois niveaux sont relativement autonomes[17] puisqu'ils ont chacun leur taxonomie et leur utilité relativement indépendantes des autres niveaux.. En ce qui concerne les symboles, cela signifie en bout de ligne que les relations sémantiques au niveau des symboles forment un niveau d'analyse distinct dont la légitimité sera avérée par sa capacité à faire apparaître des généralisations inexprimables dans les termes des niveaux inférieurs. En tant qu'états mentaux internes et représentationnels, les symboles permettent d'introduire un niveau de description sémantique ayant une portée dans la chaîne causale. Mais comment est-ce possible?

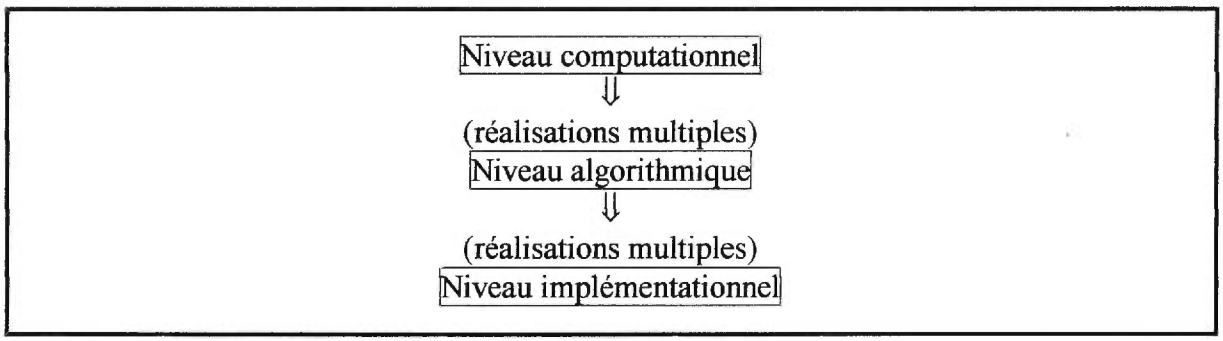

Figure 1.1.2.1.a. : La relation entre les niveaux dans le cognitivisme classique

Une façon répandue[18] d'expliquer concrètement l'usage des symboles et des liens entre niveaux d'explication est de tracer un parallèle entre le comportement cognitif des humains et le fonctionnement des ordinateurs (*the computer metaphor*). Premièrement, les

[16] Il s'agit d'une version du principe de survenance (*supervenience principle*). (Kim, 1998, 10)
[17] Il y a une très grande controverse au sujet des relations entre les niveaux d'analyse, au sujet du nombre de ces niveaux ainsi que le niveau traité par les modèles connexionnistes. Voir à cet égard : (Broadbent, 1985, Changeux et Dehaene, 1989; Paul M. Churchland, 1998c; Churchland et Churchland, 1998b; Clark, 1989; Dehaenne et Changeux, 1993; Fodor et Pylyshyn, 1988; Hofstadter, 1985; Johnson-Laird, 1988; Rueckl, 1991; Rumelhart et McClelland, 1985; Rumelhart et McClelland, 1986b; Rumelhart, Smolensky, McClelland et Hinton, 1986; Smolensky, 1992, 1986, 1988)
[18] La brève explication qui suit est tirée de Fodor (1990).

ordinateurs fonctionnent à l'aide de symboles (*tokens*). Les symboles lient les propriétés sémantiques et les propriétés causales parce que lorsqu'un ordinateur manipule un symbole (la syntaxe), il le fait en respectant ses propriétés sémantiques ayant un pendant au niveau causal. La syntaxe d'un symbole est une propriété abstraite qui se résume essentiellement à la forme (*shape*) du symbole. Or, la forme du symbole détermine le rôle causal du symbole, c'est-à-dire les effets et les causes de l'instanciation (*tokening*) d'un symbole. En outre, nous savons à partir de la logique formelle (*proof theory*) que certaines propriétés sémantiques peuvent être imitées par des relations syntactiques. Par conséquent, les ordinateurs offrent un environnement où le rôle causal d'un symbole est mis en parallèle avec son rôle inférentiel. (Fodor, 1990, 22-3) L'ordinateur offre donc une solution au problème du lien entre les propriétés causales et les propriétés sémantiques des symboles. Selon une version légèrement différente[19], celle de Pylyshyn, des « codes symboliques » (des *tokens* de symboles) incarnent les propriétés sémantiques des symboles (les *types*) d'une manière physique. Ces codes symboliques sont donc aussi des réalisations physiques de symboles, liant des propriétés sémantiques et des propriétés causales.

Deuxièmement, le fait que le cerveau fonctionne lui aussi avec des codes symboliques (Pylyshyn, 1988, 27) (ou des symboles pour Fodor, 1990) selon le cognitivisme conduit à la possibilité même d'étudier les représentations et les processus mentaux. D'une part, l'aspect symbolique des codes fait que les relations sémantiques sont respectées et que les états mentaux sont individués à un niveau d'analyse computationnel. L'individuation se fait d'après le contenu qui explique à ce niveau certaines régularités comportementales. La

[19] En gros, selon notre analyse, il y a une équivalence fonctionnelle entre les symboles (*tokens*) de Fodor et les codes symboliques de Pylyshyn. L'idée commune est qu'il existe des symboles (*types*) dont les instanciations encodées se font sous forme de symboles ou codes symboliques (*tokens*). Il y a donc une relation implicite *type-token* dans la notion de symbole. Le *type* est une entité sémantique pure tandis que le *token* est instancié physiquement (il « encode le *type*) et peut servir d'élément explicatif entrant dans la chaîne des causes et des

notion de contenu sémantique distingue les relations sémantiques ou conceptuelles des relations causales ou physiques puisque ce sont deux niveaux d'analyse distincts pour le cognitivisme classique, chacun ayant une légitimité circonscrite mais étant unis dans la notion de code symbolique ou de symbole.[20] Car, d'autre part, l'instanciation physique des symboles en tant que codes symboliques garantit leur effectivité causale. Autrement dit, ces codes sont les tokens physiques des symboles. Le cognitivisme conclut donc à la réalité psychologique des codes, ce qui fait du cerveau un système physique qui manipule des symboles. (Fodor, 1990, 23 et Pylyshyn, 1985, 40) Il faut comprendre ici les codes symboliques comme des « encodages » de symboles (*types*). Dans ce cas, les relations entre états mentaux ne seront pas exclusivement sémantiques (ou conceptuelles) comme les relations entre symboles (*types*) mais aussi causales. Alors, les codes symboliques mentaux unissent les niveaux sémantique et physique en respectant les propriétés sémantiques du niveau computationnel et les relations causales du niveau physique. (Pylyshyn, 1988, 26) Enfin, l'esprit peut être légitimement décrit et étudié dans ses opérations car les états du cerveau encodent des symboles à la manière de l'ordinateur. (Pylyshyn, 1988, 27) C'est ainsi qu'est légitimé l'usage des symboles pour définir un niveau d'analyse propre aux sciences cognitives, ce qui pave la voie à l'étude des processus mentaux dans une perspective scientifique qui respecte l'ontologie matérialiste.

---

effets. Évidemment, c'est ultimement la présence des *tokens* dans les ordinateurs et les cerveaux qui justifient l'usage des symboles comme élément d'explication de la cognition.

[20] Ceci dit, on voit bien pourquoi les modèles connexionnistes poseront problème selon le cognitivisme classique car ils tentent d'offrir une définition prototypique statistique des concepts où les relations entre prototypes seront en grande partie déterminées par leurs liens causaux. Par conséquent, les connexionnistes semblent coupables d'une confusion entre le niveau computationnel et le niveau physique car ils tentent de « réduire » les symboles (concepts) à des relations physiques.

1.1.2.2. les processus mentaux et la manipulation de symboles

Les symboles sont principalement évoqués dans le cognitivisme à titre d'éléments primitifs manipulés par les processus mentaux. Étant donné que l'esprit humain est un système symbolique physique (pour le cognitivisme classique), ce sont les « programmes »[21] ou algorithmes[22] régissant la manipulation de symboles qui forment la base des processus mentaux. (Johnson-Laird, 1988, 37; Pylyshyn, 1985, 54, 63 et 88) Ces éléments théoriques sont présentés comme permettant la vérification empirique d'hypothèses sur les processus mentaux. Prenons par exemple le cas (simplifié) où l'on supposerait qu'un raisonnement donné soit produit en quatre étapes distinctes (excluant l'entrée et la sortie), constituant ainsi une sorte d'algorithme mental. (Pylyshyn, 1985, 95) Alors, le scientifique de la cognition pourra émettre des hypothèses quant à l'augmentation du temps de réaction (temps avant qu'une réaction soit produite par l'individu) si l'une des étapes était court-circuitée. Le temps de réaction deviendrait dans ce cas un indice empirique de l'existence d'un tel programme mental. (Lemaire, 1991, 31) La figure 1.1.2.2.a. illustre cette stratégie d'explication propre au cognitivisme classique.

Nous avons vu dans cette première section que le cognitivisme classique insiste sur l'importance des symboles et de la manipulation de symboles dans l'étude des processus mentaux. Fodor poursuit cet ordre d'idées en explicitant des liens théoriques entre le cognitivisme classique et la question des concepts, ce que nous considérons plus en détail dans la section suivante.

[21] Les programmes sont des algorithmes encodés dans un langage donné. (Pylyshyn, 1985, 88)
[22] Ou même les « grammaires » telles que définies par Chomsky. (Johnson-Laird, 45)

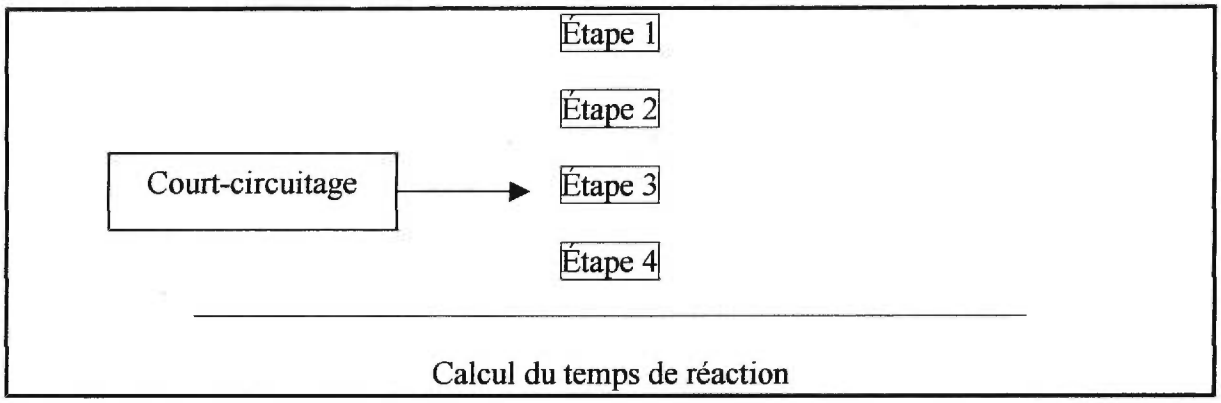

Figure 1.1.2.2.a : Le calcul du temps de réaction dans une tâche de raisonnement

1.2. Les concepts dans la perspective du cognitivisme classique : l'interprétation de Fodor[23]

Jerry Fodor est l'un des plus ardents défenseurs du cognitivisme classique. Il a accordé une attention particulière à la question des concepts. Cette deuxième section dégage sa position sur cette question en explicitant premièrement le lien entre la théorie représentationnelle de l'esprit de Fodor et la position classique sur les concepts. Un certain nombre de caractéristiques seront ensuite attribuées aux concepts, soit : (1) la compositionalité; (2) la productivité; (3) la systématicité ainsi que (4) quatre autres caractéristiques secondaires, à savoir : le fait d'être des états mentaux particuliers, d'être des catégories, d'être majoritairement appris et enfin d'être publics. Nous nous penchons ensuite brièvement sur l'hypothèse du langage de la pensée et de son lien intrinsèque avec la théorie des concepts de Fodor et nous soulignons finalement l'opposition de Fodor au holisme, au pragmatisme et à la théorie prototypique des concepts, ce qui légitime l'atomisme qu'il soutient.

[23] Nous présentons ici les éléments essentiels de la position de Fodor. Dans le cadre du débat entre Churchland et Fodor, présenté dans le prochain chapitre, section 2.3.3., nous explorons certaines implications précises de sa position plus pleinement.

1.2.1. la théorie représentationnelle de Fodor et la question des concepts dans le cognitivisme classique

La théorie représentationnelle de l'esprit (*representational theory of mind*) de Fodor, qui commande sa théorie des concepts, est une tentative d'offrir une explication rigoureuse du réalisme intentionnel. Celui-ci se définit par deux engagements théoriques, soit : (1) qu'il existe des états mentaux dont les occurrences et les relations causent des comportements et ce, en respectant les généralisations de la psychologie croyance-désir (*belief-desire psychology*)[24] et (2) que ces états mentaux causalement efficaces sont aussi sémantiquement évaluables[25]. (Fodor, 1990, 5) Fodor résume sa théorie en cinq thèses. Premièrement, l'explication psychologique est intentionnelle. (Fodor, 1998, 7) Elle fait donc référence obligatoirement à des états mentaux qui sont identifiés sous des descriptions intentionnelles. Ils ont un contenu déterminé en fonction de leur relation référentielle, leur « *aboutness* ». Or, les représentations mentales sont les porteurs primitifs du contenu intentionnel. (Fodor, 1998, 7) Elles confèrent ainsi une intentionalité aux langues naturelles. Fodor adopte donc un réalisme intentionnel et mental selon lequel l'intentionalité des représentations mentales est primitive et première dans l'économie cognitive et l'intentionalité des propositions linguistiques est dérivée. Deuxièmement, les représentations mentales ont des pouvoirs causaux et sont susceptibles d'évaluation sémantique (vrai/faux, etc.). (Fodor, 1998, 7-8) Troisièmement, la pensée est computationnelle de sorte que penser c'est computer. (Fodor, 1998, 9) Or computer c'est manipuler des symboles. Cependant, l'association des symboles n'est pas exclusivement causale comme l'associationnisme le suggère (et le connexionnisme n'est qu'une forme

---

[24] C'est-à-dire la psychologie du sens commun dont le syllogisme pratique est le paradigme.

[25] C'est-à-dire la possibilité de déterminer si une croyance correspond à un état de chose et de même à quel état de chose correspond un désir réalisé.

d'associationnisme d'après Fodor) [26] car elle se fait en vertu de relations sémantiques. (Fodor, 1998, 10)

Quatrièmement, la signification est (plus ou moins) de l'information. (Fodor, 1998, 12) Fodor adhère à une théorie sémantique informationnelle selon laquelle ce qui confère un contenu à une représentation mentale est sa relation causale avec des entités du monde extérieur. Par exemple, le contenu de « chien » est quelque chose qui provient du fait que son instanciation est causée par des chiens. Mais cette thèse amène la conséquence suivante : les représentations coréférentielles doivent être synonymes. « H_2O » et « eau » signifieraient alors la même chose parce qu'elles sont causées par le même phénomène. Bien sûr, cela est faux car les contenus de représentation sont constitués par « autre chose ». La réponse à savoir ce qu'est cette « autre chose » est traditionnellement le rôle inférentiel de la représentation de sorte qu'une personne ayant la représentation « H_2O » peut soutenir que l'H_2O contient de l'hydrogène tandis que celui qui n'a que la représentation « eau » ne peut pas soutenir cette implication. Par contre, Fodor rejette l'idée que le contenu puisse être constitué même en partie par le rôle inférentiel pour au moins deux raisons (Fodor, 1998, 13). Dans un premier temps, Fodor adhère à la thèse de Turing (Voir section 1.1.1.) selon laquelle l'inférence est une opération sur des symboles (une computation). Or, on ne peut dans un tel cadre, à moins de commettre un cercle vicieux, expliquer ce qu'est le contenu d'un symbole en référence au rôle inférentiel tout en soutenant que l'inférence est définie comme une opération sur des symboles. (Fodor, 1998, 13) Dans un deuxième temps, Fodor pense que la sémantique du rôle inférentiel (*Inferential role semantics*) a des implications holistiques inévitables et intolérables. Voici pourquoi. La théorie

[26] Notons ici un point de désaccord avec le connexionnisme sur la question des concepts. Selon Fodor, le connexionnisme ne permet pas de rendre compte de l'union des propriétés sémantiques et causales des symboles comme le permet la métaphore de l'ordinateur. Le connexionnisme ne considère que l'aspect

représentationnelle de Fodor est liée à une conception de l'explication psychologique comme subsomption sous des lois intentionnelles (thèse 1). Par contre, le holisme, selon Fodor, tend à infirmer la possibilité d'une psychologie intentionnelle, quoiqu'on en dise, parce qu'il devient impossible d'attribuer un contenu fixe à des états internes dans ce cadre.[27] (Fodor 1998, 13 et Fodor et Lepore, 1992) On pourrait aussi ajouter que Fodor défend une forme d'atomisme selon lequel avoir un concept ne nécessite jamais la connaissance d'autres concepts. Par conséquent, la théorie sémantique informationnelle de Fodor exclut la possibilité que le contenu conceptuel soit déterminé même en partie par les relations inférentielles (qui se font toujours entre plusieurs concepts). Au contraire, la théorie sémantique informationnelle rend compte du contenu exclusivement en termes de relations symbole-monde. (Fodor, 1998, 14) Contrairement à la sémantique du rôle inférentiel, Fodor soutient donc que les concepts coréférentiels sont *ipso facto* des synonymes au niveau du contenu (de l'information). Par contre, il est prêt à soutenir que ce sont des concepts différents. Autrement dit, l'individuation du contenu n'est pas l'élément unique dans l'individuation des concepts. Dans un schéma frégéen traditionnel, la référence et surtout le mode de présentation sont les éléments supplémentaires permettant d'individuer les concepts.[28] Toutefois, Fodor n'est pas convaincu que le mode de présentation est nécessairement le sens, ce qui l'éloigne des frégéens. (Fodor, 1998, 15) Cela nous conduit à présenter la cinquième thèse de Fodor selon laquelle ce qui différencie les concepts coextensifs se situe « dans la tête ».

---

causal des symboles comme les anciennes théories empiristes des Idées le faisaient, négligeant ainsi leur aspect sémantique. (Fodor, 1990, 23 et 24)

[27] Voir les sections 1.2.5., 1.2.7. et 2.3.3.2. où ce lien entre le holisme, la psychologie intentionnelle et la problématique du contenu est explorée.

[28] Par exemple, les concepts « chien » et « eau » diffèrent sur les deux points tandis que les concepts « eau » et « H_2O » diffèrent seulement quant au deuxième.

Cinquièmement, contrairement à Frege qui soutenait que le sens (l'intensionalité) distinguait des concepts avec la même extension (le même référent), Fodor affirme que l'individuation des concepts se fait par les états mentaux. Dans un premier temps, le sens ne peut pas être le mode de présentation. Le mode de présentation est quelque chose de plus « fin » que le sens lorsqu'on tente de déterminer le contenu des concepts. Prenons par exemple, une situation où un agent A affirme à un autre (B) que Jackson est un peintre et que Pollock est un peintre. Pour B, le sens des concepts « Jackson » et « Pollock » semble être le même. Cependant, même si ces deux concepts ont le même sens, il est parfaitement légitime pour B de se demander si Jackson et Pollock sont le même peintre. (Fodor, 1998, 16) Ce que l'on voit ici, c'est que le sens est quelque chose de moins précis que le mode de présentation. On semble ici admettre qu'il y a une possibilité de présenter un sens donné de différentes façons. Dans un deuxième temps, pour Frege (d'après Fodor), il semble n'y avoir qu'une façon de saisir/entretenir un mode de présentation. Cependant, si les sens sont les modes de présentation et que ceux-ci sont censés individuer les contenus selon Frege, mais que les sens ne peuvent pas individuer clairement les concepts, alors il y a un problème avec l'usage des sens (en tant que mode de présentation) pour individuer le contenu. La solution de Fodor est la suivante : les modes de présentation (et non les référents bien sûr) peuvent individuer les concepts parce que les modes de présentation sont des objets mentaux. Ce qui distingue les concepts coextensifs se situe dans les effets ou les causes proximales des processus mentaux qu'ils provoquent respectivement. Autrement dit, les modes de présentation sont fonctionnellement individués, c'est-à-dire par leur rôle dans la chaîne des causes et des effets. (Fodor, 1998, 19) Alors l'identité d'un mode de présentation est assurée par ses effets lorsque nous l'entretenons et il n'y a donc qu'une seule façon d'entretenir chaque mode de présentation dans cette optique. (Fodor, 1998, 20)

Ces cinq thèses définissent la théorie représentationnelle de l'esprit de Fodor. Mais quel est le lien entre cette théorie et le cognitivisme classique sur la question des concepts? Il est clair pour Fodor que pour rendre compte des concepts nous devons faire appel aux systèmes symboliques tels que présentés par le cognitivisme classique. Les explications psychologiques selon lui font (exclusivement) appel aux représentations mentales qui forment la base des attitudes propositionnelles. Or les représentations mentales sont, entre autres, compositionnelles, systématiques et productives. Ce sont donc là des exigences irrévocables dont toute théorie des concepts doit rendre compte. Cependant, le connexionnisme n'est pas capable de soutenir ces exigences, ce que le cognitivisme classique fait adéquatement (selon Fodor). Il s'ensuit que le cognitivisme classique est le cadre approprié pour les sciences cognitives et qu'il offre la théorie appropriée des concepts, lesquels constituent des éléments explicatifs importants pour Fodor.

1.2.2. la compositionalité conceptuelle

La principe de compositionalité est fondamental car il commande (et explique) la systématicité et la productivité des concepts et d'une façon générale l'ensemble de la théorie de Fodor.[29] (Fodor et Pylyshyn, 1988, 41) Fodor le formule de la façon suivante : *les concepts sont les éléments constitutifs des pensées et, dans un nombre indéfini de cas, les uns des autres. Les représentations mentales tirent leur contenu du contenu de leurs éléments constitutifs.* (Fodor, 1998, 25)[30] Autrement dit, le caractère compositionnel des concepts et des représentations permet de rendre compte de la production (par idéalisation, infinie) de mots et d'expressions correspondantes. La compositionalité dérive donc du

---

[29] Il sera donc impossible de présenter les principe de compositionalité, de productivité et de systématicité sans quelques recoupements.

[30] Traduction libre de : *concepts are the constituents of thoughts and, in indefenitely many cases, of one another. Mental representations inherit their contents from the content of their constituents.* (Fodor, 1998, 25)

constat suivant.[31] S'il y a un nombre infini d'expressions dans un langage donné (et que chaque expression langagière correspond à une représentation mentale plus primitive) et que les capacités représentationnelles des individus sont finies, alors le nombre de concepts, les éléments constitutifs des représentations, doit être fini. Or, pour bien saisir l'importance de cette dernière affirmation, nous devons nous rappeler que chaque représentation complexe doit son contenu à ses éléments constitutifs primitifs. Cela est possible parce que ces derniers sont individués par leur contenu et leur rôle fonctionnel. (Fodor, 1998, 95; 1987, 138) Dans ce cadre, les éléments constitutifs ont une valeur indépendante du contexte (*context independent*), ce qui rend possible leur combinaison (caractère compositionnel) car si le contenu des éléments changeait selon le contexte, alors on ne pourrait plus rendre compte de la compositionalité qui elle-même explique, à partir de moyens finis, la production infinie de représentations et d'expressions. (Fodor, 1998, 27) Il y a donc une filiation directe entre la défense de la compositionalité et la revendication d'une syntaxe et d'une sémantique combinatoires où le contenu des énoncés complexes découle du contenu des éléments constitutifs. (Fodor et Pylyshyn, 1988, 41)

L'exemple suivant illustre le développement du paragraphe précédent. Si un individu comprend *John loves the girl*, alors il doit comprendre *The girl loves John*. Cela est expliqué par le fait que la syntaxe et la sémantique sont combinatoires et que cet aspect combinatoire est causé par le sens fixe des éléments constitutifs de la représentation mentale, les concepts. Autrement dit, la structure de la syntaxe se reflète dans la sémantique car les éléments constitutifs primitifs ont un sens déterminé qu'ils contribuent à une représentation complexe. (Fodor et Pylyshyn, 1988, 43; Fodor, 1987, 138 et Fodor, 1998, 27 et 39) Mais que faire des expressions où le sens des concepts semble dépendre de

[31] Voir aussi Matthei et Roeper (1988) pour une présentation de ces thèses.

leur usage? Comment rendre compte de ce phénomène si les éléments constitutifs contribuent toujours un sens fixe et déterminé?

Prenons par exemple la différence sémantique entre *feed the chicken* et *chicken to eat*? Une stratégie plus pragmatique nous conduirait à soutenir que le sens de *chicken* dépend de l'usage. Dans un premier cas, il désigne le poulet comme animal vivant et dans le deuxième, il désigne plutôt le poulet comme met à déguster. Il y a donc une ambiguïté animal/nourriture dans ces deux expressions. Comment soutenir que l'élément constitutif, le concept *chicken* contribue alors la même chose aux deux pensées ou qu'ils signifient la même chose? Pourquoi ne pas soutenir (plus simplement) que le sens dépend ici de l'usage? Fodor refuse d'entrer dans cette dernière voie car il juge cette tendance comme une exagération de la variabilité de la signification lexicale. (Fodor et Pylyshyn, 1988, 42) Selon lui la différence sémantique entre les deux expressions est mieux expliquée par la « syncatégorématicité » (*syncategorematicity*) du terme *chicken*. Dans cette optique, il ressemblerait au terme *good* qui inclut une variable permettant d'exprimer une pluralité de sens. On peut dire en effet, *good fight*, *good book*, etc. et dans ce sens le terme *good* semble exprimer l'idée que quelque chose de bon est quelque chose qui répond à nos intérêts. (Fodor et Pylyshyn, 1988, 42-3) C'est ainsi, que l'on peut rendre compte des variations de sens à l'intérieur d'une forme d'atomisme sémantique et conceptuel selon Fodor.[32]

---

[32] À notre avis, ce dernier développement mériterait d'être précisé par Fodor car il semble affaiblir la crédibilité du principe de compositionalité selon lequel les éléments constitutifs contribuent une signification fixe aux représentations complexes. La notion de syncatégorématicité a un usage plus pertinent pour l'adjectif *good* qui comprend effectivement une sorte de malléabilité sémantique lorsqu'appliqué à différents substantifs. Cependant, il n'est pas clair que le substantif *chicken* puisse inclure à première vue cette caractéristique de malléabilité. Peut-être qu'une remise en cause de la compositionalité pourrait se faire à partir de cette brèche, mais nous ne faisons que noter ici cette difficulté.

L'intérêt pour la compositionalité devient encore plus évident si l'on ajoute une hypothèse de la psycholinguistique selon laquelle les humains utilisent le langage pour exprimer leurs pensées. La capacité d'exprimer une pensée est liée à la capacité de formuler des énoncés. Par conséquent, si la capacité de formuler certains énoncés est liée à la capacité d'en formuler certains autres, alors la capacité de penser certaines pensées (*thoughts*) dépend de la capacité d'en penser certaines autres (lorsqu'il y a un partage de certains éléments constitutifs). Dans ce cas, les représentations mentales correspondant à *John loves the girl* et *The girl loves John* doivent être elles aussi compositionnelles à l'instar des expressions linguistiques. De plus, si la pensée dépend de capacités représentationnelles comme Fodor le soutient, alors le fait que certaines pensées soient interreliées implique des représentations correspondantes qui sont elles aussi interreliées. (Fodor et Pylyshyn, 1988, 44) Mais comment les représentations mentales peuvent-elles être reliées et compositionnelles?

La réponse de Fodor est que tout comme les expressions linguistiques, les représentations complexes ont une structure interne. (Fodor, 1987, 136) Car de la même façon que les expressions linguistiques *John loves the girl* et *The girl loves John* sont reliées sémantiquement (et syntactiquement) par leur éléments constitutifs, de même en est-il des représentations mentales qui sont reliées par leurs éléments constitutifs. Cette interprétation doit se ramener à une conception atomiste des concepts car les éléments constitutifs de la pensée doivent trouver un socle sémantique dans leur individuation en tant qu'éléments constitutifs primitifs. L'exigence de compositionalité est donc posée pour les modèles connexionnistes.

1.2.2.1. le connexionnisme et la compositionalité

Fodor doute de la capacité des modèles connexionnistes à rendre compte de la compositionalité. Premièrement, ils ont de la difficulté à rendre compte de l'aspect combinatoire des représentations mentales étant donné que, pour le connexionnisme, les représentations mentales ne sont pas des éléments primitifs constitutifs. Du moins, elles ne sont pas des catégories aussi rigides et définies que dans le cognitivisme classique à cause de leur sensibilité à l'environnement externe. Deuxièmement, l'une des exigences corrolaires à la compositionalité et à l'aspect combinatoire des représentations mentales est l'identité conceptuelle car, comme nous l'avons souligné, les éléments constitutifs doivent contribuer la même chose dans la formation de représentations complexes. Or encore une fois, à cause de la grande sensibilité des modèles connexionnistes à l'environnement externe, cela pose une difficulté puisqu'ils sont incapables d'assurer une identité conceptuelle de ce genre. Au mieux, selon Fodor, on évoquera un « air de famille » (*family ressemblance*) ou la notion de similarité pour rendre compte des aspects sémantiques plus fixes[33] d'un concept. Or une telle réponse est inacceptable dans des cas paradigmatiques comme celui de l'inférence syllogistique où nous avons véritablement besoin d'une identité conceptuelle au sens strict du terme. (Fodor et Pylyshyn, 1988, 46) En effet, si l'on affirme que (1) les tortues sont plus lentes que les lapins; (2) que les lapins sont plus lents que les Ferraris; alors (3) les tortues sont plus lentes que les Ferraris. Dans une telle structure inférentielle, il serait très mal indiqué d'introduire une instabilité sémantique dans les différentes manifestations des termes tortues, lapins et Ferrari sous peine de commettre des sophismes d'ambiguïté ou des amphibologies à répétition. Évidemment cela serait inacceptable et ne rendrait pas compte de la solidité de l'inférence. Fodor s'oppose donc à

toute définition de l'identité conceptuelle à partir de la notion de similarité conceptuelle dont le connexionnisme fait abondamment usage. Ce débat est repris à la section 2.3.3.2. où la critique de Fodor est détaillée et la réponse de Paul M. Chuchland exposée.

1.2.3. la productivité conceptuelle

Ce premier regard sur la compositionalité conduit à son acceptation pour rendre compte de la productivité. Le principe de productivité est l'affirmation que l'esprit est capable de générer à partir de moyens finis un nombre infini de représentations mentales.[34] (Fodor et Pylyshyn, 1988, 33) Or, comme on l'a vu, c'est la structure interne des représentations mentales qui permet de rendre compte de leur compositionalité. Par conséquent, la productivité de l'esprit peut être comprise en référence au système de représentation en tant qu'ensemble d'expressions générées. (Fodor et Pylyshyn, 1988, 33-4) Certains s'opposent au principe de productivité en soutenant que l'idéalisation qui le sous-tend (la production d'un nombre infini de représentations mentales) est illégitime. Par contre, il permet selon Fodor d'éviter de postuler des éléments inappropriés comme des mémoires à capacité invraisemblable. En outre, la productivité a été défendue en linguistique (Chomsky) d'un point de vue empirique. L'idéalisation serait donc justifiée parce qu'elle permet de générer des hypothèses, des recherches et des résultats concluants.

1.2.4. la systématicité conceptuelle

Le principe de systématicité avance que la possibilité de produire ou de comprendre certains énoncés (et donc certaines représentations mentales, voir section 1.2.1.) se reflète dans notre capacité d'en produire et d'en comprendre certains autres. (Fodor et Pylyshyn,

---

[33] Comme par exemple le fait que certains concepts tel que celui de chaise renvoient à certaines propriétés rigides en dépit de la diversité des contextes d'apparition.

1988, 37) En fait, il s'agit d'une caractéristique propre à plusieurs fonctions cognitives telles que le langage et le raisonnement. Comme nous l'avons vu avec l'exemple de *John loves the girl* et *The girl loves John*, c'est la structure compositionnelle des représentations mentales qui expliquent la systématicité. (Fodor 1999, 97) Un autre exemple serait que si un individu peut saisir la pensée que $P \supset Q$, alors il peut aussi saisir la pensée $Q \supset P$. Ou encore si un individu comprend que $\neg(P \wedge Q)$, alors il peut saisir les pensées $\neg P$ et $\neg Q$. (Fodor, 1998, 97) Dans ces exemples, la systématicité est expliquée par l'usage des mêmes éléments constitutifs primitifs.[35] Et le recours aux mêmes éléments est garanti en retour par la structure compositionelle des représentations mentales.

1.2.5. Quatre caractéristiques secondaires des concepts

Fodor dégage quatre autres caractéristiques des concepts qui s'appuient sur les trois principes présentés précédemment.[36] Premièrement, les concepts sont des états mentaux particuliers (*mental particulars*) qui peuvent servir de causes ou d'effets mentaux. (Fodor, 1998, 23) Ce sont des états fonctionnels internes qui interviennent dans la série causale entre le stimulus et la réponse. Deuxièmement, les concepts sont des catégories, ce qui veut dire que les concepts s'appliquent à des choses dans le monde et que les choses du monde sont subsumées par des concepts. (Fodor, 1998, 24) Cela explique la possibilité de leur évaluation sémantique qui détermine si le concept s'applique ou non à une chose, s'il est vrai ou faux, s'il est correct ou incorrect, approprié ou inapproprié, etc. Troisièmement, les

---

[34] Dans cette formulation, l'argument ressemble à celui évoqué pour défendre la grammaire générative chomskienne.

[35] Fodor identifie parfois dans des cas analogues au précédent une sous-classe de systématicité, la *systématicité inférentielle*. (Fodor et Pylyshyn, 1988, 46-8) Elle est une application du principe de systématicité à l'inférence où la similarité logique implique des capacités cognitives similaires. Par exemple, une caractéristique de l'esprit est que si l'on est capable d'obtenir P à partir de $P \wedge Q \wedge R$, alors on est capable d'obtenir P à partir de $P \wedge Q$. (Fodor et Pylyshyn, 1988, 46-8)

[36] C'est en ce sens qu'elles sont « secondaires » et non pas parce qu'elles sont négligeables.

concepts sont majoritairement appris ce qui va de soi. (Fodor, 1998, 27) Quatrièmement, les concepts sont publics parce qu'ils sont des entités définies et stables dont l'identité ne peut pas être réduite à une quelconque notion de similarité. (Fodor, 1998, 30) Mais pourquoi? Étant donné que Fodor accepte le cadre de la psychologie intentionnelle, il ne peut pas accepter un relativisme conceptuel (Fodor, 1998, 29) d'après lequel nos concepts fondamentaux tels ceux d'eau, de nourriture, etc. seraient différents au risque d'infirmer la validité de l'explication intentionnelle formulée en termes de « J'ai soif pour de l'eau » et « Je crois qu'il y a de l'eau dans le puit » donc « Je vais au puit ». Car sans stabilité conceptuelle (ce que je crois être de l'eau = ce que vous croyez être de l'eau, ou ce que je crois être de l'eau au temps 1 = ce que je crois être de l'eau au temps 2), l'explication n'aurait aucune généralité et serait donc futile. (Fodor, 1998, 29) Il faut donc une identité conceptuelle entre les individus et pour le même individu (à différents temps donnés) dans la psychologie intentionnelle. Toutefois, il y a, selon Fodor, un quasi-consensus voulant que l'explication intentionnelle soit valide même s'il n'y a pas de contenu de croyance publique, l'idée étant qu'une notion de similarité conceptuelle puisse remplacer la notion d'identité conceptuelle. Évidemment, Fodor s'oppose à tout remplacement de l'identité conceptuelle par la notion de similarité conceptuelle. Car comment pourrait-on alors déterminer en vertu de quoi des concepts semblables sont semblables sans présupposer une identité conceptuelle? (Voir les sections 2.3.3.1., 2.3.3.2. et 2.3.3.3. qui présentent en plus de détail la tentative de Churchland de relever ce défi et les critiques de Fodor.)

1.2.6. le « langage de la pensée »

L'une des prémisses sous-jacentes à la position de Fodor et en général au cognitivisme classique[37] est qu'il y a un langage de la pensée (*language of thought*). (Fodor, 1987) L'hypothèse est la suivante et elle sous-tend les développements précédents. Les capacités linguistiques sont systématiques et les phrases ont une structure compositionnelle (principe de systématicité et compositionalité linguistiques). Or, les capacités cognitives sont elles aussi systématiques et les représentations mentales ont aussi une structure constitutive (principe de systématicité et de compositionalité conceptuelles et cognitives). Par conséquent, si les pensées ont une structure constitutive, alors le langage de la pensée existe car la structure du langage découle de celle de la pensée. [38] (Fodor, 1987, 150-1) Sans entrer dans les détails[39], le langage de la pensée est un langage privé[40] réglé par un code interne de représentations mentales. (Fodor, 1975) Les computations effectuées sur ces représentations sont exprimables en termes d'attitudes propositionnelles dans le cadre d'une syntaxe et d'une sémantique combinatoires. (Clark, 1989, 19) On voit donc que l'hypothèse du langage de la pensée sert d'assise à la théorie des concepts de Fodor et sous-tend sa conception représentationnelle de l'esprit.

1.2.7. le holisme, le pragmatisme, les prototypes et l'atomisme

Fodor s'oppose à toute forme de pragmatisme ou de holisme. Ces deux tendances briment, selon lui, le critère d'identité conceptuelle et compromettent donc les principes de compositionalité, de productivité et de systématicité ainsi que certaines autres propriétés

---

[37] Quoique sous une forme peut-être plus modéré. (Lepore, 1994)
[38] Pour une conclusion fort divergente au sujet des aspects structurés de la pensée, consulter Edelman (1987, 140-8) selon qui la syntaxe mentale (*presyntax*) se forme par des processus de réentrées, amenant ainsi une mise en ordre des concepts mentaux (*preconcepts*).
[39] Voir Fodor (1975) à cet égard.
[40] En opposition à Wittgenstein (1997) sur ce point.

des concepts. Elles conduisent en fin de compte à nier la viabilité de l'atomisme, c'est-à-dire l'exigence irréductible selon laquelle les éléments constitutifs sont porteurs d'une interprétation sémantique fixe. Car le pragmatisme assouplit le contenu conceptuel en fonction de son usage tandis que le holisme brime la possibilité d'une psychologie intentionnelle fondée sur la notion d'identité conceptuelle (Voir section 1.2.5.). Le fait que Fodor considère les théories prototypiques des concepts comme une forme de pragmatisme et de holisme permet de comprendre pourquoi il s'y oppose farouchement car la théorie prototypique des concepts définit le sens par l'usage et définit l'identité conceptuelle en vertu de la notion de similarité. (Fodor, 1998, 106) Cela conduit aussi Fodor à proposer sa théorie sémantique informationnelle selon laquelle le contenu conceptuel est rigide. Il suffit pour l'instant de mentionner ces éléments (qui seront repris en plus de détails dans la section 2.3.) avant de passer aux critiques connexionnistes de la conception classique des concepts.

1.3. Les critiques connexionnistes de la conception classique des concepts

La théorie classique et notamment l'interprétation de Fodor ne font pas l'unanimité. Les tenants du connexionnisme ont adressé un nombre important de critiques au cognitivisme classique, soit (1) la rigidité et la fragilité excessives des modèles classiques; (2) leur insensibilité à l'environnement externe; (3) leur implausibilité neurobiologique; (4) leur conception de l'autonomie de la psychologie; (5) leur compréhension erronée de la relation entre le langage et la pensée ainsi que (6) le lien exclusif qu'ils tentent d'entretenir avec le principe de compositionalité. Certaines des critiques présentées dans les pages suivantes ont

été formulées de façon générale contre le cognitivisme classique. Nous les dirigeons ici plus directement vers la problématique des concepts.[41]

1.3.1. la rigidité et la fragilité excessives des modèles classiques

Pour certains (Smolensky, 1992; Rumelhart, Smolensky, McClelland et Hinton, 1986, 19; Hofstadter, 1985, 639-40), les modèles classiques des concepts[42] ne permettent pas de rendre pleinement compte du processus humain de conceptualisation parce qu'ils sont trop rigides et fragiles. En réalité, les concepts ont une dimension rigide parce qu'ils ont des applications délimitées mais ils ont aussi une dimension malléable puisqu'ils peuvent s'appliquer assez spontanément à de nouvelles choses (généralisation spontanée). Cependant, les modèles classiques ne rendent compte que de la première dimension de la conceptualité. Ils ne parviennent pas à saisir adéquatement le caractère naturel et spontané des applications de concepts. La solution proposée par Rumelhart, Smolensky, McClelland et Hinton (1986) est que les concepts ne sont pas des « éléments constitutifs primitifs », mais plutôt des solutions de classification trouvées par des réseaux de neurones.[43] Cette approche considère que les concepts « émergent » de l'interaction d'éléments sous-symboliques lorsqu'une solution stable est présentée par le réseau de neurones. Le résultat est une théorie plus souple des concepts comparativement à l'approche symbolique jugée trop rigide et fragile. (Smolensky, 1992, 83; Clark, 1989, 76)

---

[41] Il est à noter que les critiques présentées ne sont pas des critiques définitives. Pour l'instant, il suffit d'en prendre acte, ce qui permettra ensuite de mieux comprendre le débat présenté au deuxième chapitre entre Churchland et Fodor (section 2.3.3.).

[42] Notons que les auteurs discutent des schèmes (*schema*), c'est-à-dire les structures informationnelles responsables de la représentation des concepts génériques. (Rumelhart, Smolensky, McClelland et Hinton, 1986, 18). En ce qui nous concerne nous ne ne traçons pas de démarcation nette entre les concepts et les schèmes entendus dans cette acception.

[43] La deuxième chapitre présente une explication de cette affirmation.

1.3.2. la sensibilité à l'environnement externe et la question de l'intentionalité

La rigidité et la fragilité des modèles classiques sont en grande partie causées par leur « insensibilité à l'environnement externe ». Ces modèles n'ont aucune façon de modifier leurs représentations en fonction de l'expérience et de l'apprentissage. Nous avons noté à ce titre que les concepts dans le cognitivisme sont des éléments constitutifs (atomes) ayant une signification rigide. Par contre, les modèles connexionnistes préservent au sein même de leur représentation une sorte de lien avec le monde extérieur parce qu'ils y puisent les éléments de leurs représentations. (Ramsey, 1992, 259[44]) En outre, ils sont sensibles aux changements se produisant dans l'environnement puisque les représentations se modifient avec l'expérience. Ils sont donc plus sensibles à l'environnement extérieur. (Smolensky, 1988, 16) Certains (Hofstadter, 1985, 651) ont même avancé que cela les rapprochait d'une résolution du problème de l'intentionalité des états mentaux dans la mesure où les représentations connexionnistes conservent en elles-mêmes une forme de lien représentationnel avec le monde extérieur. Car l'interprétation des états du réseau dépend directement des rapports entre le réseau et le monde puisque le domaine d'interprétation est en rapport direct avec le monde. En d'autres mots, les représentations connexionnistes visent quelque chose. Cela expliquerait l'usage général de la modélisation et de la conceptualisation dans la résolution de problèmes puisque ces activités passent souvent par des « représentations matérielles » (symboles logiques et mathématiques, graphiques, etc.) qui ont un rapport au monde, d'où leur efficacité. En effet, si ces dernières peuvent être vues comme une expression de la relation entre notre esprit et le monde, alors cela explique en partie leur utilité et leur efficacité.[45] On peut même alors envisager leur attribuer un

[44] Voir Smolensky (1988, 14) pour un avis contraire.
[45] Voir Edelman (1987) et Churchland (1989 et 1998b) pour des réflexions allant dans ce sens.

avantage évolutif pour la survie. (Rumelhart, Smolensky, McClelland et Hinton, 1986, 44-8)

1.3.3. la plausibilité neurobiologique et la valeur explicative des modèles classiques : le débat sur la psychologie du sens commun

Les modèles classiques ont aussi été critiqués à cause de leur manque de plausibilité neurobiologique. (Churchland, 1998a; Clark, 1989, 80) Il y aurait ici une tension entre l'utilisation d'un langage qui décrit informellement les processus cognitifs (la perspective classique) et une « explication scientifique » de ces processus par le connexionnisme. Cette tension fait intervenir le débat sur la psychologie du sens commun. D'un côté, on utilise la terminologie psychologique intuitive pour caractériser les concepts. Leur compositionalité, productivité et systématicité sont repérés en ayant recours à « l'évidence du sens commun ». D'un autre côté, on tente de s'éloigner de la typologie du sens commun. Ainsi, le connexionnisme souhaite expliquer le fonctionnement de l'esprit et la nature des concepts par le comportement de réseaux de neurones, jugés plus « scientifiques ». Par conséquent, les descriptions de la cognition faites en termes d'algèbre linéaire ont préséance sur les descriptions intuitives. Sans entrer dans les dédales de ce débat extrêmement touffu, nous pouvons au moins souligner que selon certains, les modèles symboliques ont une faible plausibilité neurobiologique parce qu'ils font usage de la psychologie du sens commun (Churchland, 1998a), ce qui aurait un impact sur leurs prétentions explicatives.[46] Le caractère « naturel » et « organique « des représentations connexionnistes est censé venir combler ce fossé. (Rumelhart, Smolensky, McClelland et Hinton, 1986, 36)

---

[46] Smolensky affirme que c'est le caractère éthéré des symboles qui fait qu'ils n'ont pas permis de contribuer significativement à la compréhension de la conceptualité humaine. (Smolensky, 1992, 82)

1.3.4. l'autonomie de la psychologie et l'interdépendance des niveaux

La perspective classique tranche le débat du statut de la psychologie en faveur de son autonomie relative par rapport aux autres sciences puisque les différents niveaux d'analyse sont relativement autonomes. (Broadbent, 1985; Fodor et Pylyshyn, 1988; Johnson-Laird, 1988) Le niveau de description de la psychologie, essentiellement le niveau computationnel, est considéré comme le domaine des manipulations de symboles effectuées dans le cadre de processus cognitifs. (Fodor et Pylyshyn, 1988, 7-10) Par contre, plusieurs connexionnistes refusent un tel isolement de la psychologie.[47] Premièrement, cette situation est indésirable étant donné les possibilités offertes par l'échange avec les autres disciplines telles que les neurosciences. Deuxièmement, cette situation ne reflète même pas la nature de l'entreprise d'explication psychologique qui a très souvent recours simultanément à plusieurs niveaux d'analyse (comportement, cognition, données physiologiques, introspection, etc.). Les connexionnistes en général perçoivent justement les réseaux de neurones comme pouvant rapprocher le niveau neurobiologique du niveau psychologique et établir fermement leur interdépendance. Cette attitude les amène, par conséquent, à critiquer sévèrement l'autonomie[48] de la psychologie soutenue par les tenants de la perspective classique, à savoir : la possibilité de limiter l'explication psychologique aux niveaux computationnel et algorithmique. Pour d'autres, (Hofstadter, 1985, 642) le manque de prise en compte des éléments sous-cognitifs (ou sous-symboliques) dans le cognitivisme classique (étant donné sa focalisation sur les symboles) restreint sévèrement sa capacité à modéliser et à comprendre les actes cognitifs, y compris la conceptualité.

---

[47] Voir, entre autres, Changeux et Dehaene, 1989; Paul M. Churchland, 1998c; Churchland et Churchland, 1998b; Clark, 1989; Dehaenne et Changeux, 1993; Hofstadter, 1985; Rueckl, 1991; Rumelhart et McClelland, 1985; Rumelhart et McClelland, 1986b; Rumelhart, Smolensky, McClelland et Hinton, 1986; Smolensky, 1992, 1986, 1988.

[48] C'est-à-dire la validité intrinsèque des niveaux d'analyse à cause de leur indépendance relative.

1.3.5. la relation entre le langage et la pensée

La stabilité des concepts dans la perspective symbolique fait d'eux des représentations foncièrement passives et statiques. (Hofstadter, 1985, 645-6) Par contre, les connexionnistes proposent des concepts plus dynamiques parce qu'ils sont activement en rapport avec l'environnement. Ils sont « actifs » car ils incorporent un rapport au monde extérieur. Bien que les deux approches adoptent une sorte de réalisme mental où le langage (ordinaire) est un véhicule de la pensée, le connexionnisme, avec ses représentations dynamiques, offre quelques pistes intéressantes insoupçonnées par l'approche classique en conférant au langage un rôle synergétique dans l'expression et la formulation de la pensée. (Rumelhart, Smolensky, Mclelland et Hinton, 1986) Dans cette optique, le langage est un outil qui canalise et précise l'expression de la pensée, ce qui fait de la relation langage/pensée une dynamique constructive.[49] Cette interdépendance explique, par exemple, pourquoi notre façon de résoudre et de penser certains problèmes est modifiée par le langage que l'on utilise pour le décrire car si la pensée et le langage sont interreliés, alors un développement dans l'un amène une précision dans l'autre. De même, elle offre une solution au problème de l'utilité et de l'efficacité des concepts dans la précision de la pensée puisqu'elle situe les concepts dans une dynamique étroite avec le langage. (Rumelhart, Smolensky, McClelland et Hinton, 1986, 44) D'une manière générale, l'approche classique n'offre pas ces pistes parce que l'hypothèse du langage de la pensée ne confère pas un rôle aussi synergétique au langage.

[49] Voir Edelman (2000, 206-7) pour une liaison synergétique similaire (*semantic bootstrapping*) entre la pensée et le langage.

1.3.6. l'approche classique et son lien exclusif avec la compositionalité, la systématicité et la productivité

Les tenants de l'approche classique sont convaincus que la productivité, la systématicité et la compositionalité des concepts ne peuvent être garanties qu'à l'intérieur de leur propre perspective. (Fodor, 1998, 1995, 1987; Fodor et Lepore, 1992; Fodor et Pylyshyn, 1988)[50] Cependant, il y a eu des tentatives pour répondre à cette affirmation. Smolensky (1991a et 1991b)[51] soutient que les réseaux de neurones peuvent produire des représentations compositionnelles ayant une certaine structure mais ils le feraient de façon différente. Là où l'approche classique postule des éléments constitutifs, le connexionnisme soutient que la structure des représentations mentales est un phénomène émergeant de la sensibilité structurelle des représentations vectorielles. Autrement dit, le macro-phénomène de la compositionalité serait expliqué par le micro-phénomène des interactions entre vecteurs d'activation. (Smolensky, 1991b, 298) La tentative de tracer un lien d'exclusivité entre la compositionalité, la productivité ainsi que la systématicité et le cognitivisme est donc contestée.

1.3.7. le connexionnisme et le langage de la pensée

Les connexionnistes rejettent en général l'une des hypothèses sous-jacentes au cognitivisme et à sa conception des concepts, soit qu'il y ait un langage de la pensée.[52] Certains (Rumelhart et McClelland, 1986b) le font parce qu'ils doutent de la validité du principe de compositionalité conceptuelle et de son application aux processus cognitifs. D'autres (Paul

---

[50] Sur ce débat plus large que nous effleurons, voir aussi Fodor et McLaughlin, 1991; Smolensky, 1991a et 1991b.

[51] Sans en dire plus Fodor et Pylyshyn (1991) rejettent cette tentative d'intégrer l'exigence de compositionalité dans le connexionnisme sous prétexte que Smolensky, au mieux, ne fait qu'implémenter un modèle classique inspiré du langage de la pensée dans un cadre connexionniste. Voir aussi Ramsey (1992, 264) sur ce point.

M. Churchland, 1998a, 1998b et 1998c) le font parce qu'ils croient que la description de l'esprit en termes d'attitudes propositionnelles est fausse étant donné que l'esprit fonctionne selon les modèles connexionnistes. Par conséquent, ils affirment que cette hypothèse est erronée et qu'elle doit être évacuée des *véritables* sciences cognitives. Au niveau des concepts, cela signifie que les descriptions en termes d'attitudes propositionnelles et de représentations mentales n'ont plus la légitimité qu'elles avaient et qu'elles doivent donc faire place aux vecteurs d'activation et aux matrices de connectivité du connexionnisme.

1.3.8 Conclusion

Dans ce permier chapitre, nous avons introduit le cognitivisme classique et sa conception des symboles et des processus mentaux. L'interprétation de Fodor a fait l'objet d'une présentation plus détaillée ainsi que de nombreuses critiques connexionnistes. Le prochain chapitre aborde le connexionnisme et sa conception des concepts. L'accent sera mis sur le débat que nous avons entamé entre le cognitivisme et le connexionnisme sur les concepts.

---

[52] Voir Davies (1991) pour une discussion générale.

Chapitre 2 : Les concepts dans le connexionnisme

Ce deuxième chapitre a pour but d'introduire à la problématique des concepts dans le connexionnisme. La première section débute avec une brève mise en contexte de l'approche connexionniste et une présentation de huit éléments structuraux et fonctionnels des réseaux de neurones[1]. Elle clôt avec quelques indications sur quatre propriétés des réseaux (généralisation, descente graduelle, tolérance à l'erreur et apprentissage) qui sont expliquées en référence à la reconnaissance de schémas (*pattern recognition*), une fonction fondamentale des réseaux de neurones. Vient ensuite dans une deuxième section une présentation du fonctionnement de la reconnaissance de schémas (ou classification) dans les réseaux. Il s'agit ici de bien comprendre comment les réseaux arrivent à classer des schémas d'activation (*activation patterns*), activité à la base des capacités conceptuelles des réseaux. Pour ce faire, nous exposons diverses techniques de classification telles que celles du « plus proche voisin » et de la distance métrique ainsi que les classifications linéaires, statistiques et non linéaires. La fonction de classification est ensuite présentée en termes de satisfaction de contraintes dans un espace d'activation à n-dimensions. Ces deux premières sections conduisent à la question des concepts dans les réseaux de neurones, laquelle est introduite avec l'interprétation de Paul M. Churchland dans une troisième section. Ce dernier soutient que les concepts sont des prototypes obtenus par des méthodes mathématiques de classification de schémas d'activation. Leur fonction est de délimiter des agglutinations (ou régions d'activation) dans un espace d'activation à n-dimensions. Il en découle un tableau où le contenu conceptuel est (1) déterminé à la fois par des éléments externes et internes, (2) holistique, (3) modérément empirique, (4) foncièrement

[1] La dénomination « réseau de neurones » ne fait pas référence, pour l'instant, à une interprétation possible des réseaux comme des modélisations de véritables réseaux de neurones mais constitue seulement une façon de nommer les réseaux inspirés d'une architecture neuronale.

pragmatique et (5) essentiellement dynamique. En dernière analyse, Churchland soutient une position éliminativiste par rapport à l'approche classique puisque les concepts sont révélés par l'analyse des agglutinations de points dans l'espace d'activation (*cluster analysis*) et non en référence à des symboles. Cette interprétation conduit à un conflit ouvert avec la position de Fodor. Celui-ci reproche à Churchland de ne pas présenter un critère rigoureux pour l'identité conceptuelle, de s'embourber dans le problème de l'information collatérale et de l'individuation des dimensions, de sombrer dans un empirisme naïf et dans une forme de holisme ainsi que de tracer une fausse adéquation entre les prototypes et les concepts.

Une remarque s'impose avant le début de l'exposé. Le connexionnisme est partiellement mathématisé et est peut-être un peu plus complexe que le cognitivisme classique. Il sera donc nécessaire de se pencher plus longuement sur des notions techniques avant d'arriver aux interprétations philosophiques portant sur la conceptualité. Car avant d'interpréter les capacités conceptuelles des réseaux de neurones, il faut absolument comprendre comment ces derniers fonctionnent. La qualité de l'analyse en dépend largement. Faire autrement risque de nous conduire dans un débat beaucoup trop large, voire à des interprétations et à des évaluations erronées. Par conséquent, nous aurons l'impression de nous éloigner quelque peu de nos interrogations philosophiques avant d'y revenir définitivement et, nous l'espérons, plus éclairés.

2.1. Présentation générale du connexionnisme

Cette première section est une présentation générale du connexionnisme. Une mise en contexte contient quelques notes historiques et introductives suivies d'une présentation des éléments structuraux et fonctionnels des réseaux ainsi que l'explication de quatre propriétés des réseaux.

2.1.1. Mise en contexte[2]

Il serait erroné de penser que le connexionisme est une création des quinze dernières années. Au contraire, il y a eu des précurseurs importants de certaines thèses du connexionnisme. Certains ont fait leur contribution il y a plus d'une cinquantaine d'années. Warren McCulloch et Walter Pitts, par exemple, ont développé en 1943 une unité computationnelle ressemblant à un neurone, le neurone formel. (McCulloch et Pitts, 1943; Bechtel et Abrahamsen, 1991; Andler, 1992) Celui-ci, fonctionnant sur une base binaire, pouvait computer des opérations logiques (« et », « ou » et « non ») à l'intérieur de réseaux à configurations déterminées. Le neurone recevait des entrées inhibitrices ou excitatrices. S'il recevait une entrée inhibitrice, il demeurait en état de repos. Alors que s'il ne recevait pas d'entrées inhibitrices, le neurone émettait une sortie si la sommation de ses entrées excitatrices dépassait un certain seuil donné. (Bechtel et Abrahamsen, 1991, 3; Clark, 1989, 84-85) Cette contribution constitue le point de départ de l'exploration des réseaux de neurones.

Dans la vingtaine d'années qui succédèrent à la publication de McCulloch et Pitts, un nombre intéressant de chercheurs s'appliquèrent à explorer les propriétés de ces neurones formels. John von Neumann, par exemple, démontra que la fiabilité des réseaux de neurones formels pouvait être améliorée de deux façons, soit par l'augmentation du nombre d'entrées pour chaque neurone individuel (en distribuant l'erreur) et en déterminant l'activation d'un neurone en fonction du schéma statistique de l'activation de ses entrées[3]. Winograd et Cowan ajoutèrent à cela en 1963 la possibilité pour un neurone d'interagir avec d'autres neurones en amont et en aval. (Bechtel et Abrahamsen, 1991, 3-4) Ces

---

[2] Il ne s'agit pas ici d'une histoire complète du connexionnisme. Au plus, les paragraphes qui suivent sont une brève mise en contexte.

[3] C'est-à-dire de façon probabiliste.

développements ainsi que ceux de McCulloch et Pitts sur la reconnaissance de schémas (*pattern recognition*) annonçaient l'idée fort importante, dans une perspective connexionniste, de la représentation distribuée.

Dans ses publications de 1958 et 1959, Frank Rosenblatt assouplit les réseaux de McCulloch et Pitts. Il introduisit la possibilité pour les neurones de modifier leur connectivité en fonction d'un apprentissage. Le réseau était constitué de deux couches. La première recevait des entrées et ses sorties formaient les entrées de la deuxième couche. Le réseau était entraîné à reconnaître un schéma d'activation donné et à produire une sortie correspondante. Le poids (*weight*) des connexions reliées à un neurone dont la réponse était inadéquate se voyait modifié afin d'arriver éventuellement à une réponse adéquate. Rosenblatt donna le nom de « perceptron » à ces systèmes. Enfin, selon son théorème de la convergence du perceptron (*perceptron convergence theorem*), Rosenblatt démontra que si un ensemble de connectivités pouvait conduire à une bonne réponse, alors un nombre fini de répétitions, lors de l'entraînement du réseau, pouvait conduire le réseau à la solution adéquate.[4] (Bechtel et Abrahamsen, 1991, 5)

Il faut dire que jusqu'au milieu des années soixante, des recherches et des progrès étaient enregistrés à la fois dans la perspective classique et dans la perspective des réseaux de neurones.[5] Par contre, la publication de *Perceptron* en 1969 par Marvin Minsky et Seymour Papert « sonna le glas » de la perspective des réseaux de neurones. Ils explorèrent rigoureusement les possibilités computationnelles offertes par les réseaux à deux couches. Ils conclurent que ces derniers ne pouvaient pas résoudre certains problèmes de classification tel que le « ou exclusif » (*xor*). Il fallait ajouter une troisième couche pour

[4] Il y a des problèmes pratiques qui peuvent pondérer cette affirmation : nombre d'unités, la dynamique entre le taux d'apprentissage et les minimum locaux, etc.

résoudre ce genre de problème. Le hic c'est que les techniques d'apprentissage de Rosenblatt ne s'appliquaient pas aux réseaux à trois couches et plus.[6] (Bechtel et Abrahamsen, 1991, 15) Ils soulevèrent aussi des doutes quant à l'utilité des réseaux car ceux-ci semblaient soumis à une croissance exponentielle du nombre d'unités en raison de la complexité des problèmes qu'ils avaient à résoudre. Le jugement de Minsky et Papert fit autorité et la communauté émergente des sciences cognitives opta pour l'approche classique au détriment des réseaux de neurones.

Cependant, un regain d'intérêt[7] dans les années quatre-vingt fit « renaître les réseaux de neurones de leur cendre ». Les recherches portant sur les réseaux se multiplièrent. Des avancées formelles, une volonté de rapprocher les sciences cognitives des neurosciences et un certain nombre de lacunes présentes dans les modèles symboliques en sont en partie responsables. (Bechtel et Abrahamsen, 1991, 17; Andler, 1992, 38) La publication en 1986 de *Parallel Distributed Processing. Explorations in the Microstructure of Cognition* par James McClelland, David Rumelhart et le groupe de recherche PDP est l'un des événements auquel on doit attribuer cette résurgence des réseaux de neurones. C'est aussi ce qui nous amène maintenant à abandonner notre perspective historique pour nous attarder à la présentation des réseaux de neurones contenue dans cet ouvrage.

2.1.2. Élements structuraux et fonctionnels des modèles connexionnistes

Les réseaux de neurones peuvent être décrits selon huit composantes : (1) un ensemble d'unités; (2) un état d'activation; (3) une fonction de sortie; (4) un schéma de connectivité entre les unités; (5) une règle de propagation; (6) une règle d'activation; (7) une règle

---

[5] D'ailleurs, le découpage entre ces deux approches n'était pas aussi clair qu'aujourd'hui.

[6] La découverte de techniques d'apprentissage plus sophistiquées (telle la règle delta) pouvant être utilisées dans des réseaux à trois couches est l'une des causes du regain d'intérêt pour les réseaux de neurones.

d'apprentissage et (8) un environnement. (Rumelhart, Hinton et McClelland, 1986, 46) La présentation en ces termes est générale étant donné que chaque modélisation tend à être une variante de ce cadre.[8] La difficulté à laquelle nous sommes confrontés ici est de conserver un niveau de généralité qui se situe au-dessus des différentes interprétations possibles des « propriétés cognitives » des réseaux. En effet, les réseaux peuvent recevoir, selon les auteurs, bon nombre d'interprétations passant carrément du niveau cognitif au niveau neuronal.[9] Si le troisième chapitre aborde cette question, nous faisons déjà face à ce problème. Alors, loin d'être superflue, la présentation plutôt formelle qui suit enrégimentera les définitions des concepts utilisés dans les interprétations des capacités cognitives des réseaux et balisera nos explications, nos évaluations et nos discussions ultérieures. En outre, elle permettra de mieux comprendre comment les réseaux parviennent à effectuer des tâches de classification, éléments essentiels pour la compréhension des capacités conceptuelles des réseaux et l'appréciation des ressources formelles leur permettant d'effectuer ces tâches. La figure 2.1.2.a., tirée de Rumelhart, Hinton et McClelland (1986, 47) illustre les éléments qui sont présentés.

---

[7] Il y avait un très petit nombre de chercheurs qui s'intéressaient aux réseaux de neurones dans les années soixante-dix.

[8] C'est aussi pourquoi il nous serait très difficile, voire impossible, de produire un exemple paradigmatique permettant d'illustrer clairement ces huit notions. En fait, il faudrait un nombre considérable d'exemples pour parvenir à illustrer chacune des notions présentées. Il faudra donc attendre la section 2.3. pour des exemples illustrant certaines de ces notions.

[9] Voir la typologie de Ramsey présentée à la section 3.2.3.1. pour une clarification plus sytématique des interprétations possibles des réseaux au niveau de la représentation mentale.

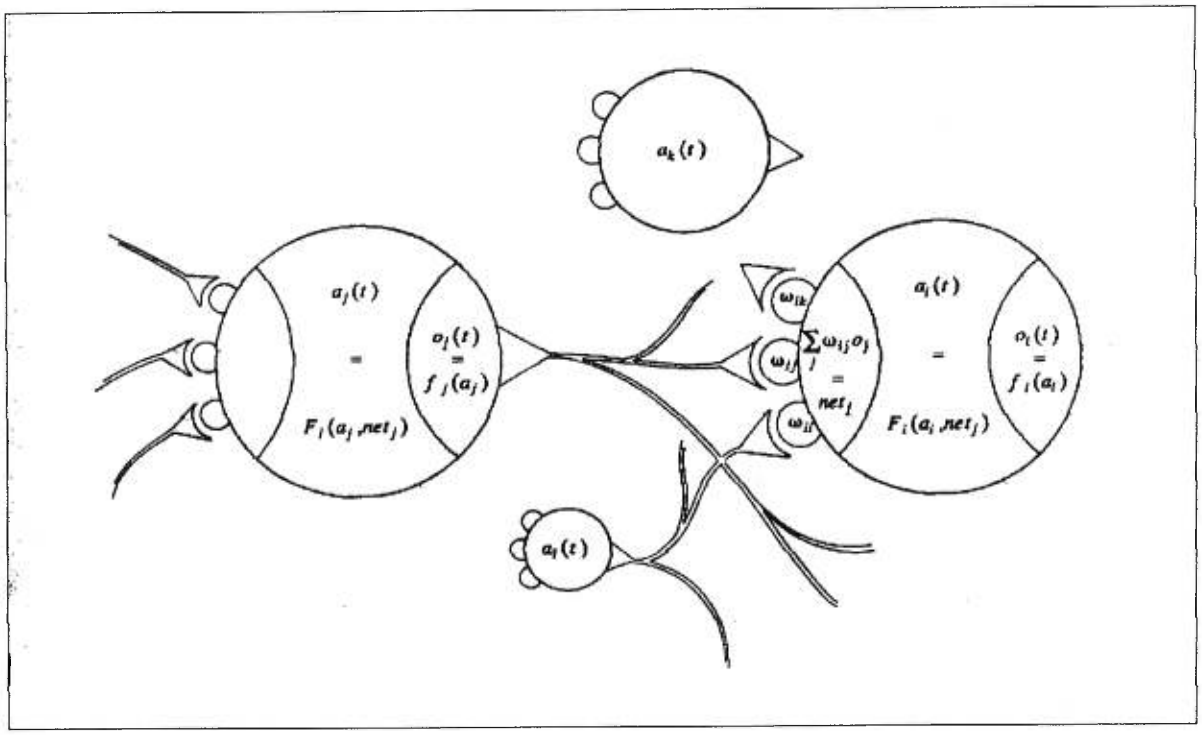

Figure 2.1.2.a. : Les éléments structurels et fonctionnels
des réseaux de neurones (Rumelhart, Hinton et McClelland, 1986, 47)

2.1.2.1. les unités

Chaque réseau connexionniste contient un ensemble d'unités (u_i) souvent représentées, par analogie,[10] sous la forme de neurones. La fonction d'une unité est de recevoir des signaux, de les traiter et d'acheminer en sortie un signal qui respecte les caractéristiques computationnelles du neurone (état d'activation, fonction de sortie, règle de propagation, règle d'activation, etc.). Il y a trois types d'unités : les unités d'entrée, les unités de sortie et

[10] Cette première analogie neurobiologique est elle-même la source de certaines interprétations où les réseaux de neurones représentent des réseaux de véritables neurones. (McCulloch et Pitts, 1943, 18-9) Cependant, l'analogie est lacunaire et superficielle sur certains points bien que révélatrice sur certains autres. Nous ne

les unités cachées (*hidden units*). Les unités d'entrée reçoivent des entrées externes au réseau. Ces entrées peuvent provenir du monde extérieur ou d'un autre réseau de neurone. Les unités de sortie transmettent un signal au monde extérieur ou à un autre réseau. Les « unités cachées » portent ce nom parce qu'elles ne sont pas visibles pour les systèmes extérieurs qui ne voient que les entrées et les sorties du réseau. L'ajout d'une couche d'unités cachées accroît la puissance computationnelle du réseau. (Rumelhart, Hinton et McClelland, 1986, 46-48; Plunkett et Elman, 1997, 2-3) L'analyse des unités cachées, comme nous le verrons plus loin, joue un rôle important (selon l'interprétation de Churchland) pour déterminer ce qu'un réseau « représente » dans ses classifications internes.

La nature des unités a des conséquences importantes. La représentation distribuée des réseaux de neurones découle du fait que les unités d'entrée « abstraient » des éléments sur lesquels des schémas significatifs d'activation peuvent être repérés. (Rumelhart, Hinton et McClelland, 1986, 47) En outre, le traitement en parallèle des signaux se fait justement parce que l'activation des unités se fait simultanément et parallèlement dans le réseau. Cela est rendu possible par la nature même des neurones formels et de leur arrangement en réseau car chaque unité traite un élément particulier[11] dans un cadre de traitement parallèle.[12] (Rumelhart, Hinton et McClelland, 1986, 47)

---

pouvons pas en discuter plus longuement ici. Voir Smolensky (1988) pour une analyse des divergences et Beale et Jackson (1991) pour des clarifications.

[11] Le fait que les éléments à traiter peuvent être des concepts, des propositions, des lettres ou des caractéristiques perceptuelles crée une sorte de confusion quant au niveau d'application des réseaux. Nous reviendrons sur le sujet dans le troisième chapitre afin de clarifier cette ambiguïté.

[12] Notons qu'il existe des réseaux connexionnistes à représentation locale (en opposition à la représentation distribuée) lesquels ont des neurones dont la représentation est fixe. Notre travail se concentre sur les modèles à représentation distribuée.

2.1.2.2. l'état d'activation

L'état d'activation[13] au temps t est ce que le système « représente »[14]. Cette mesure est donnée (la plupart du temps) sous forme vectorielle $\mathbf{a}(t)$ qui représente le schéma d'activation de l'ensemble des unités en traitement. Chaque élément du vecteur représente l'état d'un neurone particulier au temps t. Les valeurs d'activation varient selon les modèles. (Rumelhart, Hinton et McClelland, 1986, 48) L'illustration graphique de « l'espace d'activation » est une façon de représenter l'état d'activation d'un réseau. Il s'agit d'un espace à n-dimensions où chaque dimension du vecteur correspond à une dimension de l'espace et où la valeur du vecteur est représentée précisément dans cet espace.[15]

2.1.2.3. la fonction de sortie

Les unités interagissent en transmettant leur signal à d'autres neurones. La fonction de sortie, $f(a_i(t))$ lie l'état d'activation courant du neurone $a_i(t)$ au signal de sortie $o_i(t)$. Par conséquent, la fonction de sortie est $o_i(t) = f(a_i(t))$. Dans le plus simple des cas, la fonction de sortie est la fonction d'identité $f(x) = x$, mais dans la plupart des cas, la fonction de sortie est une fonction à seuil où un neurone envoie sa sortie si et seulement si l'état d'activation du neurone excède un seuil établi (une valeur déterminée telle que 0,5; 0,1, etc.).[16] Les sorties d'un réseau, à l'instar des états d'activation d'un réseau, peuvent être

---

[13] Il faut prendre garde ici du fait que la notion d'état d'activation est appliquée à la fois au système et au neurone individuel. La notion de vecteur $\mathbf{a}(t)$ peut donc être appliquée au système ou au neurone individuel.

[14] La question de savoir exactement ce que les réseaux de neurones représentent sera l'objet central du présent chapitre ainsi que du prochain. Nous l'esquivons pour l'instant.

[15] Elles peuvent être discrètes (le plus souvent binaires) ou continues (nombres réels, valeurs sur une échelle délimitée par un minimum et un maximum). Les suppositions à ce niveau conduisent à des propriétés computationnelles légèrement différentes au niveau des réseaux. (Rumelhart, Hinton et McClelland, 1986, 48)

[16] La fonction de sortie peut aussi être stochastique et dépendre de façon probabiliste de la valeur d'activation du neurone. Dans un tel cas, le réseau gagne en complexité et en non-linéarité.

etc.).[16] Les sorties d'un réseau, à l'instar des états d'activation d'un réseau, peuvent être représentées sous la forme d'un vecteur, $\mathbf{o}_i(t)$ qui présente l'ensemble de valeurs de sortie pour chaque unité au temps *t*. (Rumelhart, Hinton et McClelland, 1986, 48-9)

2.1.2.4. le schéma de connectivité

Le schéma de connectivité (*connectivity pattern*) est ce qui détermine ce que le réseau « connaît », c'est-à-dire comment le réseau traite les signaux. Un poids (*weight*) précise l'influence d'une entrée sur l'état d'activation d'un neurone. Un poids positif signale un lien excitatoire tandis qu'un poids négatif signale un lien inhibitoire. Le schéma de connectivité présente l'ensemble des poids, c'est-à-dire l'importance respective de chaque connectivité. Il peut être représenté sous la forme d'une matrice **W** dans laquelle le poids w_{ij} représente précisément comment l'unité u_j excite/inhibe l'unité u_i. La valeur absolue de w_{ij}, soit $|w_{ij}|$ représente la *force* du lien entre l'unité u_j et l'unité u_i.[17] (Rumelhart, Hinton et McClelland, 1986, 49-51)

2.1.2.5. la règle de propagation

La règle de propagation est une fonction qui prend le vecteur de sortie $\mathbf{o}_i(t)$ et le combine avec le(s) matrice(s) de connectivité pour produire le *net input* pour chaque type d'entrée dans une unité. La variable net_{ij} représente le net input de type i dirigé vers l'unité u_j. Lorsqu'il n'y a qu'un type d'entrée, la notion de net input est généralement abrégée à net_j

[16] La fonction de sortie peut aussi être stochastique et dépendre de façon probabiliste de la valeur d'activation du neurone. Dans un tel cas, le réseau gagne en complexité et en non-linéarité.

[17] De façon générale, deux matrices sont produites lorsqu'une unité a des poids positifs (excitation) et des poids négatifs (inhibition). Cependant, lorsqu'un seul type de poids est présent, une seule matrice est produite. Des cas plus complexes de connectivité sont facilement imaginables. On pourrait avoir des matrices spécifiques pour certains liens spécifiques et, à la rigueur, une matrice pour chaque lien. (Rumelhart, Hinton et McClelland, 1986, 49-51)

car nous n'avons pas à prendre en compte tous les types d'entrée dans une telle situation. En notation vectorielle, le vecteur $\mathbf{net}_i(t)$ désigne les net inputs de type i au temps t. Lorsqu'il y a deux types d'entrée, comme par exemple, des entrées excitatrices et des entrées inhibitrices, on effectue deux calculs pour déterminer le net input. Étant donné une situation (très fréquente) où le net input est simplement le produit vectoriel de la matrice de connectivité, $\mathbf{W}$ et du vecteur de sortie $\mathbf{o}$, nous avons comme net input, $\mathbf{net} = \mathbf{W}\mathbf{o}(t)$. Dans une situation où il y a deux types d'entrée (excitatrice et inhibitrice), nous avons alors $\mathbf{net}_e = \mathbf{W}_e\mathbf{o}(t)$ (pour le net input excitatoire) et $\mathbf{net}_i = \mathbf{W}_i\mathbf{o}(t)$ (pour le net input inhibitoire). Il y a bien sûr des variantes plus complexes possibles.

2.1.2.6. la règle d'activation

La règle d'activation combine le(s) net input(s) avec l'état d'activation de l'unité. C'est en quelque sorte la règle qui permet de « mettre à jour » l'état d'activation du neurone en tenant compte des nouvelles entrées afin de produire un nouvel état d'activation. Elle est une fonction $\mathbf{F}$ qui prend l'état d'activation $\mathbf{a}(t)$ et le vecteur de net input $\mathbf{net}_j$ et produit un nouvel état d'activation. Si $\mathbf{F}$ est la fonction d'identité, on peut alors écrire $\mathbf{a}(t+1) = \mathbf{W}\mathbf{o}(t) = \mathbf{net}_j$. Mais dans la plupart des cas, $\mathbf{F}$ est une fonction à seuil dans la mesure où le net input doit excéder une certaine valeur avant de contribuer à un nouvel état d'activation. Parfois, l'activation prend des valeurs continues et, dans ce cas, il est tenu pour acquis que $\mathbf{F}$ est une fonction sigmoïde. Par contre, la fonction la plus fréquente est la fonction quasi-linéaire. Dans le cas où il n'y a qu'un type d'entrée (« j » dans notre cas), on la formule ainsi : $\mathbf{a}_i(t+1) = \mathbf{F}(\mathbf{net}_i(t+1)) = \mathbf{F}(\sum_j w_{ij}o_j)$.

2.1.2.7 les règles d'apprentissage

Si la connaissance repose dans les schémas de connectivité, alors l'apprentissage ou la modification de la connaissance dans le réseau impliquera la modification des schémas de connectivité. Les règles d'apprentissage sont généralement des variantes de la règle de Hebb qui formalise l'affirmation qu'étant donné deux unités u_i et u_j, si les deux sont actives simultanément, alors le poids déterminant leur connectivité w_{ij} devrait être renforcée. La formule est la suivante : $\Delta w_{ij} = g(a_i(t), t_i(t))h(o_j(t), w_{ij})$, où Δw_{ij} est la différence dans le poids et $t_i(t)$ est un signal d'apprentissage (*teaching input*). La règle affirme que la modification de la connectivité entre u_i et u_j est donnée par le produit de la fonction $g()$ de l'activation de u_i, c'est-à-dire $a_i(t)$ et du teaching input $t_i(t)$ et d'une autre fonction, $h()$ de la sortie de u_i, c'est-à-dire $o_i(t)$ et du poids de connectivité w_{ij}. Il y a aussi une formulation plus simple où $g()$ et $h()$ sont simplement proportionnelles à leur premier argument respectif de la formulation précédente. La voici : $\Delta w_{ij} = \eta a_i o_j$, où η est une constante de proportionnalité déterminant le taux d'apprentissage, c'est-à-dire le taux selon lequel les connectivités seront modifiées. Plus le taux sera élevé, plus l'apprentissage se fera rapidement et inversement[18]. Toutefois, la règle la plus utilisée est la règle delta (ou règle de Widrow-Hoff). Elle énonce que l'apprentissage est directement proportionnel à la différence entre l'état d'activation actuel du neurone et celui souhaité. Nous avons donc : $\Delta w_{ij} = \eta(t_i(t) - a_i(t))o_j(t)$, où encore une fois η est le taux d'apprentissage, où $t_i(t)$ est l'activation visée (*target input*) et où le produit de $a_i(t)$ (l'état d'activation de l'unité i au temps t) et $o_j(t)$ (la sortie de l'unité j au temps t) est soustrait de l'activation visée.

(Rumelhart, Hinton et McClelland, 1986; 52-4; Plunkett et Elman, 1997, 13-14) De cette façon, plus l'activation de l'unité est proche de la valeur d'activation visée, moins il y aura de modification dans le poids reliant les unités u_i et u_j. Encore une fois, il y a plusieurs variantes de règles d'apprentissage possibles. Toutefois, *grosso modo* leur but est toujours de modifier la connectivité entre les neurones d'un réseau de façon à atteindre les valeurs d'activation souhaitées.[19] Elles sont donc à la base des capacités d'apprentissage des réseaux.

2.1.2.8. l'environnement

Enfin, il faut prendre note que le réseau se situe dans un environnement. De façon générale, l'environnement est conçu de façon non déterministe mais dans la pratique, il constitue une distribution stable de probabilités. Ainsi, les vecteurs d'entrée (ou vecteurs de caractéristiques) sont présentés de manière à refléter le non-déterminisme de l'environnement.

2.1.3. Quatre propriétés des réseaux de neurones

Nous venons de décrire et d'expliquer comment les réseaux de neurones sont constitués structurellement et fonctionnellement. Nous allons maintenant considérer de façon plus globale quatre propriétés[20] intéressantes des réseaux de neurones, c'est-à-dire : (1) la généralisation; (2) la tolérance à l'erreur; (3) la descente graduelle (*gradual descent*) et (4)

---

[18] En deux mots, un taux élevé peut conduire le réseau à des minimums locaux (des solutions insatisfaisantes) tandis qu'un taux plus modeste d'apprentissage signife un apprentissage plus lent quoique plus apte à éviter l'emprisonnement dans des minimum locaux. (Voir section 2.2.7)

[19] Sans insister sur les capacités d'apprentissage des réseaux, il faut dire que l'état d'activation visé (*target input*) peut être donné d'avance par le superviseur de la simulation mais il peut aussi être dérivé d'une phase précédente d'apprentissage du réseau. Dans un tel cas, le réseau apprend un schéma d'activation dans un premier temps et il évalue les schémas qui lui sont présentés à l'aulne de celui qu'il a appris dans un deuxième temps.

l'apprentissage.[21] Nous verrons que la reconnaissance de schémas est à l'origine de ces propriétés.

2.1.3.1. la généralisation

Les réseaux de neurones peuvent généraliser. Lorsqu'on leur présente des éléments en entrée, ils peuvent extirper un profil statistique ou un prototype des entrées auxquelles ils ont été exposés. Cette propriété fondamentale fait que les réseaux peuvent reconnaître des schémas d'activation et ensuite les comparer, les transformer, les compléter et les associer. (Bechtel et Abrahamsen, 1991, 106-7; Beale et Jackson, 1991, 89; McClelland, Rumelhart et Hinton, 1986; Clark, 1989, 91-6) Dans ce dernier cas, on parle souvent de « généralisation spontanée » (*spontaneous generalization*) car le réseau est capable de produire un vecteur de sortie approprié en « étendant » ses connaissances à des exemples auxquels il n'avait pas été exposé. (McClelland, Rumelhart et Hinton, 1986, 30-1; Clark, 1989, 90-1)

2.1.3.2. la descente graduelle

Étant donné que les réseaux de neurones sont des représentations distribuées et qu'ils traitent l'information en parallèle (et non de façon séquentielle comme les systèmes classiques), ils peuvent subir certains dommages tout en demeurant capables d'exécuter leur fonction. Ce que l'on observe alors est une diminution graduelle des capacités du réseau en fonction du dommage infligé. (McClelland, Rumelhart et Hinton, 1986, 29; Clark, 1989, 89) En d'autres mots, si un réseau a un grand nombre d'unités et voit l'une d'entre elles endommagée, il pourrait tout de même traiter le stimulus.

---

[20] Il ne s'agit pas d'une présentation complète des propriétés des réseaux mais seulement de celles qui ont un intérêt dans le cadre de la question des concepts.

2.1.3.3. la tolérance à l'erreur

Encore une fois, étant donné le traitement parallèle et distribué des signaux dans les réseaux de neurones, les réseaux peuvent recevoir de l'information déformée tout en offrant des réponses appropriées ou du moins plausible. (Beale et Jackson, 1991, 89-90) Puisque l'information n'est pas « localisée » dans un seul symbole ou dans une seule unité, le fait que certaines entrées soient déformées n'affectent pas la réponse du réseau autant que dans un cadre conventionnel. Par exemple, un réseau recevant des vecteurs d'entrées de grande dimension dont l'une est déteriorée serait capable néanmoins de traiter l'information.

2.1.3.4. l'apprentissage

La propriété peut-être la plus spectaculaire des réseaux de neurones est leur capacité d'apprendre. Sans intervention autre que l'ajustement des règles d'apprentissage, des taux d'apprentissage et des *target inputs*, le manipulateur peut entraîner le réseau à apprendre certaines opérations. Les réseaux ont appris un large éventail de tâches incluant, entre autres, la capacité de prononcer des textes en anglais, la capacité de former des mots et la capacité d'apprendre le passé des verbes anglais.

La reconnaissance de schémas (ou classification) est la fonction fondamentale qui sous-tend toutes les autres. Car l'apprentissage est l'acquisition d'une tâche de classification des entrées, la descente graduelle et la tolérance à l'erreur sont deux propriétés qui découlent de la méthode de classification du réseau tandis que la généralisation est une conséquence directe de la reconnaissance de schémas. Cet aspect noté, la deuxième section tentera de le démontrer tout en nous rapprochant de notre problématique, les concepts.

---

[21] On pourrait ajouter à cette liste, la plausibilité neuronale, la satisfaction des contraintes faibles (*soft constraints*) et la mémoire adressable par contenu (Bechtel et Abrahamsen, 1991, 56-65)

2.2. La reconnaissance de schémas

La reconnaissance de schémas (ou classification) est la fonction principale[22] des réseaux de neurones. Afin de l'expliquer, nous abordons dans cette deuxième section les notions de fonction discriminante ainsi que diverses techniques de classification, soit la classification du plus proche voisin et les distances métriques. L'activité des réseaux est ensuite expliquée en termes de classifications linéaire, statistique et non linéaire. Par ailleurs, le fonctionnement interne des réseaux est entrevu sous l'optique de la satisfaction de contraintes.

2.2.1. Les réseaux de neurones et la reconnaissance de schémas

Deux choses doivent être distinguées au niveau de la reconnaissance de schémas. Il doit y avoir en premier, une extraction de caractéristiques (*feature extraction*) et ensuite une classification. De prime abord, toute caractéristique dans une modélisation est extraite selon la tâche à effectuer et la nature des entrées. Par exemple, un réseau dont la fonction est de créer des mots à partir de lettres devra être sensible aux lettres qui sont les éléments constitutifs des mots. Ensuite, la classification implique une procédure par laquelle les caractéristiques sont classées. (Beale et Jackson, 1991, 16-17)

Les caractéristiques sont regroupés sous la forme d'un vecteur à n-dimensions où chaque dimension représente une caractéristique. Ce vecteur porte le nom de vecteur de caractéristiques (*feature vector*) ou vecteur d'entrée. Par exemple, on pourrait avoir un vecteur à 2 dimensions codant des individus selon deux caractéristiques, soit leur taille et leur masse respectives. Enfin, un espace de caractéristiques (*feature space*) ou espace d'activation est un espace à n-dimensions. Ses éléments correspondent à des vecteurs à n-dimensions où les vecteurs représentent des caractéristiques choisies. (Anderson, 1995,

434-6; Beale et Jackson, 1991, 17-8) Dans le cadre de notre exemple (Figure 2.2.1.a.), il s'agirait d'un plan cartésien où l'on présenterait sur l'axe des *y* la taille et sur l'axe des *x* la masse. Chaque point dans l'espace d'activation représenterait les données encodées par le vecteur pour un individu donné.

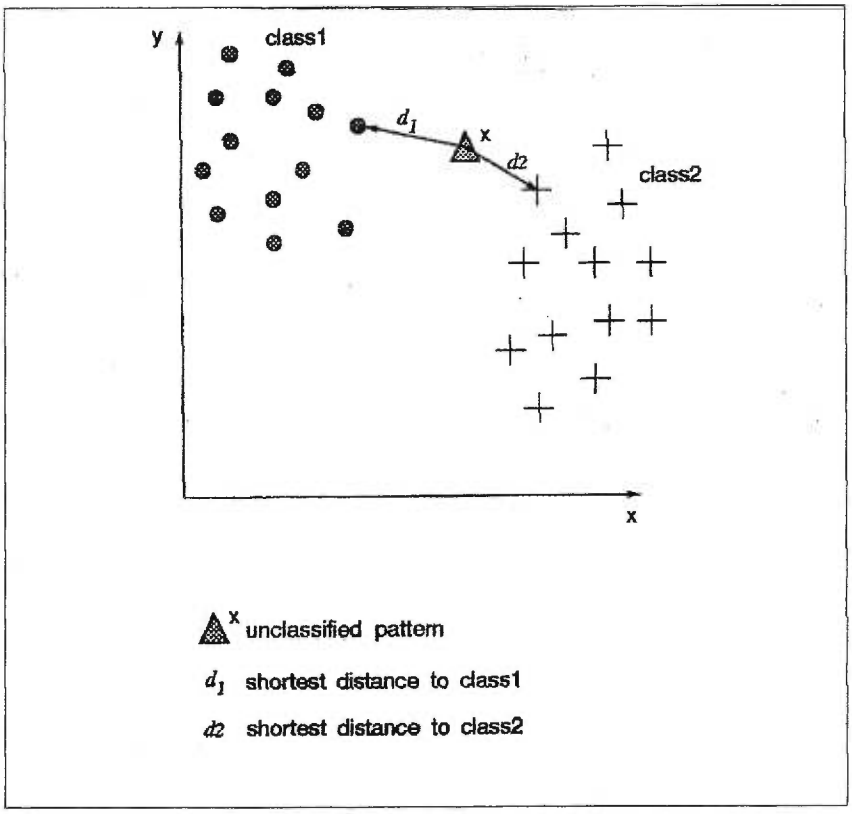

Figure 2.2.1.a. : Plan cartésien illustrant la séparation de deux classes (Beale et Jackson, 1991, 21)

2.2.2. La fonction discriminante

Dans l'exemple nous occupant, la fonction qui doit être effectuée par le réseau est une discrimination afin de former deux classes (grande taille et petite taille) à partir des individus regroupés en agglutinations (*clusters*). Le réseau doit ensuite pouvoir assigner tout nouvel individu à l'une des deux classes. Mathématiquement, une telle tâche est celle d'une fonction discriminante (*discriminant function*). (Beale et Jackson, 1991, 18-20) Bien

[22] Nous nous situons à un niveau qui est encore neutre par rapport aux applications et aux interprétations cognitives des réseaux de neurones (classifications conceptuelles). Ce qui suit nous permettra de mieux situer

évidemment l'exemple ci-dessus est simpliste car les classes sont très rarement aussi facilement départageables. Notez aussi que les représentations géométriques du vecteur de caractéristiques et de l'espace d'activation deviennent impossibles lorsqu'ils dépassent la tridimensionnalité.

2.2.3. Les techniques de classification

Plusieurs techniques de classification remplissent le rôle de la fonction discriminante. Nous en présentons deux : la classification du plus proche voisin (*nearest neighbour classification*) et la distance métrique.

2.2.3.1. la classification du plus proche voisin

La classification du plus proche voisin implique le raisonnement suivant pour déterminer à quelle classe un nouvel élément appartient. Admettons que nous avons deux classes dans un espace de caractéristiques et nous souhaitons attribuer un nouvel élément à l'une des deux classes. La technique du plus proche voisin soutient que nous devons calculer la distance entre le nouvel élément et les deux classes. Nous obtenons donc deux distances, soit la distance entre le nouvel élément (x) et la première classe – ppv(classe1) pour « plus proche voisin dans la classe 1 » – et la distance entre le nouvel élément et la deuxième classe – ppv(classe2). Une fonction discriminante possible est la suivante : $f(x) = ppv(classe1) - ppv(classe2)$. Dans le cas où $f(x)$ est négatif, x appartient à la première classe et dans le cas où $f(x)$ est positif, x appartient à la deuxième classe, ce qui permet de déterminer à quelle classe appartient le nouvel élément. Une façon de rendre cette mesure plus précise est de mesurer la distance entre le nouvel élément et plusieurs membres de chacune des classes en faisant ensuite la moyenne des distances. Cette

---

les thèses de Paul M. Churchland.

classification est la « classification du plus proche voisin K » (*"K" nearest neighbour classification*). (Anderson, 1995, 436-40; Beale et Jackson, 1991, 21-3 et Smith, 1990)

2.2.3.2. la distance métrique

Le calcul de la distance métrique (*distance metric*) est une façon de déterminer la similarité entre deux éléments.[23] Mesurer la distance Hamming (*Hamming distance measure*) permet d'évaluer la différence entre deux vecteurs. Dans le cas où **x** et **y** sont deux vecteurs différents et que chacun de leur élément respectif est noté comme étant x_i et y_i et que la distance Hamming est **H**, alors **H** = $\sum(|x_i - y_i|)$. Nous obtenons alors la somme des différences pour chaque élément entre les deux vecteurs. Notons que dans le cas de vecteurs codés en binaire, la distance Hamming devient une opération de « ou exclusif » car dans ce cas $|x_i - y_i|$ est équivalent à $x_i \vee y_i$. Il y a aussi le calcul (plus précis) de la distance euclidienne. Il s'agit de mesurer la distance euclidienne entre tous les éléments des vecteurs. Si l'on a encore deux vecteurs **x** et **y**, alors $\mathbf{d(x, y)}_{euc} = \sqrt{\left(\sum_{i=1}^{n}(x_i - y_i)^2\right)}$. Il y a aussi deux versions simplifiées de la distance euclidienne. La « mesure du pâté de maison » (*city block measure*) n'utilise pas les carrés pour mesurer les distances. Elle ne fait que calculer comme la distance Hamming la distance ou la différence entre les deux vecteurs, ce qui offre un calcul moins précis. La voici : $D_{cb} = \sum_n |x_j - y_j|$. La distance carrée (*square distance*) est d'autant plus simple à calculer. Il s'agit de calculer la distance maximale entre deux vecteurs, c'est-à-dire de calculer la distance la plus grande entre les deux éléments les plus éloignés des deux vecteurs. Nous avons donc : $D_{sq} = MAX|x_i - y_i|$.

[23] Ce qui suit servira directement d'explication de la mesure de la similarité conceptuelle par une métrique sémantique chez Churchland.

(Beale et Jackson, 1991, 22-6) Ces différentes mesures de distance métrique sont intégrés dans l'interprétation conceptuelle de Churchland sous la bannière d'une « métrique sémantique ». Le calcul de la proximité des vecteurs est alors interprété comme un calcul de la similarité de deux concepts.

2.2.4. Les classificateurs linéaires

Nous avons vu que les réseaux de neurones effectuent des tâches de classification. L'exemple illustré par la figure 2.2.1.a. est une simplification. Dans la pratique, les tâches de classification sont des opérations plus complexes où les régions à classer ne sont pas aussi facilement départageables. Une façon d'arriver à effectuer ces classifications est d'inclure un vecteur de poids (*weight vector*). Celui-ci servira à faciliter la classification en introduisant un vecteur à partir duquel le vecteur délimitant les classes est tracé. La nouvelle fonction discriminante que nous obtenons ainsi est : $f(x) = \sum_{i=1}^{n} w_i x_i$ où x_i est le i-ème élément du vecteur d'entrée; où w_i est le i-ème élément du vecteur de poids et n est la dimensionalité du vecteur d'entrée. Si l'on définit les deux appartenances possibles comme si $f(x) > 0$ = classe A et si $f(x) < 0$ = classe B, alors le défi est de trouver un vecteur de poids qui nous permettra de découper les classes d'une telle façon. Si l'on tient compte du fait que le vecteur de poids détermine la pente du vecteur de classification et que nous avons l'intersection du vecteur de classification avec l'axe des x_2, alors nous pouvons convertir notre fonction discriminante en une droite ($x_2 = mx + b$). Autrement dit, nous avons ici la possibilité d'une fonction discriminante départageant deux classes. Malheureusement, le grand problème est de trouver quel vecteur de poids nous amènera à cette solution, ce qui n'est pas une mince tâche étant donné que ce vecteur est trouvé par tâtonnements. (Beale et Jackson, 191, 27-30)

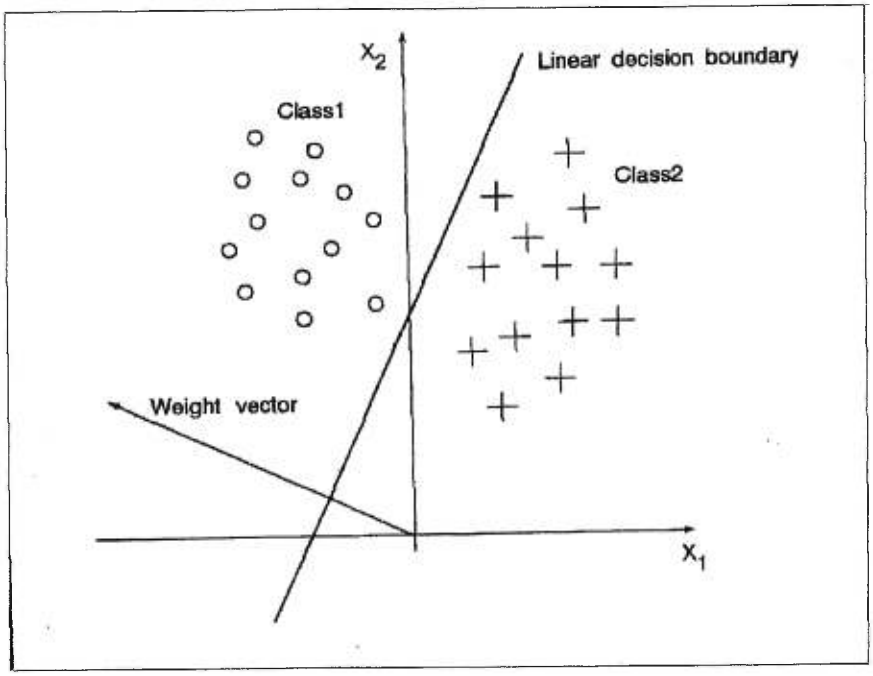

Figure 2.2.4.a. : Une tâche de classification produite
à l'aide d'un vecteur de poids (Beale et Jackson, 1991, 28)

2.2.5. la classification statistique

Les classificateurs linéaires sont des méthodes de classification déterministes. Il y a aussi des méthodes statistiques [24] de classification. La classification bayesienne notamment est utilisée pour formaliser l'idée que plus nous avons de connaissances au sujet d'un élément à classer, plus nous avons de chance de déterminer sa classe d'appartenance. Il s'agit donc de probabilités conditionnelles. (Beale et Jackson, 1991, 33) La formalisation est la suivante : $P(G_i|X) = \frac{P(X|G_i) \bullet P(G_i)}{\sum_j P(X|G_j) \bullet P(G_j)}$ où $P(G_i|X)$ = la probabilité que G_i étant donné X (probabilité conditionnelle). La formule permet de déterminer la probabilité qu'un

[24] La distinction entre méthodes déterministes (rigides) et méthodes statistiques sera importante pour notre troisième chapitre parce que les méthodes probabilistes peuvent parfois être converties en méthodes déterministes sans perte d'information ou de précision. Nous avons alors une transformation non réductionniste d'un système statistique en un système déterministe.

élément donné appartienne à la classe G_i lorsque nous savons déjà que $P(G_i|X) > P(G_j|X)$ pour $i \neq j$.

2.2.6. les classificateurs non linéaires : le vrai potentiel des réseaux de neurones

Le véritable potentiel des réseaux de neurones se fait voir dans les classifications non linéaires, c'est-à-dire celles qui ne peuvent pas être représentées par une simple droite. Déjà lorsqu'on a un réseau avec deux neurones d'entrée, on peut créer des classifications qui ne sont plus des lignes droites mais des espaces de classification formés par l'entrecroisement de deux droites que l'on nomme « régions convexes » (*convex regions*, *convex hulls*). (Voir figure 2.2.6.a.) Plus il y a d'unités dans la première couche, plus il est possible de tracer d'arêtes à la région convexe. Cependant, lorsqu'on ajoute une troisième couche d'unités[25], celle-ci peut prendre comme entrée les régions convexes des unités antérieures/inférieures. L'enchevêtrement des régions convexes peut ensuite conduire à la classification de n'importe quelle classe[26], c'est-à-dire que n'importe quel type de région convexe peut être créé. Un réseau à trois couches peut donc effectuer toute tâche de classification (théorème de Kolmogorov). (Beale et Jackson, 1991, 86) De plus, l'usage d'unités avec des fonctions de sortie sigmoïdes permet aux premières unités de produire des régions convexes avec des surfaces courbes. Cela ajoute à la précision de la classification. La figure 2.2.6.a., reprise de Beale et Jackson (1991, 87), compare la puissance de classification des réseaux à une, deux et trois couches.

[25] C'est-à-dire lorsque l'on passe d'un réseau à deux couches à un réseau à trois couches.
[26] Cette puissance computationnelle est bien sûr limitée par le nombre d'unités.

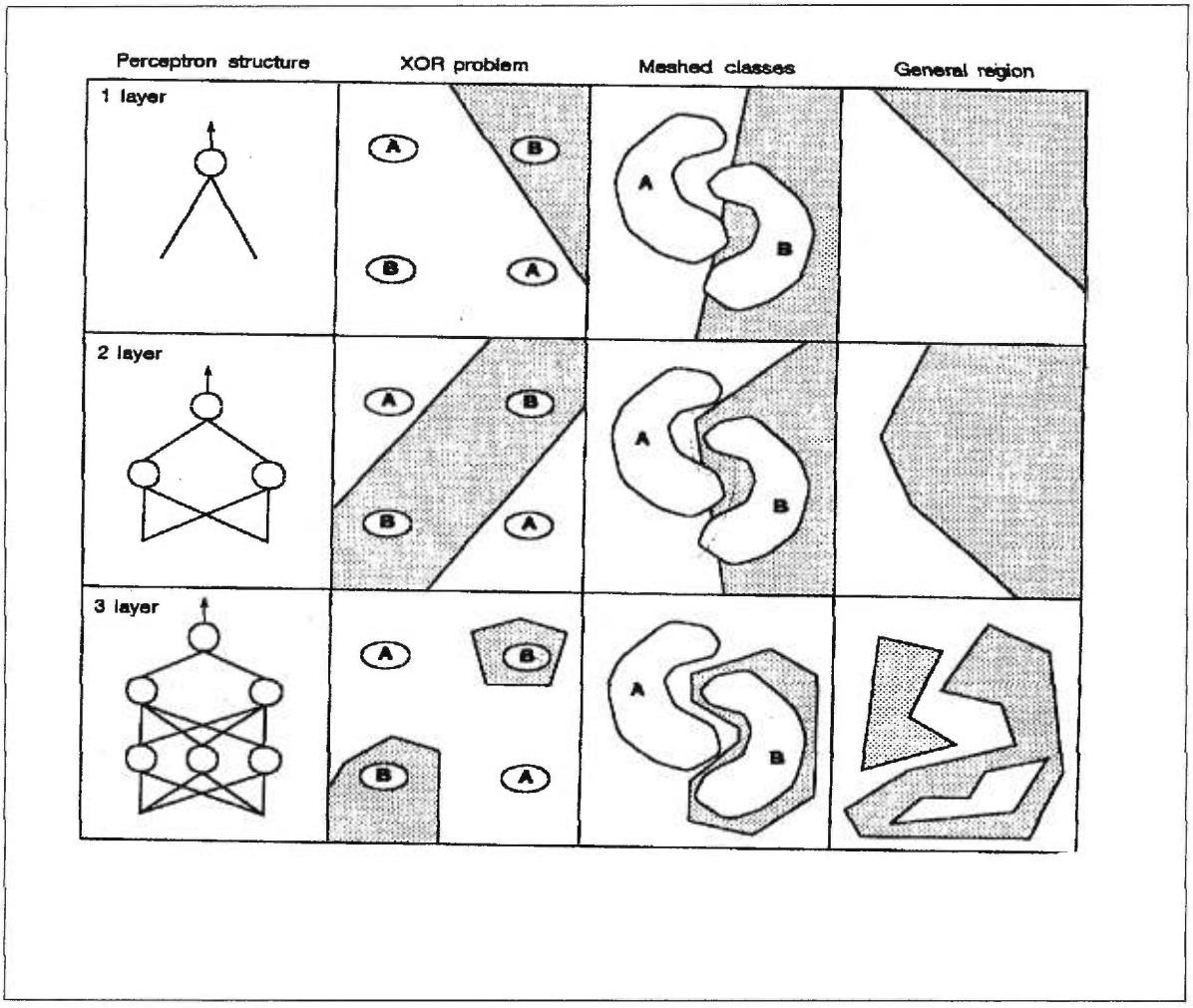

Figure 2.2.6.a. : Les régions convexes des réseaux à une, deux et trois couches selon Beale et Jackson (1991, 87)

2.2.7. Le fonctionnement interne des réseaux : la classification comme satisfaction de contraintes

Une façon répandue de concevoir les tâches de classification des réseaux de neurones consiste à dire que la classification des réseaux est la recherche d'une solution à un problème de satisfaction de contraintes dans un espace d'activation à n-dimensions. (Anderson 1995, 401-32; Smolensky, 1992, 92-100; Beale et Jackson, 1991, 79-83 ainsi que Rumelhart, Smolensky, McClelland et Hinton, 1986) Brièvement, l'idée est la suivante. La réponse au problème de classification est représentée comme la recherche d'une solution où il y a satisfaction maximale de contraintes. Admettons un réseau où deux poids peuvent varier. Nous avons alors un graphique à trois dimensions, deux pour représenter les poids

et une troisième pour représenter « l'énergie », c'est-à-dire le taux d'erreur[27] du réseau. Plus le taux d'erreur est bas, moins il faut d'énergie au réseau pour produire les réponses adéquates. L'espace d'activation est représenté sous la forme d'un paysage (*landscape*) à trois dimensions où l'on trouve des bassins et des sommets. Les bassins correspondent à la solution optimale, l'endroit où le plus grand nombre de contraintes sont satisfaites. (Beale et Jackson, 1991, 80) Autrement dit, les bassins représentent la situation où le réseau déploie le moins d'énergie pour produire les bonnes réponses (en produisant moins d'erreur). Or, la règle d'apprentissage permet au réseau de « trouver » les bassins en réduisant le taux d'erreur, c'est-à-dire en favorisant la production de réponses appropriées. Cette façon de concevoir le comportement (interne) des réseaux révèle que si un réseau classe correctement des vecteurs, alors cela signifie qu'il a trouvé les vrais minimum (les bassins) pour chaque entrée. La règle d'apprentissage a en quelque sorte permis l'exploration du terrain jusqu'à ce que les « oasis » soient repérées! Il faut aussi noter que cette façon d'expliquer le fonctionnement interne des réseaux note adéquatement le problème des minimum locaux, c'est-à-dire des situations où le réseau « croit » à tort avoir trouvé la bonne réponse, ce qui l'amène à se figer dans de faux minimum d'énergie. Une fois le réseau coincé dans une telle situation, l'apprentissage est en quelque sorte bloqué par une fausse réponse que l'on ne peut pas remplacer par une meilleure.[28]

---

[27] Le taux d'erreur est la différence entre le target input et l'état d'activation que nous avions notée dans la section 2.1.2.7. dans la formule $\Delta w_{ij} = \eta(t_i(t) - a_i(t))o_j(t)$. L'apprentissage permet de réduire le taux d'erreur à proportion inverse de l'éloignement de l'état d'activation du target input.

[28] Il y a des façons de remédier à ce problème que nous laissons de côté. Voir Beale et Jackson (1991, 91-7)

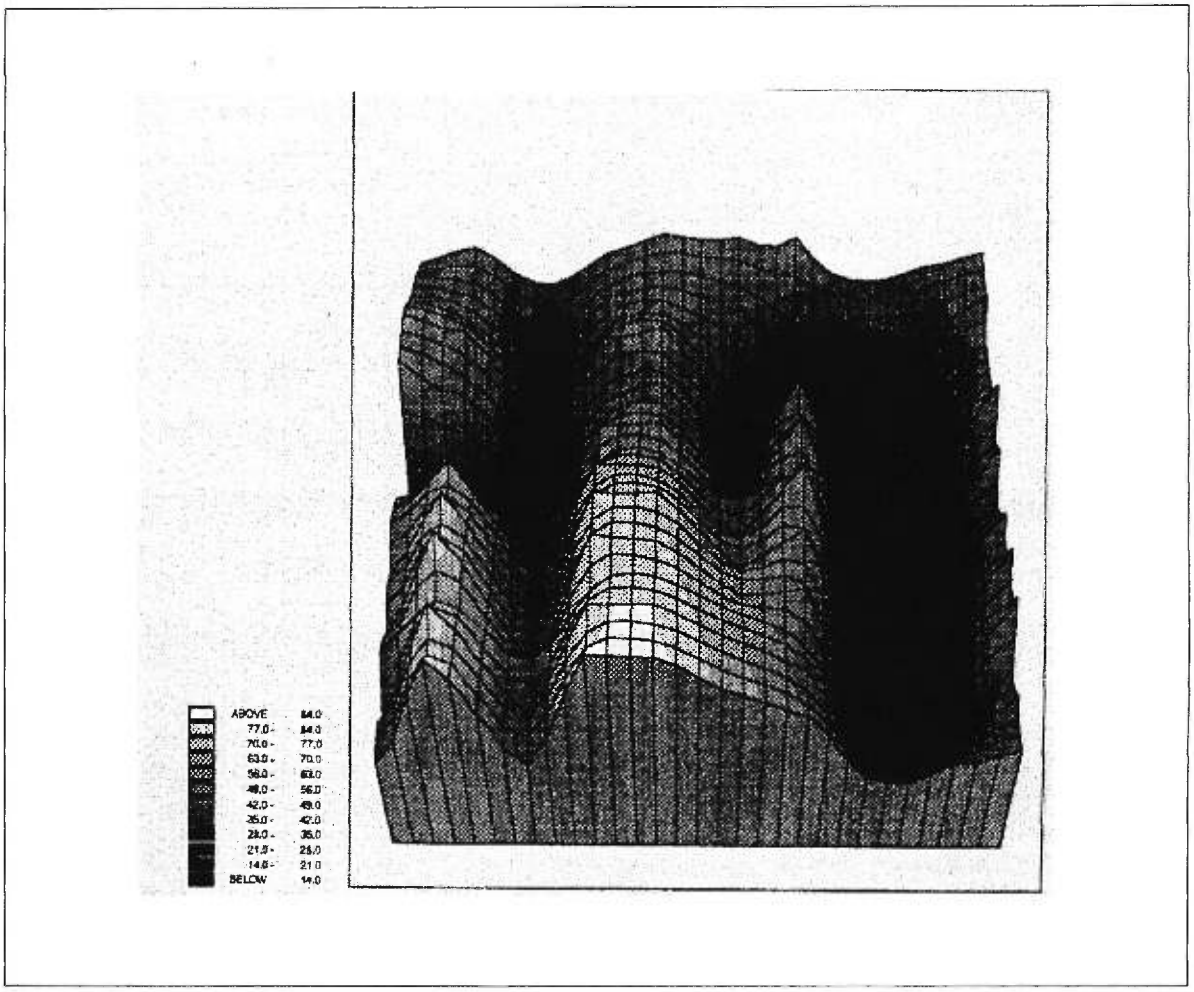

Figure 2.2.7.a. : L'espace de solution en termes de satisfaction de contraintes (Beale et Jackson, 81)

Rumelhart, Smolensky, McClelland et Hinton (1986) ont utilisé une version de cette technique d'illustration pour étudier les schèmes, des « structures de data représentant les concepts génériques ». (Rumelhart, Smolensky, McClelland et Hinton, 1986, 18) Leur simulation visait à créer des concepts de pièce (par ex. : une cuisine, un salon, etc.) à partir d'une quarantaine d'éléments trouvés dans des pièces (par ex. : un plafond, un foyer, etc.). La simulation se déroulait comme suit. Le réseau se voyait présenté un vecteur de caractéristiques et il devait l'associer à un type de chambre. À la différence du paragraphe précédent, le fait de trouver une solution était représenté par l'atteinte des sommets d'une

figure (au lieu des bassins). Le réseau de la figure 2.2.7.b. présente un espace de solution pour la simulation de Rumelhart et *al.*

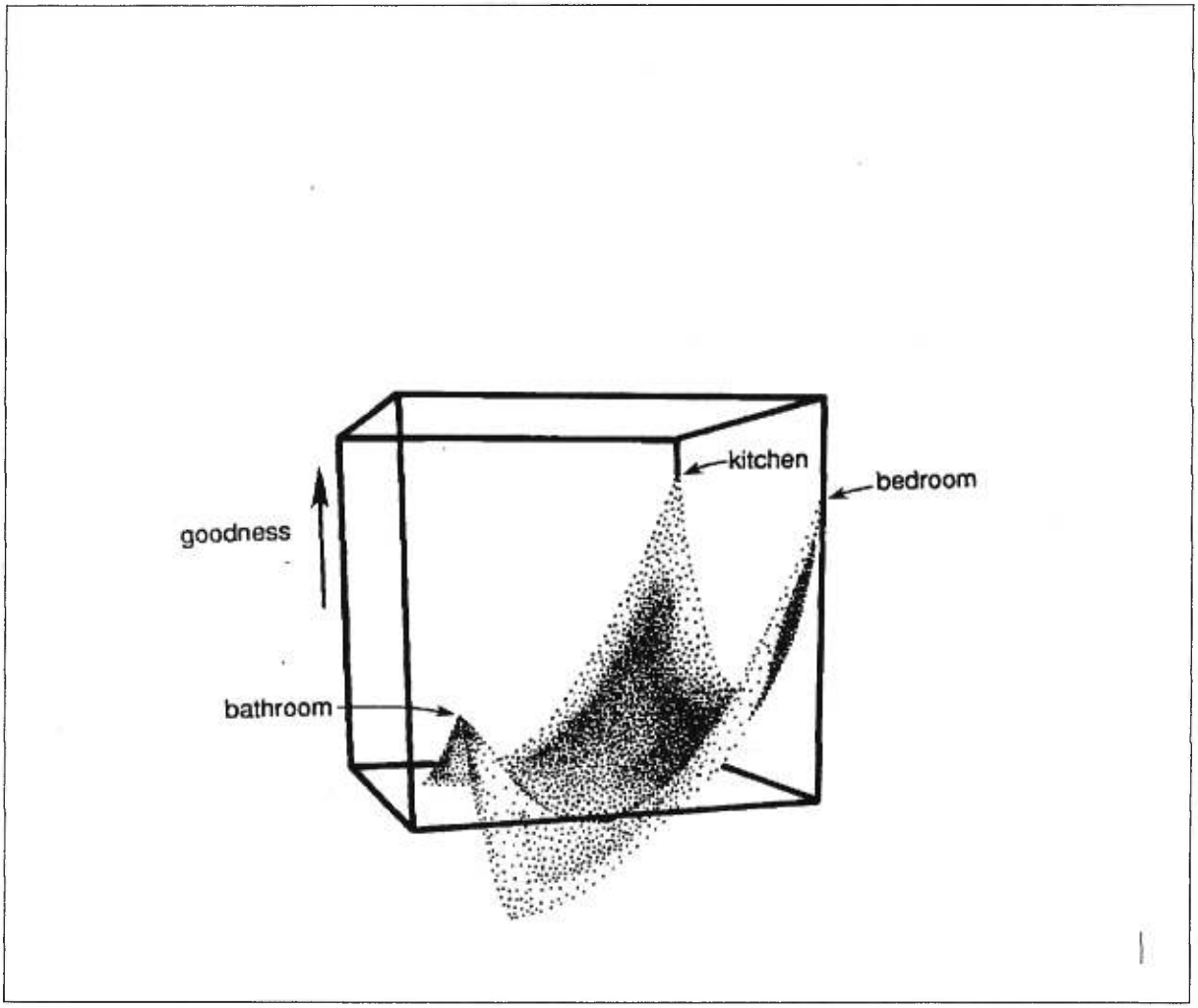

Figure 2.2.7.b. La satisfaction de contraintes pour la production
de schèmes de chambres (Rumelhart, Smolensky, McClelland et Hinton, 1986, 31)

2.3. Paul Churchland et la conceptualité des réseaux de neurones

Cette troisième et dernière section expose l'interprétation des capacités conceptuelles des réseaux de neurones formulée par Churchland.[29] Nous présentons ensuite le débat sur la question des concepts qu'il entretient avec Fodor, représentant du cognitivisme classique dont la position a été présentée dans le premier chapitre.

2.3.1. l'interprétation de Churchland

Paul Churchland est l'un de ceux parmi la communauté philosophique qui tentent avec le plus d'ardeur de dégager les conséquences philosophiques des réseaux de neurones. La philosophie de l'esprit est un domaine d'attention privilégiée puisque Churchland fait appel aux développements connexionnistes pour proposer un modèle computationnel de l'esprit.[30] (Churchland, 1998a; 1998b; 1995; 1993; 1989)[31]

2.3.1.1. l'hypothèse connexionniste computationnelle générale de Churchland

Churchland soutient que l'esprit est computationnel et sur ce point précis il est d'accord avec les tenants du cognitivisme classique.[32] (Smolensky, 1994) Par contre, les computations du cerveau[33] se font non pas par l'intermédiaire de manipulations de symboles mais plutôt par des opérations sur des vecteurs d'activation. Son hypothèse générale est donc que le cerveau fonctionne d'après les principes des réseaux de neurones et cela l'amène à développer une théorie des concepts,[34] radicalement différente de celle retrouvée dans le cognitivisme classique.

---

[29] À moins d'avis contraire nous désignons par « Churchland », Paul M. Churchland.

[30] Peut-être pourrait-on parler d'une sorte de connexionnisme méthodologique ou philosophique à l'instar du béhaviorisme philosophique de Hempel qui s'attardait à dégager des conséquences philosophiques du béhaviorisme qui avait un pendant empirique.

[31] Voir aussi Paul M. Churchland et Patricia S. Churchland (1998a; 1998c et 1996), Patricia S. Churchland (1986) ainsi que Terence J. Sejnowski et Patricia S. Churchland (1992).

[32] Ils sont aussi d'accord sur le caractère représentationnel de l'esprit bien que la représentation se fasse de façon divergente, notamment au niveau conceptuel.

[33] On parle plus volontiers du cerveau dans le connexionnisme plutôt que de l'esprit comme dans une perspective classique.

[34] Pour des raisons qui sont probablement dues aux propriétés mêmes des réseaux (voir section 2.1.3.1. sur la généralisation), la question des concepts a été traitée de façon soutenue dans le connexionnisme. Voir Rumelhart, Smolensky, McClelland et Hinton (1986) pour des précision historiques sur l'abord fondamental de la problématique des concepts dans le connexionnisme.

2.3.2. le connexionnisme et les concepts

La divergence entre le connexionnisme de Churchland et le cognitivisme classique est particulièrement notable au niveau de la problématique des concepts car Churchland soutient que le cerveau se représente le monde à l'aide de vecteurs d'activation à très haute dimension et non à l'aide de symboles.[35] (Churchland, 1998a, 41) Dans le cadre d'une analogie entre les réseaux connexionnistes et le fonctionnement du cerveau, Churchland soutient que les concepts sont des régions où s'agglutinent des éléments dans des espaces d'activation à n-dimensions.[36] Des méthodes[37] permettent ensuite de circonscrire et d'interpréter ces agglutinations afin d'en dégager des prototypes, c'est-à-dire des représentations typiques d'une chose donnée (par ex. : un chien prototypique, une salle de bain prototypique, un rouge prototypique, etc.).[38] Il faut noter que la généralité du prototype est fondée sur une cueillette statistique, une induction qui permet de dégager les caractéristiques principales des objets subsumés sous le concept. Somme toute, dans la perspective de Churchland un concept est une classe de caractéristiques et une classe est une région de classification dans un espace d'activation.[39] (Paul M. Churchland, 1998b) Notons ici que Churchland intègre la théorie prototypique des concepts au connexionnisme, ce qui va presque de soi. Autrement dit, pour Churchland, le connexionnisme offre une

---

[35] Notons ici que ce n'est pas le cerveau qui « uilise » ces vecteurs. Les vecteurs sont plutôt une description du fonctionnement du cerveau. La nature de l'explication diffère de la chose à expliquer.

[36] Voir section 2.2. pour des explications

[37] Nous exposons ces méthodes dans le cadre du débat entre Fodor et Churchland afin de faire valoir leur pertinence dans ce contexte.

[38] En peu de mots, la théorie prototypique des concepts soutient que les concepts sont des prototypes. Les prototypes sont des représentations typiques d'éléments donnés tels une chaise, un arbre, une pomme, etc. La généralité du prototype, en tant qu'indicateur exemplaire d'un type donné, est fondé sur une cueillette visant à déterminer quelles sont les caractéristiques fondamentales d'un concept. Une fois qu'un prototype est fixé, des calculs de similarité peuvent être effectués afin de déterminer dans quelle mesure l'instanciation d'un concept donné correspond au prototype approprié. (Par exemple dans quelle mesure un dalmatien correspond-il au chien prototypique?) (Smith, 1990)

[39] Ce sont les classes d'éléments que l'on peut délimiter avec les méthodes exposées dans la section 2.2. précédente.

implémentation convaincante de cette théorie.[40] Cependant, la théorie prototypique est une théorie indépendante du connexionnisme et on peut la défendre à partir d'autres positions.

Par ailleurs, le processus de conceptualisation où un « élément tombe sous un concept » est complètement réinterprété en termes de relations de proximité à un concept, une région de classification dans un espace d'activation. Une métrique sémantique (section 2.2.3.) permet de déterminer objectivement dans quelle mesure un nouvel élément ressemble au concept prototypique, ce qui détermine ensuite sa classe d'appartenance. Par conséquent, un élément tombe sous un concept lorsque ses propriétés (déterminées par son encodage vectoriel) le mettent en relation de proximité avec le prototype dans l'espace d'activation.

Prenons l'exemple hypothétique de la figure 2.3.2.a. afin de bien saisir cette thèse. On y retrouve le profil d'un chat prototypique ainsi que deux nouveaux éléments à classer, le premier représentant un chien ($élément_1$) et le deuxième un oiseau ($élément_2$). Le profil d'un élément est codé sous la forme d'un vecteur binaire à huit dimensions. Le calcul simple de la distance métrique Hamming (voir section 2.3.2.) révèle la distance entre deux vecteurs et donc leur ressemblance ou dissemblance. En comparaison avec le chat prototypique, le chien montre une distance de H = 4 tandis que l'oiseau a une distance H = 6. Dans une représentation à l'intérieur d'un espace d'activation (ici impossible à cause des huit dimensions), le vecteur chat se situerait dans le coin où toutes les arêtes indiquent un « 1 », c'est-à-dire la présence de chaque caractéristique. Le chien serait en quelque sorte à mi-chemin entre la jonction des « 0 » et la jonction des « 1 ». Et l'oiseau serait à mi-chemin entre le chien et la jonction des « 0 ». En d'autres mots, le vecteur chien est plus près du vecteur chat que le vecteur oiseau. Le chat ressemble donc plus au chien qu'à l'oiseau. Mais le chien n'est pas assez semblable au chat pour être considéré comme un

[40] Par conséquent, pour Churchland, un concept = un prototype.

chat selon la répartition des régions de classification (c'est-à-dire les concepts) dans l'espace d'activation.

Caractéristiques	Profil d'un chat prototypique	Nouvel élément$_1$ à classer	Nouvel élément$_2$ à classer
poilu	1	1	0
miaule	1	0	0
agile	1	0	1
carnivore	1	1	0
mammifère	1	1	0
griffu	1	1	1
oreilles triangulaires	1	0	0
yeux oblongs	1	0	0

Figure 2.3.2.a. : La classification vectorielle de prototypes

Les capacités conceptuelles (et cognitives en général) sont obtenues à la fin temporelle et computationnelle du processus d'entraînement du réseau. Nous obtenons alors une hiérarchisation de l'espace d'activation puisque le réseau parvient à classer les éléments avec l'apprentissage d'un schéma de connectivité (voir sections 2.1.2.4., 2.2.4. et 2.2.7) permettant de départager l'espace d'activation adéquatement. La hiérarchie ainsi obtenue reflète les classifications que le réseau a appris à effectuer. Dans l'exemple précédent, cela voudrait dire que le réseau terminerait son entraînement avec le découpage suivant : chat = 8; chien = 4 et oiseau = 2. Formulé autrement, le chat répond à 100% (8/8) des critères établis, le chien à 50% (4/8) et l'oiseau à 25% (2/8).

Churchland (1995, 24) offre un deuxième exemple plus réaliste de classification par région occupée dans l'espace d'activation. Dans ca cas, le goût est représenté selon un espace à

trois dimensions.[41] Une arête représente l'activation des cellules réceptrices sensibles au goût sucré, une autre au goût sûr et une dernière au goût salé. Chaque point dans l'espace représente une combinaison unique des trois types de goût représentés. Les éléments prototypiques d'une catégorie sont regroupés autour d'agglutinations respectives. On retrouve ainsi une région typiquement sucrée, une région typiquement salée et une autre typiquement sûre. Si l'on émet l'hypothèse assez réaliste de Churchland selon laquelle les humains sont capables de distinguer dix degrés à l'intérieur de chacun des types de goût[42], alors nous sommes capables de distinguer avec la combinaison des quatre types de goût (en rétablissant les récepteurs de l'amer) $10 \times 10 \times 10 \times 10$, soit 10^4 goûts. Cependant, les capacités linguistiques limitées de l'humain font que la communication ne reflète pas une telle complexité. (Churchland, 1995, 21-4)

En bref, l'interprétation de Churchland soutient que le processus de conceptualisation et les concepts eux-mêmes peuvent être expliqués par les classifications opérées par des réseaux de neurones qui modélisent des fonctions cognitives. Bien que sommaire, la présentation précédente nous prépare à considérer de plus près quelques caractéristiques des concepts que l'on peut dégager de l'interprétation de Churchland, à savoir : (1) le caractère interne et externe du contenu conceptuel, (2) le caractère holistique du contenu conceptuel; (3) le caractère modérément empirique du contenu conceptuel, (4) le caractère pragmatique de la conceptualité et (5) le caractère dynamique de la conceptualité. Regardons un à un ces éléments.

---

[41] Il va s'en dire qu'une dimension du goût (l'amer) a été abandonnée pour des fins d'illustration.
[42] Les degrés représentent le niveau d'activation des réseaux de récepteurs d'un goût donné.

2.3.2.1. le caractère interne et externe du contenu conceptuel

Churchland soutient une forme mixte d'internalisme et d'externalisme des concepts (*concept internalism* et *concept externalism*) selon lequel le contenu d'un concept est déterminé par son contenu étroit et large (*narrow* content et wide *content*)[43]. (Churchland, 1998b, 108) Le contenu étroit (*narrow content*) est fixé en vertu de ce qui se passe (causalement) à l'intérieur de la personne ayant le concept tandis que le contenu large d'un concept est fixé par les interactions causales avec le monde extérieur. (Kim, 1998, 194) Les modèles connexionnistes rendent compte de cette dimension étroite et large du contenu d'une façon originale. En effet, au niveau du contenu large et externe, le résultat de leur classification est d'abord une réponse à des vecteurs d'entrée représentant le monde (par ex. : la codification prototypique des animaux ou du goût). (voir les sections 2.1.2.1 et 2.1.2.8.) Ces vecteurs, comme nous l'avons vu, présentent des éléments significatifs[44] du monde à partir desquels le réseau classera, c'est-à-dire produira des agglutinations où les objets semblables sont regroupés. Les concepts découlent d'un découpage des régions d'agglutination qui respecte ces paramètres. Churchland intègre donc une forme d'externalisme dans sa théorie des concepts. Qu'en est-il de l'internalisme?

Il faut ajouter que Churchland tient compte aussi des relations internes entre concepts puisque dans les réseaux connexionnistes, l'économie cognitive interne est structurée de façon à refléter les relations de similarité entre les concepts. La structure représentationnelle interne est un reflet de son rapport avec le monde. (Les choses sucrées sont rassemblées autour d'un pôle, les sûres d'un autre, etc.) Selon lui, ce phénomène de

---

[43] Voir Kim (1998, 184-210)

[44] Il est à noter que la fonction représentationnelle du vecteur peut souffrir d'un certain arbitraire car qui déterminera quelles seront les caractéristiques dont on doit tenir compte? Voir la section 2.3.4.3. pour les critiques de Fodor à ce sujet.

structuration interne devient d'autant plus important que l'organisme est complexe. Par conséquent, dans le cas des humains, il y aurait même plus d'accent à mettre sur les relations internes entre les concepts que sur les rapports avec le monde au niveau du contenu. (Churchland, 1998b, 108) La position de Churchland est donc complexe dans la mesure où il intègre des éléments de l'externalisme et des éléments de l'internalisme. Deux propriétés importantes des concepts découlent de cette position nuancée. Les concepts sont définis à la fois par : (1) leurs relations causales par rapport aux macro-caractéristiques stables et objectives de l'environnement extérieur (l'externalisme de Churchland) et (2) leurs positions dans l'espace (d'activation) relativement aux autres points qui ont un contenu représentationnel (l'internalisme de Churchland). (Paul M. Churchland, 1998a, 84-5).

2.3.2.2. le caractère holistique du contenu conceptuel

Un autre point important est le fait que le contenu conceptuel pour Churchland est holistique. (Paul M. Churchland, 1993, 667) Cela veut dire précisément que le contenu conceptuel n'est pas déterminé de façon atomique en référence à des symboles individuels ayant une interprétation fixe comme chez Fodor. En fait, le holisme de Churchland est double. Premièrement, les concepts entretiennent une relation complexe au monde. Ils ne sont pas des éléments constitutifs donnés mais des regroupements de caractéristiques extirpées du monde. Leur contenu est délimité dans un environnement dynamique incluant du bruit et de la confusion sensorielle. (Churchland, 1993, 669) Par exemple, la reconnaissance d'un chaton implique une multitude de sous-caractéristiques qui coopèrent dynamiquement pour former une structure plus stable à un niveau d'abstraction supérieur. Autrement dit, les caractéristiques sont rassemblées pour ensuite donner lieu à un niveau plus abstrait à un concept qui regroupe les caractéristiques. Le contenu conceptuel est donc

déterminé dans un très large cadre interactif de sous-caractéristiques et n'est donc pas « donné » avec une interprétation standard comme dans le cognitivisme classique. (Churchland, 1993, 670) En ce sens, le contenu externe est holistique. Deuxièmement, le contenu conceptuel est déterminé par le rôle inférentiel. (Churchland, 1993, 671) Churchland soutient même que le contenu conceptuel est déterminé en grande partie par le rôle global qu'un concept joue dans l'économie cognitive interne[45] et le comportement moteur d'un organisme. (Churchland, 1998b, 109; Churchland et Churchland, 1993, 671) Il est donc évident que Churchland défend une forme de holisme du contenu conceptuel interne, ce qu'il avoue d'ailleurs ouvertement. (Churchland et Churchland, 1993, 667 et 671) On pourrait donc conclure que le holisme de Churchland est double. Car il est présent à la fois dans l'individuation du contenu dans les rapports avec le monde externe (confusion, discrimination sélective des caractéristiques, etc.) mais aussi dans l'individuation du contenu d'après le rôle global (au niveau de l'inférence et des interactions internes) d'un concept.

2.3.2.3. le caractère modérément empirique du contenu conceptuel

L'empirisme des concepts (*concept empiricism*) est présent dans l'idée externaliste que le contenu est déterminé en partie par les relations causales extérieures de l'organisme avec le monde. Le connexionnisme a une façon originale de rendre compte de la relation entre la conceptualité et le monde comme nous l'avons souligné à la section 1.3.2. parce que les concepts intègrent dans leur organisation interne (l'espace d'activation) une sorte de topographie du monde extérieur. Cependant, tout contenu conceptuel ne se réduit pas à des

[45] Il s'agit du départage de l'espace d'activation effectuée par les couches cachées, des relations entre les couches, des relations causales entre les premières et les dernières couches, des « champs récepteurs » des neurones individuels, etc. (Churchland, 1993, 671)

relations causales avec le monde extérieur.[46] Churchland laisse (beaucoup) de place aux éléments causaux internes (par ex. : croyances et les rôles inférentiels) qui viennent modifier la position finale d'un concept par rapport à la hiérarchie des autres concepts. Il faut donc voir les classifications internes pour ce qu'elles sont : moins des reflets de la structure externe de l'environnement que la façon interne de codifier les schémas observés dans le monde. En d'autres mots, le résultat d'une classification tient autant sinon plus compte des éléments internes que des éléments externes. (Churchland et Churchland, 1996, 278-9) Une preuve de cette affirmation est que peu importe la modalité sensorielle (le nombre et la nature des unités), les réseaux sont capables de produire des codifications internes hautement similaires. Les concepts ressemblent donc plus aux formes abstraites platoniciennes qu'aux concepts empiriques de Hume. (Churchland et Churchland, 1996, 281-2) Ceci dit, il n' y a pas que des « concepts empiriques » admis dans le schéma d'explication de Churchland. D'ailleurs, il insiste pour souligner que les couches cachées ne sont pas que de simples reflets de relations causales extérieures mais des codifications internes d'une dynamique entre les différents éléments internes du réseau. (Churchland et Churchland, 1996, 279) En outre, certains concepts peuvent être situés dans des réseaux qui sont reliés indirectement au monde, c'est-à-dire par l'intermédiaire d'autres réseaux. (Churchland, 1998b, 110; 1993, 671) Ainsi, Churchland soutient une position empiriste, bien sûr, mais elle est modérée parce qu'elle admet que le contenu conceptuel n'est pas déterminé exlusivement par l'expérience directe avec le monde.

2.3.2.4. le caractère pragmatique de la conceptualité

Le caractère pragmatique de la conceptualité peut se faire sentir dans la dépendance du contenu d'un concept à son environnement. Le contenu n'est pas isolé comme un atome

[46] Contrairement aux accusations de Fodor, voir section 2.3.4.4.

réseaux connexionnistes au monde extérieur, ils sont directement arrimés à se qui passe dans le monde. Cela est particulièrement vrai des contenus conceptuels plus près de la sensation et de la perception car ils ont une signification directement branchée sur ce qui se passe dans le monde. Deuxièmement, les relations internes entre concepts présentés sous forme de transformation matricielle de vecteurs d'activation et de vecteurs de sortie permettent de rendre compte des liens mutuels entre concepts. Autrement dit, ce portrait tient compte de l'économie interne de la conceptualité et il admet la possibilité d'opérations sur des concepts car les concepts peuvent servir d'entrée à un réseau d'ordre supérieur. Des contraintes (plus internes) telles que le rôle inférentiel peuvent donc venir jouer un rôle dans la détermination du contenu conceptuel à l'intérieur de ce schéma d'explication. (Churchland, 1993, 671) Une forme de pragmatisme est donc présente à deux niveaux. Premièrement, il y un pragmatisme plus empirique dans le rapport que la conceptualité entretient avec le monde. Et deuxièmement, il y a un pragmatisme plus abstrait à l'oeuvre dans l'idée que le contenu conceptuel est déterminé par ses relations aux autres concepts et non pas en simple relation causale directe. Le rôle commande donc le sens à ces deux niveaux.

2.3.2.5. le caractère dynamique de la conceptualité

Le dynamisme de la conceptualité découle à la fois des quatre caractéristiques précédentes et des capacités d'apprentissage des réseaux. D'une part, le contenu conceptuel a une dimension externe et interne, il est holistique, il a des origines empiriques et il est régi par un certain pragmatisme. D'autre part, les réseaux peuvent modifier leurs schémas de connectivité de façon à intégrer les données de l'expérience dans leurs processus de classification. On peut en conclure qu'il y a une dynamique double inhérente à la conceptualité qui parachève la théorie connexionniste des concepts. La conceptualité est

dans une dynamique directe avec le monde puisqu'elle puise son contenu en partie dans ses relations au monde. À ce niveau si le monde change, la conceptualité saisissant le monde change aussi. Par contre, il y a aussi une dynamique qui régit l'organisation interne de la conceptualité. D'un côté, certains concepts ont un contenu rigide devenu indépendant de l'expérience (par ex. : les nombres[47]). D'un autre côté, certains concepts indirectement empiriques peuvent avoir une influence sur d'autres concept indirectement empiriques (par ex. : une addition sur un nombre) ce qui peut en dernier ressort influer sur le rapport direct avec le monde. La dynamique conceptuelle des réseaux de neurones est donc double car elle est influencée par le monde extérieur et le monde intérieur. De là découle aussi la sensibilité à l'environnement externe et la relative souplesse des réseaux que nous avions indiquées aux sections 1.3.1. et 1.3.2.

2.3.2.6. la thèse éliminativiste de Churchland

Churchland est amené à rejeter l'usage de la psychologie intentionnelle comme cadre d'explication scientifique de la cognition. Il prône un éliminativisme par rapport aux attitudes propositionnelles essentiellement parce que cette approche n'a pas connu de progrès et n'est pas susceptible d'intégration avec les neurosciences. (Paul M. Churchland, 1998a; 1998c; 1989 Churchland et Churchland, 1998b) Au niveau des concepts, cela signifie que Churchland rejette le cadre du cognitivisme classique (tel que présenté par Fodor) car celui-ci s'appuie sur la psychologie intentionnelle puisqu'il y a une correspondance entre les concepts de cette perspective et la terminologie du sens commun que nous utilisons pour caractériser les représentations mentales. Par exemple, le concept de table du cognitivisme classique correspond à ce que nous appelons communément « table ». Pour Churchland, le connexionnisme offre un cadre plus rigoureux et plus

[47] Voir Dehaene et Changeux (1993) ainsi que Dehaene (1997) à ce sujet.

plausible neurobiologiquement pour traiter la conceptualité. Évidemment, de telles affirmations ne plaisent pas à la plupart des tenants du cognitivisme classique. Fodor a formulé de nombreuses objections que nous allons considérer dès maintenant.

2.3.3. Le débat Fodor-Churchland

Le débat entre Churchland et Fodor devrait maintenant se présenter clairement. Churchland défend une conception holiste, pragmatique et dynamique de la conceptualité liée à un empirisme modéré et à un mélange d'internalisme et d'externalisme. Fodor, quant à lui, soutient une forme d'atomisme dans sa théorie sémantique informationnelle. Il s'insurge contre le holisme (fondé essentiellement sur le rôle inférentiel) et le pragmatisme puisqu'il défend une conception plus statique de la conceptualité. (voir section 1.2.6.) Les critiques de Fodor visent donc à faire valoir les éléments problématiques dans la position de Churchland relativement à ces points.

Selon Fodor, la position de Churchland comprend au moins cinq lacunes.[48] (1998, 1995, 1992 et 1993) Premièrement, elle ne permet pas de rendre compte de l'importante notion d'identité conceptuelle. Étant donné la complexité du fonctionnement du cerveau, il n'est pas évident qu'une chose telle que l'identité conceptuelle, voire même la similarité conceptuelle, soit possible dans un cadre connexionniste. Car si chaque concept correspond à l'état d'activation d'un réseau ou d'un ensemble de réseaux du cerveau, alors il n'est pas clair que deux personnes puissent jamais avoir le même concept. (Fodor et Lepore, 1992, 190-3) Le nombre de neurones dans nos réseaux (c'est-à-dire la dimensionalité des états d'activation et des sorties) ainsi que la diversité de leurs schémas d'activation semblent exclure effectivement cette possibilité. Pourtant peu importe l'agencement et les

productions idiosyncratiques de nos pensées, nous avons (et devons avoir pour rendre compte de la communication et du raisonnement) bel et bien dans les faits une identité conceptuelle d'après Fodor. Ce premier reproche est une sorte de reprise du problème de la réalisation multiple des états mentaux[49] (Poirier, 2000; Kim, 1998, 73-6) mais adapté à la question des concepts. Deuxièmement, le problème de l'identité conceptuelle conduit à celui de l'individuation des dimensions et de l'information collatérale. L'idée ici est que s'il est impossible de définir rigoureusement l'identité des concepts, alors il sera difficile de préciser comment on détermine quelle caractéristique compte au niveau du contenu d'un concept (problème de l'individuation des dimensions) et il sera difficile de déterminer quelle information compte pour déterminer le contenu (problème de l'information collatérale). Pour Fodor, cela revient au problème de déterminer quel est le critère qui détermine quand deux concepts sont identiques. Troisièmement, Fodor accuse Churchland de soutenir une forme d'empirisme conceptuel naïf (naive *concept empiricism*) selon lequel le contenu sémantique proviendrait exclusivement de l'expérience. Quatrièmement, le contenu conceptuel selon Churchland est holiste et, comme nous l'avons vu dans le chapitre précédent, Fodor défend une forme d'atomisme conceptuel dans lequel le contenu des concepts est suffisamment rigide pour ne pas avoir à tenir compte du rôle inférentiel ou d'autres éléments d'une théorie pragmatique de la signification. Et cinquièmement, Churchland trace une adéquation trop rapide entre prototypes sensoriels et concepts selon Fodor. Les concepts ne peuvent pas être des prototypes puisque les concepts sont compositionnels mais les prototypes ne le sont pas. (Fodor et Lepore, 1993; Fodor et Pylyshyn, 1988) Les prochains paragraphes présentent plus clairement ces objections et

---

[48] Pour une autre critique importante du connexionnisme mais de nature plus empirique et moins philosophique, voir Pinker et Price (1988).

[49] *Grosso modo* cet argument est évoqué par les fonctionnalistes contre les théories de l'identité psychophysique afin de défendre la possibilité de la réalisation d'un état mental dans des états physiques

ébauchent la « solution » de Churchland. L'accent sera mis sur la première et la quatrième critiques étant donné que la deuxième et la troisième se rattachent en grande partie à la première, laquelle est plus fondamentale. Cependant, avant d'aborder ces critiques, nous allons exposer la notion de similarité conceptuelle de Churchland.

2.3.3.1. la similarité conceptuelle

Les représentations dans les espaces d'activation (*activation spaces*) sont des moyens d'illustrer les vecteurs et les vecteurs sont des moyens de saisir l'état d'activation du réseau à un temps donné. (Voir section 2.2.1.) Chaque vecteur à n-dimensions peut être représenté dans un espace d'activation à n-dimensions. Par conséquent, tout vecteur donné peut être représenté comme un point dans un espace d'activation. (Paul M. Churchland, 1989, 102) Cette technique a l'avantage de « conserver » les relations métriques entre les différentes positions à l'intérieur même de l'espace d'activation. Il en découle que la relation de similarité (conceptuelle ou représentationnelle) est représentée comme une relation de proximité dans l'espace d'activation. Plus deux représentations sont proches l'une de l'autre, plus elles partagent de contenu et inversement plus elles sont éloignées, moins elles le font. Cette façon de concevoir la conceptualité et la représentationalité permet de définir la similarité conceptuelle et par extension l'identité conceptuelle par une métrique sémantique (voir la section 2.2.3.). Tous ces éléments sont contenus dans l'interprétation de Churchland, telle que nous l'avons exposée au début de cette section.

---

différents. Par exemple, une croyance est un état mental pouvant être réalisé différemment, la douleur aussi, etc.

2.3.3.2. le problème de l'identité conceptuelle

Cependant, Fodor reproche à Churchland de ne pas pouvoir tenir compte de l'identité conceptuelle et même jusqu'à un certain point de la similarité conceptuelle.[50] (Fodor et Lepore, 1993, 681; Fodor et Lepore, 1992, 190-7) Car pour déterminer si deux concepts sont semblables, il faut, selon Fodor, déterminer à la fois en vertu de quoi ils sont semblables et quel critère permettra de déterminer leur similarité. Le premier aspect de l'argument est repris à la section suivante (2.3.3.3.) portant sur les problèmes d'individuation des dimensions et de l'information collatérale. Le deuxième pose l'exigence d'un critère d'identité conceptuelle et il nous conduit directement à la problématique du holisme sur laquelle Fodor insiste longuement.

La critique du holisme est essentielle pour Fodor car il défend l'idée selon laquelle la psychologie repose sur l'explication en termes d'attitudes propositionnelles. (Fodor et Lepore, 1992, 187) Or, le holisme invalide en partie les explications de cette nature car il ne leur attribue pas de contenu solidement identifiable. Il conduit à les considérer comme une simple façon de parler et amenuise le critère d'identité conceptuelle nécessaire à ce genre d'explication.[51] Par conséquent, selon Fodor, le premier défi du holisme est de fournir un critère d'identité conceptuelle pour garantir la viabilité de l'explication psychologique par excellence. (Fodor et Lepore, 1993, 681) Toutefois, il y a aussi le lien intrinsèque entre la compositionalité et l'identité conceptuelle qui est brisé selon Fodor.

Nous avons mentionné à la section 1.2.4. que la compositionalité implique l'identité conceptuelle. Nous avons illustré cette notion en faisant référence au raisonnement

---

[50] Voir la section 1.2.2. pour l'élaboration de l'exigence d'identité conceptuelle et son lien à la compositionalité conceptuelle.

syllogistique qui doit nécessairement intégrer cette exigence. Nous aimerions maintenant revenir sur ce point en précisant deux implications de ce lien. Premièrement, si les concepts sont compositionnels, alors ils doivent être insensibles au contexte interne et externe de leur apparition. (Fodor, 1995, 145) Cette exigence est requise puisque selon le principe de compositionalité, un concept (élément constitutif) a le même contenu dans toutes les représentations complexes dans lesquelles il apparaît. (Fodor, 1995, 145) Le concept de chien dans *Rover is a dog* et *Rover is a brown dog* doit contribuer la même chose au niveau du contenu. Cela permet en fait d'expliquer la compositionalité. Deuxièmement, le principe de compositionalité exige que le contenu d'une représentation complexe soit déterminé par les contributions des éléments constitutifs. Par exemple, le contenu du concept « chien brun » doit être déterminé par le contenu des concepts « brun » et « chien » et l'appareillage combinatoire (*combinatorial apparatus*) qui les lie ensemble. (Fodor, 1995, 145) Autrement, la saisie des concepts « chien » et « brun » ne permettrait pas la saisie de « chien brun ». Selon le principe de compositionalité, le tout ne doit pas être plus que la somme des parties. (Fodor, 1998, 106-7; 1995, 146) Or, si le contenu des concepts dépend trop de l'environnement externe, alors on ne pourra pas les considérer comme les éléments primitifs à partir desquels on peut rendre compte du contenu des représentations complexes. Ces deux remarques font mieux valoir pourquoi la compositionalité requiert l'identité conceptuelle des éléments constitutifs et donc, en retour, pourquoi Fodor attache autant d'importance à la notion d'identité conceptuelle.

Par contre, Churchland rétorque qu'il est possible de définir l'identité conceptuelle non pas en tenant seulement compte des représentations idiosyncratiques d'un réseau en tant que produit d'un réseau particulier mais en ajoutant les deux conditions mentionnées à la fin de

---

[51] Voir section la section 1.2.7. pour des explications supplémentaires. On pourrait ajouter à cela la position

la section 2.3.1.3, soit les éléments externes et internes du contenu conceptuel. De cette façon, même si deux concepts son réalisés différemment, c'est-à-dire dans deux réseaux différents avec des architectures différentes[52], des unités réglées différemment, etc., il est quand même possible de mesurer la similarité de leur concept pour un phénomène donné en tenant compte à la fois de leur relation au monde externe et de leur configuration interne à l'aide d'une métrique sémantique. (Paul M. Churchland, 1998a, 85)

Clarifions quelque peu ce dernier énoncé. L'analyse des unités cachées des réseaux permet de dégager des représentations prototypiques.[53] Les classes sont identifiées comme des agglutinations (*clusters*) où des relations de similarité entre différentes représentations trouvent un parallèle topographique dans leurs relations de proximité (dans l'espace d'activation). Et les classes sont les concepts d'un réseau. Lorsque l'on représente les solutions de classification trouvées par deux réseaux différents, il y aura de légères divergences dans la position relative des points à l'intérieur de l'espace d'activation. Cependant, l'analyse des réseaux en fonction de leur rapport avec le monde externe et de la configuration de leur économie interne externe révélera une isomorphie des solutions de classification à certaines conditions. (Voir l'exemple de la page suivante.) Par conséquent, les stratégies de classification sont différentes mais la position relative des points est presque identique (Paul M. Churchland, 1998a, 87). Cela signifie que face à des phénomènes objectifs stables, deux réseaux codent de façon légèrement différente les signaux. Mais la position des vecteurs d'activation dans l'espace d'activation est essentiellement la même. La seule différence est l'orientation de la géométrie interne dans l'espace d'activation. Alors, il n'y aura qu'une translation et/ou une rotation à opérer sur la

---

éliminativiste générale de Churchland quant aux attitudes propositionnelles. (Voir section 2.3.2.6.)

[52] On pourrait ajouter que le problème est le même pour un individu à deux temps différents.

[53] C'est-à-dire des classifications opérées par le réseau. Voir la section 2.2 pour des explications à ce sujet.

figure de classification pour révéler l'isomorphie de la géométrie interne des classifications et donc l'équivalence du contenu de la conceptualisation. (Churchland, 1998a, 94)

Churchland donne l'exemple d'une tâche de classification de visages (en fonction de quatre familles) effectuée par deux réseaux pour illustrer cette idée. (Voir la figure 2.3.3.2.a.) Les deux réseaux forment des prototypes pour les autres familles. Bien que leurs stratégies de classification diffèrent légèrement, leur géométrie interne est la même. Par conséquent, ils illustrent la possibilité « d'une identité conceptuelle en dépit de la diversité neurale ». (Paul M. Churchland, 1998a, 88) Churchland utilise une mesure de distance métrique particulière. (voir les sections 2.2.3.2. à 2.2.3.5. à ce sujet) Elle permet de calculer la différence entre chaque arête de la figure contenue dans l'espace d'activation. Une moyenne est ensuite calculée en prenant en compte la différence entre chaque arête des deux figures et elle est ensuite soustraite à « 1 », la valeur d'une similarité parfaite. Il s'agit de[54] : $\text{similarité} = 1 - \text{avg.}\frac{|AB - A'B'|}{(AB + A'B')}$ où « 1 » est la valeur d'une similarité parfaite; AB et $A'B'$ sont les arêtes (variables) de deux figures et avg. signale le calcul de la moyenne de la différence entre toutes les arêtes des deux figures. Prenons par exemple le cas de deux triangles (ABC et $A'B'C'$) représentant les espaces d'activation de deux réseaux différents où AB = 1; BC = 1,5; AC = 1,6 et $A'B' = 0{,}98$; $B'C' = 1{,}54$ et $A'C' = 1{,}57$ alors le calcul des différences après le calcul de la moyenne serait de 0,989, soit presqu'une identité parfaite.[55] Il y a aussi une variante tenant compte des différences de taille entre les figures dans l'espace d'activation car un défaut de la première formulation est de ne pas

[54] C'est une variante de la mesure du pâté de maison (*city block measure*). Voir la section 2.2.3.2.

[55] On peut bien entendu envisager de simples modification qui conduiraient à des formules plus exigentes pour le calcul de la similarité.

pouvoir tenir compte de la grandeur relative des figures.[56] Voici donc une formulation tenant compte de cette contrainte où l'on introduit une constante « *c* » obtenue en calculant la somme des arêtes de chaque figure et en divisant ensuite la somme pour la première figure par la deuxième. Les vecteurs de la deuxième figure sont en quelque sorte normés sur ceux de la première. Nous avons ainsi[57] : $c = \sum_{1}^{n}(AB) \Big/ \sum_{1}^{n}(A'B')$ où n = le nombre d'arêtes des figures (dans le cas où les figures ont un nombre différents d'arêtes, il faudrait introduire n1 et n2 et ensuite trouver une nouvelle constante pour établir une comparaison plausible de la proportion entre les deux figures. C'est en fait le problème de comparer des figures à dimensionalité différente.). La nouvelle formule devient, en intégrant cette constante de proportionnalité : $sim. = 1 - avg.\frac{|AB - c(A'B')|}{(AB + c(A'B'))}$. (Churchland, 1998a, 89 et 112)

[56] La comparaison d'une grande figure à une très petite conduira inévitablment à une grande dissemblance des figures.

[57] Nous ajoutons le nombre n d'arrêtes dans l'opération de sommation afin de mettre en perspective la comparaison entre figures. Il ne s'agit que d'une réécriture ne modifiant aucunement la pensée et les conclusions de Churchland.

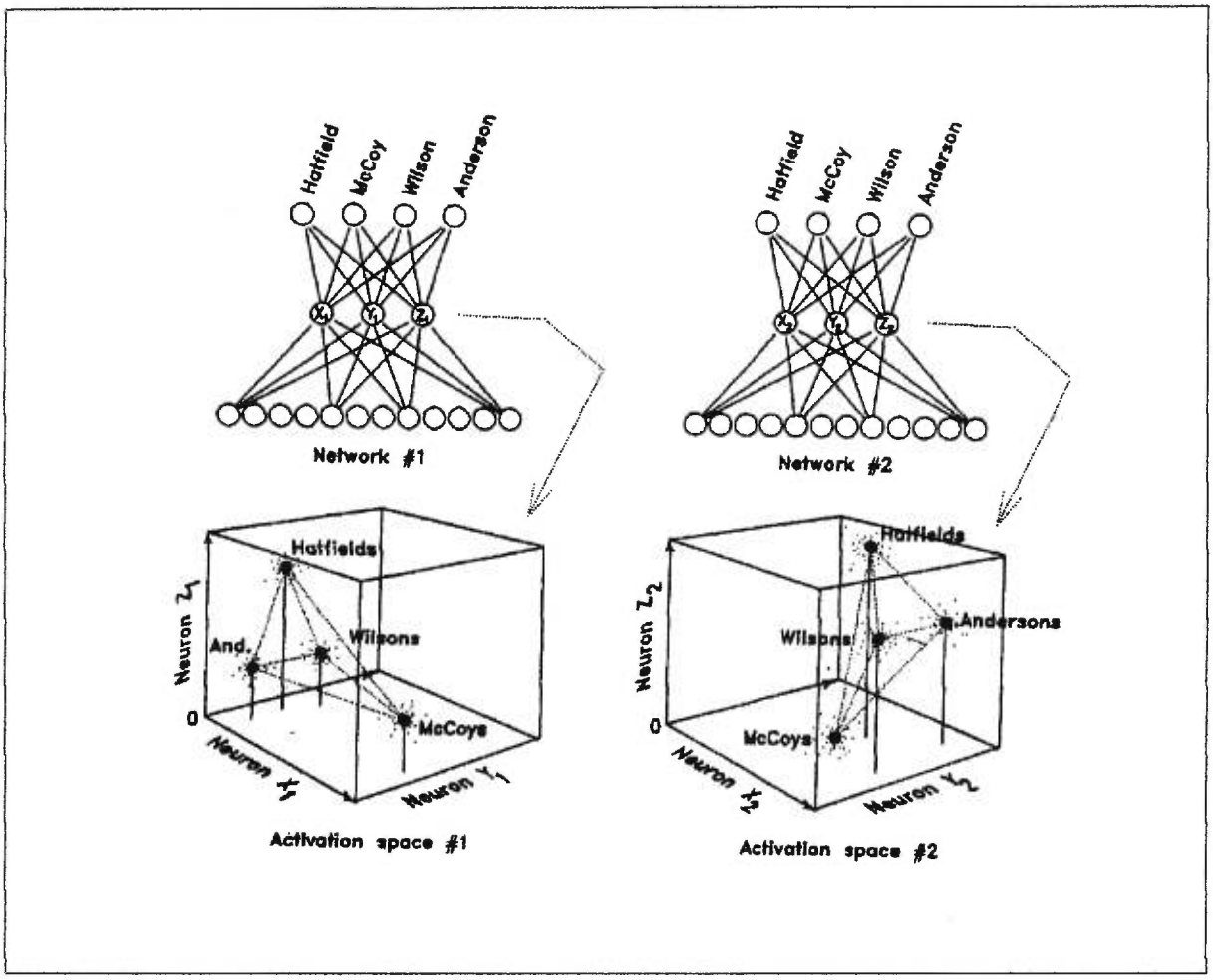

Figure 2.3.3.2.a. : La représentation de quatre types de visage dans des espaces d'activation légèrement différents (Churchland, 1998b, 86)

Churchland évoque aussi deux enquêtes empiriques où on a utilisé des mesures de distance métrique afin d'analyser les relations de similarité conceptuelle. La première est l'analyse de la couche cachée du célèbre réseau NETtalk où l'usage d'un dendogramme (*dendogram*) par ses créateurs, Rosenberg et Sejnowski, a permis d'identifier des relations de similarité entre les soixante-dix-sept catégories qui ont émergé pendant l'entraînement. L'avantage du dendogramme sur la formule précédente est son apparence graphique où les relations sont visuellement accessibles mais aussi son indifférence quant à la dimensionalité des vecteurs d'activation car le dendogramme est toujours en deux dimensions. (Figure 2.3.3.2.b.) Par contre, ce dernier avantage implique une perte de précision avec la

simplification des dimensionalités dans l'analyse des vecteurs. Il s'agit donc d'une mesure de similarité passe-partout mais imprécise. La deuxième enquête vient des modélisations de la reconnaissance des couleurs (ou classification) de Laasko et Cottrell.[58] Ces derniers ont déployé un arsenal de réseaux avec un nombre diversifié de neurones et avec des batteries d'entraînements différentes. Leurs résultats démontrent que lorsque confrontés à la classification des couleurs, ces divers réseaux le font d'une façon extrêmement similaire. En effet, un réseau avec quatre-vingt-seize éléments et un autre avec douze éléments ont obtenu une similarité moyenne de 0,95. (Paul M.Churchland, 1998a, 99) Leur méthode d'analyse des couches cachées (GPA pour *Gutman point alienation measure*) fait appel à une formulation de la mesure de la distance euclidienne (voir la section 2.2.3.2.). Nous la reprenons ici schématiquement (voir Paul M. Churchland, 1998a, 111 pour plus de détails).

$$\mu_{X,Y} = \frac{\sum_{r=1}^{m}\sum_{p=1}^{m}\sum_{q=1}^{m}(d_{X_r}x_p - d_{X_rX_q})(d_{Y_rY_p} - d_{Y_rY_q})}{\sum_{r=1}^{m}\sum_{p=1}^{m}\sum_{q=1}^{m}\left|d_{X_r}x_p - d_{X_rX_q}\right|\left|d_{Y_rY_p} - d_{Y_rY_q}\right|}$$

où $\mu_{X,Y}$ est la mesure de la similarité des couches cachées de deux réseaux X et Y; $_X$ et $_Y$ sont deux matrices; m est le nombre de rangée dans les deux matrices; $_{X_r}$ est le r-ième rangée de la matrice $_X$ et $d_{X_rX_p}$ est la distance entre la r-ième rangée et la p-ième rangée de la matrice $_X$. La distance euclidienne $d_{X_pX_q}$ entre deux vecteurs de n-éléments est calculée selon la mesure de la distance euclidienne exposée dans la section 2.2.3.2. L'idée est simplement d'attribuer une matrice (représentant un schéma de connectivité) à la couche cachée et ensuite de comparer les deux couches cachées pour des réseaux différents. Le résultat est une mesure objective de la

[58] Nous sommes incapables de préciser si cette recherche a été publiée car la référence de Paul M. Churchland (1998a) indique qu'elle était en processus d'examen pour une publication.

similarité de la configuration interne des deux réseaux. (Paul M. Churchland, 1998a, 110-2)[59] Car la formule permet de comparer deux familles de points dans un espace global d'activation en déterminant la distance (euclédienne) dans l'ordre de chaque famille de points des espaces d'activation propres aux deux réseaux (X et Y). Les limites de cette formulation sont donc de ne pas tenir compte de la métrique sémantique des deux réseaux contrairement à la formule utilisée dans la première enquête pour comparer les deux triangles. Churchland répond ainsi selon lui à l'un des défis de Fodor, soit de fournir un critère rigoureux d'identité conceptuelle en dépit de la « diversité neurale » de deux systèmes différents. Mais ce n'est pas tout car Fodor soutient que Churchland est confronté au problème de l'individuation des dimensions et de l'information collatérale.

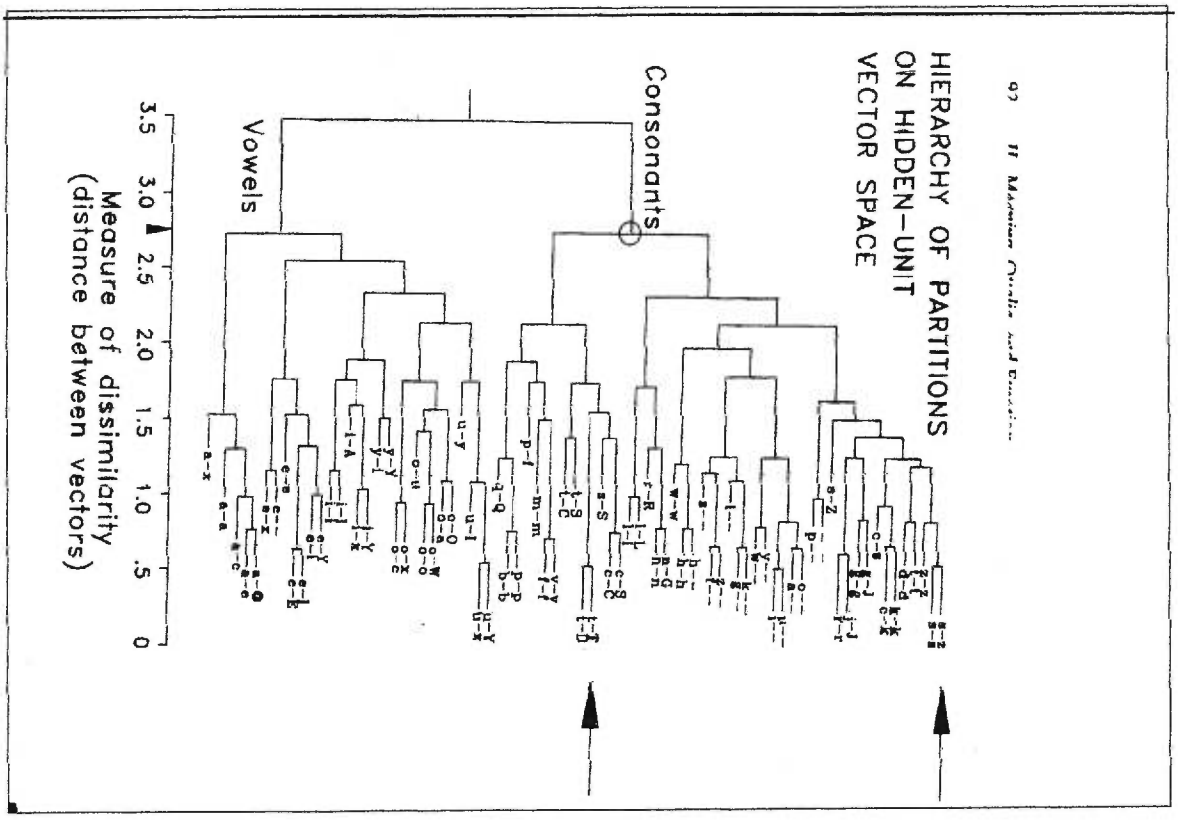

Figure 2.3.3.2.b. : L'usage d'un dendogramme pour illustrer
la similarité dans le contenu de couches cachées (Churchland, 1998b, 92)

[59] Il faut noter ici que le premier et le troisième exemple sont plus convaincants parce qu'il s'agit plutôt de « réprésentation conceptuelles », mais dans le deuxième plutôt de représentations phonétiques classés par NETtalk. Ils sont donc des exemples plus probants d'une « métrique sémantique » (Poirier et Fisette, 2000).

2.3.3.3. le problème de l'individuation des dimensions et de l'information collatérale

Nous revenons maintenant à la deuxième partie de la question introduite dans la section précédente, à savoir : en vertu de quoi peut-on déterminer l'identité conceptuelle dans la perspective de Churchland? Selon Fodor, Churchland, en définissant l'identité conceptuelle en termes de similarité sous-entend ce qu'il est censé démontrer. Car si l'on soutient qu'une région d'espace d'activation représente un degré de F*éité* (*F-ness*), alors comment parvenons-nous à établir que c'est de F*éité* et non de la G*éité* sans présupposer ce qu'est un F et un G? D'après Fodor, il ne suffit pas d'attribuer des étiquettes à certaines régions d'un espace d'activation pour régler le problème de l'individualtion du contenu conceptuel parce qu'il faut expliquer pourquoi ces dimensions sont individuées de cette façon. (Fodor et Lepore, 1992, 198-9) Si Churchland rétorque que l'individuation des dimensions se fait selon des lois psychophysiques, alors il sera aux prises avec un autre problème car, dans une théorie des concepts, les concepts doivent être individués d'après leur contenu et non d'après leurs causes. (Fodor et Lepore, 1992, 199-200) Car au niveau des couleurs, par exemple, le cerveau se représente les choses comme étant rouge et non comme ayant certaines propriétés psychophysiques. (Fodor et Lepore, 1992, 201) Il y aurait donc une confusion dans les prétentions de Churchland à offrir un critère de similarité conceptuelle car d'un côté son entreprise est interprétée dans un cadre psychophysique selon la nature physique des stimuli mais d'un autre elle est interprétée comme une entreprise sémantique dans laquelle le contenu permet l'individuation. (Fodor et Lepore, 1992, 201) Le problème de l'identité conceptuelle rejoint donc le problème de l'individuation des dimensions. De plus, il y a le problème corrolaire de l'information collatérale.

Les prototypes reposent sur certaines intuitions pour leur caractérisation sémantique. Car si la relation de similarité est définie par un partage de caractéristiques, alors il faut établir quelles

caractéristiques sont les plus importantes pour définir un concept donné.[60] (Fodor, 1998, 91)

Il nous faudrait une théorie des caractéristiques (*theory of features*) pour éclairer quels sont éléments principaux définissant un concept donné. Comment déterminer quelle information est pertinente afin d'identifier le concept d'un individu à celui d'un autre? Nous risquons de nous confronter à une masse d'information collatérale importante. Pourtant tout ce qu'un individu connaît sur les chiens ne doit pas nécessairement être connu par un autre pour qu'il partage le même concept. Cette critique peut être éclairée par l'exemple suivant tiré de Fodor (1998, 103 et suivantes). Admettons que nous avons un prototype de pomme. Si l'on se voit présenté une pomme mauve, alors l'importance de la couleur (à titre de caractéristique) sera augmentée pour son traitement et pour définir ce qu'elle est. Dans ce cas une caractéristique se voit accordée plus d'importance dans cette situation étrange que dans l'usage normal. En outre, la banalité de certaines autres caractéristiques fait qu'elles sont peu informatives pour définir une chose. Par exemple, la masse d'une pomme est une caractéristique plus ou moins négligeable mais elle pourrait être très importante pour définir un diamant. Il y a donc ici une sorte de relativité des caractéristiques et de leur importance pour la classification, d'où une difficulté pour la théorie des prototypes.

Dans un réseau de neurones, c'est le vecteur de caractéristiques qui se charge de regrouper et de comptabiliser la présence de certaines caractéristiques. Néanmoins, il reste à se demander comment on arrive à fixer les paramètres qui seront présents sur le vecteur. Évidemment c'est l'expérimentateur qui le fait. Mais comment peut-on arriver à baliser cette première démarche de sélection de façon à ne pas biaiser la classification? C'est le sens du problème soulevé par Fodor. Une réponse connexionniste provisoire serait de soutenir qu'il y a des réseaux qui s'occupent de déterminer l'importance (*salience*) de certains éléments des objets d'un point de vue descendant. Mais il s'agit d'une chose à démontrer.

---

[60] Voir aussi Smith (1990) pour des remarques semblables et pour un exposé d'une tentative de rendre compte de

2.3.4.4. le reproche d'empirisme naïf

Nous avons vu que Fodor reproche à Churchland de réduire tous les concepts à des sensations (Fodor et Lepore, 1992, 191-3; 1993, 679-80) et la combinaison des concepts à des associations statistiques (Fodor et Lepore, 1992, 205; 1993, 680-1), deux tendances perversement huméennes. Cet « empirisme huméen sophistiqué » d'après Fodor est en proie aux arguments de Berkeley selon lesquels les concepts abstraits ne proviennent pas (exclusivement) de l'expérience. Par contre, Churchland évoque un peu implicitement pour se défendre l'hypothèse de différents niveaux d'abstraction selon laquelle les concepts abstraits seraient des opérateurs sur d'autres réseaux de neurones ou bien des vecteurs d'activation à plus haute dimensionalité situés plus haut dans la hiérarchie des concepts. (Churchland, 1998b, 110; voir la section 2.3.2.1. pour le souci internaliste de Churchland) Cela signifierait qu'il y aurait des concepts plus primitifs (plus sensoriels) sur lesquels d'autres concepts (plus abstraits) pourraient venir agir. Par exemple, on pourrait avoir le concept d'un nombre sur lequel viendrait agir le concept d'addition. Cependant, il faut le noter, cet argument demeure très spéculatif, mais a une plausibilité dans une perspective computationnelle connexionniste.

2.3.4.5. le reproche de holisme et de sémantique du rôle inférentiel

Fodor s'insurge contre toute forme de holisme et de pragmatisme selon lesquels le rôle inférentiel d'un concept déterminerait son contenu. (Fodor, 1998 et 1993; Fodor et Lepore, 1992; voir aussi la section 1.2.7.) Sans reprendre toute sa critique, il est important ici de noter les éléments pertinents pour la problématique des concepts. Premièrement, une sémantique du rôle inférentiel (SRI)[61] conduit à une forme de holisme insoutenable. (Fodor, 1998, 13) Pour Fodor, une théorie atomiste des concepts (*atomistic theory of concepts*) selon laquelle les éléments constitutifs (atomes) transportent de l'information permet de valider le cadre de la

---

la « *salience* » de certaines caractéristiques des concepts.

[61] Inferential role semantics (IRS).

1998, 13) Pour Fodor, une théorie atomiste des concepts (*atomistic theory of concepts*) selon laquelle les éléments constitutifs (atomes) transportent de l'information permet de valider le cadre de la psychologie intentionnelle. Par contre, les SRI invalident et discréditent la psychologie intentionnelle. De plus, les SRI soutiennent que le contenu d'un concept est toujours déterminé par ses relations avec d'autres concepts. (Il faut plus d'un concept pour former une inférence.) Il s'ensuit que pour qu'un concept ait un contenu, d'autres doivent en avoir. Par contre, la sémantique informationnelle de Fodor selon laquelle un symbole obtient son contenu par ses relations avec le monde (Fodor, 1998, 14) et aucunement par des relations inférentielles constitutives (Fodor, 1998, 108) permet une défense du contenu individuel et est donc hautement compatible avec le principe de compositionalité. Cette approche respecte aussi les exigences d'une théorie computationnelle inspirée de Turing où les computations sont définies en termes d'opérations sur les symboles.[62] Par contre, le connexionnisme, selon Fodor, adhère à une SRI. Les théories prototypiques définissent le contenu d'un concept par ses relations statistiques à d'autres concepts. Par exemple, que les concepts « bicylette » et « auto » soient subsumés par le concept véhicule est une question statistique. En outre, l'importance des inférences qui accordent le contenu est mesurée statistiquement. (Fodor, 1998, 92) Il s'ensuit que la théorie prototypique des concepts adhère à une forme de SRI et donc à une forme de holisme et de pragmatisme insoutenables. (Fodor, 1998, 13; voir aussi la section 1.2.7.)

[62] Fodor évite ici un cercle vicieux que nous avons déjà noté, soit de définir la computation (incluant les inférences) par la manipulation de symboles et de définir à leur tour les symboles par leur rôle inférentiel, c'est-à-dire une manipulation de symbole. (Fodor, 1998, 13)

2.3.3.6. le reproche de fausse adéquation entre les prototypes et les concepts

Selon Fodor, les concepts sont compositionnels mais les prototypes ne le sont pas. (Fodor, 1998; Fodor, 1995; Fodor et Lepore, 1993; Fodor et Pylyshyn, 1988). Ce qui déplaît à Fodor dans ce tableau, c'est que les prototypes ne permettent pas de rendre compte du caractère compositionnel des concepts et donc de la systématicité et de la productivité. (Fodor 1998, 1995, 1993, 1992, 1991, 1988 et 1987) Or nous avons vu dans le chapitre précédent (section 1.3.) que ces caractéristiques sont essentielles pour Fodor. Par conséquent, les prototypes ne peuvent pas rendre compte pleinement de la conceptualité. Or, Churchland tente d'identifier les modèles connexionnistes des concepts aux modèles prototypiques. Il s'ensuit qu'il ne peut pas rendre compte de la compositionalité des concepts. Regardons de plus près l'argument contre les prototypes.

L'argument principal contre les prototypes est premièrement « qu'ils ne composent pas » parce qu'ils ne rendent pas compte de l'identité conceptuelle.[63] (Fodor, 1998, 94) Deuxièmement, les prototypes ne sont pas toujours compositionnels parce que la classe des prototypes ne recoupe pas parfaitement la classe des concepts. En d'autres mots, il y a des concepts (compositionnels) qui ne sont pas des prototypes. Prenons par exemple « le problème du non-chat » (*the uncat problem*). Le concept « non chat »[64] (*not a cat*) est parfaitement intelligibe dans la mesure où il exprime une propriété que nous pouvons parfaitement attribuer. Cependant, ce concept n'a pas de prototype. Qu'est-ce qu'un non-chat? Cela pourrrait être une tranche de pain, un chien, etc., mais quel en serait l'exemplaire prototypique? On ne pourrait pas dire que plus une chose est non chat, alors plus elle est une tranche de pain ou un chien. Selon Fodor, les concepts entendus comme

[63] Voir les sections 1.2.4. et 2.3.4.2. au sujet du lien entre la compositionalité et l'identité conceptuelle.
[64] C'est-à-dire le prédicat « ne pas être un chat » (*'is not a cat'*).

prototypes ne peuvent pas rendre compte de cette situation. Il faut plutôt faire appel à la fonction booléenne ¬(chat) et cela implique l'abandon des prototypes pour rendre compte de la productivité et de la systématicité des prédicats booléens complexes. (Fodor, 1998, 102-3)

Il y a aussi *le Pet Fish Problem* qui illustre comment les prototypes ont de la difficulté à « composer ». Un poisson rouge est un piètre exemple d'un poisson ou d'un animal domestique. Par contre, c'est un excellent exemple d'un poisson-animal domestique (*pet fish*). (Fodor, 1998, 102) Par conséquent, dans un cadre prototypique, connaître les éléments constitutifs de poisson-animal domestique (c'est-à-dire le prototype poisson et le prototype animal domestique) ne permet pas de saisir le contenu du prototype poisson-domestique. Ainsi, comme l'illustre ces deux exemples, les concepts ne sont pas des prototypes parce que les prototypes ne rendent pas compte de la compositionalité. Car le principe de compositionalité exige que le sens des expressions complexes provienne de ses éléments constitutifs, ce que l'exemple précédent infirme. Il s'agit selon Fodor (1998, 106) d'une question de principe et les prototypes ne pourront jamais être des concepts.

2.3.4. Conclusion : qui a raison?

Nous avons exposé dans ce deuxième chapitre certains éléments théoriques fondamentaux du connexionnisme afin de préparer un exposé plus pointu sur la reconnaissance de schémas, une fonction essentielle des réseaux de neurones et particulièrement importante pour comprendre les capacités conceptuelles des réseaux. L'interprétation de Churchland nous a conduit à la question des concepts. Churchland soutient qu'un concept est une région de classification dans un espace d'activation. Sur ce point, Fodor et Churchland entretiennent un désaccord fondamental qui se reflètent sur plusieurs points centraux tels

que l'identité conceptuelle, le caractère dynamique et pragmatique de la conceptualité, etc. Il reste maintenant à trouver une manière de clarifier le débat, ce que le troisième et dernier chapitre tentera.

Chapitre 3 : Les concepts, symboles ou connexions?

Nous avons présenté dans les deux chapitres précédents deux positions sur les concepts. D'un côté, les concepts sont des symboles, des éléments constitutifs de la pensée. D'un autre côté, les concepts sont des régions de classification dans des espaces d'activation. Notre question est maintenant de déterminer si les concepts sont mieux rendus en termes de symboles ou de connexions. Trois pistes peuvent être explorées pour préciser le rapport entre ces deux perspectives (reprises ici sous forme d'affirmation générale) : (1) les concepts sont des symboles; (2) les concepts sont des connexions et (3) les concepts revêtent des caractéristiques permettant de les rattacher à la fois au cognitivisme et au connexionnisme. Nous optons pour la troisième option et nous nous rattachons à certains auteurs (Smolensky et Clark) qui tentent de faire valoir la complémentarité des deux perspectives au niveau de la compréhension des concepts.

Notre stratégie dans ce troisième chapitre est la suivante. La première section récapitule, schématise et compare brièvement les deux positions quant aux concepts. Ce premier exercice permet de dégager une grille de base à partir de laquelle nous analysons le débat. Dans une deuxième section, nous introduisons la distinction de Smolensky entre les symboles et les sous-symboles et la distinction de Clark entre les sciences cognitives descriptives et les sciences cognitives causales. Ensuite, la typologie de Ramsey présentant quatre types de représentations mentales (dans les modèles connexionnistes) est exposée. Ces trois éléments de clarification seront analysés et évalués successivement. Une troisième section tente ensuite de reprendre ces éléments de clarification dans un cadre plus global et systématique. Nous arrivons alors à dégager quatre dimensions d'une explication (fonctionnelle-causale / fonctionnelle-descriptif / relationnelle-causal / relationnelle-descriptive) pouvant être appliquées à chaque niveau d'analyse. Enfin, une quatrième

section a comme objectif la clarification du débat entre le cognitivisme et le connexionnisme sur la question des concepts à l'aide de ce cadre global. Le fil conducteur de cette analyse est que le cognitivisme et le connexionnisme deviennent complémentaires sur la question des concepts suite à une double clarification, à savoir : le niveau d'analyse auquel on a affaire et la dimension d'explication que nous utilisons. Le résultat est une complémentarité nuancée entre le cognitivisme et le connexionnisme sur la question des concepts.[1] . Cela signifie en fait qu'aucune des positions n'a droit à une prétention d'exclusivité au sujet de la conceptualité. Cependant, pour arriver à cette fin nous devons réinterpréter les prétentions de Fodor et de Churchland. Car en dernière analyse la modélisation des capacités conceptuelles est une entreprise plus pragmatique que ces deux interprétations le laissent entendre.

3.1. Analyse des points de divergence entre le cognitivisme classique et le connexionnisme sur les concepts

Cette première section tente de clarifier les points de divergence entre le cognitivisme classique et le connexionnisme sur les concepts. Une comparaison est introduite sous la forme d'un tableau suivi d'une brève analyse qui tente de dégager les forces et les faiblesses respectives de chacune des approches.

3.1.1. Comparaison entre le cognitivisme classique et le connexionnisme

Le tableau 3.1.1.a. présente une série de propositions et de caractéristiques attribuables aux deux approches.[2] Il récapitule les éléments présentés antérieurement pour chaque approche ainsi que les critiques respectives qui leur sont adressées dans les deux chapitres

[1] Il ne s'agit ici que d'une tentative de clarifier le débat et non une prétention à offrir le dernier mot.
[2] Bien sûr, l'attribution de certaines caractéristiques fait l'objet d'un débat. Nous devons rester à un niveau plus superficiel pour caractériser moindrement les deux approches.

précédents. La comparaison est effectuée dans les deux prochaines sections (3.1.2. et 3.1.3.).

Les concepts selon le cognitivisme classique	**Les concepts selon le connexionnisme**
les concepts sont compositionnels au sens fort du terme et sont systématiques	les concepts sont compositionnels mais seulement dans un sens faible (leurs sous-éléments le sont) et sont faiblement systématiques
les concepts sont productifs à cause de leur nature compositionnelle	les concepts sont productifs à cause de leurs relations causales et à cause du fait qu'ils peuvent servir d'entrée à d'autres réseaux de neurones
les concepts sont des états mentaux particuliers bien définis pouvant servir comme effets ou causes dans la chaîne des stimuli et des réponses	les concepts sont des états mentaux plus ou moins bien définis selon leur nature prototypique et insérés dans un continuum avec la sensation et pouvant servir comme effets ou causes dans la chaîne des stimuli et des réponses
les concepts sont des catégories sous lesquelles « tombent » les objets du monde	les concepts sont des catégories dont le rattachement au monde est codé sous forme de vecteurs de caractéristiques
les concepts sont publics et doivent souscrire à un critère rigoureux d'identité conceptuelle	les concepts sont publics mais ils souscrivent à un critère de similarité conceptuelle
les concepts sont des atomes (états mentaux particuliers) statiques, passifs et rigides	les concepts sont des rassemblements plus ou moins bien définis de sous-caractéristiques ayant des caractéristiques dynamiques, actives et souples
les concepts sont rigides et font partie de systèmes ayant une certaine fragilité, par conséquent, les concepts subissent une descente abrupte au niveau de leur application	les concepts sont souples et font partie de systèmes robustes, par conséquent, les concepts subissent une descente graduelle au niveau de leur application
les concepts ont un contenu représentationnel non distribué	les concepts ont un contenu représentationnel distribué
il y a une interprétation sémantique directe des concepts et les concepts ont donc un contenu insensible à l'environnement externe	l'interprétation sémantique des concepts pose problème dans les réseaux distribués (non locaux) et les concepts ont donc un contenu sensible à l'environnement externe

l'étude des concepts fait partie d'une explication située à un niveau relativement autonome (autonomie de la psychologie)	l'étude des concepts fait partie d'une explication en relation d'interdépendance avec d'autres niveaux d'analyse
la catégorisation est discrète, avec une désignation rigide car les concepts sont « donnés »	la catégorisation est l'évolution continue d'un système dynamique (apprentissage de solutions dans un modèle de satisfaction de contraintes) car les concepts émergent
il y a un rôle passif attribué au langage dans l'expression de la conceptualité	il y a un rôle synergétique attribué au langage dans l'expression de la conceptualité
cette position offre une explication conceptuelle des relations sémantiques entre les concepts puisque la classification conceptuelle se fait d'après le contenu sémantique informationnel	cette position offre une explication causale des relations sémantiques entre concepts (proximité dans un espace d'activation) puisque la classification conceptuelle se fait d'après la similarité (métrique sémantique)
cette position n'offre pas d'explication pour la généralisation spontanée	cette position offre une explication vectorielle de la généralisation spontanée selon la reconnaissance de schéma

Tableau 3.1.1a. : Comparaison des deux positions sur les concepts

3.1.2. Les forces et les faiblesses du cognitivisme classique

Selon le tableau 3.1.1.a., le cognitivisme est particulièrement apte à rendre compte de la compositionalité, de la productivité et de la systématicité de la conceptualité. Il offre un critère rigoureux d'identité conceptuelle, il tient compte des interactions entre les concepts selon leur contenu sémantique et il offre aussi une interprétation sémantique des concepts. Par contre, ces caractéristiques conduisent à une fragilité et à une rigidité excessives (insensibilité au contexte), à une dégradation abrupte, à une représentation passive et non distribuée où la conceptualité doit être explicite afin d'être un point de départ des computations (satisfaction de contraintes dures). En outre, l'accent mis sur la dimension

sémantique amène la thèse de l'autonomie relative de la psychologie et une certaine indifférence au niveau de l'origine et des interactions causales des concepts[3].

3.1.3. Les forces et les faiblesses du connexionnisme

De son côté, le connexionnisme rend compte de l'emprise empirique et causale des concepts. Il offre une explication de la généralisation spontanée car les concepts sont implicites, souples, dynamiques, statistiques, émergents, distribués et robustes. Les concepts sont classés par leur similarité et sont en évolution continue avec la sensation. Ce cadre général fait que l'explication des concepts doit se faire dans un contexte d'interdépendance entre les différents niveaux d'analyse. Par contre, les relations sémantiques entre les concepts et leur interprétation sémantique ainsi que les aspects compositionnel, productif et systématique des concepts sont moins bien rendus. Maintenant que nous avons établi une comparaison entre les deux approches, comment devons-nous articuler leurs différences afin de clarifier ce qu'est un concept? La prochaine section aborde cette question en présentant trois éléments de clarification.

3.2. Analyse et clarification de l'opposition sur la question des concepts

Déterminer si les concepts sont des symboles ou des connexions implique une clarification de la relation entre ces positions sur les concepts. Cette deuxième section présente trois éléments allant dans ce sens. Premièrement, Paul Smolensky (1988, 1992) considère que les concepts dans la perspective classique sont des « approximations » des concepts tels que présentés par le connexionnisme. Dans le premier cas, ils sont des symboles, dans le deuxième des sous-symboles. Deuxièmement, Andy Clark (1989) estime de son côté que

[3] Si l'on définit les concepts comme des symboles du niveau sémantique autonome par rapport aux autres niveaux d'analyse, alors on est en proie au problème de la causalité mentale (mental causation). (Kim, 1997,

les symboles du cognitivisme ne sont pas de simples approximations et que nous devons les prendre littéralement. Il présente une distinction entre les sciences cognitives descriptives et les sciences cognitives causales pour légitimer les deux approches en jeu. Troisièmement, Ramsey (1992, 1996) introduit une typologie définissant quatre types de représentation mentale présents dans les réseaux de neurones. Chacun de ces éléments de clarification sera suivi de commentaires visant à nuancer leur valeur et à préparer leur articulation.

3.2.1. L'interprétation de Smolensky

Smolensky tente de trouver une interprétation appropriée de la relation entre le cognitivisme classique et le connexionnisme. Il fait partie de ceux qui croient à une certaine complémentarité des deux perspectives. (1988, 22) Cette affirmation est expliquée en référence au « paradoxe de la cognition ».

3.2.1.1 le paradoxe de la cognition

Smolensky (1991b) soutient qu'il y a un paradoxe dans la modélisation de l'esprit. D'une part, la cognition est rigide (*hard*) car elle fonctionne avec les règles de la logique et du langage. Par contre, elle est souple (*soft*) car les sytèmes trop rigides ont de la difficulté à coopérer avec les situations réelles et ils ont tendance à flancher. Les deux dimensions de la cognition nous amènent à deux types de représentation : d'une part aux représentations symboliques et d'autre part à des représentations statistiques sous-symboliques. (Smolensky, 1991b) Selon que l'on met l'accent sur l'un ou l'autre de ces deux aspects de la cognition, on sera attiré respectivement par le cognitivisme classique ou le connexionnisme. Une manière de clarifier le débat dans cette optique est donc de

---

125-54) L'enjeu semble conduire à l'acceptation d'une forme de dualisme.

considérer que le cognitivisme traite des symboles et le connexionnisme des sous-symboles.

3.2.1.2. les symboles et les sous-symboles

Les symboles dans le cadre du cognitivisme sont littéralement les éléments que nous utilisons dans notre langage courant (par ex. : chien, table, etc.). Ils reflètent la psychologie du sens commun. Ils sont « sémantiquement pénétrables » ce qui signifie qu'ils ont une interprétation naturelle et évidente. Un programme de recherche propre au cognitivisme et qui découle de cette position est de « trouver » comment le système nerveux implémente les symboles et leur manipulation. La perspective du connexionnisme s'écarte sensiblement de ce tableau. Premièrement, on ne parle pas de symboles dans le connexionnisme mais de sous-symboles se situant à un niveau d'analyse inférieur. (Smolensky, 1988, 3) Les concepts sont alors des éléments d'une représentation distribuée et sont « sémantiquement impénétrables ».[4] (Smolensky, 1992, 83-5) Ce sont plutôt les schémas d'activation qui peuvent être interprétés comme représentant quelque chose. Un programme de recherche important pour le connexionnisme est donc de tenter de combler le fossé entre les données neuronales et les niveaux supérieurs en mettant à jour les éléments constitutifs de la cognition humaine, soit les sous-symboles. (Hanson et Olson, 1990) Quelle est précisément la relation entre les symboles et les sous-symboles? Entre le cognitivisme et le connexionnisme?

3.2.1.3. les modèles symboliques comme approximation conditionnelle

L'hypothèse générale de Smolensky est que les descriptions de niveau supérieur du cognitivisme sont des approximations des niveaux inférieurs offerts par les réseaux

connexionnistes. (Smolensky, 1992 et 1988) En d'autres termes, les concepts dans les modèles classiques sont des approximations des concepts dans les modèles connexionnistes. Cela vient du fait que les modèles connexionnistes sont « plus près » du niveau neuronal bien qu'ils soient, toute proportion gardée, plus près du niveau symbolique que du niveau neuronal. (Smolensky, 1998, 8) Cependant, Smolensky (1988) mais non Smolensky (1992) nuance cette hypothèse générale. Il distingue deux niveaux de modélisation, soit le niveau symbolique et le niveau sous-symbolique et deux types de tâches cognitives à modéliser, soit les tâches fondées sur des règles conscientes et les tâches intuitives. Si nous avons affaire à une tâche fondée sur des règles, alors le niveau symbolique sera une description appropriée de la tâche alors que si nous avons affaire à une tâche intuitive, le niveau symbolique ne sera qu'une approximation (et le niveau sous-symbolique sera plus approprié). Le caractère approximatif du niveau symbolique est donc conditionnel au type de tâche que l'on tente de modéliser. Le tableau 3.2.1.3.a récapitule ce raisonnement. On y voit clairement que Smolensky croit que les réseaux de neurones se rapprochent du niveau neuronal et reflètent adéquatement le niveau sous-symbolique tandis que le niveau symbolique est une bonne approximation des tâches impliquant des règles conscientes et une moins bonne approximation des tâches intuitives. D'après Smolensky, le cognitivisme est au connexionnisme ce que la théorie newtonienne est à la théorie quantique. Dans certains cas les premières fourniront des approximations commodes et foncièrement justes mais lorsqu'il faudra plus de précision, il faudra faire appel aux deuxièmes.

---

[4] Exclusion faite des réseaux à représentation locale. (Smolensky, 1992, 86) Voir la distinction entre réseau à repsentation locale et réseau à représentation distribuée à la section 2.1.2.1.

Niveau	Processus modélisé	Système cognitif : le cerveau	Système cognitif sous-symbolique	Système cognitif symbolique
symbolique	intuitif	?	approximation grossière	(+/-) exact
symbolique	règle consciente	?	bonne approximation	un peu plus exact
sous-symbolique	intuitif et règle consciente	bonne approximation	(+/-) exacte	---
neural	---	exact	---	---

Tableau 3.2.1.3.a : Les niveaux d'analyse et leurs relations avec certains processus et avec certains systèmes cognitifs (Smolensky, 1988, 11)

3.2.1.4. quelques réserves quant à l'interprétation de Smolensky.

On pourrait formuler quelques réserves quant à l'interprétation de Smolensky, notamment à l'égard de son hypothèse que la distinction entre les symboles et les sous-symboles permet de départager les applications appropriées du cognitivisme et du connexionnisme respectivement. Premièrement, Smolensky porte assez peu attention aux types de réseaux et à leurs types correspondants de représentation mentale. (Voir section 3.2.3. avec la typologie de Ramsey) Cela l'amène à supposer que tous les réseaux connexionnistes (à représentation distribuée) modélisant des fonctions cognitives s'adressent à un niveau d'analyse sous-symbolique. Or il n'est vraiment pas clair que l'on puisse départager les choses aussi nettement. Il y a des modèles connexionnistes de fonctions cognitives qui ne font pas usage de sous-symboles comme c'est le cas du modèle propositionnel évoqué par Ramsey (1992, 263-4). Par conséquent, le tableau de Smolensky est incomplet et demande à être rectifié étant donné qu'il ne tient pas compte du nombre de modèles possibles et de la complexité de leur interprétation. Deuxièmement, il n'est pas évident que le niveau symbolique soit une simple approximation du niveau sous-symbolique. La perspective sous-symbolique offre bien sûr des indices sur la façon dont la manipulation de symboles

Niveau	Processus modélisé	Système cognitif : le cerveau	Système cognitif sous-symbolique	Système cognitif symbolique
symbolique	intuitif	?	approximation grossière	(+/-) exact
symbolique	règle consciente	?	bonne approximation	un peu plus exact
sous-symbolique	intuitif et règle consciente	bonne approximation	(+/-) exacte	---
neural	---	exact	---	---

Tableau 3.2.1.4.a : Les niveaux d'analyse et leurs relations avec certains processus et avec certains systèmes cognitifs (Smolensky, 1988, 11)

3.2.1.4. quelques réserves quant à l'interprétation de Smolensky.

On pourrait formuler quelques réserves quant à l'interprétation de Smolensky, notamment à l'égard de son hypothèse que la distinction entre les symboles et les sous-symboles permet de départager les applications appropriées du cognitivisme et du connexionnisme respectivement. Premièrement, Smolensky porte assez peu attention aux types de réseaux et à leurs types correspondants de représentation mentale. (Voir section 3.2.3. avec la typologie de Ramsey) Cela l'amène à supposer que tous les réseaux connexionnistes (à représentation distribuée) modélisant des fonctions cognitives s'adressent à un niveau d'analyse sous-symbolique. Or il n'est vraiment pas clair que l'on puisse départager les choses aussi nettement. Il y a des modèles connexionnistes de fonctions cognitives qui ne font pas usage de sous-symboles comme c'est le cas du modèle propositionnel évoqué par Ramsey (1992, 263-4). Par conséquent, le tableau de Smolensky est incomplet et demande à être rectifié étant donné qu'il ne tient pas compte du nombre de modèles possibles et de la complexité de leur interprétation. Deuxièmement, il n'est pas évident que le niveau symbolique soit une simple approximation du niveau sous-symbolique. La perspective sous-symbolique offre bien sûr des indices sur la façon dont la manipulation de symboles

se fait. Par contre, même si l'on a une description adéquate du niveau sous-symbolique, est-ce que cela invalide la description au niveau symbolique? Quel est le rôle exact des symboles? Sont-ils de simples approximations ou des entités ayant une utilité propre? Nous verrons que Clark offre une réponse qui remet en question l'interprétation à tendance éliminativiste de Smolensky en proposant une relation différente entre ces deux niveaux d'analyse.

3.2.2. L'interprétation de Clark

Clark (1989) tente de concilier le cognitivisme et le connexionnisme mais d'une façon légèrement différente de Smolensky. Mettant moins l'accent sur une relation dynamique entre les deux perspectives, il confère de prime abord une certaine légitimité à deux types d'entreprise distincts et légitimes en sciences cognitives : les sciences cognitives descriptives et les sciences cognitives causales. De cette façon, il veut éviter toute supposition d'uniformité (*uniformity assumption*) où seulement l'une des perspectives aurait de la valeur. (Clark, 1989, 128-9)

3.2.2.1. les sciences cognitives descriptives et les sciences cognitives causales

Les sciences descriptives ont pour but d'offrir une théorie formelle ou un modèle de la structure abstraite du domaine de la pensée en utilisant le programme informatisé comme outil ou médium. (Clark, 1989, 153) Les sciences cognitives causales, quant à elles, tentent d'expliquer les causes computationnelles internes des comportements intelligents que nous désignons comme des états mentaux. (Clark, 1989, 154) Clark donne l'exemple de travaux en sciences cognitives traitant de la « physique naïve »[5] (*naive physics*) pour illustrer cette distinction. Du côté descriptif, on aurait une mise à jour de la structure de l'ensemble des

connaissances requises pour fonctionner dans le monde. Tandis que du côté causal, on aurait la mise à jour d'un mécanisme cérébral permettant à un individu de fonctionner dans le monde. De même, au niveau de l'étude de la pensée, on aurait d'une part l'hypothèse du langage de la pensée et d'autre part les explications causales connexionnistes des mécanismes de la pensée. Les deux entreprises sont différentes mais légitimes. Par conséquent, lorsque Fodor défend une forme de réalisme intentionnel, il ouvre la boîte de Pandore car si des attitudes propositionnelles ne sont pas trouvées dans le cerveau, alors son entreprise risque de basculer. Il s'expose à une forme d'éliminativisme à cause de ses prétentions à décrire ce qui se passe réellement dans le cerveau. Mieux vaudrait alors adopter, selon Clark, une prétention descriptiviste plutôt que réaliste (au sens causal) de sorte que l'entreprise de Fodor devienne compatible (et non en concurrence) avec les explications causales connexionnistes de la pensée. Par conséquent, selon Clark, il ne faudrait pas confondre de façon générale la description d'un phénomène de la pensée et son explication causale. Ces démarches sont relatives à l'adoption de deux stratégies : le *Mind'eye view* et le *Brain's eye view*. Grossièrement, du côté de l'esprit, on décrit les choses tandis que du côté du cerveau, on les explique. Clark invite donc à une exploration des combinaisons possibles entre les deux types de modélisation en tenant compte de la spécificité de leurs objectifs.

3.2.2.2. la validité intrinsèque de l'approche symbolique

L'interprétation de Clark légitime solidement les deux perspectives de façon plus intégrale que Smolensky. Un point majeur de désaccord est donc la question du caractère approximatif des modèles cognitivistes. (Clark, 1989, 187) Comme nous l'avons vu, pour Smolensky (1992 et 1988) le niveau symbolique est une description utile et plus ou moins

---

[5] C'est l'étude de nos croyances intuitives sur les phénomènes physiques nous permettant de fonctionner dans

exacte de ce qui se passe au niveau sous-symbolique. Cependant, Clark défend l'intégrité du niveau symbolique parce qu'un peu à la manière de Dennett (1987 et 1981), il affirme que certaines choses seraient « invisibles » sans recours à ce niveau de description. (Clark, 1989, 197) Le niveau abstrait dans ce cas permet de regrouper des phénomènes qui ne pourraient l'être autrement. (Clark, 1989, 200) Le langage de la pensée, par exemple, peut servir à regrouper des systèmes dont le comportement et la vie mentale seraient autrement fortement disparates. (Clark, 1989, 201) Par conséquent, loin d'être de simples approximations, les descriptions en termes de symboles peuvent être exactes et puissantes puisqu'elles génèrent des regroupements d'explication. (Clark, 1989, 202) Clark suggère même que dans l'analyse des agglutinations (*cluster analysis*), comme par exemple celle de Paul M. Churchland, les étiquettes conceptuelles attribuées aux représentations font intervenir une description d'ordre supérieur qui n'est pas disponible du simple point de vue connexionniste. (Clark, 1989, 203)

3.2.2.4. quelques réserves quant à l'interprétation de Clark

Comme nous l'avons vu, Clark offre une interprétation plus généreuse de l'approche symbolique car il ne la soumet pas aux pressions de l'éliminativisme (contrairement à Smolensky). Sa solution consiste à départager deux types d'entreprise en sciences cognitives : la description et l'explication causale. Nous sommes en principe d'accord[6] avec un tel schéma mais nous aimerions formuler deux commentaires visant à le complexifier légèrement. Premièrement, il n'est pas évident qu'il faille toujours accorder une validité intrinsèque au niveau symbolique. Dans certains cas le niveau symbolique

---

la vie de tous les jours. (Clark, 1989, 157)

[6] La section 3.3.2. présente une formulation plus nuancée de la distinction entre les sciences cognitives descriptives et les sciences cognitives causales.

pourrait être réduit/reconstruit sans que rien ne soit « éliminé ».[7] Il s'agirait alors d'une explication connexionniste d'une propriété symbolique sans réduction. Par conséquent, l'explication causale d'un niveau symbolique n'implique pas nécessairement un éliminativisme réducteur car il y a la possibilité d'une reconstruction qui prendrait en compte tous les éléments du niveau supérieur à partir d'un niveau d'explication inférieur. Cela voudrait dire que le niveau supérieur ne ferait pas toujours « voir » des éléments que l'on ne peut voir d'un niveau inférieur. Deuxièmement, il n'est pas clair que la dichotomie entre la description et l'explication causale est bien utile telle que formulée par Clark. Peut-on véritablement départager ces deux entreprises clairement? Comment décrire quelque chose scientifiquement sans faire appel à certaines suppositions quant à l'explication causale du phénomène? Une description sans assises causales solides est-elle une véritable description? Les hypothèses qui guident la description de la pensée telle l'hypothèse du langage de la pensée présupposent-elles certaines choses au sujet du fonctionnement (l'explication causale) de la pensée? Cette question est importante pour les sciences cognitives parce qu'elle fait intervenir la question du lien entre les niveaux d'analyse. Une solution possible est d'attribuer à la description supérieure un rôle heuristique comme Dennett le fait. (Dennett, 1987) Nous développerons ce point plus loin dans la section 3.3.6. mais l'idée est de répéter la distinction descriptif/causal à l'intérieur de chacun des niveaux d'analyse. De cette façon, la distinction générale descriptif/causal est introduite dans des cadres d'explication précis. Mais avant de passer à cette étape plus constructive, nous devons préciser les diverses manifestations de la représentation mentale dans les réseaux connexionnistes telles que répertoriées par Ramsey.

---

[7] Il est à noter que Clark (1993) semble admettre cette possibilité plus ouvertement.

3.2.3. L'interprétation de Ramsey

Ramsey tente de clarifier quels types de représentation les réseaux connexionnistes peuvent instancier. Il dégage quatre types de représentation auxquels correspondent quatre types de réseaux différents. Cela tend à complexifier les applications possibles des modèles connexionnistes dans la modélisation de l'esprit.

3.2.3.1. les quatre sortes de représentation dans le connexionnisme

Ramsey (1992 et 1996) offre une classification qui permet de distinguer quatre types de représentations dans les modèles connexionnistes. Le tableau 3.2.4.1.a présente la classification de Ramsey. Elle est fondée sur ce qui est représenté par le réseau au niveau des schémas d'activation des unités cachées (concept ou proposition) et au niveau des unités individuelles (micro-caractéristiques ou rien).

Type de représentation	**Contenu du schéma d'activation dans les unités cachées**	**Contenu des unités individuelles**
Type 1	concept	micro-caractéristiques
Type 2	concept	rien (aucune interprétation)
Type 3	proposition	rien (aucune interprétation)
Type 4	proposition (séries de schémas d'activation)	rien (aucune interprétation)

Tableau 3.2.4.1.a : La typologie des représentations mentales dans les réseaux connexionnistes selon Ramsey (1992)

Le premier type de représentation est obtenu lorsque les unités individuelles représentent des micro-caractéristiques, alors que des concepts sont contenus dans les schémas d'activation. Un exemple de ce type de représentation est le modèle de Rumelhart, Smolensky, McClelland et Hinton (1986)[8] où les unités représentent différents éléments de

[8] Voir section 2.2.7. et la figure 2.2.7.b. à cet effet.

cinq types de chambres tandis que le contenu du schéma d'activation est la représentation d'une chambre donnée. Les unités ont des liens excitatoires et inhibitoires déterminés qui font que certaines caractéristiques (par ex. : avoir un lavabo) entretiennent des relations d'inhibition avec certaines autres caractéristiques (par ex. : avoir un foyer). Par conséquent, le réseau est capable de classer les entrées selon des « chambres prototypiques » et peut aussi produire des concepts hybrides à partir des éléments qu'il connaît.

Dans le deuxième type de représentation des concepts sont encore obtenus en considérant le schéma d'activation des couches cachées. Par contre, il n'y a pas d'interprétation sémantique possible pour les unités individuelles. Des méhodes d'analyse (analyse des agglutinations dans l'espace d'activation) permettent de « dégager » des représentations acquises par apprentissage (le plus souvent). C'est le type de modèle auquel Paul M. Churchland (1998b, 1995, 1993 et 1989) fait généralement référence lorsqu'il traite de la question des concepts dans un cadre connexionniste.[9]

Le troisième type de représentation implique des propositions et non des concepts comme dans les deux cas précédents. Cependant, comme dans le deuxième cas, les unités individuelles n'ont pas d'interprétation sémantique. La forme de la proposition est déterminée par le schéma d'activation. Le réseau reçoit des propositions (par ex. : Les chiens ont de la fourrure) et il les classe selon certains de leurs éléments (« Chien » et « Fourrure »). Les éléments isolés ne peuvent être reçus par le réseau.

Tout comme le deuxième et le troisième types de représentation, les unités du quatrième ne sont pas interprétables. Toutefois, à la différence du troisième, il faut dans ce cas une série

[9] Voir la section 2.3.2. à cet égard.

de schémas d'activation afin d'obtenir une proposition. Les réseaux faisant usage de ce type de représentation tentent de capturer la structure propositionnelle par de légères variations dans les schémas d'activation. Le niveau d'activité de chaque unité détermine alors le rôle d'un élément dans une structure plus grande telle qu'une proposition.[10] Ce type de réseaux rapproche le connexionnisme d'une solution pour rendre compte des principes de compositionalité, de productivité et de systématicité.

3.2.4.2. quelques réserves quant à l'interprétation de Ramsey

La typologie de Ramsey nous semble lacunaire parce qu'elle ne tient pas compte de la souplesse des réseaux et de leur possibilité d'être appliqués à différents niveaux d'étude de la cognition. Elle ignore la capacité des modèles à refléter différents niveaux de phénomènes. Par exemple, le deuxième modèle manipule des unités sémantiquement ininterprétables afin de produire des concepts. Cependant, on pourrait très bien appliquer ce modèle à un niveau neuronal où des activations sont les éléments d'activation de neurones et où la sortie finale est un schéma complexe d'activation d'un groupe neuronal. Les pouvoirs computationnels demeureraient les mêmes, mais les interprétations seraient différentes. Comme on le voit, les réseaux sont « aveugles » ou indifférents à ce qu'ils

[10] Ramsey tente de répertorier les conséquences philosophiques de chaque type de modèle connexionniste, advenant que l'un d'eux soit « vrai ». Mais étant donné nos réserves quant à sa typologie, il est inutile de la commenter ici. Le tableau 3.2.4.1.b. récapitule son analyse.

Type de représentation	**Conséquences philosophiques**
Type 1	Validation de l'empirisme
Type 2	Validation d'un mélange d'empirisme et de nativisme
Type 3	Validation du holisme structurel et abandon de la sémantique compositionnelle
Type 4	Validation de l'associationnisme

Tableau 3.2.4.1.b : Les conséquences philosophiques
des quatre types de représentation selon Ramsey (1992)

modélisent. Il faudrait donc prendre en considération la possibilité d'appliquer les réseaux à différents niveaux d'analyse. Par conséquent, ce n'est pas véritablement le type de réseaux qui déterminera quel est le type de représentation mais (1) le niveau cognitif auquel le modèle sera appliqué et (2) les capacités computationnelles du réseau. Le scénario est donc plus complexe que la typologie de Ramsey ne le présente.

On voit aussi qu'il y a une multiplicité d'applications des réseaux à la modélisation de l'esprit. Cela est dû au fait qu'avant d'être des modélisations de l'esprit, les réseaux de neurones sont des systèmes computationnels. Et ces systèmes ont la propriété de pouvoir modéliser à peu près n'importe quelle tâche pouvant être réduite ou expliquée en référence à une computation qu'il s'agisse de la reconnaissance de catégories, du traitement de la parole ou de calculs en économie. (Beale et Jackson, 1991, 97-104) Les réseaux de neurones ne sont donc pas « intrinsèquement cognitifs » même si certaines de leurs applications sont des modélisations cognitives. On doit tenir compte de cette ambiguïté lorsqu'on traite de la question de la représentation dans les modèles connexionnistes, ce que Ramsey ne fait pas pleinement. Mais pourrait-on remédier à ces lacunes en complétant la typologie de Ramsey avec la distinction entre les niveaux symbolique et sous-symbolique de Smolensky?

3.3.4. Une tentative de systématiser l'interprétation cognitive des réseaux de neurones

Une tentative de complexifier l'analyse pourrait s'inspirer de Ramsey et de Smolensky en soutenant qu'il y a quatre types de représentations mentales dans les réseaux de neurones et que ces quatre types entretiennent des relations divergentes avec les trois niveaux d'analyse relevés par Smolensky (neuronal, sous-symbolique et symbolique). L'idée est que chaque type de réseau avec une représentation mentale particulière aurait une relation spécifique avec les trois niveaux d'analyse. Le tableau 3.3.4.a. systématise cette intuition. Ainsi, le

premier type entretient une relation exacte avec le niveau sous-symbolique mais approximative avec le niveau neuronal, le deuxième serait une bonne approximation du niveau neuronal et une description exacte du niveau sous-symbolique, etc. Par contre, cette tentative est lacunaire parce qu'elle ne tient pas compte de la distinction entre sciences cognitives descriptives et sciences cognitives causales que nous avons vue chez Clark. Il faudrait donc complexifier cette intuition pour y inclure cette dichotomie descriptif/causal et la rendre pleinement accessible à tous les niveaux de modélisation. La prochaine section tentera de dégager plus clairement un tel cadre à partir des trois éléments de clarification que nous venons de présenter et de commenter.

Type de rep. mentale*	Relation avec le niveau neuronal**	Relation avec le niveau sous-symbolique	Relation avec le niveau symbolique
Type 1	+/- approximation	exacte	---
Type 2	bonne approximation	exacte	---
Type 3	---	bonne approximation	(+/-) exacte
Type 4	---	(+/-) approximation	exacte

Tableau 3.3.4.a. : Une tentative de systématisation inspirée de Smolensky et Ramsey

* les types de représentation mentale sont tirés de Ramsey (1992 et 1996)
** les relations avec les différents niveaux (neuronal, sous-symbolique et symbolique) sont tirées de Smolensky (1988)

3.3. Éléments théoriques pour l'analyse du débat entre le cognitivisme classique et le connexionnisme sur la question des concepts

Cette troisième section tente de rassembler les éléments de clarification précédents et de les systématiser afin de parvenir à un schéma de clarification plus complet. Nous introduisons les niveaux d'analyse, les types d'analyse et les types d'entreprise, lesquels ont été entrevus dans les clarifications de la section précédente.[11] Ces deux derniers éléments débouchent

[11] Notre prétention n'est pas une tentative de développer une thèse mais seulement de rassembler et d'organiser quelques éléments de clarification pour la question des concepts.

sur quatre dimensions d'explication découlant de leur jumelage. La stratégie de l'interprète de Dennett est ensuite prise comme un exemple de description du mental cadrant avec l'esprit de nos distinctions entre les dimensions d'explication.

3.3.1. Les niveaux d'analyse

Les niveaux d'analyse[12] sont les niveaux auxquels s'adressent une modélisation ou une explication. On peut distinguer à l'instar de Dehaene et Changeux (1989, 67-71 et Changeux et Dehaene, 1993, 124-6) ainsi que Lycan (1987, 38)[13] les niveaux moléculaires, cellulaires, des réseaux et des réseaux de réseaux. L'idée fondamentale est que le système nerveux, base biologique de l'esprit et donc de la conceptualité, a une organisation hiéarchique qui se reflète dans ses niveaux d'organisation. Le clivage de chacun de ses niveaux correspond à l'apparition phylogénétique de nouvelles fonctions. (Dehaene et Changeux, 1989, 71-2; Changeux et Dehaene, 1993, 124) Par conséquent, il est important de distinguer dans la mesure du possible à quel niveau nous avons affaire lorsque nous tentons de modéliser une tâche cognitive.[14] Cependant, cette clarification assez simple du niveau d'analyse impliquera la distinction supplémentaire et plus abstraite entre deux types d'analyse, soit l'analyse fonctionnelle et l'analyse des propriétés relationnelles que nous puisons dans Bunge et Mahner (1997).

---

[12] Nous nous rattachons ici à un débat qui implique notamment, Changeux et Deheane, 1989; Paul M. Churchland, 1998c; Churchland et Churchland, 1998b; Clark, 1989; Dehaene et Changeux, 1989 et 1993; Dennett, 1981 et 1987; Edelman, 1987, 1989 et 1991; Hofstadter, 1985, Lycan 1987; Rueckl, 1991; Rumelhart et McClelland, 1985; Rumelhart et McClelland, 1986b; Rumelhart, Smolensky, McClelland et Hinton, 1986; Smolensky, 1992, 1986, 1988.

[13] Voir aussi Rosenzweig et Leiman (1991, 13) pour cette distinction dans un contexte psychophysiologique.

[14] On peut aussi préciser quel est le mode transition entre niveaux. Dehaene et Changeux (1989, 71-3 et Changeux et Dehaene, 1993, 125-6) optent pour un schéma darwinien généralisé où l'on a une variabilité à un niveau inférieur (générateur de diversité) et une sélection à un niveau supérieur de propriétés ayant des propriétés relationelles avantageuses. Edelman, avec sa théorie des systèmes somatiques sélectifs (1987, 1989 et 1991) va dans le même sens.

3.3.2. les types d'analyse

Nous dégageons deux types d'analyse à partir de la distinction entre la fonction et le rôle biologiques de Bunge et Mahner (1997). Clarifions d'abord ces deux notions. Admettons un système *b* (par ex. : un organisme) ayant un sous-sytème *a* (par ex. : un système digestif). *L'analyse fonctionnelle* de *a* est la considération de ce que ce sous-système fait. En d'autres mots, la *fonction biologique* c'est le fonctionnement ou l'ensemble des processus qui ont lieu dans un sous-système.[15] (Bunge et Mahner, 1997, 155) *L'analyse des propriétés relationnelles* (ou l'analyse relationnelle) est la considération du rôle biologique d'un sous-sytème dans un super-système *e* (par ex. : un système environnemental). Le *rôle biologique* est l'ensemble des relations liant le sous-sytème *a* au supersystème *e*. Le *rôle biologique spécifique* d'un sous-système est l'ensemble des relations pouvant être exclusivement attribuées au sous-système *a* et qui lient *a* à *e*. (Bunge, 1997, 155) Nous reprenons cette distinction pour produire deux types d'analyse, soit l'analyse fonctionnelle et l'analyse relationnelle. Le premier s'attarde à la considération de la fonction biologique et le deuxième au rôle biologique.

Cette distinction permet d'affirmer avec Lycan ainsi que Deheane et Changeux (1989 et Changeux et Dehaene, 1993) que la distinction traditionnelle entre la structure et la fonction[16] transperce tous les niveaux d'organisation du vivant. Par conséquent, une distinction cloisonnée entre le *hardware* (structure) et le *software* (fonction) ou entre les niveaux computationnel, algorithmique et implémentationnel est trop rigide pour l'étude de la cognition et de la conceptualité humaines. Cette prise de contact avec la distinction

[15] Cet usage du terme fonction s'écarte donc d'une conception téléologique où la fonction est ce à quoi sert un sous-système (son but).

[16] Il faut prendre note que la distinction entre fonction et rôle de Bunge et Mahner n'équivaut pas complètement à la distinction traditionnelle entre la structure et la fonction. Il s'agit d'une reformulation plus

fonction/rôle permettra de complexifier plus loin l'analyse des propriétés respectives des modélisations classiques et connexionnistes de la conceptualité. Nous allons maintenant introduire une distinction supplémentaire entre deux types d'entreprise, l'entreprise descriptive et l'entreprise causale.

3.3.3. les deux types d'entreprise

Nous allons maintenant dégager deux types d'entreprises à partir de la distinction de Clark (1989) entre les sciences cognitives descriptives et les sciences cognitives causales. La première est *l'entreprise descriptive* où l'on tente de décrire à un niveau abstrait la cognition. La deuxième est *l'entreprise causale* où l'on tente de dégager les mécanismes causaux expliquant la cognition. Cette distinction débouche sur une distinction correspondante entre la dimension descriptive et la dimension causale d'une explication plus globale. La deuxième a une base ontologique certaine puisqu'elle s'appuie sur les faits et leurs interactions causales. La première a un statut plus difficile à cerner dans une ontologie matérialiste. Son but est de dégager des généralisations qui sont au-dessus de la description causale et qui s'expriment en termes d'algorithme. Par conséquent, étant donné sa nature, l'explication descriptive semble un peu plus approximative parce qu'elle abstrait littéralement à partir des données. Encore une fois cette distinction entreprise causale/descriptive est présente à plusieurs niveaux d'analyse. Elle n'est donc pas liée exclusivement au niveau cognitif (par ex. : algorithme sémantique) mais pourrait aussi bien s'appliquer au niveau neuronal (par ex. : algorithme d'activation).

---

précise. Par contre, notre rapprochement est légitimé par le fait que les auteurs ébauchent cette distinction dans le but de clarifier le débat portant sur la relation structure/fonction.

3.3.4. Les quatre dimensions d'explication

Si l'on combine les deux types d'analyse (section 3.3.2.) avec les deux types d'entreprise (section 3.3.3.), on obtient alors quatre dimensions d'explication, reflétant ainsi les quatre combinaisons possibles, ce que le tableau 3.3.4.a. reprend.

fonctionnelle-descriptive (F-D)	relationnelle-descriptive (R-D)
fonctionnelle-causale (F-C)	relationnelle-causale (R-C)

Tableau 3.3.4.a. : Les quatre dimensions d'explication

En gros, la dimensions F-C serait la dimension causale du fonctionnement d'un sous-sytème. La dimension R-C serait la dimension causale d'une relation entre un sous-sytème et un super-système. La dimension F-D serait une description abstraite (ou algorithmique) du fonctionnement d'un sous-système. Et enfin la dimension R-D serait une description abstraite (ou algorithmique) des relations d'un sous-sytème par rapport à un super-système. Prenons un exemple tiré de la neurologie de la vision pour illustrer comment ces distinctions permettent de clarifier les dimensions d'une explication plus globale. Considérons un bâtonnet de la rétine comme un sous-sytème d'un système (c'est-à-dire le système phototopique, un réseau sensible à la couleur et aux intensités plus élévées) lui-même retrouvé à l'intérieur d'un super-système, un réseau de réseaux instanciant une fonction cognitive complexe, la vision. On pourrait expliquer (F-C) le fonctionnement causal du potentiel d'action au niveau neurochimique, ensuite expliquer (F-D) l'algorithme (la règle d'activation) qui régit l'émission du potentiel d'action, ensuite expliquer (R-C) causalement comment le déclenchement du potentiel d'action contribue au niveau du super-

système à la perception d'une luminosité et enfin expliquer (R-D) l'algorithme qui régit la relation entre ce neurone[17] et le système visuel.

Ces nuances dans l'explication permettent de saisir la multiplicité des dimensions possibles mais aussi l'interdépendance des dimensions d'explication et des niveaux d'analyse. Car on voit bien que les quatre dimensions d'explication ont une certaine relativité d'application puisque dans notre exemple le neurone est le sous-sytème et le système visuel le super-système mais ce dernier aurait bien pu être un sous-système et l'environnement, le super-système. L'idée est donc que les quatre dimensions d'explication sont relatives à chaque niveau, ou répétables à plusieurs niveaux. Il est donc d'importance fondamentale de préciser rigoureusement quels sont les sous-système, système et super-système d'ancrage pour un cadre d'explication donné. Le tableau 3.3.4.a. reprend schématiquement les quatre types d'analyse et leur application à tous les niveaux d'analyse.

Niveau d'analyse	**Type d'analyse**	**Type d'entreprise**	**Dimensions d'explication**
Réseaux de réseaux *	F ou R***	C ou D****	F-C/F-D/R-C/RD
Réseaux de neurones**	F ou R	C ou D	F-C/F-D/R-C/RD
Neurones	F ou R	C ou D	F-C/F-D/R-C/RD
Molécules	F ou R	C ou D	F-C/F-D/R-C/RD

Tableau 3.3.4.b. : Les quatre dimensions d'explication appliquées à chaque niveau d'analyse

* une capacité cognitive sous-tendue par divers réseaux
** une fonction cognitive élémentaire
*** c'est-à-dire fonctionnelle ou relationnelle
**** c'est-à-dire causale ou descriptive

[17] Bien sûr , il y a la difficulté pratique d'isoler la contribution d'un neurone au super-système.

3.3.5. Les dimensions d'explication et la question de l'éliminativisme

Il est très important de noter que toutes les dimensions d'explication ne sont pas également légitimes à un niveau donné. Certaines ne sont pas ou peu utilisées (utilisables). Cela peut, entre autres, être dû à notre connaissance lacunaire des processus causaux en question (dans le cas des dimensions F-C et R-C) ou bien simplement au faible gain d'information lorsque l'analyse descriptive est employée (dans le cas des dimensions F-D et R-D) étant donné notre grande connaissance des assises causales d'un phénomène. Par exemple, il est moins bien indiqué d'utiliser la dimension de type R-D lorsque nous connaissons la dimension F-C d'un processus. L'une ne fait plus ou moins que décrire abstraitement ce que l'autre explique causalement. Cependant, sur le débat du statut approximatif des niveaux supérieurs, nous demeurons prudent sans trancher définitivement sur la possibilité d'un découpage général de cette question.

Pour des processus plus difficilement traduisibles en termes de dimension F-C, telle que la formation des concepts ou d'autres processus cognitifs abstraits, il est plus essentiel d'avoir recours aux macro-descriptions abstraites de type R-D et F-D pour saisir certaines régularités qui autrement ne seraient pas « visibles » d'un micro-point de vue F-C. Étant donné l'état de nos connaissances sur la cognition, il est inévitable d'avoir recours à la dimension R-D. Maintenant que nous avons clarifié la question des dimensions d'explication et de leur application aux différents niveaux d'analyse, nous allons revenir à la question des concepts. Notre stratégie est de caractériser plus pleinement le caractère et l'usage de la dimension R-D appliquée à la cognition en la mettant en relation avec la stratégie de l'interprète de Dennett entrevue ici comme dimension de type R-D.

3.3.6. La stratégie de l'interprète comme dimension d'explication relationnelle descriptive (R-D)

Nous allons maintenant rapprocher la dimension R-D (appliquée au macro-niveau de la cognition humaine) et la stratégie de l'interprète de Dennett. La stratégie de l'interprète est une stratégie se définissant en rapport à la stratégie physique (*physical stance*) et à la stratégie de la conception (*design stance*). Tandis que la deuxième stratégie explique un système par sa composition physique et la troisième par sa conception (*design*), la stratégie de l'interprète explique l'activité d'un organisme en tant que système intentionnel.[18] Un système intentionnel est tout système dont le comportement peut être adéquatement décrit par l'attribution d'attitudes propositionnelles. On lui attribue alors des désirs, des croyances et une rationalité minimale, ce qui permet de conclure à une action dans le cadre d'un syllogisme pratique. Cette stratégie est vue comme un moyen de saisir à un niveau plus abstrait des généralisations qui ne seraient pas apparentes au niveau physique ou au niveau de la conception. (Dennett, 1981) L'adoption de la stratégie est gratuite mais les schémas (*patterns*) auxquels elle renvoie sont objectifs. C'est donc une stratégie commode qui dégage un nouveau niveau d'analyse : celui que l'on nomme mental ou cognitif. Mais quel est le rapport entre cette stratégie et la dimension d'explication de type R-D de nature plus générale?

Nous allons maintenant avancer que la stratégie de l'interprète est une dimension d'explication de type R-D de la cognition. La stratégie de l'interprète renvoie plus indirectement aux faits car elle se situe à un niveau d'abstration plus élevé. Elle décrit les algorithmes, des relations abstraites entre « entités sémantiques » telles que les croyances,

---

[18] Nous sommes conscients que la division tripartite de Dennett ne cadre pas avec notre découpage des quatre dimensions d'explication notamment parce que les quatre dimensions d'explication s'appliquent à tous les niveaux. Notre intérêt pour Dennett est essentiellement pour sa caractérisation de la stratégie de l'interprète que nous rattachons à la dimension d'explication R-D. Notre interprétation tend aussi à intégrer la stratégie de l'interprète à l'explication par la conception.

les désirs, etc. Son caractère heuristique se fait valoir dans ce cadre car elle est une sorte de modèle d'optimalisation. Elle va au-delà des faits pour guider la recherche empirique sur le mental (notamment en éthologie cognitive), pour générer des questions, pour suggérer des hypothèses et organiser l'observation et la collecte de données. Elle suggère des algorithmes mentaux que les organismes instancient et suivent dans leur comportement (son aspect descriptif). (Dennett, 1987, 260) Elle permet aussi de soutenir une sorte de réalisme au sujet des propriétés relationnelles (son aspect relationnel) sans pour autant tomber dans le piège de prendre une analyse relationnelle pour une analyse fonctionnelle et une entreprise descriptive pour une analyse causale. Autrement dit, la stratégie de l'interprète est une dimension d'explication R-D de la cognition qui est consciente de ses limites. Elle constitue une caractérisation du mental ouverte sur les autres dimensions d'explication (F-C, R-C et F-D).[19] Sa nature descriptive et relationnelle fait qu'elle peut sembler contrefactuelle et approximative[20] puisqu'elle est peu vérifiable directement. Cette analyse de la stratégie de l'interprète en tant que dimension d'explication R-D nous permet de mieux entrevoir l'utilité des distinctions entre dimensions d'explication. La prochaine section exploite la grille d'analyse que nous avons systématisée afin de clarifier le débat entre le cognitivisme et le connexionnisme sur la question des concepts.

3.4. Clarification du débat entre le cognitivisme et le connexionnisme sur la question des concepts

Cette quatrième section aura pour but la clarification du débat entre Churchland et Fodor sur la question des concepts en utilisant la distinction entre les quatre dimensions d'explication. Nous notons premièrement le caractère pragmatique, heuristique et continu

[19] D'où les reproches d'instrumentalisme que l'on formule à l'égard de Dennett car les nuances retrouvées dans sa position quant à la dimension d'explication R-D (le caractère provisoire et ouvert de la caractérisation du mental en termes R-D) sont novatrices.

des quatre dimensions d'explication dans l'étude de la cognition, ensuite nous les appliquons aux positions de Fodor et de Churchland. En peu de mots, le premier réduit l'étude de la cognition à la dimension R-D, le deuxième à la dimension F-D. Nous suggérons ensuite un cadre plus généreux dans lequel les quatre dimensions d'explication auraient un rôle à jouer afin de parvenir à une explication plus globale de la conceptualité.

3.4.1. L'aspect pragmatique, heuristique et continu des quatre dimensions d'explication

Nous aimerions faire valoir avant de poursuivre, le caractère pragmatique, heuristique et continu des quatre dimensions d'explication. Ces dimensions sont des idéalisations théoriques permettant d'effectuer des clarifications qui n'ont valeur que dans la mesure où elles peuvent éclairer le débat. Le découpage est heuristique et recouvre une certaine continuité. Cela revient à dire que certains éléments d'explication pourraient être difficilement classables. La figure 3.4.1.a. illustre cet aspect de continuité entre les dimensions d'explication.

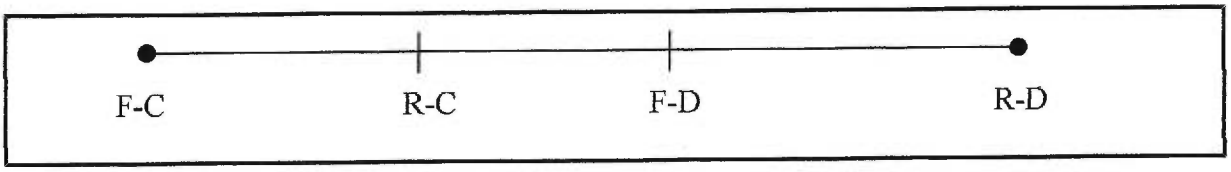

Figure 3.4.1.a. : La continuité entre les quatre dimensions d'explication

Certains pourraient s'opposer à la distinction entre ces quatre dimensions d'explication sous prétexte qu'elle est simpliste. Une brève réponse consisterait à soutenir que premièrement, elle est plus complète et plus souple qu'une distinction rigide hardware/sotware (Johnson-Laird, 1988, Pylyshyn, 1985) ou la trilogie computation/algorithme/implémentation (Marr, 1982). Et deuxièmement, son utilité

---

[20] On voit ici que la dimension R-D appliquée au macro-niveau de la cognition souffre des lacunes de nos connaissances, d'où son aspect approximatif et instable.

découle du fait, comme nous le verrons, qu'elle permet de clarifier plus précisément la complicité possible entre différentes modélisations. Elle est donc une hypothèse de travail qui pourrait servir éventuellement de cadre de classification pour l'intégration de nouveaux paradigmes (Kuhn, 1996) de modélisation de l'esprit humain.

3.4.2. L'étude de la cognition et les dimensions d'explication

Si l'on considère l'étude de la cognition, à l'aide des dimensions d'explication, on peut alors saisir sa perplexité. Une question telle que « À quel niveau doit-on s'adresser pour étudier la cognition? » devient excessivement complexe. La réponse doit en quelque sorte refléter une sorte de pragmatisme étant donné le caractère partiel de nos connaissances scientifiques sur les processus cognitifs. Nous allons analyser les positions de Fodor et de Churchland selon les dimensions d'explication et constater qu'ils n'adhèrent pas à ce pragmatisme nécessaire. D'une part, Fodor se blottit dans une forme de réalisme rigide par rapport à la dimension d'explication R-D de la cognition humaine. D'autre part, il en va de même pour Churchland au niveau de la dimension F-D (selon notre interprétation). La ligne d'analyse que nous poursuivons à la suite de Clark (1989) conduit à l'abandon de la supposition d'uniformité (*uniformity assumption*) faite par les interprètes rigides du cognitivisme et du connexionnisme. Nul ne peut avoir le monopole de la vertu. Les prétentions à l'exclusivité doivent être abandonnées.

3.4.2.1. les limites de l'interprétation réaliste de la dimension d'explication R-D de Fodor

Comme nous l'avons souligné dans le premier chapitre, Fodor défend l'autonomie de la psychologie. Il soutient qu'elle doit se situer aux niveaux computationnel (surtout) et algorithmique. Par conséquent, sa théorie des concepts se situe à un niveau élevé d'abstraction. Cependant, il défend une forme de réalisme par rapport à ce niveau ce qui

fait qu'il n'envisage pas la possibilité de son explication par les niveaux inférieurs (opposition principielle). Enfin, il est conduit à soutenir que la dimension d'explication R-D est exclusivement appropriée pour étudier les concepts et la cognition. (Voir tableau 3.4.2.1.a.) L'étude des concepts, par conséquent, doit se limiter à considérer les relations sémantiques entre états mentaux dans le cadre d'algorithmes (processus mentaux) décrivant leurs relations. Les principes de compositionalité, de productivité et de systématicité, comme nous l'avons noté à la section 3.1.2., forment une caractérisation abstraite des concepts. Cette caractérisation a son utilité. Cependant, elle demeure fortement lacunaire car elle n'intègre à peu près rien des dimensions d'explication causales (F-C et R-C).

fonctionnelle-descriptive (F-D)	relationnelle-descriptive (R-D)
fonctionnelle-causale (F-C)	relationnelle-causale (R-C)

Tableau 3.4.2.1.a. : La dimension d'explication appropriée pour l'étude de la cognition selon Fodor

L'attitude de Fodor par rapport à l'usage de la dimension d'explication R-D pour étudier la cognition est donc radicalement différente de celle de Dennett. Bien que les deux soient d'accord sur un certain usage de la psychologie intentionnelle, ils diffèrent radicalement quant à l'interprétation que l'on doit faire de cette activité. Le premier défend un réalisme de la dimension d'explication R-D et de la psychologie intentionnelle, le deuxième une sorte d'instrumentalisme heuristique. L'usage conscient de la dimension d'explication R-D pour étudier la cognition ressemble plus à la position de Dennett. Celui-ci ne définit pas a priori ce qui est cognitif. En fait, il adhère à une forme de pragmatisme dans la définition du niveau cognitif ou proprement mental. Sa position est aussi ouverte à la reconstruction progressive du mental par les sciences empiriques. Cette ouverture n'existe pas pour Fodor, étant donné son réalisme (excessif) quant à la dimension d'explication R-D. Mais qu'en est-il de Churchland?

3.4.2.2. les limites de l'interprétation réaliste de la dimension d'explication F-C de Churchland

Un éliminativiste comme Churchland soutient que nous devons nous limiter à l'utilisation de la dimension d'explication F-C[21], qui serait le paradigme de l'explication scientifique matérialiste, pour étudier la cognition. (Ou du moins exclure les éléments provenant du niveau R-D, voir section 2.3.2.6). Cela explique pourquoi sa théorie des concepts rend bien compte des assises plus empiriques de la conceptualité. Cependant, cela explique aussi ses lacunes au niveau supérieur de la conception comme, par exemple, la problématique de l'interprétation sémantique des espaces et des schémas d'activation, la compositionalité, la productivité, la systématicité et l'identité conceptuelle. L'éliminativisme de Churchland le conduit à une impasse parce qu'il utilise lui-même ce niveau lorsqu'il attribue une conceptualité (au niveau sémantique) à des régions de classification dans des espaces d'activation (problème de l'individuation des dimensions et de l'information collatérale, section 2.3.4.3.). Le tableau 3.4.2.2.a. illustre l'interprétation de Churchland selon laquelle la cognition doit être étudiée essentiellement selon la dimension d'explication F-C.

fonctionnelle-descriptive (F-D)	relationnelle-descriptive (R-D)
fonctionnelle-causale (F-C)	relationnelle-causale (R-C)

Tableau 3.4.2.2.a. : La dimension d'explication appropriée pour l'étude de la cognition selon Churchland

En outre, Churchland a tendance à soutenir de façon simplificatrice que tous les modèles connexionnistes sont du deuxième type de Ramsey, soit des réseaux qui parviennent à des représentations à partir d'éléments non interprétables de leurs unités. Or il n'y a pas de

[21] Il n'est pas du tout clair que le connexionnisme permettrait d'expliquer causalement le fonctionnement de l'esprit-cerveau. À notre avis, il se situe plutôt dans le cadre de dimensions d'explication de type F-D. lesquelles proposent des algorithmes de fonctionnement pour des sous-systèmes. De façon générale, on pourrait soutenir que les modèles classiques et les modèles connexionnistes sont des dimensions d'explication de type F-D et R-D respectivement puisqu'ils servent à *modéliser des algorithmes*. Leur rapport à l'observation est moins direct que les dimensions d'explication F-C et R-C à cause de leur nature descriptive.

telle homogénéité dans les modélisations connexionnistes. Certaines sont vouées à la modélisation de réseaux de neurones, d'autres à la manipulation de concepts et d'autres encore au maniement de propositions. L'interprétation monolithique de Churchland révèle un programme de recherche, celui de modéliser les fonctions supérieures du cerveau à partir du niveau neuronal. Or il n'est pas évident que tous les modèles connexionnistes soient intégrables à un tel cadre ou rattachables à un tel programme. Autrement dit, l'idée de réduire le connexionnisme à la dimension d'explication F-C est elle-même erronée. À notre avis, plusieurs modélisations connexionnistes se situent plutôt au niveau F-D, contrairement à l'interprétation de Churchland puisqu'elles offrent des algorithmes du fonctionnement de sous-systèmes. Le tableau 3.4.2.2.b. illustre cette réinterprétation que nous faisons de sa position. Comment devons-nous alors penser la relation entre le cognitivisme et le connexionnisme sur la question des concepts?

fonctionnelle-descriptive (F-D)	relationnelle-descriptive (R-D)
fonctionnelle-causale (F-C)	relationnelle-causale (R-C)

Tableau 3.4.2.2.b. : La dimension privilégiée par Churchland pour l'étude de la cognition

3.4.2.3. une tentative de tracer une plus grande complicité entre le cognitivisme et le connexionnisme sur la question des concepts

Le cognitivisme et le connexionnisme sont en proie à des difficultés pour une raison qui devrait être maintenant très claire, soit leur prétention à expliquer la cognition à partir d'une seule dimension d'explication, négligeant ainsi la nécessité d'une explication plus globale. La solution est donc de valoriser les dimensions d'explication délaissées soit, F-C et R-C. Il faut rétablir une interdépendance entre les différents niveaux d'analyse et les différentes dimensions d'explication. En un mot, les prétentions à l'exclusivité de R-D de Fodor et de F-D de Churchland doivent être abandonnées. Le tableau 3.4.2.2.a. présente les dimensions d'explication à valoriser. Néanmoins, le problème ne s'arrête pas là puisqu'il faut aussi

tenir compte des différents niveaux d'analyse (en plus des dimensions d'explication). Par conséquent, il y a une double complexité. Celle de déterminer quelle dimension d'explication nous utilisons et à quel niveau elle s'applique. Une grande difficulté de l'étude de la cognition est justement de préciser concrètement ces deux éléments théoriques. Peut-être pourrions-nous avancer avec Deheane et Changeux (1989) que cette clarification constitue une composante importante sinon fondamentale pour une théorie scientifique de l'esprit?

fonctionnelle-descriptive (F-D)	relationnelle-descriptive (R-D)
fonctionnelle-causale (F-C)*	relationnelle-causale (R-C)

Tableau 3.4.2.2.a. : Les dimensions d'explication à valoriser pour une étude plus complète de la cognition

* Dimension à laquelle s'appliqueraient les modélisations connexionnistes selon l'interprétation unilatérale de Churchland.

3.4.2.4. la compatibilité du cognitivisme et du connexionnisme sur la question des concepts

La compatibilité que nous traçons entre le cognitivisme et le connexionnisme est sûrement schématique mais elle offre une vision plus complexe de l'interprétation cognitive et surtout conceptuelle des modélisations. La consigne est donc la suivante, avant de prétendre qu'une modélisation ou un type de modélisation a le dernier mot pour expliquer une fonction cognitive, telle que la conceptualité, il faut déterminer à quel niveau nous avons affaire et quelle dimension d'explication nous utilisons. De cette façon, on peut tracer une plus grande complicité entre le cognitivisme et le connexionnisme sur la question des concepts. Par exemple, une modélisation connexionniste visant à expliquer les différents concepts du goût (Churchland, 1995) en s'inspirant de l'organisation neuronale n'est pas nécessairement incompatible avec une description de l'usage et des relations sémantiques de ces concepts selon les principes de compositionalité, de productivité et de systématicité. (Fodor et Pylyshyn, 1988). Cependant, cela se fait moyennant l'adoption

(déflationniste) d'une attitude pragmatique quant aux prétentions des deux types de modélisation. Si l'on rattache la première démarche à une dimension d'explication F-D[22] et la deuxième à une dimension d'explication R-D, alors nous avons un tableau plus complet des aspects de la conceptualité. (Voir aussi la section 3.1.1. pour les apports possibles de chacune des approches.) Le tableau 3.4.2.4.a. offre un portrait de la collaboration entre les différentes dimensions des concepts, ici les concepts du goût.

fonctionnelle-descriptive (F-D) :	**relationnelle-descriptive (R-D) :**
Description abstraite (ou algorithmique) du fonctionnement du sous-système du goût. Identification formelle des types de catégorisations produites par le goût (régions d'agglutination dans un espace d'activation) à l'aide de schémas d'activation, ce qui permet de cerner les sous-éléments du contenu conceptuel. (Churchland, voir section 2.3.2.)	Description abstraite (ou algorithmique) de l'usage des concepts et des relations sémantiques des concepts décrivant le goût selon les principes de compositionalité, de productivité et de systématicité de façon à intégrer la relation entre le sous-sytème du goût et un super-système cognitif. (Fodor)
fonctionnelle-causale (F-C) :	**relationnelle-causale (R-C) :**
Explication de la dimension causale du goût (en tant que sous-système sensoriel) à partir d'une étude neurophysiologique et neurochimique détaillée des neurones sensoriels et de leur activité.	Explication de la dimension causale des processus permettant de relier la dimension sensorielle du goût (en tant que sous-système) aux descriptions sémantiques des concepts du goût (en tant que super-système). Par exemple, comment les catégorisations du goût sont-elles traitées par les régions du cerveau à vocation plus cognitive?

Tableau 3.4.2.4.a. : La complicité des dimensions d'explication dans l'étude des concepts du goût

Il n'y a donc pas de réponse absolue à savoir quel type de modèle est approprié pour la modélisation des concepts. Comme l'exemple précédent l'illustre, l'usage d'une

[22] Notons ici que nous interprétons, contrairement à Churchland, les modélisations connexionnistes comme fournissant une description plutôt abstraite (formelle ou algorithmique) du goût (dimension F-D) plutôt qu'une dimension F-C. Pour rendre compte de la dimension F-C de façon convaincante, il faudrait intégrer encore plus de données physico-chimiques venant de la neurophysiologie et de la neurochimie que les modèles connexionnistes le font. Les modélisations proposées par Churchland (1989 et 1995) se situent plutôt au niveau des descriptions abstraites du fonctionnement (F-D) du goût. Cela revient à notre

modélisation classique ou connexionniste dépend du niveau auquel s'adresse la modélisation ainsi que de la dimension d'explication utilisée. L'aspect symbolique ou l'aspect connexionniste des concepts est donc relatif sans être relativiste puisqu'ils s'insèrent dans un cadre d'explication plus global.

3.4.3. Remarques sur la clarification proposée

Maintenant que nous avons développé un cadre de clarification pour le débat entre le cognitivisme et le connexionnisme sur la question des concepts, quelques remarques finales s'imposent sur les limites de cette tentative de clarification sur la portée de ce cadre.

3.4.3.1. les limites de la clarification

Distinguer quatre dimensions d'explication à l'oeuvre dans l'étude de la cognition ne règle pas tout le débat. Au mieux, cette clarification, si elle est avérée, constitue un cadre général à partir duquel nous pouvons envisager une intégration possible du cognitivisme et du connexionnisme, notamment sur la question des concepts. Elle constitue donc une sorte de macro-perspective sur le débat, une analyse générale. Par contre, il faudrait déterminer avec plus de précision comment le cadre proposé articule la complémentarité des deux approches sur des débats plus précis. Par exemple, comment pouvons-nous penser le débat sur l'identité conceptuelle à l'aide de ce cadre? Comment nous permettrait-il de mettre en valeur les progrès méthodologiques et conceptuels des deux perspectives? Pourrait-il amener une interaction entre les deux « paradigmes »? Si oui, de quelle façon? Ces questions intéressantes et importantes nécessiteraient sûrement un traitement complet et rigoureux, mais nous pouvons quand même proposer quelques éléments de réponse.

---

commentaire précédent que Churchland a tendance à interpréter de façon unilatérale le niveau auquel s'adressent les modélisations connexionnistes

À titre d'hypothèse provisoire, nous pouvons avancer que la compatibilité des deux approches serait possible moyennant la capacité de reconnaître éventuellement un cadre commun aux deux perspectives où l'on précisera plus exactement ce que *sont* les concepts, ce qui permettra ensuite de développer des modélisations qui tiennent comptent de certains aspects particuliers des concepts. Cela voudrait dire que l'on serait capable de situer où se trouvent certaines caractéristiques des concepts dans l'optique des quatre dimensions d'explication proposées. Cependant, cela semble peu probable puisque le désaccord persiste sur la nature des concepts car les deux approches semblent « instituer » des objets différents. L'ambiguïté persiste sur l'identité de l'objet à étudier, mais peut-il en être autrement dans ce débat très « abstrait »?

À défaut de préciser clairement et de façon absolue ce qu'est un concept, une façon de clarifier davantage le débat serait de distinguer différents types de concepts comme, par exemple, les catégorisations sensorielles, les catégorisations perceptuelles, les catégorisations conceptuelles, etc. Cela permettrait de cerner avec un plus haut degré de précision et de certitude ce dont les modélisations tentent de rendre compte. Verrait-on alors un départage précis de l'étude de la conceptualité entre les deux approches? Si on ne peut espérer dans cette optique un découpage rapide, ce dernier pourrait tout au moins se faire selon l'évolution et la progression des perspectives. Le débat serait alors réglé au cas par cas à l'intérieur d'un cadre plus général qui lui-même se développerait et serait déterminé progressivement.

Un travail corollaire de clarification s'imposerait au niveau des distinctions entre les concepts et les catégorisations. Quel phénomène ces notions désignent-elles? Quelles sont les différents niveaux de conceptualisation et de catégorisation? Cette voie serait peut-être d'un secours éventuel pour clarifier les micro-débats entre le cognitivisme et le

connexionnisme sur la question des concepts. Enfin, il suffit pour l'instant de prendre note que le schéma proposé souffre d'une certaine généralité et qu'il faudrait éprouver sa capacité à éclairer des débats plus précis sur la conceptualité. Car déterminer dans la mesure du possible à quel niveau une modélisation s'applique et quel est sont but (modélisation R-D ou F-D) ne clôt pas définitivement le débat sur les questions précises.[23] Ce n'est tout au plus qu'un début de clarification. En fait, le cadre de clarification développé, en laissant voir ses limites, soulèvent d'autres questions pertinentes qui appellent en retour des clarifications plus fines.

3.4.3.2. la portée de la clarification

La portée du cadre de clarification présenté est déterminée en partie par la façon dont nous concevons le rapport entre la conceptualité et les deux types de modèles que nous avons examinés. Si, comme c'est souvent le cas dans ce travail, les concepts sont considérés comme ce que les deux types de modélisations permettent de « confirmer » au sujet de la conceptualité, alors nous aurons tendance à soutenir que les concepts sont ce que le cognitivisme et le connexionnisme nous permettent d'affirmer au sujet des concepts. Par contre, si nous pensons que la théorie des concepts est implémentée par deux types de modèle, alors nous aurons tendance à soutenir que les concepts sont modélisés en partie par les modèles et que, par conséquent, les concepts sont sous-déterminés par les deux approches. D'un côté on part des modèles pour arriver aux concepts, de l'autre, on part des concepts pour aller aux modèles. Si ce travail a eu tendance à emprunter la première voie, il ne faut pas perdre de vue que les modèles sont des modèles. Ils ne nous disent pas tout sur la conceptualité, mais néanmoins ils peuvent contribuer à nous faire découvrir certaines

[23] En outre, les ambiguïtés qui persistent au sujet du niveau auquel s'applique les modèles connexionnistes n'aident en rien à clore le débat. Si les modèles connexionnistes sont appliqués au niveau symbolique, alors

facettes de la conceptualité en dévoilant peu à peu sa complexité. Ainsi, il faudrait développer un cadre intermédiaire tenant compte de la dynamique concepts/modèles afin de mieux saisir cette relation problématique. La portée de la clarification que nous avons apportée pourrait donc gagner en acuité et en étendue si un travail de fond était entamé sur le rapport entre les concepts et les modélisations.

3.4.4. Conclusion

Dans ce troisième et dernier chapitre, nous avons comparé deux approches quant à la question des concepts. Nous avons tenté de clarifier l'opposition entre les modélisations classiques et connexionnistes des concepts en nous inspirant de Smolensky, Clark et Ramsey. L'analyse de ces contributions nous a conduit à la systématisation de quatre dimensions d'explication. Ceux-ci nous ont permis de clarifier l'opposition entre le cognitivisme et le connexionnisme tout en faisant voir la complexité du débat sur le traitement approprié des concepts et la difficulté de proposer un cadre général d'analyse. De façon générale, le débat sur les concepts est complexifié par une double saisie, soit celle du niveau d'analyse et de la dimension d'explication privilégiés dans le cadre d'une modélisation donnée.

---

ils entrent en concurrence avec les modèles classiques, ce qui aliment la controverse et met en péril la perspective d'une colloboration des deux approches dans une étude intégrée de la conceptualité.

Conclusion

La question des concepts nous a conduit à considérer deux positions divergentes en sciences cognitives, soit celle du cognitivisme classique et celle du connexionnisme. Pour le cognitivisme classique, l'approche traditionnelle en sciences cognitives, les concepts sont des symboles dont l'usage est justifié par leur rôle fondamental dans l'explication des processus mentaux. Fodor précise cette thèse dans le cadre de sa théorie représentationnelle de l'esprit. Selon lui, les concepts sont compositionnels, productifs et systématiques en plus d'être des états mentaux particuliers. Ce sont des catégories s'appliquant au monde et pouvant donc être évaluées. En outre, ils sont majoritairement appris et publics. Cette position s'insère dans l'hypothèse du langage de la pensée et un rejet du holisme, du pragmatisme et de la théorie prototypique des concepts. Les connexionnistes rejettent un tel cadre pour traiter des concepts. Ils condamnent sa rigidité et sa fragilité; son insensibilité au contexte; son implausibilité neurobiologique et son lien avec la psychologie (douteuse) du sens commun; sa dépendance sur l'autonomie relative de la psychologie; sa conception peu synergétique du lien entre le langage et la pensée; son lien faussement exlusif avec la compositionalité ainsi que son recours à l'hypothèse du langage de la pensée.

Le connexionnisme, « paradigme émergent » en sciences cognitives, offre une autre vision des concepts. Selon l'interprétation de Churchland, les concepts sont des régions de classification dans des espaces d'activation de réseaux de neurones. L'analyse des agglutinations à l'intérieur des espaces d'activation à l'aide de techniques de classification (« plus proche voisin » et distance métrique) permet de dégager des concepts. Ces tâches de classification faites par les réseaux sont interprétées comme des recherches de satisfaction de contraintes dans un espace à n-dimensions. On peut dégager de

l'interprétation de Churchland que les concepts ont un contenu interne et externe; un contenu holistique; un contenu modérément empirique ainsi qu'un caractère pragmatique et dynamique. En fin de compte, Churchland, propose une forme d'éliminativisme par rapport à la psychologie du sens commun, la source même de l'approche classique. Ce tableau déplaît considérablement à Fodor. Ce dernier rétorque que le connexionnisme ne peut pas rendre compte de l'identité conceptuelle; de l'individuation des dimensions et de l'information collatérale. Il accuse de surcroît le connexionnisme d'empirisme naïf, de holisme et de tracer une fausse adéquation entre les concepts et les prototypes. Qui a raison? Les concepts sont-ils des symboles ou des connexions?

Le départage de ces deux positions sur les concepts nous a conduit à établir une comparaison entre les deux types de modélisation. *Grosso modo*, la première rend mieux compte des comportements sémantiques des concepts (compositionalité, productivité, systématicité, etc.) tandis que la deuxième rend mieux compte des assises empiriques de la conceptualité (souplesse, dynamisme, caractère émergent, distribué et robuste, etc.). Trois éléments de clarification ont été sous-tirés des analyses de Smolensky, Clark et Ramsey, soit respectivement la distinction entre les symboles et les sous-symboles, la distinction entre les sciences cognitives descriptives et les sciences cognitives causales ainsi que quatre types de représentation mentale instanciés par les réseaux connexionnistes. Par contre, ces éléments étaient individuellement lacunaires, d'où la nécessité de les rassembler afin de parvenir à une clarification plus convaincante. Nous avons donc introduit de multiples niveaux d'analyse et quatre dimensions d'explication (fonctionnelle-causale / fonctionnelle-descriptive / relationnelle-causale / relationnelle-descriptive) formant un cadre plus global pour l'étude de la cognition. Cette clarification a permis de rattacher l'interprétation de Fodor au modèle R-D et celle de Churchland au modèle F-D. Nous avons ensuite tracé une

voie de complémentarité sur la question des concepts en témoignant d'une certaine forme de pragmatisme. Notre conclusion est donc que les concepts ne sont pas de façon absolue des symboles ou des connexions mais seulement de façon nuancée et relative dans un cadre d'explication plus global.

La question de la nature des concepts est au coeur de la philosophie. Le philosophe réfléchit à l'aide des concepts et il produit des analyses et des distinctions conceptuelles. Bien qu'une théorie scientifique des concepts est un programme de recherche, comme nous l'avons entrevu, tout développement sur cette question devrait intéresser la philosophie au plus haut point. Peut-être pourrait-elle même en espérer une base plus solide pour sa propre entreprise.

BIBLIOGRAPHIE

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

Andler, Daniel (1992) : « Calcul et représentation : les sources », *in Introduction aux sciences cognitives*, Daniel Andler (sous la dir. de), Gallimard, Saint-Amand, 9-46.

Beale, R. et Jackson, T. (1991) : *Neural Computing*, Adam Hilger, New York.

Bechtel, William (1994) : « Connectionism », *in A Companion to the Philosophy of Mind*, Blackwell, Cambridge, MA, 200-10.

Bechtel, William *and* Abrahamsen, Adele (1991) : *Connectionism and the Mind: An Introduction to Parallel Processing in Networks*, Cambridge, MA, Basil Blackwell.

Boole, George (1992) : *Les lois de la pensée*, traduction de Souleymane Bachir Diagne, Vrin, Paris.

Boolos, George S. et Jeffrey, Richard D (1996) : *Computability and Logic*, Cambridge University Press, Cambridge.

Bradshaw, D.E. (1991) : Connectionism and the Specter of Representationalism », *in* Terence Horgan et John Tienson, *Connectionism and the Philosophy of Mind*, Boston, Kluwer Academic Publishers, 417-36.

Broadbent, Donald (1985) : « A Question of Level : Comment on McClelland and Rumelhart », *Journal of Experimental Psychology*, vol. 114, No. 2, 189-92.

Bunge, Mario et Mahner, Martin (1997) : *Foundations of Biophilosophy*, Springer, Heidelberg.

Changeux, Jean-Pierre et Dehaene, Stanislas (1989) : « Neuronal models of cognitive functions », *Cognition*, 33, 63-109.

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

Churchland, Paul M. (1998a) : « Activation Vectors vs. Propositional Attitudes : How the *Brain* Represents Reality », *in On the Contrary. Critical Essays 1987-1997*, Paul M. Churchland et Patricia S. Churchland, MIT Press, Cambridge, MA, 39-44.

Churchland, Paul M. (1998b) : « Conceptul Similarity across Sensory and Neural Diversity : The Fodor-Lepore Challenge Answered », *in On the Contrary. Critical Essays 1987-1997*, Paul M. Churchland et Patricia S. Churchland, MIT Press, Cambridge, MA, 81-112.

Churchland, Paul M. (1998c) : « Folk Psychology », *in On the Contrary. Critical Essays 1987-1997*, Paul M. Churchland et Patricia S. Churchland, MIT Press, Cambridge, MA, 3-15.

Churchland, Paul M. (1989) : « Some Reductive Strategies in Cognitive Neurobiology », *in A Neurocomputational Perspective. The Nature of Mind and the Structure of Science*, MIT Press, Cambridge, MA, 77-110.

Churchland, Paul M. (1993) : « State-Space Semantics and Meaning Holism », *Philosophy and Phenomenological Research*, vol. 53, No. 3, 667-72.

Churchland, Paul M. (1995) : *The Engine of Reason, the Seat of the Soul. A Philosophical Journey into the Brain*, MIT Press, Cambridge, MA.

Churchland, Paul M. et Churchland, Patricia S. (1998a) : « Could a Machine Think? », *in On the Contrary. Critical Essays 1987-1997*, Paul M. Churchland et Patricia S. Churchland, MIT Press, Cambridge, MA, 47-63.

Churchland, Paul M. et Churchland, Patricia S. (1998b) : « Intertheoric Reduction : A Neuroscientist's Field Guide », *in On the Contrary. Critical Essays 1987-1997*, Paul M. Churchland et Patricia S. Churchland, MIT Press, Cambridge, MA, 65-79.

Churchland, Paul M. et Churchland, Patricia S. (1998c) : « Recent Work On Consciousness : Philosophical, Theoretical, and Empirical », *in On the Contrary. Critical Essays 1987-1997*, Paul M. Churchland et Patricia S. Churchland, MIT Press, Cambridge, MA, 159-76.

Churchland, Paul M. et Churchland, Patricia S. (1996) : « Second Reply to Fodor and Lepore », *in The Churchlands And Their Critics*, Robert N. McCauley (éd.), Blackwell, Cambridge, MA, 278-83.

Clark, Andy (1993) : *Associative Engines. Connectionism, Concepts, and Representational Change*, MIT Press, Cambridge, MA.

Clark, Andy (1989) : *Microcognition : Philosophy, Cognitive Science, and Parallel Distributed Processing*, Cambridge, MA, MIT Press.

Clark, Andy (1991) : « Systematicity, Structured Representations and Cognitive Architecture : A Reply to Fodor and Pylyshyn », *in* Terence Horgan et John Tienson, *Connectionism and the Philosophy of Mind*, Boston, Kluwer Academic Publishers, 198-218.

Cummins, Robert (1991) : « The Role of Representation in Connectionist Explanations of Cognitive Capacities », *in* William Ramsey, Stephen Stich et David E. Rumelhart (éd.), *Philosophy and Connectionist Theory*, Lawrence Erlbaum Associates, New Jersey, 91-114.

Davies, Martin (1991) : « Concepts, Connectionism, and the Language of Thought », *in* William Ramsey, Stephen Stich et David E. Rumelhart (éd.), *Philosophy and Connectionist Theory*, Lawrence Erlbaum Associates, New Jersey, 229-57.

Dehaene, Stanislas (1997) : *La bosse des maths*, Odile Jacob, Paris.

Dehaene, Stanislas et Changeux, Jean-Pierre (1993) : « Pensée logico-mathématique et modèles neuronaux des fonctions cognitives : l'exemple des capacités numériques », *in Pensée logico-mathématique : Nouveaux objets interdisciplinaires*, Olivier Houdé et Denis Miéville (éd.), PUF, Paris, 123-46.

Dennett, Daniel C. (1998) : *Brainchildren. Essays on Designing Minds*, MIT Press, Cambridge, MA.

Dennett, Daniel C. (1981) : *Brainstorms. philosophical Essays on Mind and Psychology*, MIT Press, Cambridge, MA.

Dennett, Daniel C. (1987) : *The Intentional Stance*, MIT Press, Cambridge, MA.

Dyer, Michael G. (1991) : « Connectionism Versus Symbolism in High-Level Cognition », *in* Terence Horgan et John Tienson (1991), 382-416.

Edelman, Gerald M. (2000) : *Biologie de la conscience*, traduction de Ana Gerschenfeld, Odile Jacob, Paris.

Edelman, Gerald M. (1987) : *Neural Darwinism : The Theory of Neuronal Group Selection*, Basic Books, New York.

Edelman, Gerald M. (1989) : *The Remembered Present. A Biologivcal Theory of Consciousness*, Basic Books, New York.

Engel, Pascal (1994) : *Introduction à la philosophie de l'esprit*, Éditions la découverte, Paris.

Fetzer, James H. (1992) : « Connectionism and Cognition : Why Fodor and Pylyshyn Are Wrong », *in Connectionism in Context*, Andy Clark et Rudi Lutz (éd.), Springer-Verlag, New York, 37-56.

Fodor, Jerry A. (1990) : *A Theory of Content and Other Essays*, MIT Press, Cambridge, MA.

Fodor, Jerry A. (1995) : « Concepts : a potboiler », *in Cognition on Cognition*, Jacques Mehler et Susanna Frank (éd.), MIT Press, Cambridge, MA.

Fodor, Jerry A. (1998) : *Concepts : Where Cognitive Science Went Wrong*, Clarendon Press, Oxford.

Fodor, Jerry A. (1987) : *Psychosemantics. The Problem of Meaning in the Philosophy of Mind*, MIT Press, Cambridge, MA.

Fodor, Jerry A. (1975) : *The Language of Thought*, Harvard University Press, Cambridge, MA.

Fodor, Jerry A. et Lepore, Ernest (1992) : *Holism : A Shopper's Guide*, Basil Blackwell, Cambridge, MA.

Fodor, Jerry A. et Lepore, Ernest (1993) : « Reply to Critics », *Philosophy and Phenomenological Research*, vol. 53, No. 3, 673-82.

Fodor, Jerry A. et McLaughlin, Brian P. (1991) : « Connectionism and the Problem of Systematicity : Why Smolensky's Solution Doesn't Work », *in* Terence Horgan et John Tienson (1991), 331-54.

Fodor, Jerry A. et Pylyshyn, Zenon W. (1988) : « Connectionism and cognitive architecture : A critical Analysis », *in Connections and Symbols*, Steven Pinker et Jacques Mehler (éd.), MIT Press, Cambridge, MA, 3-71.

Goschke, Thomas et Koppelberg, Dirk (1991) : « The Concept of Representation and the Representation of Concepts in Connectionist Models », *in* William Ramsey, Stephen Stich et David E. Rumelhart (éd.), *Philosophy and Connectionist Theory*, Lawrence Erlbaum Associates, New Jersey, 129-61.

Hahon, Jean-Paul et Hahon, Marie-Christine (1993) : *L'intelligence artificielle*, PUF, Paris.

Hanson, Stephen Hosé et Olson, Carl R. (1990) : « Connectionism and Neuroscience », *in Connectionist Modeling and Brain Function : The Developing Interface*, Stephen Hosé Hanson et Carl Olson (éd.), MIT Press, Cambridge, MA, 1-4.

Hebb, D.O. (1949) : *The Organization of Behavior*, réimp. *in Neurocomuting : Foundations of Research*, J. Anderson et R. Rosenfield (éd.), MIT Press, Cambridge, MA.

Hinton, Geoffrey E., McClelland, James L. et Rumelhart, David E. (1986) : « Distributed Representations », in David E. Rumelhart, James L. McClelland *and* The PDP Research Group, *Parallel Distributed Processing. Explorations in the Microstructure of Cognition*, vol. 1, Cambridge, MIT Press, MA, 77-109.

Hofstadter, D.R. (1985) : « Waking up from the Boolean Dream, or, Subcognition as Computation », *in Metamagical Themas*, Basic Books, New York, 631-65.

Horgan, Terence et Tienson, John (1991) : *Connectionism and the Philosophy of Mind*, Boston, Kluwer Academic Publishers.

Imbert, Michel (1992) : « Neurosciences et sciences cognitives », *in Introduction aux sciences cognitives*, Daniel Andler (sous la dir. de), Gallimard, Saint-Amand, 49-76.

Johnson-Laird, P.N. (1988) : *The Computer and the Brain. An Introduction to Cognitive Science*, Harvard University Press, Cambridge, MA.

Jordan, M.I. (1986) : « An Introduction to Linear Algebra in Parallel Distributed Processing », in David E. Rumelhart, James L. McClelland *and* The PDP Research Group, *Parallel Distributed Processing. Explorations in the Microstructure of Cognition*, vol. 1 Cambridge, MIT Press, MA, 365-422.

Kim, Jaegwon (1998) : *Philosophy of Mind*, Westview Press, Oxford.

Kuhn, Thomas (1996) : *La structure des révolutions scientifiques*, traduction de Laure Meyer, Flammarion, Manchecourt.

Lemaire, Patrick (1999) : *Psychologie cognitive*, De Boeck Université, Paris.

Lepore, Ernie (1994) : « Cognitive Psychology », *in A Companion to the Philosophy of Mind*, Blackwell, Cambridge, MA, 167-76.

Lycan, William G. (1987) : *Consciousness*, MIT Press, MA.

Matthei, Edward et Roeper, Thomas (1988) : *Introduction à la psycholinguistique*, traduction de Ranka Bijeljac, Dunod, Paris.

Marr, David (1982) : *Vision. A Computational Investigation into the Human Representation and Processing of Visual Information*, W.H. Freeman and Company, San Francisco.

McClelland, James.L., Rumelhart, David E. et Hinton, Geoffrey E. (1986) : « The Appeal of Parallel Distributed Processing », in David E. Rumelhart, James L. McClelland *and* The PDP Research Group, *Parallel Distributed Processing. Explorations in the Microstructure of Cognition*, vol. 1, Cambridge, MIT Press, MA, 3-44.

McCulloch, Warren S. et Pitts, Walter (1943) : « A logical calculus of the ideas immanent in nervous activity », réimp. *in Neurocomuting : Foundations of Research*, J. Anderson et R. Rosenfield (éd.), MIT Press, Cambridge, MA, 18-27.

Minsky, Marvin et Papert, Seymour (1969) : *Perceptron. An Introduction to Computational Geometry*, MIT Press, Cambridge, MA.

Newell, Allen (1982) : « The Knowledge Level », *Artificial Intelligence*, 18, 87-127.

Newell, Allen (1980) : « Physical Symbol Systems », *Cognitive Science*, 4, 135-83.

Pinker, Steven et Prince, Alan (1988) : « On language and connectionism : Analysis of a parallel distributed processing model of language acquisition », *Cognition*, 28, 73-193.

Plunkett, Kim et Elman, Jeffrey L. (1997) : *Exercices in Rethinking Innateness. A Handbook for Connectionist Simulations*, MIT Press, Cambridge, MA.

Poirier, Pierre et Fisette, Denis (2000) : *L'esprit en causes*, Vrin, Paris. À paraître.

Pylyshyn, Zenon W. (1985) : *Computation and Cognition. Toward a Foundation for Cognitive Science*, MIT Press, Cambridge, MA.

Ramsey, William (1996) : « Conceptual Analysis and the Connectionist Account of Concepts », *in Philosophy and Cognitive Science. Consciousness and Reasoning*, Andy Clark et *al.* (éd.), Kluwer Academic Publishers, Netherlands, 35-57.

Ramsey, William (1992) : « Connectionism and the Philosophy of Mental Representation », *in Connectionism : Theory and Practice*, Steven Davis (éd.), Oxford University Press, Oxford, 247-76.

Ramsey, William, Stich, Stephen P. et Rumelhart, David E. (1991) : *Philosophy and Connectionist Theory*, Lawrence Erlbaum Associates, New Jersey.

Rey, Georges (1994) : « Concepts », *in A Companion to the Philosophy of Mind*, Blackwell, Cambridge, MA, 185-93.

Rosenzweig, Mark R. et Leiman, Arnold L. (1991) : *Psychophysiologie*, traduction de David Bélanger, Décarie Éditeur, Ville Mont-Royal.

Rueckl, Jay G. (1991) : « Connectionism and the Notion of Levels », *in* Terence Horgan et John Tienson, *Connectionism and the Philosophy of Mind*, Boston, Kluwer Academic Publishers, 74-89.

Rumelhart, David E., Hinton, Geoffrey E. et McClelland, James L. (1986) : « A General Framework for Parallel Distributed Processing », in David E. Rumelhart, James L. McClelland, *and* The PDP Research Group, *Parallel Distributed Processing. Explorations in the Microstructure of Cognition*, vol. 1, Cambridge, MIT Press, MA, 45-76.

Rumelhart, David E. et McClelland, James L. (1985) : « Levels Indeed! A Response to Broadbent », *Journal of Experimental Psychology*, vol. 114, No. 2, 193-7.

Rumelhart, David E. et McClelland, James L. (1986a) : « On Learning the Past Tenses of English Verbs », in David E. Rumelhart, James L. McClelland *and* The PDP Research Group, *Parallel Distributed Processing. Explorations in the Microstructure of Cognition*, vol. 2, Cambridge, MIT Press, MA, 216-71.

Rumelhart, David E. et McClelland, James L. (1986b) : « PDP Models and General Issues in Cognitive Science », in David E. Rumelhart, James L. McClelland *and* The PDP Research Group, *Parallel Distributed Processing. Explorations in the Microstructure of Cognition*, vol. 1, Cambridge, MIT Press, MA, 110-49.

Rumelhart, David E., McClelland, James L. *and* The PDP Research Group (1986) : *Parallel Distributed Processing. Explorations in the Microstructure of Cognition*, vol. 1 et 2, Cambridge, MIT Press, MA.

Rumelhart, David E., Smolensky, Paul, McClelland James L. et Hinton, Geoffrey E. (1986) : « Schemata and Sequential Thought Processes in PDP Models », in David E. Rumelhart, James L. McClelland *and* The PDP Research Group, *Parallel Distributed Processing. Explorations in the Microstructure of Cognition*, vol. 2, Cambridge, MIT Press, MA, , 7-57.

Sejnowski, T.J., Koch, C et Churchland, P.S. (1990) : « Computational Neuroscience », *in Connectionist Modeling and Brain Function : The Developing* Interface, Stephen Hosé Hanson et Carl Olson (éd.), MIT Press, Cambridge, MA, 5-35.

Shastri, L. et Ajjanagadde, V. (1993) : « From simple associations to systematic reasoning : A connectionnist representation of rules, variables and dynamic bindings using temporal synchrony », *Behavioral and Brain Sciences*, 16 (3), p. 417-94.

Smith, Edward E. (1990) : « Categorization », *in Thinking. An Invitation to Cognitive Science*, vol. 3, Daniel N. Osherson et Edward E. Smith (éd.), MIT Press, Cambridge, MA, 33-53.

Smolensky, Paul (1994) : « Computational Models of Mind », *in A Companion to the Philosophy of Mind*, Blackwell, Cambridge, MA, 176-85.

Smolensky, Paul (1991a) : « Connectionism, Constituency, and the Language of Thought », *in Meaning and Mind*, Loewer and Rey (éd.), Blackwell, Cambridge, MA.

Smolensky, Paul (1991b) : « The Constituent Structure of Connectionist Mental States : A Reply to Fodor and Pylyshyn », *in* Terence Horgan et John Tienson (1991), 281-308.

Smolensky, Paul (1992) : « IA connexionniste, IA symbolique et cerveau », in *Introduction aux sciences cognitives*, Daniel Andler (sous la dir. de), Saint-Amand, Gallimard, 77-106.

Smolensky, Paul (1986) : « Neural and Conceptual Interpretation of PDP Models », in David E. Rumelhart, James L. McClelland *and* The PDP Research Group, *Parallel Distributed Processing. Explorations in the Microstructure of Cognition*, vol. 2, Cambridge, MIT Press, MA, 390-431.

Smolensky, Paul, (1988) : « On the Proper Treatment of connectionism », *Behavioral and Brain Sciences*, 11, 1-75.

Turing, A.M. (1936) : « On Computable Numbers, With An Application to the *Entscheidungsproblem* », *Proceedings of the London Mathematical Society*. vol. 42, 230-65.

Turing, A.M. (1950) : « Computing Machinery and Intelligence », *Mind*, vol. 59, No. 236, 433-60.

Van Gelder, Timothy (1991a) : « Classical Questions, Radical Answers : Connectionism and the Structure of Mental Representations », *in* Terence Horgan et John Tienson, *Connectionism and the Philosophy of Mind*, Boston, Kluwer Academic Publishers, 355-81.

Witgenstein, Ludwig (1997) : *Tractatus logico-philosophicus suivi de Investigations philosophiques*, traduction de Pierre Klossowski, Gallimard, Saint-Amand.