Intellectica, 1995/2, 21, pp. 175-197

Pierre LIVET*

Connexionnisme et fonctionnalisme

Connectionism and functionalism

Abstract : Connectionist networks are valued by eliminativists, but they can be interpreted according to the functionalist views. Criticisms against their lack of compositionality and systematicity are to be moderated as they exhibit compositionality limited to their training context, and systematicity dependent on some domain of relevance. We can use them to define ways from physicalist descriptions to functional (categorial) descriptions, when we select as relevant physical properties only the ones which make contextual emergence possible.

Ceux qui voient dans le fonctionnalisme une manoeuvre pour éviter le réductionnisme à bon compte, et pour maintenir vaille que vaille une certaine autonomie du mental par rapport au physique, ont pu voir dans le connexionnisme un moyen de contrer cette manoeuvre. Ainsi nombre de philosophes «éliminativistes», c'est-à-dire qui refusent de considérer des analyses en termes de contenus mentaux comme explicatives, qui veulent éliminer la psychologie populaire avec ses désirs et ses croyances, ont prôné les modèles connexionnistes comme solution alternative (en particulier Patricia et Paul Churchland). Inversement, si on en reste à la problématique de l'Intelligence Artificielle classique, qui définit des procédures pour transformer des séquences de symboles en d'autres séquences de symboles, on ne s'intéresse pas particulièrement à la réalisation physique de ces symboles et de ces procédures (on peut avoir des ordinateurs électroniques, d'autres optiques, ou de simples machines à calculer avec roues dentées) et on s'intéresse davantage à la liaison entre la syntaxe de ces séquences de symboles et une sémantique. Si bien que dans la mesure où l'on peut espérer que les contenus mentaux puissent se réduire à des contenus sémantiques (des conditions de vérité), eux-mêmes écrits dans une syntaxe, le problème du réductionnisme est éclipsé pour le

* Pierre Livet, Professeur à l'Université de Provence, Aix-Marseille I, La Muande, Résidence Beaufort, Chemin des Cruyes, 13090, Aix-en-Provence.
e mail: livet@univ-aix.fr

philosophe par le problème de doter ces traitements de symboles de sémantique et d'intentionnalité, et le rapport de nos contenus mentaux à nos neurones n'est pas à l'ordre du jour.

Mais le connexionnisme n'est peut-être pas l'idéal comme contre-modèle éliminativiste. En effet, les systèmes connexionnistes sont simulés par des ordinateurs classiques, et leur rapport avec les neurones physiologiques est parfois des plus lointains. Et surtout, on peut se demander quelles catégories sémantiques ils sont capables de posséder. Certes, l'avantage des systèmes connexionnistes sur les systèmes symboliques classiques, c'est qu'ils ne présupposent pas qu'on leur donne les symboles, les opérations, et une bonne partie des catégories (c'est-à-dire des liaisons entre séquences de symboles qu'il faut assurer, et des liaisons qu'il faut éviter). Ils sont censés fournir des classifications dans des domaines où les inputs ne sont pas forcément symboliques, et où on ne connaît pas toujours les opérations qui structurent le domaine, ni bien sûr les catégories pertinentes. Ils font «émerger» ces classifications du fonctionnement de leurs unités, reliées par des connexions dont les poids changent au cours de l'apprentissage[1] (chaque unité fait la somme des produits des poids de chaque connexion par l'activation qui parvient à l'unité par cette connexion, et elle transforme cette somme par une fonction à seuil, qui donne l'activation résultat de cette unité, activation qui est transmise à son tour vers d'autres unités, éventuellement les unités de sortie). La plupart du temps, aucune unité ne réalise à elle seule la classification des informations qui parviennent en entrées, mais le réseau tout entier y parvient. Cette classification «émerge» donc du fonctionnement des unités de base. Cette «émergence» permet le réductionnisme, puisque le fonctionnement du réseau est déterministe (l'ajout de déterminations stochastiques n'est dans ce domaine qu'une variante).

Jusque là tout va bien pour l'éliminativiste. Mais les classifications des réseaux sont-elles de bons candidats pour rendre compte de capacités cognitives? Oui quand il s'agit simplement de reconnaître des formes, donc de faire du triage de l'information. La réponse est plus mitigée quand il s'agit de faire des inférences ou en général de relier les classifications produites selon des opérations réglées, elles-mêmes reconnaissables, ce qui revient, dans le vocabulaire philosophique, à être capable de catégorisations. Les attaques de Fodor et Pylyshyn, soutenant que les

---

[1] Nous utiliserons sans trop de scrupules ce terme d'"apprentissage", mais il faut bien reconnaître que le terme d'évolution serait préférable, qu'on parle d'une évolution dirigée, quand on impose au réseau de donner une sortie prédéterminée, ou d'une évolution non dirigée, quand ce n'est pas le cas.

réseaux ne sont capables ni de systématicité, ni de compositionnalité[2], ont suscité beaucoup de réactions et de travaux connexionnistes, mais pour l'instant on n'a pas prouvé que les systèmes connexionnistes sont capables d'apprendre des catégories au sens philosophique du terme, ni d'ailleurs qu'ils n'en sont pas capables.

Il semble que les réseaux n'aient pas forcément partie liée avec l'éliminativisme, et donc qu'ils soient compatibles avec le fonctionnalisme. Nous allons montrer qu'il en est bien ainsi. Encore faudrait-il, pour que les réseaux soient intéressants dans cette perspective, qu'ils manifestent quelque penchant pour des vertus conceptuelles comme la systématicité ou la compositionnalité, et en général pour la catégorisation au sens fort du terme. Nous verrons que si ce n'est pas vraiment le cas, on peut cependant trouver des formes de compositionnalité et de systématicité restreintes compatibles avec les réseaux, qui ont l'avantage et l'inconvénient de nous proposer une cognition dépendante du contexte d'apprentissage. On est alors amené, en analysant les réseaux, à reformuler et les notions de catégories, et celle de «survenance» du mental sur le physique, et à tenter de définir un intermédiaire entre l'éliminativisme et le fonctionnalisme, l'émergence contextuelle.

I. QUAND LE CONNEXIONNISME MIME LE FONCTIONNALISME

Rappelons brièvement quelles sont les thèses fonctionnalistes, pour pouvoir déterminer si les modèles connexionnistes sont compatibles avec elles[3]. Les états mentaux sont reconnus comme des états internes définis par leurs rôles fonctionnels, c'est-à-dire leur capacité de causer des réponses comportementales, et de causer d'autres états mentaux, et cela en

---

[2] La systématicité, selon Evans, consiste en ce que, si je maîtrise le concept F et que je sais l'appliquer à un individu x, et que je maîtrise le concept G et sais l'appliquer à un individu y, alors je peux aussi comprendre et évaluer le résultat de l'application du concept F à l'individu y et du concept G à l'individu y. La compositionnalité consiste en ce que, si je peux utiliser la formule F(x) et évaluer sa valeur de vérité, et de même pour la formule G(y), alors je peux relier les deux par un connecteur $*$ qui les compose, et la valeur que j'attribuerai au composé $F(x) * G(y)$ dépendra seulement des valeurs des composants et de la règle liée au connecteur, qui combine ces valeurs. Il m'est donc à la fois possible et nécessaire de pouvoir retrouver dans le composé les valeurs des composants, sans qu'elles aient été modifiées par cette composition. Autrement dit, les propriétés des composants restent invariantes par la composition. Et la composition n'est en principe pas limitée.

[3] On trouve un exposé très clair et précis des différents problèmes du fonctionnalisme dans Elizabeth Pacherie, *Naturaliser l'intentionnalité*, PUF 1993.

liaison avec des stimuli. Ces rôles fonctionnels peuvent être réalisés selon une liste ouverte de modalités physiques. Le réductionnisme est donc réduit à la notion de «survenance» : il n'y a pas de différences mentales sans différences physiques, mais les types utilisés dans les explications mentales n'ont généralement pas un unique type physique qui leur corresponde, mais une liste ouverte de types physiques. Dans un premier temps le modèle du fonctionnalisme était la Machine de Turing Universelle, dont les fonctions ou rôles computationnels sont réalisables dans des machines très différentes. On pouvait ainsi amalgamer les deux sens du mot «fonction», fonction au sens mathématique, et fonction au sens téléologique, ou téléonomique au sens de Monod, dans la mesure où on considère un programme comme devant aboutir à un terme qui est sa fin. Mais une Machine de Turing a l'inconvénient de faire des différences entre des états mentaux qui produisent des sorties différentes de manière contingente, alors qu'ils sont par hypothèse identiques. Par ailleurs, la multiréalisabilité de la machine deTuring est trop étendue : une foule pourrait réaliser le programme d'une Machine de Turing, cela ne voudrait pas dire que cette foule en tant que collectif a des états mentaux. Fodor a donc proposé de prendre comme niveau d'analyse des rôles fonctionnels un langage de la pensée, qui assure une parfaite correspondance entre sa structure syntaxique et sa stucture sémantique. Les rôles syntaxiques des formules de ce langage permettent donc de définir les rôles sémantiques. Cette proposition se heurte au problème de l'externalisme : même en supposant que moi et mon jumeau nous ayons tous deux les mêmes états internes, (psychologiques ou physiologiques) nous pouvons cependant habiter deux environnements différents, et donc renvoyer à deux référents qui diffèrent sinon par leur apparence, du moins par leur essence, donc par leurs conditions de vérité. Cette référence externe ne peut pas figurer dans les formules du langage de la pensée. Le contenu d'un état mental, c'est-à-dire ce qui de cet état mental peut se penser, se croire, se vérifier ou se falsifier, doit donc être un contenu «large», incluant la référence à l'environnement. Mais alors on ne peut plus définir le rôle fonctionnel, l'entité fonctionnaliste, en isolant un niveau fonctionnel cognitif ou mental.

Fodor a proposé de sauver partiellement le fonctionnalisme de cette contamination, en distinguant le contenu «large» qui permet de définir des conditions de vérité à partir de référents, du contenu «étroit», qui est une fonction d'un contexte dans un contenu «large». Ce contenu «étroit» va devenir une entité fonctionnaliste quand il prend la forme d'un «véhicule» cognitif, ce qui veut dire qu'il a à la fois des propriétés syntaxiques, sémantiques, et causales. Ces relations causales vont lui permettre de

passer une épreuve classique pour un état intentionnel, à savoir celui de la «méprise».

On doit pouvoir distinguer une représentation de «chamois» d'une représentation de «chamois ou rocher », et pour traiter la seconde de méprise, il faut pouvoir éliminer l'hypothèse que nous avions affaire dans ce cas à un détecteur qui ne commet aucune méprise, mais qui est activé par la disjonction «chamois ou rocher ». Pour ce faire, Fodor utilise la notion de dépendance causale asymétrique : la relation causale par laquelle telle propriété cause en nous la représentation «chamois couché ou rocher » quand nous avons affaire à un rocher dépend causalement de la relation causale par laquelle les propriétés des chamois couchés causent en nous des représentations de chamois couchés, au lieu que la relation causale par laquelle les propriétés des chamois couchés produisent des représentations de chamois couchés dépende de la relation causale par laquelle sont causées en nous des représentations de "chamois couché ou rocher". Un état qui passe ce test, qui donc présente cette dépendance causale asymétrique, et qui par ailleurs présente la propriété plus générale de robustesse au sens de Fodor, à savoir que, justement, un rocher peut parfois provoquer une représentation de chamois, et donc qu'un état intentionnel peut signifier autre chose que sa cause effective, sans perdre sa capacité à signifier, est un bon candidat comme état mental intentionnel (Fodor, 1987 et 1990).

Il nous semble que, si l'on fait momentanément abstraction du problème de la systématicité et de la compositionnalité, les systèmes connexionnistes satisfont les définitions fonctionnalistes de Fodor. On retrouve le double sens de la notion de fonction. Les réseaux sont bien des êtres mathématiques qui approximent des fonctions. Par ailleurs leurs sorties sont comparées à des sorties souhaitées, et leur apprentissage les aligne sur cette finalité fonctionnelle. On retrouve aussi la «multiréalisabilité». Pour une fonction donnée, la liste des réseaux qui l'approximent est ouverte.

On peut aussi dans un réseau distinguer entre des états dispositionnels, et des états occasionnels. En effet, la propriété intéressante des réseaux, c'est qu'ils peuvent remplir plusieurs fonctions, parce qu'il suffit de superposer (dans le cas le plus simple d'additionner) les différents poids de connexion qui permettent de réaliser les diverses fonctions correspondant à différentes entrées pour que le réseau puisse calculer chacune de ces fonctions quand on lui donne l'entrée correspondante. Les états dispositionnels sont donc définis par l'architecture du réseau et les poids sur ses connexions, les états occasionnels par le fonctionnement du réseau quand on lui injecte tel vecteur en entrée. Dans une représentation

en termes de Machine de Turing, les états dispositionnels ne sont pas seulement les transitions entre états de la Machine, mais aussi les symboles inscrits sur la bande, qui peuvent jouer et le rôle de valeurs occasionnelles et celui de commandes. Dans un réseau, il faut ajouter un processus d''apprentissage'', pour que les vecteurs de sortie puissent déclencher des changements de poids, donc de fonctionnement. Un autre point peut nous faire espérer des capacités intentionnelles (si l'intensionnalité est un premier pas vers l'intentionnalité) : l'individuation d'un réseau n'est pas purement extensionnelle. Deux réseaux qui calculent la même fonction peuvent cependant différer par la structure de leurs opérations (leur architecture, le nombre d'unités, le nombre de couches d'unités cachées, et bien sûr leurs poids). Cette individuation par les opérations peut être qualifiée d'intensionnelle. Mais par ailleurs les réseaux permettent de retrouver des équivalences extensionnelles (les fonctions calculées, la liste des couples d'entrées et de sorties) au travers de leurs différences intensionnelles. Et si d'aventure les sorties de calcul d'une même fonction différaient d'une manière contingente parce que du bruit se serait introduit, les réseaux sont justement des machines à filtrer ce bruit et à restaurer l'équivalence originelle.

Les réseaux constituent bien des «véhicules» au sens de Fodor. Ils ont en effet des propriétés syntaxiques (la règle de la superposition des poids, le calcul des fonctions à seuils, etc.), des propriétés sémantiques (ce sont, nous y reviendrons de manière plus précise, les structures d'évolution des partitions faites par le réseau dans l'espace de ses états), et des propriétés causales, puisqu'ils produisent certaines sorties qui peuvent être reliées à des effecteurs, et qu'ils produisent par ailleurs certains effets sur d'autres réseaux qui leurs sont reliés. On vante souvent la ''robustesse'' des réseaux, c'est-à-dire leur capacité à répondre par l'état correspondant à la propriété F alors que le signal d'entrée présente des traits qui sont bruités, lacunaires, et qui ne sont pas vraiment ceux d'un objet doté de F. La notion de ''robustesse'' de Fodor semble plus forte : l'état peut signifier alors même qu'il signifie autre chose que la chose qui le cause. Mais si l'on associe la robustesse connexionniste, et la capacité de mémoire associative d'un réseau, on a bien une propriété voisine de la robustesse de Fodor, puisque l'entrée qui évoque dans le réseau la réponse désignant la chose C possédant la propriété F peut par hypothèse être lacunaire concernant la propriété F. Cette réponse est donc induite par d'autres objets et propriétés, et ce parce la causalité de ces objets et propriétés est relayée par les patterns mémorisés par le réseau.

Enfin, les réseaux semblent satisfaire la propriété de dépendance asymétrique. Soit le réseau n'est pas capable de séparer F de G, et il répond donc seulement à la disjonction ''F ou G''. Mais s'il en est capable,

alors il a dû modifier les poids de ses connexions de manière à répondre à F par un état interne "A" (et à répondre à G). Examinons alors le fait que l'état interne «A» soit causé par une disjonction présentant la propriété «F ou G». Cela exige que le vecteur d'entrée présente soit la propriété F, soit la propriété G, soit les deux. C'est évidemment le fait que le réseau ait "appris" à répondre à la propriété F, donc sa dépendance causale par rapport aux exemples de son ensemble d'apprentissage présentant la propriété F, qui explique qu'il puisse répondre "A" alors qu'éventuellement on ne lui présente pas d'F, (mais une disjonction F ou G) et non le fait qu'il ait été causalement réorienté en cours d'apprentissage selon la disjonction F ou G qui explique qu'il réponde "A" quand on lui présente F. Certes, on peut faire commencer l'apprentissage du réseau par celui de la disjonction F ou G pour ensuite lui faire distinguer F, si bien que la disjonction aura un effet causal (comme origine de l'apprentissage) sur la capacité de distinguer F. Mais c'est seulement une fois F distingué que l'on pourra parler de méprise quand le réseau donnera une réponse "A" à la disjonction F ou G, ce qui satisfait bien la relation de dépendance asymétrique.

Les réseaux nous permettent même de répondre mieux que Fodor à une objection qu'il fait lui-même à cette notion de dépendance asymétrique, celle de la dépendance des lois de niveau supérieur par rapport aux lois de niveau inférieur. Les propriétés syntaxico-sémantiques vont évidemment dépendre des lois inférieures, par exemple physiques, mais à ce niveau inférieur ces propriétés ne sont pas pertinentes. Fodor pose en principe que les *types* des deux niveaux sont différents (ce qui est présupposer une forme forte de fonctionnalisme). Dans un réseau on voit très bien quelle est la différence, et on peut l'expliquer selon les principes constitutifs mêmes du réseau : au niveau inférieur, nous trouvons les unités et leurs connexions, au niveau supérieur nous trouvons le comportement global du réseau (l'ensemble de ses états) et les relations entre ses entrées et ses sorties. Dans les réseaux, la syntaxe est pour l'instant au niveau inférieur, la sémantique au niveau supérieur, et la relation entre les deux fait tout l'intérêt d'un système connexionniste.

Ainsi les réseaux satisfont les contraintes que Fodor, leur adversaire, propose comme solutions aux problèmes que rencontre le fonctionnalisme quand il veut rendre compte de l'intentionnalité! Bien entendu, cela n'enlève rien aux objections de Fodor contre les réseaux. Mais de deux chose l'une : ou Fodor a raison dans sa théorie du contenu étroit, mais alors il faut qu'il accepte que les réseaux satisfassent les requisits de cette théorie ; ou bien il a raison dans sa critique des réseaux, mais alors il doit considérer le fait que les réseaux satisfont les requisits de sa théorie du contenu étroit comme un indice de ce que cette théorie n'est pas satisfaisante. Il reste d'ailleurs la possibilité que les critiques de

Fodor contre le connexionnisme ne soient pas suffisantes, et que les requisits en question ne le soient pas non plus.

Davantage, les partisans de l'externalisme, c'est-à-dire de la thèse qu'il n'existe pas de contenu étroit, mais seulement un contenu large, voient les réseaux d'un oeil favorable. Supposons que notre fonctionnement cognitif interne soit celui des réseaux de nos neurones. Aucun des états internes du réseau n'est par lui-même digne de l'appellation de «contenu». C'est seulement quand on lie les sorties du réseau au contexte que présente l'environnement de référence (qui fournit aussi au réseau ses entrées) qu'on peut assigner des conditions de vérité et donc définir un contenu. Cependant, cette position ne se préoccupe pas de savoir si les états de sortie du réseau permettront d'un contexte à l'autre de transposer des représentations ou des modes de traitement de l'information, donc s'ils seront capables d'intégrer des références différentes dans des catégories identiques.

Avant de revenir sur le problème de la compositionnalité, notons que les réseaux nous permettent aussi de clarifier les choses si l'on se place dans l'autre perspective «fonctionnaliste», celle de Millikan et de Dretske, qui prend la notion de fonction au sens biologique du terme, et y voit la structure causale, propre à un organisme, et qui est la cause de son avantage sélectif dans l'évolution[4]. Ainsi Dretske pense que les magnétosomes que possèdent certaines bactéries sont la cause de ce qu'elles se tournent vers le nord et le bas, qui est la zone de leur océan où il y a le moins d'oxygène (quand ce sont des bactéries de l'hémisphère nord) (Dretske, 1986). Ces bactéries sont sélectionnées par l'évolution pour cette cause, puisqu'elles ne peuvent survivre que dans un milieu pauvre en oxygène. Le problème qui se pose alors est celui de l'indétermination fonctionnelle : les magnétosomes ont-ils pour fonction de désigner la zone pauvre en oxygène, ou bien simplement la direction du pôle magnétique?

---

[4] Signalons ici seulement pour mémoire les différences entre la conception de Millikan et celle de Dretske. Selon Millikan un organisme O ayant le caractère C a la fonction propre F ssi si C a eu un rôle causal pour ses ancêtres, et que le fait que l'organisme O existe dépend causalement de ce rôle causal (soit parce que le caractère C permet la reproduction de O, soit parce qu'il a donné à ses ancêtres un avantage sélectif). Millikan ne ne s'intéresse donc qu'à l'histoire de l'évolution pour définir la fonction d'un organe. Dretske au contraire s'intéresse aussi à l'apprentissage de l'organisme. Il définit une fonction d'indication d'un état E dans un organisme O, supposant une corrélation statistique (et non causale) proche de 1 entre une propriété F et cet état de l'organisme, état interne qui contrôle un comportement C. Un apprentissage par renforcement du comportement C permettrait de recruter l'état E à cause de son contrôle de ce comportement .

Or ce problème de l'indétermination fonctionnelle se pose aussi pour les réseaux. On peut imaginer qu'une distribution de poids sur les connexions d'un réseau puisse avoir pour effet de calculer ce qui est dans un cas de figure une classification grammaticale, et dans un autre cas de figure une reconnaissance de forme. Quelle est alors la fonction du réseau? Pour la déterminer, il faut faire varier soit la distribution de poids, soit les vecteurs d'entrée. Si en faisant varier le vecteur d'entrée, mais de telle façon qu'il propose toujours un problème d'analyse grammaticale, le réseau ne peut plus donner cette analyse, alors qu'il peut toujours faire de la reconnaissance de forme si on fait varier le vecteur d'entrée qui proposait ce problème de formes, alors la fonction du réseau est d'abord de reconnaître des formes, et accessoirement et par coïncidence, de faire de l'analyse grammaticale. Si de même, en faisant varier légèrement les poids, on détruit les performances du réseau dans un domaine mais pas dans l'autre, c'est dans ce dernier domaine que se définit sa fonction[5].

La conclusion à en tirer semble avoir une portée plus générale. Supposons que ces fonctions définissent une sémantique et permettent ainsi de naturaliser l'intentionnalité (c'est la thèse de Dretske). Puisque nous ne pouvons lever (partiellement) l'indétermination fonctionnelle qu'en procédant à des variations des entrées et du fonctionnement du système, c'est qu'une sémantique ne peut pas se définir de manière statique, mais qu'elle suppose des capacités de variations dirigées. C'est simplement une autre manière de dire que l'intentionnalité, qui consiste à ne viser un référent que sous un aspect donné et pas sous un autre pourtant éventuellement coextensif, ne va pas sans apprentissage. Cet énoncé abrupt semble entaché de circularité, puisque qu'un apprentissage consiste justement à réussir à reconnaître un objet ou une situation sous un certain aspect intentionnel. Mais les systèmes connexionnistes nous ouvrent une piste pour définir un apprentissage de manière pré-intentionnelle, un apprentissage*, et pour ensuite définir l'intentionnalité à partir de cet apprentissage*.

L'apprentissage* consisterait à faire varier la fonction satisfaite par le réseau dans sa mise en corrélation d'entrées et de sorties, tout en continuant à maintenir des dispositions à satisfaire des approximations d'autres fonctions. Ces dispositions sont simplement les poids des

---

[5] Cette possibilité de variation semble sous-jacente dans la notion de dépendance causale asymétrique de Fodor, mais, bien qu'il recourre à une théorie des contrefactuels, qui fait appel à des mondes possibles et à une fonction de sélection qui permet de choisir parmi les mondes possibles les plus proches du monde actuel, ce qui va de pair avec des possibilités de variations qui nous font passer d'un monde possible à un monde possible proche, il ne développe pas cette idée de variation.

connexions, qui, grâce à la superposition des poids, peuvent activer d'autres fonctions quand d'autres entrées se présenteront. Dans le processus nommé apprentissage par les connexionnistes, le réseau va peu à peu raffiner sa classification des entrées de son ensemble d'apprentissage, ce qui revient à réaliser une fonction caractéristique, qui envoie les entrées appartenant à telle classe sur une valeur de sortie et celles appartenant à telle autre sur une autre valeur. Supposons que le réseau apprenne à réaliser cette fonction. Apprendre se réduit donc d'abord ici, par exemple, à appliquer un algorithme de rétro-propagation. Le réseau peut le faire en réalisant une opération parmi toute une classe d'opérations qui réalisent la même fonction. Ces opérations (les poids de ses connexions, ses fonctions à seuils, etec.) permettent au même réseau de réaliser d'autres fonctions : on peut considérer en effet que la généralisation de la classification à un autre ensemble d'entrées que celles de l'ensemble d'apprentissage consiste à réaliser une autre série de correspondance entre entrées et sortie, donc une autre fonction. La superposition des poids permet en général d'envisager la réalisation d'autres fonctions. Certaines autres fonctions peuvent exiger des modifications des poids supplémentaires. Ces modifications sont plus ou moins importantes, et elles altèrent plus ou moins la capacité du réseau à réaliser de manière approchée ses premières fonctions. On pourrait donc sélectionner parmi les opérations possibles réalisant une fonction celles qui exigent le moins de modifications pour réaliser une autre fonction. Cet apprentissage de second degré est un apprentissage*. Car le réseau qui en serait doté pourrait corriger l'apprentissage de la première fonction en fonction des contraintes imposées par la seconde. Il identifierait une première classe d'entrées en fonction d'une seconde classe. Il "qualifierait" cette première classe par la seconde. Il imposerait donc les contraintes de la seconde classe comme "aspect", ou "modalité" aux items de la première classe. Or imposer des "aspects", "corriger", c'est être fidèle à un "mode de présentation" plutôt qu'un autre. Or dès qu'il y a mode de présentation et correction, il y a intentionnalité. Mais ces corrections là ont pu être induites sans présupposer l'intentionnalité, simplement par l'ajustement réciproque de deux modes de variations.

II. LES RESEAUX ET LEURS CATEGORISATIONS

Les réseaux semblent nous permettre de développer des thèses plus riches que le fonctionnalisme. Mais pour le faire, il nous faut d'abord lever l'hypothèque de la critique de Fodor : les réseaux ne présentent ni systématicité, ni compositionnalité, ils ne peuvent donc produire de sémantique satisfaisante.

La communauté connexionniste a imaginé bien des réponses à cette attaque. Les plus intéressantes nous semblent être celle de Van Gelder et celle de Sharkey, bien qu'elles ne nous satisfassent pas complètement. Ces auteurs vont prendre acte du fait que les réseaux ne réussissent pas à totalement satisfaire les exigences de la compositionnalité entre symboles. La pierre d'achoppement, en l'occurrence, n'est pas la composition, mais l'exigence que la décomposition, une fois la composition faite, retrouve les unités constituantes. À cet égard, les tentatives de Smolenski (1990) et de Shastri (1993), par exemple, ont des limites. Smolenski a proposé de représenter le croisement des rôles et des teneurs de rôles (des *fillers*) en utilisant le formalisme du produit tensoriel. Il suffit ensuite de multiplier le tenseur par le vecteur rôle pour retrouver le vecteur des teneurs de rôles et réciproquement. Mais cela ne marche que si les vecteurs sont linéairement indépendants, ce qui interdit au réseau d'apprendre des dépendances intéressantes. De plus, la multiplication des unités qu'exige la réalisation connexionniste du tenseur est très lourde. Shastri a proposé de simuler la liaison de variables par la mise en synchronie de patterns représentés par différentes phases, chaque structure de phase étant assignée à des rôles différents (rôles de prédicats ou rôles d'arguments). Mais le réseau ne fait pas lui-même la distinction entre ces rôles, et donc la structure n'est pas apprise par le réseau, mais importée.

Ce que proposent Van Gelder et Sharkey, c'est de définir une notion plus large de compositionnalité, et de montrer que si les réseaux ne satisfont pas les exigences de la compositionnalité par concaténation de symboles, ils satisfont les exigences d'autres compositionnalités. Van Gelder définit ainsi une composition «fonctionnelle», qui n'est pas forcément concaténative : il faut et il suffit qu'on puisse disposer de procédures fiables pour produire une expression une fois ses constituants donnés, et pour retrouver à partir de l'expression ses constituants (Van Gelder, 1993). Il a simplement le tort, dans un premier article (Van Gelder, 1990), de soutenir que la numérotation de Gödel est une composition fonctionnelle et non pas concaténative [6]. Elle est en fait à la

---

[6] Cette numérotation consiste effectivement à partir de nombres premiers qui sont les codes de chaque symbole de base, et à les multiplier entre eux, en s'appuyant pour assurer la décomposition sur l'unicité de la décomposition en facteurs premiers. Mais cela seul ne permettrait pas de retrouver la structure de l'expression (on pourrait avoir des parenthèses disposées au hasard, par exemple), et il faut donc bien tenir compte dans cette numérotation de l'ordre concaténatif des symboles, chaque code de symbole étant par exemple porté à une puissance dont le degré indique sa place dans la séquence.

fois fonctionnelle et concaténative[7]. Mais il reste qu'il peut exister des compositions fonctionnelles non concaténatives.

Sharkey propose une expérience pour montrer que les réseaux permettent composition et décomposition (Sharkey et Jackson, 1994). Qu'ils permettent des compositions, c'est évident, puisqu'ils le permettent doublement. D'abord en superposant différents poids sur la même connexion, ils permettent à l'unité cachée de répondre à un type d'input et à un autre type, ce qui est une première composition. Mais surtout, les données de chaque unité input sont envoyées vers chaque unité cachée, si bien que chaque unité cachée compose les activations qui lui viennent de toutes les unités inputs (elle les compose d'abord en les additionnant, puis en imposant à cette somme une fonction à seuil). Cela permet à un réseau d'être «sensible au contexte» de toute la suite de signaux qui lui parvient sur un vecteur d'entrées. Le problème qui se pose est donc toujours celui de la décomposition. Sa difficulté tient justement à ce que, pour permettre des classifications intéressantes, les réseaux doivent être dotés d'unités qui calculent non seulement une somme, mais une fonction à seuil sur cette somme. On introduit ici une non linéarité. Cela veut dire que l'influence d'une information en provenance d'une unité d'entrée va être très différente selon qu'elle intervient près du seuil – elle va faire basculer l'unité au delà de son seuil et donc faire transmettre l'activation vers les sorties – ou bien loin du seuil. Et cela dépend bien entendu des contributions des activations en provenance des autres unités d'entrée. Autrement dit, la contribution d'une unité d'entrée au comportement de l'unité cachée dépend non linéairement de la contribution des autres unités. Si bien que le résultat donné par l'unité cachée fusionne ces contributions, et que la décomposition semble bien hasardeuse.

Malgré ce handicap, Sharkey trouve un moyen de décomposition. On prend pour une unité cachée la somme pondérée des activations en

---

[7] Oden fait quelques remarques plus incisives (Oden, 1994). Il montre que l'aspect «distribué» de ce qu'on peut tenir pour des «représentations» dans un réseau (par exemple les corrélations entre les activations des unités cachées et les sorties correctes pour des entrées données) n'est pas propre aux réseaux. La numération décimale, ou encore la numération binaire, qui ont indubitablement un aspect concaténatif, sont en un sens distribuées. Les 1 et les 0 sont comme les activations des unités cachées du réseau, ils n'ont pas de sens en eux-mêmes, mais seulement dans l'ensemble que constitue par exemple 101 (5). On pourrait avoir une représentation localiste, dans laquelle le 1 désignerait toujours l'arrêt du comptage, et chaque 0 désignerait toujours une application supplémentaire de la fonction successeur. Ainsi 5 se lirait 000001. A contrario, on ne peut pas d'emblée prétendre que l'aspect «distribué» des représentations connexionnistes exclut toute compositionnalité.

provenance des unités d'entrée. On calcule la même somme pondérée, en fixant à 0 la contribution d'une unité d'entrée donnée. La différence entre les deux sommes nous donnera la contribution effective de cette unité d'entrée. Sharkey améliore le procédé en sommant toutes les contributions individuelles ainsi repérées, et en calculant la différence de cette somme avec la somme effective propre à l'unité cachée. Cette différence sert de «base». On va alors procéder à des compositions de ces décompositions, pour vérifier si, en imposant aux unités cachées des sommes de ces nouvelles contributions individuelles recombinées, on obtient toujours des sorties correctes. Pour calculer ces nouvelles sommes sur les unités cachées, on recombine donc les contributions des unités d'entrées, de manière variée, et à chaque fois on ajoute la «base» à leur somme. Et on observe effectivement que le réseau a des performances classificatoires qui sont très proches des performances précédentes. Autrement dit, on a pu réaliser une décomposition des «représentations» connexionnistes, et les recomposer de telle manière que leurs propriétés fonctionnelles (leur sensibilité à la structure des entrées) se conservent.

Cette compositionnalité reste cependant limitée, et n'a pas la généralité de la compositionnalité symbolique. En effet si le procédé fonctionne, c'est qu'on reste soigneusement à l'intérieur du même domaine sur lequel a été précédemment réalisé l'apprentissage du réseau. L'ensemble des poids sur les connexions du réseau est celui qui a finalement permis au réseau, après apprentissage, de faire les bonnes classifications par rapport aux structures présentées en entrées. On ne crée pas de nouvelles structures, mais simplement des combinaisons différentes, voisines de celles qui étaient déjà présentes dans l'ensemble d'apprentissage. Dès lors on peut considérer que les activations que chaque recombinaison envoie sur l'unité cachée sont additionnables ensemble, sans que l'on sorte des domaines de classification que produit la conjugaison des unités cachées et des connexions vers les unités de sortie. Cela, on ne l'a pas changé, et on a même tenu compte de la contribution «distribuée» propre à ce domaine d’apprentissage, en calculant cette «base» qui est maintenue dans les nouvelles combinaisons. Donc, tout en reconnaissant que le résultat de Sharkey est surprenant, son procédé même implique que les décompositions et recompositions se produisent toujours à l'intérieur d'un domaine d'apprentissage déjà stabilisé. Ses recompositions ne sont au mieux que des tests pour la capacité de généralisation du réseau sur des exemples qui n'auraient pas figuré dans l'ensemble d'apprentissage, mais qui en seraient voisins. Et on sait que les réseaux peuvent généraliser, dans une certaine mesure. Cela ne permet pas la générativité illimitée de la compositionnalité classique.

Mais cette réserve doit-elle même être atténuée. La compositionnalité classique ne reste elle-même valide que dans un contexte d'apprentissage bien déterminé : celui de la maîtrise d'un langage symbolique. Or le langage naturel va plus loin que ce domaine là. En effet, nous ne pouvons pas garantir que n'importe quelle combinaison syntaxiquement correcte de constituants aura encore un sens. Pour ce qui est du sens, le domaine de validité semble toujours limité à un domaine stabilisé d'apprentissage. Certes, ce domaine de validité semble bien plus large dans un langage symbolique qu'il ne l'est pour un seul réseau connexionniste, mais la même restriction semble s'imposer. Nous ne pouvons être assurés que toute information qui nous parviendra sous forme perceptive pourra être intégrée dans une catégorisation signifiante déjà prête. Il nous faut tenter l'apprentissage qui va essayer de construire cette catégorisation avant d'être assurés que nous y parviendrons.

III. UNE SEMANTIQUE POUR LES RESEAUX ?

Cependant, il n'est pas évident que nous devions prendre au pied de la lettre le procédé de recombinaison proposé par Sharkey, et faire de ses contributions d'unités d'entrées à une unité cachée les unités représentationnelles, les constituants propres aux réseaux. Cela nous ramène au problème de savoir ce qui mérite le titre de «représentation » dans un réseau. Les propositions varient dans la littérature. Bien souvent, on considère que les représentations d'un réseau sont les activations de ses unités cachées. On procède alors à une analyse par *clusters*, par regroupement deux à deux des activations les plus proches, et on les met en rapport avec les réponses données en sortie. On procède aussi à des analyses en composants principaux de ces mêmes activations. L'idée de base est qu'une unité cachée qui est très active pour une corrélation entre entrée et sortie et peu pour d'autres «représente» cette corrélation. Bien entendu, ces représentations sont «distribuées», et il est rare qu'une unité soit à elle seule le détecteur de telle propriété. Le problème de ces analyses, c'est qu'elles ne tiennent compte que des activations et pas des poids sur les connexions, qui sont pourtant les variables de contrôle du réseau. On peut prétendre que l'efficace des poids se manifeste dans les activations des unités cachées. Pour être rigoureux, il faudrait donc tenir compte des activations et de toutes les connexions, y compris celles qui mènent des unités cachées aux unités de sortie[8]. Mais même alors on n'aurait pas identifié ce que sont les opérations catégorisatrices d'un

---

[8] Sharkey ne tient compte que des connexions des unités d'entrées aux unités cachées.

réseau. Les réseaux nous obligent en effet à passer du vocabulaire statique des représentations au vocabulaire dynamique des opérations.

En effet, l'intérêt d'un réseau, c'est l'évolution qu'il subit au cours de son apprentissage. Un article de Harnad, Hanson et Lubin montre très clairement la chose[9]. Dans ce travail, les auteurs entraînent d'abord leur réseau à faire de l'auto-association, c'est-à-dire à redonner en sortie le pattern d'entrée. Puis dans une seconde phase, le réseau doit donner aussi une catégorisation, c'est-à-dire une partition en deux classes, par exemple, des signaux d'entrée. Durant chaque phase, on observe l'évolution des activations des unités cachées. Quand le codage est aléatoire, le réseau part d'activations des unités cachées qui sont proches les unes des autres, pour évoluer vers des activations aussi différentes que possibles (si on restreint le schéma à trois unités, on atteint les sommets du cube qui représente les activations pour chacune de ces unités, ou encore les milieux des arêtes, une fois que les sommets sont occupés). Quand on donne comme entrées au réseau des signaux de codage non aléatoire, qui respectent un isomorphisme, la phase d'auto-association est déjà contrainte, et bien que l'évolution des activations des unités cachées conduise à leur différenciation, ces différenciations n'explorent pas toutes les valeurs de l'espace des états, et restent en l'occurrence consignées sur trois faces seulement du cube. La phase de catégorisation, au contraire, force certaines de ces activations très différenciées à se rapprocher l'une de l'autre, de manière à ce que l'on puisse faire couper le volume du cube par des plans qui séparent les différentes catégories. On peut donc prétendre que les réseaux représentent une classification non pas simplement par leurs états, mais par l'évolution des variations de leurs états, et plus précisément par les biais, les contraintes imposées à ces variations.

Un des intérêts de cette présentation, c'est de rapprocher le problème de savoir ce que sont les représentations pertinentes d'un réseau comme celui que nous venons d'évoquer, un réseau *feed forward*, où l'activation monte des unités d'entrée en passant par les unités cachées vers les unités de sortie (les réseaux dits récurrents ajoutent simplement à cette montée une boucle retour vers d'autres unités d'entrées), de ce même problème, mais posé à propos des réseaux de Hopfield, où toutes les unités jouent à

---

[9] Même s'il est centré sur un problème adjacent, celui de l'augmentation des distances inter-catégorielles et de la diminution des distances intra-catégorielles au cours de l'apprentissage (Harnad et al, 1994). On pourrait aussi relire dand cette perspective l'article de Elman, (1991) "Representation and structure in connectionnist models" in G. Altmann (ed.) *Cognitive Models of Speech Processing*, Cambridge, Mass., MIT Press.

la fois le rôle d'entrées, d'unités cachées et de sorties, puisqu'il y a des boucles entre toutes les unités. La mise en branle d'un tel réseau le fait évoluer vers un attracteur, c'est-à-dire un état global du réseau dans lequel un cycle d'opération de plus le maintient. Comme un réseau, grâce à la superposition de différents poids sur ses connexions, a plusieurs attracteurs, on peut faire une analogie entre les attracteurs des réseaux de Hopfield et les regroupements différenciés des activations du réseau *feed forward*.

Dans les deux cas, on voit que ce qui nous intéresse, ce ne sont pas les états finaux des deux réseaux, leurs attracteurs ou les regroupements de leurs activations, mais bien l'évolution de ces états. En effet, les réseaux, face à des entrées aléatoires, vont de toute façon aboutir à des états différenciés. L'existence de ces états différenciés ou de ces attracteurs n'a donc pas grande pertinence sémantique, si nous posons que la sémantique doit être coordonnée aux propriétés structurelles des informations reçues par le réseau. Cette exigence de coordination des structures formelles des réseaux avec les structures qu'ils reçoivent, disons, pour simplifier, du monde réel, correspond à ce que la littérature baptise du nom de «grounding», ou d'«enracinement» (enracinement des significations dans les données perceptives et motrices). Ce qui est donc pertinent, c'est le «gauchissement» des différenciations des réseaux par rapport aux différenciations qu'ils produisent à partir de signaux aléatoires. Ce qu'un réseau apprend, la façon dont il représente la structure présente dans l'information qu'on lui donne, ce sont des biais de sa classification par rapport à une classification qui assigne à chaque item différent en entrée une position de différenciation maximale dans l'espace des états du réseau. Cette caractéristique des réseaux correspond d'ailleurs à une de leurs limites quand on les considère comme des outils d'analyse statistique : ils ne peuvent à la fois rendre compte de toute la variance de l'information, et repérer les biais structurels dans cette information. Rendre compte de la variance, c'est différencier au maximum (comme si les signaux en entrée étaient aléatoires), repérer les biais, c'est rapprocher certaines différenciations les unes des autres (Geman et al, 1992).

Ces biais d'évolution des réseaux sont donc de bons candidats pour définir les «représentations» d'un réseau, et même pour assurer l'enracinement de ces évolutions «syntaxiques» du réseau dans une structuration des données qui assurerait une «sémantique» aux états du réseau (plus exactement, aux évolutions orientées de ces états). De même, dans les réseaux de Hopfield, nous ne devrions pas simplement considérer les attracteurs en eux-mêmes, mais leurs variations quand on change les entrées ou qu'on change les poids du réseau. Les représentations ne seraient pas les attracteurs, mais les évolutions des frontières qu'il

dessinent dans le paysage des états du réseau. On retrouve ici les idées de Petitot, ainsi que certains modèles de la perception des formes (Petitot, 1992). Ainsi nous parcourons du regard la surface d'un rectangle en pointant précisément sur les lignes de rencontre des fronts de diffusion à partir des côtés du rectangle (son *cut locus*). Les évolutions de ces attracteurs doivent mettre en évidence (et respecter) des structures associées aux formes perçues.

Mais si l'on peut à bon droit qualifier ces évolutions d'états de «représentations», peut-on leur associer une sémantique? Quelqu'intérêt que l'on porte aux grammaires cognitives et à leur hypothèse localiste, qui consiste à soutenir que les relations mises en jeu dans une sémantique ont comme origine les relations spatiales, nos exigences en sémantique semblent aller au delà des propriétés que nous devons jusqu'à présent reconnaître aux représentations des réseaux. Peut-être est-ce une illusion, mais nous pensons que les catégories sémantiques doivent pouvoir être transportables d'un contexte d'application à un autre (du moins ce serait le cas d'un bon nombre de verbes, et d'un bon nombre de prédicats abstraits), même si cette transposition n'est pas illimitée. Or rien ne garantit que les catégorisations apprises par un réseau sur un contexte d'apprentissage donné, c'est-à-dire les biais introduits dans ses différenciations de base, soient réutilisables dans un autre contexte. Plusieurs cas peuvent alors être envisagés : soit ces biais permettent, sans autre évolution, de donner des sorties qui tiennent compte d'une structure présente dans les données. Soit ces biais ne le permettent pas, et deux cas sont possibles : le réseau peut les faire évoluer sans difficulté pour apprendre de nouveaux biais, qui, eux permettent de rendre compte de la structure; ou bien les biais du premier apprentissage empêchent le réseau de pouvoir faire évoluer les biais du second de façon satisfaisante. Le second cas de figure ne fait que reporter le problème qui se reposera dans les mêmes termes lors de l'apprentissage suivant. Or rien ne nous permet d'assurer que le réseau satisfera l'un des deux deux premiers cas de figure, quel que soit le contexte d'apprentissage. Les capacités de transposition des représentations du réseau, et donc ses capacités de systématicité et de compositionnalité restent restreintes à un domaine limité, que l'on qualifiera de contextuel.

Mais c'est là nous semble-t-il le lot de toute sémantique qui veut satisfaire des contraintes de pertinence. Un des effets de la prise en compte de la pertinence, c'est par exemple qu'est mise en défaut la propriété de monotonie des logiques classiques, c'est-à-dire qu’il n'est plus toujours vrai qu'on puisse ajouter une prémisse «B» à la prémisse «A», donc de «A» passer à «A & B» tout en continuant à tirer de «A» (et donc de «A &B») la conclusion «C». Cela se produit quand «B» est un

nouveau contexte, et que l'inférence qui concluait à «C» était dépendante du contexte. «B» change la pertinence de l'inférence de «A» à «C». On voit clairement sur cet exemple qu'introduire des contraintes de pertinence, c'est réduire et limiter la compositionnalité (ici, réduire la possibilité de composition par «&»). Cette dépendance de la validité des inférences par rapport à un contexte qui doit conserver la pertinence conduit à des logiques non monotones, c'est-à-dire qui peuvent réviser leurs conclusions. Or les réseaux connexionnistes (une sous-classe d'entre eux) peuvent vérifier des propriétés des axiomes proposés pour l'opération de révision (Gärdenfors, 1992). Et de fait on peut considérer tout apprentissage d'un réseau comme une opération de révision. Les représentations des réseaux ne sont que les différentes orientations des révisions possibles de la tendance foncière du réseau à obtenir un maximum de différenciation, en gauchissant la classification qu'il fait initialement sur les entrées. Les capacités d'enracinement d'un réseau dans la structure de ses entrées vont donc de pair avec sa sensibilité à un contexte, sa capacité à induire des biais qui orientent ses révisions. Mais une première révision peut en biaiser une seconde, et réduire la liberté des transformations et des recombinaisons. Et on ne peut pas satisfaire pleinement et à la fois les deux exigences de sensibilité au contexte, donc de pertinence, et celle de compositionnalité, de combinaison *ad libitum.* Les systèmes formels symboliques qui satisfont pleinement l'exigence de compositionnalité sans entraves (autres que celles des règles de combinaison) ne satisfont pas l'exigence de pertinence[10].

Mais si l'exigence de compositionnalité et celle de pertinence ne sont pas totalement compatibles, il reste l'exigence de systématicité : elle implique seulement que je peux toujours juger de la validité d'une nouvelle composition, donc du plongement d'une représentation dans un nouveau contexte, y compris le rejeter quand le résultat n'est pas pertinent. Or un réseau risque tout simplement de ne pas donner de réponse utilisable quand on plonge ses représentations dans un nouveau contexte. On devrait donc tenter d'assurer sinon la compositionnalité, du moins la systématicité des opérations ou représentations dans les réseaux. Il n'est sans doute pas possible de les doter d'une systématicité universelle, qui permette d'estimer le résultat du plongement d'une représentation dans un contexte, quel que soit le contexte et quelle que soit la représentation. Pour assurer une systématicité même limitée, il faut pouvoir corriger les biais propres à l'apprentissage d'un réseau dans un contexte par d'autres

---

10 Une autre preuve en seraient les «paradoxes» de l'addition logique dans les logiques déontiques, où la compositionnalité nous amène à admettre que s'il est obligatoire de poster la lettre, il est obligatoire de la poster ou de la brûler.

biais. Et il faut s'assurer que sur un domaine de variation des entrées vaste, mais évidemment toujours limité, les chaînes de biais que l'on propose reconduisent toujours à des sorties compatibles avec la structure des entrées. Si on considère les biais d'évolution du réseau comme ses opérations, il faut par exemple s'assurer de propriétés de triangulation entre les opérations d'un même réseau dans des apprentissages sucessifs et des contextes qui varient. Par triangulation, nous entendons que l'opération O1 combinée à l'opération O2 nous donne des résultats très proches de l'opération O3 qui part des entrées de O1. On peut considérer cela comme une transitivité approchée, ou bien relier cette exigence aux contraintes d'inégalités triangulaires nécessaires pour disposer d'une distance (qu'elle soit métrique ou ultra-métrique)[11]. On peut voir dans les biais liés à chaque opération d'un réseau des propositions de distances. Bien entendu imposer cette triangulation, c'est juger de la distance proposée par O1 et O2 dans les termes de la distance proposée par O3. Or on pourrait vouloir faire l'inverse, et rien initialement ne garantit que ces jugements soient cohérents. Trouver le moyen de rendre ces jugements relatifs cohérents entre eux, cela revient à en faire de vraies distances. Les réseaux ne peuvent pas satisfaire cet objectif dans l'absolu, mais seulement parvenir sur un domaine limité d'enchaînements entre opérations à une cohérence qui nécessite des approximations. Le domaine de pertinence des opérations classificatoires d'un ensemble de réseaux (ou des différentes versions d'un réseau dans le temps de ses apprentissages) est donc simplement défini par le domaine dans lequel on peut établir des enchaînements entre les opérations du réseau qui respectent des propriétés voisines de celles qui établissent une notion de distance. Si l'on préfère, dans ce domaine, les conjugaisons des opérations s'ordonnent entre elles de façon à ce que, de manière approchée, le domaine soit pavé par un réseau de triangulations entre opérations, et donc que l'on retrouve de manière approchée les résultats d'une suite d'opérations en procédant par un autre enchaînement d'opérations.

La combinaison de plusieurs réseaux semble être utile pour parvenir à de telles configurations de quasi-distances ou de biais compatibles entre elles. En effet, si on se borne à faire évoluer un seul réseau, les révisions ou les biais induits par chaque nouvel apprentissage risquent fort d'oblitérer les propriétés de sensibilité structurelle des opérations passées. Il faut donc disposer d'autres réseaux qui maintiennent en mémoire ces opérations passées, et qui ne soient donc pas en communication constante avec le premier réseau, mais seulement quand

---

[11] d(x,y) = d(x,z) + d(z,y) pour une distance métrique, d(x,y)= max(d(x,z), d(z,y)), pour une distance ultra-métrique.

ils sont eux-mêmes activés par certains patterns d'entrées. Il serait donc sans doute nécessaire de pondérer les biais d'un réseau par ceux d'un autre pour assurer cette cohérence, ce qui peut se faire en établissant des connexions transversales entre réseaux[12] , et en n'activant ces connexions que quand leur inhibition est levée pour certaines conditions. On retrouve ici des idées de Von der Malsburg et Bienenstock, qui proposent de construire des représentations non pas à partir des opérations ou activations d'un seul réseau, mais par développement de connexions transversales entre sous-réseaux, ces connexions transversales se construisant dès lors que les unités des sous-réseaux présentent une synchronie entre les chaînes temporelles d'activation de chaque sous-réseau. On retrouve aussi les constructions de Grossberg, qui relie deux sous-réseaux par des connexions récurrentes mais en imposant des «portes» sur ces connexions, des inhibitions qui ne sont levées que temporairement. Il est alors possible qu'une activation d'un sous-réseau par ses entrées déclenche une activation partielle du réseau transversal, qui à son tour induira une activation partielle d'autres sous-réseaux. Si de plus un autre réseau transversal, stimulé par d'autres patterns d'un sous-réseau, déclenche une autre activation partielle sur un autre sous-réseau, on pourra ainsi considérer l'ensemble des deux sous-réseaux contrôlés comme l'équivalent connexionniste d'une composition à partir d'opérations décomposées. Et l'on pourra contrôler l'évolution des biais d'un sous-réseau de telle manière qu'ils restent compatibles avec d'autres opérations, assurant ainsi une systématicité, même si elle est réduite à un contexte limité d'entrées et d'opérations.

IV. CONNEXIONNISME ET SURVENANCE

La survenance ou dépendance systématique est évidemment une composante du fonctionnalisme. On peut distinguer une survenance faible, une survenance forte, et une survenance «globale». La survenance forte implique que toutes les fois qu'une propriété mentale M intervient, non seulement une propriété physique P intervient, mais cette propriété P donne nécessairement lieu à la propriété M. Le problème est qu'il peut y avoir des propriétés physiques qui ont des relations contingentes avec des propriétés mentales (ainsi, l'accroissement du taux d'adrénaline fait bien changer nos émotions, mais sans déterminer quelle émotion nous allons avoir, ni être nécessaire à la perception de cette émotion). La survenance faible pose simplement qu'il est nécessaire que pour tout changement de

---

[12] Cf. ma communication aux Entretiens Jacques Cartier, "Catégorisation et connexionnisme", 2 décembre 1994, Lyon.

propriété M, il se produise un changement de propriété P (sans que le lien entre P et M soit forcément nécessaire). Mais cela semble trop faible. Le lien entre les propriétés physiques et les propriétés mentales est alors tellement lâche que l'on risque de tomber sous la critique de Kim : des deux explications, l'explication physique et l'explication mentale, il y en a une de trop, et il faut nous dire celle qui est en trop. Or dans le cas d'une survenance faible, ce peut être l'explication physique qui n'apprend rien, parce que son lien avec le phénomène mental n'est pas assez fort. On peut donc envisager une survenance dite «globale». Il vaudrait mieux en fait la dire « régionale », puisque cela revient à limiter la nécessité de la relation de survenance entre P et M à un contexte. La propriété P n'implique nécessairement la propriété M que dans un sous-ensemble de mondes, et non dans tous les mondes possibles (mais en général dans plus d'un monde, ce qui la différencie de la survenance faible). Dans les mondes pertinents, les mondes qui sont indiscernables quant à P ne peuvent être discernables quant à M. Si dans un monde on possède P, on y possède aussi M. En revanche, dans d'autres mondes, P n'est pas pertinent par rapport à M, ce qui implique que pour deux mondes discernables quant à P, il se peut qu'ils soient indiscernables quant à M, ou bien discernables, et alors aussi différents mentalement que l'on veut. On pourrait soutenir que la pertinence implique la relation à un contexte : dans ce cas, si on admet des propriétés P relationnelles à un environnement, alors deux individus identiques physiquement en eux-mêmes sont différents s'ils sont placés dans des environnements différents.

On peut rapprocher cette notion de survenance « globale ou régionale » du quantificateur que Schlechta a proposé dans les logiques de la révision, et dont l'interprétation est : «la plupart des x» ou «les x normaux ». Contrairement à un filtre, qui requiert que l'intersection de deux filtres soit aussi un filtre, on exige ici seulement que l'intersection de deux ensembles « normaux » soit non vide. Ce quantificateur ne s'applique donc à un seul individu (ou ici, monde) que si cet individu figure dans l'intersection de tous les autres ensembles comprenant «la plupart» des individus ou les individus « normaux ». Ce quantificateur permet de dire qu'une conclusion affirmant une propriété M d'un x possédant la propriété P est valide (d'une validité révisable) si «la plupart des x» ou «les x normaux» qui sont des P possèdent bien la propriété M. Cette conclusion est évidemment révisable : si on ajoute à la propriété P une autre propriété P', elle peut rompre l'exigence de pertinence, si P' n'est pas possédée par « les x normaux ». Il faut donc tenir compte de l'environnement ou du contexte de P, et s'assurer que nulle P' n'est présente. Or ce qui nous permet d'identifier ces P' « impertinentes », c'est notre identification de la propriété mentale M. On voit que la survenance, pour conserver quelque

pertinence explicative, doit filtrer les propriétés physiques en fonction des propriétés mentales, et donc tenir compte de la pertinence des propriétés physiques par rapport à la conclusion qu'on veut en tirer au niveau mental. Cette pertinence est définie sur les propriétés physiques (entre P et P') mais elle est jugée sur la propriété mentale M. Nous admettons un cercle épistémique — c'est à partir de notre identification de M que nous repérons les propriétés impertinentes P', et donc que nous identifions le contexte de validité des propriétés P subvenantes sous les propriétés M; mais il n'y a pas là de cercle ontologique, puisque nous pouvons n'utiliser la propriété M que comme une propriété heuristique, et nous ramener à des propriétés physiques.

Les réseaux connexionnistes rentrent parfaitement dans ce schéma, et sur toute la ligne, si l'on peut dire : tout d'abord, leurs catégories sont révisables par définition, puisque ce sont des révisions de différenciations, et donc leur conclusion, c'est-à-dire la classification qu'ils font en sortie, est dépendante des propriétés physiques de leurs entrées et de leur mécanisme, la pertinence de ces propriétés étant définie par rapport à la stabilité de cette classification, et par rapport à la conservation d'une certaine structure des propriétés d'entrée. Ensuite, leur capacité à catégoriser dépend de manière révisable de la pertinence des propriétés de leurs entrées et des chemins d'évolution de leur mécanisme, elle cesse donc d'être définie quand on sort du domaine où les structures des entrées ne sont pas conservées par les successions de biais introduits par le mécanisme connexionniste. Enfin, leur systématicité est contextuelle, et dépend de la pertinence des propriétés physiques du réseau et de ses entrées par rapport à cette capacité de pouvoir transposer les catégories apprises par le réseau dans un contexte local à d'autres tâches, tant que l'on demeure dans le même domaine. Davantage, les réseaux nous permettent de définir la pertinence des propriétés physiques sans circularité par rapport aux catégories «mentales». Les propriétés structurelles au niveau physique peuvent en effet être définies pour elles-mêmes. Même si l'apprentissage amène à changer de catégorisation, donc à donner des propriétés « émergentes » différentes pour les mêmes entrées, si l'évolution de la catégorisation peut mettre en évidence dans des entrées une pluralité de structures, chacune de ces structures est définissable dans des termes non mentalistes, de même que la conservation, dans les classifications faites en sortie par le réseau, de la sensibilité à cette structure.

Ainsi, à condition de renoncer à l'impossible, c'est-à-dire à satisfaire à la fois les trois exigences de systématicité, de compositionnalité illimitée et d'enracinement (cette condition de « groundness » pouvant être généralisée comme condition de pertinence), et à condition de se

satisfaire d'une compositionnalité locale et d'une systématicité limitée à des domaines, il est possible de rétablir des relations plus satisfaisantes entre le niveau de la description fonctionnaliste et le niveau de la description physicaliste. Les réseaux connexionnistes sont un des dispositifs qui le permettent, un des dispositifs capables d'«émergence contextuelle» : «émergence», parce que les propriétés de catégorisation ne sont le propre d'aucun élément physique de base, que ce soient les données d'entrées, ou les mécanismes locaux du réseau, mais seulement de la co-évolution du réseau et de son ensemble d'apprentissage ; «contextuelle», parce que les catégorisations stables pour un domaine de cette co-évolution ne le sont plus forcément quand on passe à un autre domaine, pour lequel l'apprentissage est à reprendre.

Pierre LIVET
Professeur à l'Université de Provence, Aix-Marseille I
13090, Aix-en-Provence

Bibliographie

Barsalou L. W. et Hale C. R. (1993) , Component of conceptual representation: from features lists to recursive frames, in Van Mechelen I., Hampton J., Michalski R. S., Theuns P. (eds.) *Categories and Concepts*, Academic Press, New York .

Bienenstock E. Notes on the Growth of a "Composition machine" , Interdisciplinary Workshop on Compositionality in Cognition and Neural Networks, Royaumont, May 27-28, 1991.

Carpenter G.. A. and Grossberg S., (1991) *Pattern Recognition by Self-Organizing Networks*, MIT Press.

Clark A. (1993) *Associative engines*, Cambridge, Mass. A Bradford Book, MIT Press.

- (1992) "The presence of a symbol" *Connection Science*, Vol 4, n° 3-4, pp 193-205.

Christiansen M. and Chater N. (1992) "Connectionism, learning and meaning", *Connection Science* Vol 4, n° 3-4, pp. 227-252.

Courrieu Pierre, (1995) (mimeo 1994) "Connectionist expert systems, what is guaranteed and what is not", à paraître dans Caverni J.P. and Nisbett, R. *Psychology of expertise*, Amsterdam, North Holland.

Dretske F. (1986) Misrepresentation, in Lycan (ed.) *Mind and Cognition*, Cambridge, Mass, Basil Blackwell.

- *Explaining Behavior*, Cambridge, Mass. MIT Press.

Elman J. (1991) Representation and structure in connectionnist models" in G. Altmann (ed.) *Cognitive Models of Speech Processing*, Cambridge, Mass., MIT Press.

French R. (1992) "Semi-distributed Representations and Catastrophic Forgetting in Connectionist Networks", *Connection Science* Vol 4, n° 3-4, pp. 365-377.

Gärdenfors P. (1991) Nomonotonic inference, expectations and neural networks, in Kruse et Siegel (ed.), *Symbolic and Quantitative Approaches to Uncertainty*, Springer-Verlag, Lecture Notes in Computer Science n° 548.

Geman S., Bienenstock E., Doursat R. (1992) Neural Networks and the Bias/Variance Dilemma, *Neural Computation*, 4, pp. 1-59.

Giles C. L. et Omlin C. W.(1993) Extraction Insertion and Refinement of Symbolic Rules in Dynamically Driven Recurrent Neural Networks, *Connection Science*, Vol 5, n° 3-4.

Fodor J. A. (1987) *Psychosemantics*, Cambridge, Mass. MIT Press.

- (1990) *A Theory of Content*, Cambridge, Mass. , Bradford Book , MIT Press.

Fodor J., Pylyshyn Z. (1988) Connectionism and cognitive architecture; a critical analysis, *Cognition*, 28, pp. 3-71.

Fodor J. and McLaughlin B. P. (1989) Connectionism and the problem of systematicity: why Smolensky's solution doesn't work. *Cognition*, 35, pp. 183-204.

Harnad S., Hanson S.J., Lubin J. (1994) Learned categorical perception in neural nets: implications for symbol grouding, in Honavar Vasant, Uhr Leonard, (eds.) *Artificial Intelligence and Neural Networks*, San Diego, Academic Press, pp. 191-205.

Honavar V., Uhr L., (eds.) (1994) *Artificial Intelligence and Neural Networks*, San Diego, Academic Press.

Kim J. (1990) Explantory Exclusion and Mental Causation, in Villanueva E. (ed) *Information, Semantics and Epistemology*, Oxford, Blackwell, pp. 36-56.

Lycan W.G. (ed.) (1990) *Mind and Cognition*, Cambridge, Mass, Basil Blackwell.

Millikan R. (1984) *Language, Thought, and Other Biological Categories*, Cambridge, Mass. MIT Press.

Oden G. C.(1994) Why the difference between connectionism and anything else is more than you might think but less than you might hope, in Honavar V., Uhr L., (eds.) *Artificial Intelligence and Neural Networks*, San Diego, Academic Press, pp. 93-103.

Pacherie E., (1993) *Naturaliser l'intentionnalité*, Paris, PUF.

Petitot J. (1992) *Physique du sens*, Paris, Editions du CNRS.

Schlechta K. (1994) A reduction of the theory of confirmation to the notions of distance and measure, mimeo.

Schwarz G. (1992) Connectionism, Processing, Memory, *Connection Science*, vol 4, n° 3-4, pp. 207-226.

Sharkey N. A. , Jackson S.A., (1994) Three horns of the representational trilemma, in Honavar V., Uhr L., (eds.) *Artificial Intelligence and Neural Networks*, San Diego, Academic Press, pp. 155-189.

Shastri L. et Ajjanagadde V.,(1993) From simple association to systematic reasoning: a connectionist representation of rules, variables and dynamic bindings using temporal synchrony, *Behavioral and Brain Sciences*, 16, pp. 417-451.

Smolensky P., (1990) Tensor product variable binding and the representation of symbolic structures in connectionist networks, *Artificial Intelligence*, 46, pp. 159-216.

Van Gelder T., (1990) Compositionality: a connectionist variation on a classical theme, *Cognitive Science*, 14, pp. 335-384.

-(1992) Defining distributed representations, *Connection Science*, vol 4, n° 3-4, pp. 175-191.

-et Port R. (1994) Beyond symbolic: toward a kama-sutra of compositionality, in Honavar Vasant, Uhr Leonard, (eds.) *Artificial Intelligence and Neural Networks*, San Diego, Academic Press, pp. 107-125.

Von der Malsburg Ch. (1981) The correlation theory of brain function, Internal report 81-2 Max Planck Institute for Biophysical Chemistry, Göttingen.