CONNEXIONNISME ET COGNITION :
À LA RECHERCHE DES BONNES QUESTIONS

Lorsque les réseaux de neurones formels apportèrent la bonne nouvelle connexionniste aux quatre coins du petit monde des sciences cognitives[1], ils se présentaient sous des dehors fort différents des systèmes classiques, ceux que l'intelligence artificielle construisait et à l'image desquels les sciences cognitives tentaient de façonner leurs modèles[2]. Si

1. Le connexionnisme d'aujourd'hui est la résurgence d'un courant qui a précédé, et dont est issu le mouvement « cognitiviste » majoritaire depuis une trentaine d'années dans les sciences cognitives et l'Intelligence Artificielle. Le premier connexionnisme est issu des travaux de Warren McCulloch et Walter Pitts et constitue l'une des plus durables contributions de la cybernétique (cf. « A Logical Calculus of the Ideas Immanent in Nervous Activity », leur article de 1943 repris *in* Warren S. McCULLOCH, *Embodiments of Mind,* Cambridge, MA, MIT Press, 1965/1988). L'un de ses prolongements les mieux connus est l'invention par F. Rosenblatt du perceptron (cf. Frank ROSENBLATT, *Principles of Neurodynamics,* New York, Spartan, 1962). On s'accorde pour voir dans le livre de Marvin MINSKY et Seymour PAPERT, *Perceptrons,* Cambridge, MA, MIT Press, 1969 (nouv. éd. 1989), le certificat de décès sociologique du premier connexionnisme. Des éléments d'histoire de ce mouvement sont fournis dans les *Cahiers du CREA, 6* et *7,* nov. 1985 (épuisés mais consultables au CREA, 1, rue Descartes, 75005 Paris). L'essor du cognitivisme est ponctué de nombreux écrits ; on en trouvera une description partielle ci-dessous ; des exposés plus détaillés mais accessibles se trouvent notamment dans John HAUGELAND, ed., *Mind Design,* Cambridge, MA, MIT Press, 1981 ; Daniel ANDLER, « Les sciences de la cognition », in *La Philosophie des sciences aujourd'hui,* sous la dir. de Jean HAMBURGER, Paris, Gauthier-Villars, 1986 ; ID., « Progrès en situation d'incertitude », *Le Débat, 47,* nov-déc. 1987, p. 213-234 ; ID., article « Cognitives (sciences) », *Encyclopaedia Universalis,* Paris, nouv. éd., 1989. Le (néo) connexionnisme s'est développé à partir de la fin des années 1970, et a acquis une notoriété considérable grâce à la publication en 1986 d'un ouvrage en deux forts volumes, *Parallel Distributed Processing. Explorations in the Microstructure of Cognition,* par David RUMELHART, James McCLELLAND et le PDP Research Group, Cambridge, MA, MIT Press. (Un chapitre en a partiellement été traduit dans le n° 47 du *Débat.*) L'anthologie de James ANDERSON et Edward ROSENFELD, *Neurocomputing - Foundations of Research,* Cambridge, MA, MIT Press, 1988, permet de suivre très précisément le développement du connexionnisme, des origines à nos jours. Une très éclairante analyse des fondements de l'approche PDP se trouve *in* Andy CLARK, *Microcognition. Philosophy, Cognitive Science, and Parallel Distributed Processing,* Cambridge, MA, MIT Press, 1989.

2. On peut prendre très au sérieux une conception de l'esprit inspirée du modèle de l'ordinateur sans pour cela estimer que l'Intelligence Artificielle contribue notablement au progrès des sciences cognitives. Telle est, par exemple, la position de Jerry FODOR, in *The*

différents, de fait, qu'il fallait se frotter les yeux plusieurs fois avant de pouvoir admettre que les connexionnistes et leurs concurrents cognitivistes (nous les appellerons aussi, suivant l'exemple de leurs champions J. Fodor et Z. Pylyshyn[3], « les classiques ») se proposaient d'expliquer *les mêmes phénomènes.* Les débats qui s'ensuivirent modifièrent à ce point la perspective qu'on put penser que non seulement les uns et les autres parlaient de la même chose, mais qu'ils en proposaient finalement *le même genre d'explication* — que les réseaux n'étaient que des systèmes classiques, d'un genre sans doute un peu particulier, décrits d'une manière inhabituelle. Et il fallut se frotter les yeux à nouveau pour repérer malgré tout des différences. Aujourd'hui Fodor et Pylyshyn, défenseurs et illustrateurs du classicisme, prétendent placer le connexionnisme devant l'alternative suivante : ou bien les réseaux sont foncièrement différents des systèmes classiques, mais ils sont et demeureront alors inaptes à modéliser des aspects essentiels de la cognition ; ou bien ils peuvent surmonter leurs faiblesses actuelles, mais ils ne sont en fait alors que des systèmes classiques vus par le petit bout de la lorgnette.

Le débat oscille depuis le début entre deux extrêmes : les connexionnistes radicaux prétendent que leur *sujet* d'étude est le même que celui des classiques, mais que ceux-ci ont une conception erronée de la *nature* des phénomènes, conception qui les amène à un choix malheureux du niveau de description, donc à une caractérisation (et le cas échéant une simulation) inadéquate de la cognition ; les classiques intransigeants maintiennent que les connexionnistes se trompent de sujet (à la manière de géologues, par exemple, qui ne voudraient parler que de quanta), ce qui prive naturellement leurs explications et leurs simulations de toute véritable pertinence. Bref, nous serions sommés de trancher entre deux hypothèses : celle d'un sujet unique et de deux explications (dont une bonne et l'autre mauvaise) ; celle de deux sujets (dont l'un correspond au domaine pré-théoriquement visé et l'autre pas), chacun muni de son explication propre (et dont l'une seulement nous intéresse).

Peut-être est-il possible d'échapper à cette alternative. Le débat n'a, en tout cas, visiblement pas pour issue prochaine la désignation d'un vainqueur. Il permettra sans doute en revanche de faire la part de ce qui tient, entre les deux « camps »[4], du *désaccord,* et ce qui tient du *malentendu.*

Modularity of Mind, Cambridge, MA, MIT Press, 1983 ; tr. fr. par Abel GERSCHENFELD, *La Modularité de l'esprit,* Paris, Minuit, 1986.

3. Jerry FODOR, Zenon PYLYSHYN, « Connectionism and Cognitive Architecture : A Critical Analysis », *Cognition, 28,* 1988, p. 3-71. Repris *in* Steven PINKER, Jacques MEHLER, eds, *Connections and Mind,* Cambridge, MA, MIT Press, 1988.

4. Nombreux sont ceux qui rejettent l'idée de deux camps hostiles, pour diverses raisons qui ne tarderont pas à apparaître.

L'enjeu pour l'épistémologue est de discerner comment s'associent une *conception* de la nature véritable du domaine pré-théoriquement visé sous l'appellation « cognition », et un *type d'explication* des phénomènes qui le constituent.

I. — MACHINES

Le cognitivisme et le connexionnisme se définissent en partie (mais non, on le verra, exclusivement) par la référence à un type de machine. Pour les défenseurs du premier, il s'agit de l'ordinateur dit de von Neumann, pour ceux du second, du réseau de neurones formels (dit aussi réseau neuromimétique — en anglais, *neural net*). Nous présupposons la familiarité avec l'ordinateur, en tant qu'objet théorique, nous décrivons rapidement à présent le réseau[5].

Bien qu'il existe de nombreuses variétés de réseau, présentant de très importantes différences (alors que la machine de von Neumann est essentiellement unique), on peut, pour les besoins de la discussion, dresser une sorte de « portrait-robot » du réseau connexionniste.

Il s'agit d'un ensemble d'*automates* très simples interconnectés. Les *connexions* permettent à un automate tel que i de transmettre à un automate j une stimulation déterminée par l'état d'activité u_i de i et modulée par un *poids synaptique* w_{ji} ne dépendant que du canal. Le poids est affecté d'un signe : s'il est positif, la stimulation est positive (excitatrice) ; s'il est négatif, elle est négative (inhibitrice). Les automates (ou *unités*) sont en général tous identiques — ce sont souvent des automates *à seuil,* dont l'activité est soit 0 soit 1, et qui sont capables seulement de comparer la somme pondérée des stimulations afférentes $\Sigma_i u_i w_{ji}$ à un seuil s_j et de se mettre ou se maintenir en état d'activité ($u_i = 1$) si ce seuil est dépassé, de s'éteindre ou rester inactif ($u_i = 0$)

5. Toutes les références de la note 1 postérieures à 1985 contiennent des présentations plus ou moins détaillées des réseaux de neurones formels (nous dirons simplement désormais « réseaux »). En français, on dispose d'un manuel de Gérard WEISBUCH, *Dynamique des systèmes complexes. Une introduction aux réseaux d'automates,* Paris, Interéditions/Ed. du C.N.R.S., 1989 ; on pourra aussi consulter le numéro spécial de *Intellectica,* la revue de l'Association pour la recherche cognitive, de février 1990, dirigé par Daniel Memmi et Yves-Marie Visetti. En anglais, citons l'ouvrage collectif dirigé par Geoffrey E. HINTON et James A. ANDERSON, *Parallel Models of Associative Memory,* Hillsdale, NJ, Erlbaum, 1981, qui marque la renaissance du connexionnisme, et le magistral traité de Daniel AMIT, *Modeling Brain Function. The World of Attractor Neural Networks,* Cambridge, Cambridge University Press, 1989.

sinon. Le système est donc caractérisé, à chaque étape de son évolution, qui est discrète, par un *vecteur d'activation* u = (u_1, ..., u_n) ; la transition d'une étape à la suivante résulte d'une mise à jour, soit par tous les automates simultanément, soit par un seul choisi par exemple au hasard, de leur activité.

Lorsque le réseau est utilisé pour transformer une famille d'entrées en certaines sorties spécifiées ou spécifiables (en d'autres termes, lorsque l'on choisit de le faire fonctionner comme une fonction incarnée), le processus commence par l'imposition d'un certain vecteur d'activation u_0, qui peut être considéré comme la donnée ou *input*, se poursuit par itération de la règle de transition, et se termine (dans les cas favorables) lorsque le système atteint un équilibre, caractérisé par un vecteur $u_N = u_\infty$, résultat ou *output* de ce qu'on peut appeler le calcul effectué par le réseau. Ce n'est pas là le seul usage possible d'un réseau, ni peut-être même le plus intéressant, mais la discussion n'en exige pas davantage pour le moment[6].

En tant que calculateur abstrait (ou, si l'on veut, de système de traitement de l'information — mais il n'a pas encore été question d'information), le réseau présente, en première analyse, d'importantes différences avec la machine de von Neumann :

1. Dans celle-ci, le processus est « séquentiel », en ce sens que les opérations élémentaires sont effectuées l'une après l'autre ; dans un réseau, un grand nombre d'entre elles sont faites simultanément et indépendamment les unes des autres.

2. Un réseau est foncièrement homogène (même s'il n'est pas totalement connecté, c'est-à-dire que chaque unité n'influe que sur certaines autres), en ce sens qu'on n'y distingue pas, comme dans l'ordinateur classique, une hiérarchie de sous-systèmes spécialisés dans des tâches interdépendantes de complexité croissante.

3. En particulier, le processus n'est dirigé dans le réseau par rien qui ressemble à une unité centrale de contrôle — c'est tout le contraire de l'ordinateur.

4. Ce qui fait la spécificité d'un réseau, outre le nombre de ses unités, c'est la matrice des poids synaptiques. C'est de ce vecteur que dépend, à *input* égal, le comportement du réseau et le résultat de son calcul — en ce sens il constitue sa « compétence ». C'est donc l'équivalent du programme de l'ordinateur, mais c'est tout différent — aussi différent, par exemple, que l'ensemble des forces agissant sur une bulle de savon peut l'être d'une suite d'instructions pour résoudre le système d'équations différentielles déterminant la position d'équilibre de la bulle.

---

6. Les termes de cette présentation élémentaire sont largement empruntés à mon article de l'*Encyclopaedia Universalis*, *art. cit. supra* n. 1.

5. Enfin, un réseau peut, en théorie, traiter des grandeurs continues, alors que l'ordinateur se nourrit d'entités discrètes (il est *« digital »*, dit l'anglais, ce qu'on nous oblige à traduire par « numérique » — il faudrait dire « chiffral », ce qui éviterait bien des confusions...).

Nous nous interrogerons bientôt sur la solidité de ces contrastes, mais il nous faut d'abord comprendre comment, dans chacun des deux « paradigmes » en présence, la machine devient un système cognitif en puissance.

II. — SYSTÈMES COGNITIFS

a. Systèmes classiques

Le cognitivisme classique a pour postulat fondamental le *fonctionnalisme.* Ce terme recouvre en fait un écheveau de doctrines aux ramifications souvent subtiles, élaborées dans le cadre général du problème corps-esprit[7]. Il est nécessaire ici de n'en retenir que le principe d'une *double description* des systèmes cognitifs. Ils doivent être vus à la fois comme des systèmes matériels et comme des systèmes informationnels[8], et seule importe la possibilité de principe d'un passage d'une description à l'autre. Cette possibilité une fois établie, la tâche des sciences cognitives se borne à caractériser les systèmes informationnels capables d'exhiber un comportement conforme aux principaux aspects observables ou inférables de notre vie mentale. Tout le reste est question d'intendance, c'est-à-dire d' « implémentation ». Symétriquement, la stratégie de l'Intelligence Artificielle classique est de modeler la machine de von Neumann, pour laquelle le passage entre les deux descriptions est assez bien balisé, en sorte d'en tirer des effets d'intelligence (ou plus généralement peut-être, mais plus obscurément aussi, des effets cognitifs ou mentaux).

Un système informationnel classique est composé de deux parties. La partie *variable* est un ensemble de représentations. La partie *fixe* (invariable) est un ensemble de dispositifs capables d'effectuer sur les représentations certaines transformations. En première approximation, les représentations constituent les « connaissances » du système, et les transformations font évoluer ces connaissances, en les combinant entre elles

7. Une excellente introduction à la question est fournie par Pierre JACOB, « Le problème du rapport du corps et de l'esprit aujourd'hui. Essai sur les forces et les faiblesses du fonctionnalisme », à paraître dans *Approches de la cognition,* sous la dir. de D. ANDLER, Paris, Gallimard, « Folio ».

8. J'utilise ce terme ici dans un sens aussi neutre que possible, sans vouloir le distinguer par exemple de « système cognitif ».

d'une part, en les modifiant en fonction d'informations nouvelles d'autre part.

Les représentations sont des expressions d'un langage interne du système, langage du genre de ceux de la logique formelle[9]. Elles désignent notamment des entités, des situations, des événements singuliers, ainsi que des relations générales entre entités, situations, événements. Ce qui est décrit appartient au monde extérieur — à l'environnement au sens le plus large —, mais aussi le cas échéant au monde intérieur du système, qui peut avoir de lui-même une certaine représentation.

Les transformations sont pour l'essentiel des inférences : elles partent typiquement d'un ensemble de représentations de la forme (A, A → B) et en font (B), ou encore (A, A → B, B). Ce sont naturellement des opérations formelles, qui peuvent aussi bien être vues comme des fonctions récursives de nombres entiers codant les formules — ce codage se compose avec la représentation —, en sorte que tel fait est maintenant représenté par un nombre entier plutôt que par une formule. A cette légère modification de point de vue ne correspond aucun changement matériel : en dernière analyse, une machine de Turing ne fait qu'inscrire et effacer des croix dans les cases d'un ruban, croix qui ne sont intrinsèquement pas davantage des nombres que des formules ou que des propositions ! Mais on comprend facilement alors qu'une machine de Turing universelle[10], capable par définition d'effectuer toute opération réalisable par n'importe quelle autre machine de Turing, puisse effectuer toute opération « qui peut être décrite de manière exhaustive et dépourvue d'ambiguïté, tout ce que des mots peuvent exprimer complètement et sans ambiguïté »[11], qu'il s'agisse de nombres ou de formules. On comprend aussi du même coup que toute opération cognitive puisse être vue comme une suite d'étapes élémentaires dont chacune est une inférence au sens faible d'étape d'un calcul correct[12].

Pour essentielle et apparemment simple que soit sur le plan abstrait la

9. On parle souvent dans ce contexte de langages *symboliques*, comme s'il en était qui ne le sont pas ! Le terme « symbole » renvoie ici d'une part à la logique dite symbolique, d'autre part à la matérialisation au moins potentielle dans un système de transformations de systèmes de symboles. On est donc assez loin de l'usage courant.

10. Et donc, moyennant les idéalisations habituelles (temps et mémoire illimités), une machine de von Neumann.

11. Selon les termes de von Neumann au Hixon Symposium, le 29 septembre 1948, rapportés par Hermann GOLDSTINE, *The Computer from Pascal to von Neumann*, Princeton, Princeton University Press, 1972, p. 276. Von Neumann parlait, en fait, des réseaux de McCulloch et Pitts.

12. Il n'est donc pas question de réduire la cognition à la logique déductive classique *au niveau des représentations*. Ce point délicat est la source de malentendus sans fin et de critiques sans fondement.

distinction entre partie fixe ou processuelle et partie variable ou factuelle, elle soulève une difficulté. Pour l'apercevoir, il suffit de substituer aux termes inhabituels qui viennent d'être utilisés ceux, plus courants, de « programme » et de « données ». En effet, si l'on se réfère à la spécification physique d'un ordinateur, seul son plan de câblage est fixe, et correspond à la partie fixe du système informationnel qu'il constitue ou sous-tend. Tout le reste est variable : on sait que le trait de génie des inventeurs de l'ordinateur moderne fut de donner aux règles de fonctionnement le même statut qu'aux entités sur lesquelles elles opèrent : la différence entre programme et données, en ce qui concerne les systèmes matériels qui correspondent — ou devraient correspondre — aux systèmes cognitifs que l'on considère, n'existe que dans le regard de l'observateur ; les machines matérielles qui correspondent aux systèmes cognitifs *intéressants* sont déjà elles-mêmes virtuelles. Entre le niveau de la machine « réellement » matérielle (l'ordinateur non encore programmé) et celui de la machine cognitive s'insèrent donc des niveaux intermédiaires au statut moins clair sur le plan cognitif que sur le plan informatique[13]. Cette difficulté a pour effet, on le verra, de compliquer la comparaison entre systèmes classiques et connexionnistes.

Mais venons-en à la question centrale du rapport entre les niveaux matériel et informationnel ou cognitif. Ce rapport résulte d'une *double* articulation. La première est constituée par le rapport entre la syntaxe et la sémantique du langage formel dans lequel s'expriment les représentations internes du système. Les énoncés que sont ces représentations ont, on le sait, deux visages : leur structure morphologique détermine les transformations syntaxiques auxquelles ils sont sujets en vertu des règles d'inférence du langage, tandis que leur interprétation dans un univers donné (par exemple, l'environnement du système) leur donne une valeur sémantique qu'on peut assimiler en première analyse à une proposition portant sur les objets qui sont l'interprétation des termes du langage. Le parallélisme entre syntaxe et sémantique, que garantit le théorème de complétude de Gödel, explique que lorsque le système passe, en vertu de la syntaxe, d'une représentation R à une représentation R', il passe en même temps d'une proposition vraie (ou supposée telle par le système) à une autre. Ainsi s'explique que le système reste « en contact » avec la réalité externe, tout en ne se guidant que sur ses représentations internes[14].

---

13. Entre les différents niveaux de description de l'ordinateur programmé le rapport est de « compilation ». Dans « Le paradigme de la compilation », in *Approches de la cognition, op. cit. supra* n. 7, Jean-Pierre DESCLÉS plaide précisément pour une réduction du rapport entre niveau matériel et niveau informationnel au rapport de compilation.

14. Il reste à comprendre comment il se fait que le système parte « du bon pied », c'est-à-dire avec des représentations atomiques fidèles. C'est là l'un des points les plus faibles du

Cette première articulation ne permet pas cependant de quitter le niveau informationnel ; elle n'assure pas à elle seule le passage au niveau matériel. Une seconde articulation est nécessaire, reliant cette fois la syntaxe et la physique, ou si l'on veut l'inférence et la cause. Cette articulation est réalisée par l' « incarnation » du calcul inférentiel dans un calculateur matériel — les détails étant sans pertinence. Le modèle réduit en est la calculatrice de poche qui « réalise » les lois de l'arithmétique ; le modèle en vraie grandeur, l'ordinateur, qui « réalise » les lois de la machine de Turing, ou encore un ensemble d'opérations permettant d'engendrer toute fonction récursive. Le principe est de modéliser les traits syntaxiques des formules par des propriétés physiques particulières de constituants d'une machine réelle, et de faire coïncider les lois de transition de la machine telles qu'elles s'expriment au niveau de ces propriétés particulières avec les règles de la syntaxe. C'est ainsi finalement qu'une machine concrète fonctionnant selon les lois de la physique peut être, d'une part, une « machine syntaxique » (peut se conformer aux règles d'une syntaxe formelle), et, d'autre part, selon l'expression de D. Denett, une « machine sémantique ».

Il ne reste plus qu'à souligner que lorsque les cognitivistes classiques caractérisent la cognition comme calcul sur des représentations, ils font référence à une notion parfaitement circonscrite de calcul, celle de Church et de Turing : une fonction calculable est — moyennant codage — identique à une fonction récursive sur les nombres entiers.

b. Systèmes connexionnistes

Il existe, on l'a dit, non pas un seul type de réseau connexionniste, mais une grande variété. Mais ce n'est pas là la seule ni la plus importante raison de distinguer plusieurs formes de connexionnisme. Des différences plus déterminantes encore se situent dans la façon dont ces réseaux sont vus comme systèmes cognitifs. Aussi sera-t-il difficile, parfois impossible, d'opposer à chaque aspect de la conception cognitiviste une position connexionniste unique et précise.

Pour fixer les idées du lecteur profane, voici deux exemples de systèmes connexionnistes. Le premier est un produit typique de l'approche dite PDP[15]. Dû à D. Rumelhart et J. McClelland[16], il est capable

---

cognitivisme classique, comme le reconnaît notamment J. Fodor ; cf. « Fodor's Guide to Mental Representation : The Intelligent Auntie's Vade-Mecum », *Mind*, vol. 44, 1985, p. 76-100, notamment les dernières lignes ; et sa tentative de solution dans *Psychosemantics. The Problem of Meaning in the Philosophy of Mind*, Cambridge, MA, MIT Press, 1987.

15. Pour *Parallel Distributed Processing*, cf. *supra* n. 1.

16. *Op. cit. supra* n. 1, vol. 2, chap. 18, p. 216-271 : « On learning the past tenses of English verbs ».

d'apprendre à former le prétérit de tout verbe anglais à partir de l'infinitif. Les entrées comme les sorties sont des représentations phonémiques ; l'apprentissage est « supervisé » : un corpus d'exemples formés de couples infinitif/prétérit (*eat/ate* ; *be/was* ; *love/loved,* etc.) est d'abord montré au système — plus précisément, l'infinitif est présenté, puis la réaction spontanée du système est graduellement corrigée jusqu'à ce qu'elle soit conforme. Cette rectification progressive suit un algorithme relativement complexe, mais dont le principe général est celui formulé par Donald Hebb[17] : renforcer les poids synaptiques entre unités qui doivent être actives ou inactives simultanément, et réduire les poids dans la situation inverse. L'algorithme est indépendant de la fonction que le système doit apprendre, et son application n'exige pas l'intervention du modélisateur (c'est vraiment un algorithme[18] !). Quant au corpus, il est vaste, mais non exhaustif : une proportion non négligeable de verbes, tant réguliers qu'irréguliers, n'y figure pas. Le système est capable de maîtriser le corpus, au terme d'une (longue) période d'apprentissage ; après quoi il conjugue aussi presque infailliblement tout autre verbe anglais. Il est essentiel de remarquer qu'aucune règle n'est enseignée ou indirectement fournie au système par le modélisateur (en revanche, celui-ci est entièrement responsable du « pré-traitement » conduisant à la représentation phonémique, ainsi que du choix du corpus et du protocole d'apprentissage[19]).

Bon nombre de modèles de l'école PDP partagent avec celui-ci les caractéristiques suivantes. La tâche consiste à compléter une configuration dont l'environnement ne fournit qu'une partie. Dans un cas particulier fréquent, l'environnement fournit un vecteur x et le système doit compléter par f(x). Au contact d'un grand nombre d'exemples de configurations complètes (dans l'exemple, de points du graphe de la fonction f), le système s'adapte aux régularités de l'environnement en ajustant ses poids synaptiques, ce qui lui permet d'une part de réagir sans aucune erreur aux exemples qui lui ont été présentés au cours de l'apprentissage, d'autre part de réagir « intelligemment » à d'autres configurations incomplètes — soit en les assimilant à des parties de configurations connues, soit en y discernant un mélange de configurations connues, et en les complétant en conséquence. Bref, le système se comporte en détecteur de régularités statistiques multidimensionnelles.

---

17. Donald HEBB, *The Organization of Behavior,* New York, Wiley & Sons, 1949.

18. Ce qui laisse toutefois entier le problème de savoir quel genre d'ordinateur naturel l'exécuterait. Ce n'est pas (dans l'état actuel) le réseau lui-même.

19. Ce qui a valu à ce modèle de graves critiques, notamment de la part de Steven PINKER et Alan PRINCE, « On Language and Connectionism. Analysis of a Parallel Distributed Processing Model of Language Acquisition », *Cognition,* vol. 28, 1988, p. 73-193 ; et *in* S. PINKER, J. MEHLER, *op. cit. supra* n. 3.

Parmi les autres courants qui coexistent au sein du connexionnisme, celui qu'animent les physiciens joue un rôle privilégié ; l'un d'entre eux, D. Amit, a proposé de le désigner par le sigle ANN (pour *attractor neural network*)[20]. Notre deuxième exemple est le modèle de mémoire associative « adressable par le contenu »[21] proposé par John Hopfield dans un article mémorable[22] qui a donné au connexionnisme renaissant une impulsion décisive, et marque la naissance du courant ANN. Les réseaux de Hopfield, contrairement aux réseaux PDP, qui sont des perceptrons généralisés de type « feed forward » dans lesquels l'information se propage de manière unidirectionnelle, ménagent des connexions dans toutes les directions : ils sont complètement connectés (ou presque). Ils sont ainsi dotés d'une dynamique autonome déterminée par les poids synaptiques et caractérisée par des attracteurs. Chacun de ces attracteurs peut être considéré comme une « mémoire » (un souvenir) du réseau : toute stimulation suffisamment proche d'un attracteur place le système sur une orbite qui se stabilise en cet attracteur. Le problème est de savoir à quelles conditions le système peut, par un choix approprié de poids synaptiques, adopter pour attracteurs un certain ensemble prédéterminé de points de son espace d'états. Hopfield détermine de telles conditions (elles ont été depuis très fortement étendues). Lorsqu'elles sont réalisées, on dispose d'un système *autonome,* sans entrée ni sortie distinguées, et qui *peut* cependant servir de système *input/output,* en un sens particulier : un ébranlement initial l'amène dans un nouvel état d'équilibre — l'ébranlement est alors l'*input,* et le nouvel équilibre l'*output.*

Les divergences dans le « camp » connexionniste apparaissent d'entrée de jeu, dès qu'est posée la question du rapport entre les caractérisations matérielle et informationnelle des systèmes cognitifs. Les chercheurs qui se rattachent au courant PDP tendent à adopter le fonctionnalisme classique : ils situent leur modélisation non pas au niveau de la réalisation physique dans le tissu du système nerveux central, mais à celui du traitement de l'information. Que le composant de base soit un neurone formel et l'architecture réticulée facilitera, selon eux, l'implémentation

---

20. Cf. Daniel AMIT, *op. cit. supra* n. 5. « ANN » dénote malheureusement aussi depuis quelque temps *« artificial neural network »,* ce qui tend à faire perdre au sigle sa spécificité.

21. Dans une mémoire informatique classique (en l'absence de dispositifs *ad hoc* d'indexation ou autre), le stimulus ne peut être qu'une « adresse » déterminée une fois pour toute pour chaque item stocké, et la réponse attendue l'information logée à cette adresse. Dans une mémoire adressable par le contenu, le stimulus est une partie du contenu, et la réponse la totalité complétée du contenu mémorisé. La mémoire humaine est généralement rapportée à ce dernier type.

22. John J. HOPFIELD, « Neural Networks and Physical Systems with Emergent Collective Computational Abilities », *Proc. Natl. Acad. Sc. USA,* vol. 79, 1982, p. 2554-2558 ; et *in* J. ANDERSON, E. ROSENFELD, eds, *op. cit. supra* n. 1, p. 460-464.

des réseaux connexionnistes dans des systèmes cérébraux constitués d'assemblées de neurones réels — mais le rapport n'est pas de l'ordre de la ressemblance ou de la simplification[23]. Au contraire, certains connexionnistes considèrent que leur démarche est celle d'une neurophysiologie théorique (ou « computationnelle »), partie intégrante de la biologie théorique ; l'idée d'une séparation radicale de deux niveaux d'explication privilégiés est explicitement rejetée par certains, qui se gaussent du « rêve de Marr »[24]. Plus proches, sur ce point, de ces connexionnistes « biologisants » que de la branche PDP, les connexionnistes de tendance ANN estiment prématuré de se prononcer sur la nature des rapports entre théorie neurobiologique et modèles connexionnistes. La situation ne leur semble pas foncièrement différente de celle qui prévaut dans toute tentative pour attaquer, par les méthodes abstraites de la physique mathématique, un phénomène complexe.

A la distinction bipartite dans les systèmes classiques entre partie fixe et partie variable répond une division tripartite dans les réseaux. La partie vraiment fixe est constituée par l'architecture (nombre des unités et connectivité) et par la loi de transition des unités ; elle correspond en un sens à l'architecture et aux opérations câblées de l'ordinateur non encore programmé, mais elle est sensiblement plus pauvre que la partie fixe du système cognitif classique, qui incorpore les capacités générales de la machine de von Neumann (beaucoup plus complexes que celles du réseau « brut ») et les capacités spécifiques d'un programme. A l'autre extrême, la partie vraiment variable du réseau est constituée par les unités actives au moment considéré ; elle correspond elle aussi à quelque chose dans le modèle classique, à savoir le contenu de la mémoire de travail de l'ordinateur, ou de la mémoire à court terme de maint modèle de la psychologie cognitive. Mais dans le système classique, ces contenus ne sont pas de nature différente des autres représentations, celles, plus stables, de la mémoire à long terme : ensemble, elles sont la partie variable du système, partie beaucoup plus riche, par conséquent, que celle du réseau. Il y a donc, de part et d'autre, un résidu qui dans le réseau se localise dans les connexions — résidu dont le slogan premier

23. Une défense approfondie de ce point de vue est présentée par Paul SMOLENSKY dans un article retentissant, « The Proper Treatment of Connectionism », *The Behavioral and Brain Sciences*, vol. 11, 1988, p. 1-74.

24. David MARR a défendu avec beaucoup de force et de rigueur le principe d'une distinction de niveaux (il en dégage trois) de caractère fonctionnaliste. Voir son célèbre ouvrage posthume, *Vision*, San Francisco, Freeman, 1982 ; ou, pour les passages pertinents, son article dans *Mind Design*, *op. cit. supra* n. 1. Ce sont Patricia CHURCHLAND et Terrence SEJNOWSKI qui s'en prennent à cette conception dans « Neural Representation and Neural Computation », *in* L. NADEL, L. COOPER, P. CULICOVER, R. HARNISH, eds, *Neural Connections and Mental Computation*, Cambridge, MA, MIT Press, 1988.

du connexionnisme dit l'importance : « Toute la connaissance réside dans les connexions », c'est-à-dire dans l'ensemble des poids synaptiques w_{ji}. Comme le programme, les connexions sont fixes au cours d'un processus, mais comme les données représentées, elles sont acquises et reflètent directement des aspects de l'environnement dont la connaissance est indispensable au bon fonctionnement du réseau.

Malgré de notables différences, on discerne donc sur les deux premiers points de doctrine (la distinction entre niveaux et entre parties fixe et variable) une parenté entre les approches, classique et connexionniste. Il en va tout autrement sur les autres points, à savoir le système de représentation, les principes tant physiques que cognitifs régissant l'évolution des systèmes et, enfin, le fondement de la cohérence entre le système et son environnement.

Le système de représentation, tout d'abord, n'est pas un langage formel ; ses « formules » sont des unités, ou des vecteurs d'unités, selon que les représentations sont « localistes » (le support d'une entité sémantique étant une unité du réseau) ou « distribuées » (le support étant une suite ordonnée d'unités, et chaque unité étant inversement impliquée dans la représentation de plusieurs entités)[25]. Dans l'approche PDP, il y a bien une combinatoire des représentations individuelles, mais elle est de type ensembliste, non concaténatoire : les ensembles activés se superposent et s'intersectent — c'est une combinatoire fruste. Dans l'approche ANN, il n'y a pas pour l'heure de combinatoire du tout, mais certains envisagent d'en développer une qui soit fondée sur la théorie des bifurcations[26], ou plus généralement, sur une temporalisation des représentations.

L'évolution d'un réseau obéit non à des calculs figurant des inférences qui s'enchaînent linéairement, mais à un système d'équations différentielles (en général cependant discrétisées) ; c'est un processus de « relaxation », c'est-à-dire de recherche d'une position d'équilibre contrainte par un grand nombre d'interactions simultanées. Sur le plan cognitif, ces interactions correspondent à des associations de force variable, ou encore à des « micro-inférences » non impératives se compensant partiellement, et concourant à un effet global non réductible à une force unique.

N'est-il pas cependant abusif de parler d'inférence, serait-ce à l'abri du préfixe « micro », avant d'avoir précisé de quelle logique il s'agit, et à quoi elle s'applique ? Que serait l'équivalent, dans un réseau, du niveau

25. Cf. D. RUMELHART, J. MCCLELLAND *et al., op. cit. supra* n. 1, vol. 1, chap. 3.

26. Cf. Jean PETITOT, « Hypothèse localiste, modèles morphodynamiques et théories cognitives : remarques sur une note de 1975 », *Semiotica*, vol. 77, 1989, p. 65-119, et son article dans le présent numéro.

syntaxique ? Contentons-nous pour le moment d'observer que la réponse n'a rien d'évident, contrairement à ce que certains écrits connexionnistes voudraient faire croire ; ils emploient, en effet, le terme « syntaxe » au sens de niveau des déterminations physiques commandant les transitions du système. C'est là un abus de langage dangereux, provenant du paradigme classique, dans lequel, on l'a vu, c'est par la médiation de la syntaxe que s'établit le pont entre inférence et cause, ou entre niveau cognitif et niveau physique. Or, dans un réseau, rien ne s'offre immédiatement au regard qui puisse jouer le rôle de syntaxe *indépendamment* des lois de transition[27].

Rien, par conséquent, d'analogue au parallélisme entre syntaxe et sémantique d'un langage formel ne semble pouvoir constituer la garantie du maintien de l'adhésion du système à l'environnement. Mais cette garantie n'est justement pas nécessaire. Dans l'approche PDP, la possibilité d'écarts entre la réponse du système et la bonne réponse est constitutive et censée correspondre à la faillibilité caractéristique des systèmes cognitifs biologiques. On exige toutefois une conformité statistique des résultats obtenus par le réseau, et celle-ci est assurée par l'application d'un principe général de minimisation des écarts[28]. Ce principe a pour effet, étant donné un stimulus s proche d'un exemple s_0 pour lequel le système a préalablement appris la réponse correcte r_0, de pousser le système à fournir, selon les cas, soit une réponse r proche de r_0 et qui en diffère approximativement comme s diffère de s_0, soit tout simplement la réponse r_0. Que le réseau soit capable, par apprentissage au contact de l'environnement, d'ajuster ses paramètres en sorte de pouvoir appliquer le principe général dans cet environnement-là en fait essentiellement un capteur de régularités statistiques ; et qu'il l'applique en complétant un *input* conformément au principe de continuité ou de stabilité qu'on vient d'énoncer en fait essentiellement une machine associative.

Il n'y a donc ici, pour articuler le niveau matériel au niveau cognitif, qu'une seule médiation : les représentations, tant explicites et fugaces (ce sont les unités actives au début du processus), qu'implicites et permanentes (ce sont les poids synaptiques ajustés au cours de l'apprentissage).

---

27. Remarquons qu'il en est de même d'une machine de Turing ou de von Neumann. Contrairement à ce que semblent par moments croire J. Fodor et Z. Pylyshyn, ainsi que maints autres acteurs (qui vont jusqu'à parler du « paradigme de von Neumann », auquel s'opposerait un « paradigme "non-von" »), l'architecture de von Neumann ne suffit pas à imposer la conception classique : celle-ci n'émerge précisément que lorsqu'on fait opérer cette architecture sur la syntaxe d'un langage formel. Lorsqu'on contemple un système classique, il est donc nécessaire de chausser des lunettes « cognitives » (de le considérer comme un système de traitement de l'information) pour y discerner une syntaxe.

28. Ce que P. Smolensky appelle, de manière évocatrice, le principe de maximisation de l'harmonie ; cf. D. RUMELHART, J. MCCLELLAND *et al., op. cit supra* n. 1, vol. 1, chap. 6.

Les premières représentent les entités considérées par le système, et sont traitées non pas conformément à une syntaxe interne, mais au principe d' « harmonie » qui n'est que la description ramassée de la loi de transition du système, déterminée par les traces laissées par l'environnement dans le système sous la forme des poids.

Dans l'approche ANN, on peut envisager de se dispenser complètement de garantie d'adhésion ou de correction. Reprenons l'exemple de la mémoire ; il est caractérisé par l'absence de contrainte sur les réponses : on demande seulement au système de se stabiliser après avoir été exposé à un stimulus significatif. Il est vrai que l'on cherche à obtenir de lui une taxinomie raisonnable : on lui demande de ne pas tout confondre, de ne pas non plus tout distinguer. Mais il n'est plus question alors de correction — seulement de commodité pour l'utilisateur, et le cas échéant, s'il s'agit d'un organisme, pour le système lui-même.

On ne saurait conclure cette présentation des systèmes connexionnistes sans s'interroger sur leur caractérisation comme systèmes « computo-représentationnels ». Ils prétendent, en effet, illustrer une conception de la cognition selon laquelle celle-ci se ramène à des processus calculatoires sur des représentations — et ne se distinguer des systèmes classiques que par la nature des calculs mis en œuvre, ou encore celle des représentations, voire les deux. On dira, par exemple, que les calculs sont parallèles, ce qui les rend foncièrement différents des calculs séquentiels de l'approche classique. Le malheur est que tout calcul au sens d'une manipulation effectuable de signes discrets peut être exécuté de manière séquentielle sur une machine de Turing ou de von Neumann : cela résulte de la célèbre thèse de Turing-Church, mais il n'est pas nécessaire de l'invoquer, puisque les réseaux peuvent (visiblement) être simulés sur des ordinateurs classiques[29], et le sont effectivement. On dira encore que les réseaux manipulent de l'information *numérique* et non *symbolique ;* que la cognition s'explique donc comme un calcul sur des représentations numériques — et non symboliques ! On entend par là que les entités sur lesquelles s'effectuent les calculs ne sont pas les éléments d'un langage ou système formel. Mais toute représentation est par définition symbolique, en sorte que si l'on veut que les nombres représentent, on ne peut les empêcher de le faire exactement comme les symboles classiques. Si, inversement, l'on veut que les calculs des réseaux ne se ramènent pas principiellement à la notion canonique de calcul, il faut les empêcher d'opérer sur des représentations — faire en sorte qu'ils ne soient que l'hypostase descriptive de *processus,* dans le sens où les calculs qui

29. La réciproque est vraie, abstraction faite de la limitation de mémoire.

donnent les trajectoires des planètes ne sont que notre manière de les décrire, les planètes elles-mêmes n'en ayant cure. La suite du présent article apportera peut-être quelques indications utiles sur ce difficile problème, mais il semble dès à présent que *si* les connexionnistes veulent se démarquer à ce très haut niveau de généralité des classiques, il leur faut renoncer soit aux représentations, soit au calcul — sacrifices presque impossibles dans le contexte actuel, puisque aussi bien la notion de représentation est quasiment constitutive des sciences cognitives[30], et que le terme « calculatoire » ou « computationnel » est devenu quasi synonyme de « scientifique » ou « sérieux ».

III. — LA VRAIE NATURE DES PHÉNOMÈNES

L'introspection n'est pas un bon instrument d'investigation scientifique — c'est sur ce rejet[31] que le béhaviorisme prit appui, et les sciences cognitives n'y reviennent pas. Au contraire, bon nombre de leurs succès consistent à mettre au jour des effets invisibles à l'œil introspectif, voire inacceptables pour l'intuition. Cependant, elles réhabilitent les états mentaux — quelque chose, donc, que l'introspection permet de deviner, si partiellement et si trompeusement parfois que ce puisse être. Ce que nous pensons penser, ce que nous croyons croire fournit au moins une indication sur ce que nous pensons et croyons, et notamment sur la manière dont nos pensées et croyances s'enchaînent. Et si l'analyse scientifique ou la simulation révèlent des processus que nous ne parvenons pas à faire coïncider au moins partiellement avec les enchaînements de nos pensées, ce hiatus appelle à son tour une explication.

a. Le connexionnisme sauve les phénomènes

Or, que nous montre le spectacle de nos pensées — ou, si l'on préfère, des événements mentaux dont nous avons conscience ? Prenons d'abord le cas des pensées « rapides » : celles qu'accompagnent la compréhension ou l'émission d'une phrase courante dans notre langue maternelle ;

30. Il existe un courant antireprésentationniste, incarné notamment par J.J. Gibson, W. Freeman et Ch. Skarda, B. Shanon, F. Varela, mais il est très minoritaire (ce qui ne signifie pas, bien entendu, qu'il s'égare nécessairement).

31. Cf., par ex., George SPERLING, « The Magical Number Seven : Information Processing Then and Now », *in* William HIRST, ed., *The Making of Cognitive Science. Essays in Honor of George A. Miller*, Cambridge, Cambridge University Press, 1988 ; William LYONS, *The Disappearance of Introspection*, Cambridge, MA, MIT Press, 1988.

la reconnaissance d'un visage, d'un lieu, d'un objet, d'une voix, d'une situation ou d'un air de musique familier ; la résolution d'un problème simple d'un type parfaitement connu ; une décision de routine ; un déplacement sur le court de tennis pendant un échange de balle ; la traversée en voiture d'un carrefour rencontré des milliers de fois, etc. L'introspection ne nous dit certes pas grand-chose sur ces pensées-là, précisément peut-être parce qu'elles viennent trop vite[32]. Intuitivement, en tout cas, elles se ressemblent, et ressemblent donc toutes aux mieux identifiées d'entre elles, à savoir les perceptions, alors que la tradition philosophique les distingue soigneusement, séparant perception de raisonnement et d'action, et que le cognitivisme veut tout ramener à l'inférence.

Peut-être faut-il donc chercher du côté des pensées plus lentes, celles qui comportent un délai perceptible, indicateur d'un processus complexe nous fournissant une information, une connaissance ou une compréhension qui ne nous sont pas immédiatement accessibles. Examinons donc la manière dont nous parvenons à une conclusion à partir d'informations nouvelles et à la lumière de nos connaissances et de notre expérience ; observons ce qui semble se produire en nous lorsque nous prenons une décision, lorsque nous résolvons un problème, lorsque nous identifions un visage, un objet peu connu, ou encore l'auteur d'un morceau de musique que nous entendons sans doute pour la première fois, lorsque nous lisons un texte difficile, lorsque nous comprenons ce qu'on vient de nous dire dans une langue qui n'est pas la nôtre, lorsque nous nous frayons un passage à travers une foule dense ou pilotons une voiture sur un itinéraire accidenté. Le processus que nous observons en nous-mêmes s'apparente-t-il dans ces cas à une suite d'inférences ? ou bien à la perception d'une scène mal éclairée ? Notre flux mental ressemble-t-il à l'explication que Sherlock Holmes fournit au pauvre Watson après avoir conclu son enquête, ou est-il plus proche de l'*Aha-Erlebnis* que nous vivons au moment précis où, nos yeux s'habituant à la pénombre, nous saisissons comme un tout organisé et chargé de sens les fragments de scène que nous discernions l'instant d'avant ? Sans doute des étapes se sont-elles esquissées fugitivement ; peut-être même avons-nous eu conscience d'essayer d'assembler certaines pièces du puzzle à l'aide d'inférences (« Si ceci est X, alors... », ou « Si je fais d'abord ceci, alors... »). Mais le rôle qu'elles jouent dans la stabilisation finale de nos pensées n'a rien de clair : elles orientent notre recherche (parfois vers une impasse), mais notre impression est généralement qu'elles ne constituent

32. Il s'écoule, par exemple, quelque chose comme un dixième de seconde entre l'apparition d'un visage familier et sa reconnaissance.

pas le tissu même du processus, qui s'est déroulé pour l'essentiel comme une révélation progressive menant à une cristallisation finale.

Des pensées assez lentes aux pensées très lentes, le passage est sans rupture. Nous arrivons dans le domaine du raisonnement mathématique et scientifique, de la délibération, du jeu d'échecs, du diagnostic complexe, de l'expertise. Et nous y trouvons, dressé pour nous depuis longtemps, depuis que les savants, les écrivains, les joueurs d'échecs et les experts nous ont forcés à distinguer le contexte de la découverte de celui de la justification, le même constat : les étapes, les inférences, l'assemblage réfléchi de fragments, jouent un rôle, souvent important, jamais suffisant pour rendre compte du processus. En fait l'expert à son affaire rejoint l'homme ordinaire à la sienne : sa pensée n'est pas lente, elle est au contraire de l'ordre de la perception et du réflexe, comme y insistent Hubert Dreyfus et Stuart Dreyfus[33]. Le grand maître d'échecs voit presque immédiatement ce qu'il va faire, et ne passe son temps qu'à des vérifications, des auto-explications, des comparaisons réfléchies avec les précédents, des spéculations sur les réactions de l'adversaire, sur son état psychique, etc. Ayant en mémoire quelque 50 000 configurations, il reconnaîtrait, selon ces auteurs, celle qu'il affronte en l'assimilant, globalement et sans calcul, à l'une d'elles, et dans le même mouvement choisirait la réaction appropriée.

Admettons la validité de cette description : admettons que, vus ou vécus par nous-mêmes, examinés même par le psychologue réglant son instrument de façon à ne pas perdre complètement ce dont témoigne l'expérience consciente des sujets, nombre de nos processus mentaux tiennent plus de la perception, fût-elle différée, que de l'enchaînement d'inférences. En quoi cela fait-il pencher la balance du côté connexionniste ?

La réponse est évidente, même si elle est, on s'en doute, loin d'être définitive. D'une part, en effet, ce que modélisent, en première analyse, les réseaux vus comme systèmes cognitifs, ce sont des formes, plus ou moins généralisées, de perception (ou, si l'on préfère, songeant notamment à la mémoire associative de Hopfield, de reconnaissance). D'autre part, l'évolution d'un réseau au cours d'une tâche correspond bien aux descriptions que l'on vient de donner des processus mentaux : elle est parfois ponctuée d'étapes — passage d'une hypothèse à la suivante, élimination d'une possibilité, etc. ; elle est parfois orientée par une inférence — si l'on choisit d'interpréter ainsi, par exemple, l'activation de telle sous-population d'unités provoquée par l'activation antérieure de telle autre ; mais elle se déroule pour l'essentiel dans le désordre apparent

33. Cf. *Mind over Machine. The Power of Human Intuition and Expertise in the Era of the Computer*, New York, The Free Press, 1986, p. 32-35.

des interactions multiples et ne prend un sens clair que lors d'étapes éventuelles et surtout lors de la stabilisation finale. A chaque instant intermédiaire, des hypothèses sont inégalement actives, des liens se font sentir simultanément avec des forces variables, positives ou négatives. Et ces « micro-événements » ne se laissent qu'exceptionnellement sommer en des « macro-inférences ».

D'autres phénomènes sont (en première analyse), sinon sauvés, du moins respectés par le connexionnisme. Il en est ainsi du rôle reconnu depuis les travaux d'Eleanor Rosch aux prototypes dans les tâches de catégorisation[34], où selon les classiques intervenaient des définitions par conditions nécessaires et suffisantes. Il en est ainsi de l'aptitude singulière de l'homme à traiter une information incomplète, incertaine, voire contradictoire : contrairement aux systèmes classiques, les réseaux, comme l'homme, fournissent des réponses raisonnables dans des conditions informationnelles loin de l'optimal, et sans faire intervenir de mécanisme particulier : ce sont les mêmes mécanismes, et les mêmes raisons qui expliquent leur excellent comportement dans de bonnes conditions et leur comportement convenable dans de mauvaises. Il en est ainsi, enfin, de certaines erreurs systématiques chez le sujet normal, et de certains syndromes observés en neuropsychologie : erreurs et déficits humains prennent souvent des formes bien spécifiques que reproduisent les réseaux, le cas échéant lésés artificiellement par exemple par suppression de certaines unités.

Le dernier grand phénomène qui selon certains donnerait un avantage décisif aux réseaux est celui du *contexte.* Il est reconnu par tous que le contexte joue un rôle crucial dans tous les domaines de la cognition, de la communication, notamment verbale, jusqu'au raisonnement, à l'action, voire à la perception. La « sensibilité au contexte » est une qualité prisée chez les humains et poursuivie, tel le Graal, par les spécialistes d'Intelligence Artificielle et plus généralement par les concepteurs de logiciels — c'est la marque même de la véritable intelligence (et son absence la marque de la bêtise, et l'une des principales sources du comique).

Là s'arrête l'unanimité. Il n'est d'abord pas clair que tous les phénomènes auxquels les uns et les autres font référence lorsqu'ils évoquent, ou brandissent, le contexte aient en commun un ensemble significatif de propriétés. Ensuite, tout un spectre de positions se dessine sur l'ampleur de la « contamination » par le contexte : intervient-il aux tout derniers stades des processus, ou le trouve-t-on dès le départ, affectant les primitives mêmes ? Certains stades sont-ils à l'abri, ou bien le contexte intervient-il à tout moment ? Si Fodor, par exemple, défend l'idée de modula-

34. Cf., par ex., Eleanor ROSCH, « Principles of Categorization », *in* E. ROSCH, B. B. LLOYD, eds, *Cognition and Categorization,* Hillsdale, NJ, Erlbaum, 1978.

rité[35], c'est pour protéger certains processus (notamment perceptuels) des excès de la psychologie *« new-look »* qui voit les influences *« top-down »* s'immiscer à tous les stades, comme si le contexte (le *« top »*) pouvait tout faire et tout empêcher au niveau *« down »* : nous faire voir dans la cage du zoo un tigre en l'absence du tigre, ou inversement nous cacher le véritable tigre qui s'apprête, contre toute vraisemblance, à nous dévorer place de l'Opéra.

L'espace logique des positions possibles continue longtemps de se ramifier. Ce n'est pas le lieu de l'explorer exhaustivement. Une partie de l'arbre peut se parcourir rapidement, à partir d'une position classique, de la façon suivante : l'influence du contexte est-elle formalisable ? Si oui, si le contexte est donc constitué d'informations générales complémentaires homogènes aux informations particulières de premier plan, ces informations sont-elles de l'ordre de faits ou de l'ordre de règles ? S'il s'agit de faits, ont-elles pour effet d'enrichir les conclusions (logique monotone) ou de les infléchir (logique non monotone, régime des exceptions) ? S'il s'agit de règles, sont-elles de même niveau que les règles de premier degré, ou d'un niveau supérieur (métarègles) ? Dans ce dernier cas, guident-elles les inférences (par exemple, en imposant un ordre de priorité sur les règles de premier niveau), ou bien les modifient-elles (par exemple, en invalidant certaines règles ou en les modifiant) ? Retournant à la racine de notre arbre, plaçons-nous dans le cas d'une réponse négative : l'influence du contexte n'est pas formalisable. Avant de jeter l'éponge, nous pouvons nous demander si elle est modélisable. Répondre positivement, c'est imaginer soit, du côté des faits, un système de pondération, soit, du côté des règles, un infléchissement du régime inférentiel. (Dans les deux cas, ce qui empêche la modélisation d'être une formalisation est l'absence de règles formelles de pondération ou d'infléchissement[36].) Les prolégomènes pour une logique située, de Jon Barwise[37], et la théorie de la pertinence de Sperber et Wilson[38] sont deux tentatives (inégalement développées) pour rendre raison du contexte en caractérisant son influence sur le régime inférentiel, sans la formaliser au sens étroit du terme.

---

35. Cf. J. FODOR, *op. cit. supra* n. 2.

36. La distinction modélisation/formalisation appelle une analyse plus sérieuse qui nous entraînerait trop loin de notre propos.

37. Cf. « Information and Circumstance », *Notre-Dame Journal of Formal Logic,* vol. 27, *3,* July 1986, p. 324-338, et « Unburdening the Language of Thought », *Mind and Language,* vol. 2, *1,* Spring 1987, p. 82-96. Barwise répond à Fodor, qui l'attaque avec une vivacité signalant l'importance de l'enjeu.

38. Dan SPERBER, Deirdre WILSON, *Relevance : Communication and Cognition,* Oxford, Basil Blackwell, 1986 ; trad. fr. par A. GERSCHENFELD et D. SPERBER, *La Pertinence,* Paris, Minuit, 1989.

Dans la perspective connexionniste, le contexte est en un sens si bien incorporé au processus qu'on peut se demander s'il demeure quelque chose de la distinction entre « texte » et contexte, ou entre premier plan et arrière-plan. Ce qui marque, en effet, la place du contexte dans le fonctionnement du réseau, c'est que les conséquences de l'activation d'une unité ou d'un groupe d'unités U sont fonction de l'état d'activation de l'ensemble R des autres unités : le fait, la situation, l'événement ou l'entité e que représente U est donc interprété(e) différemment selon que son environnement v, représenté par R, est dans un état plutôt qu'un autre. Mais il n'existe en vérité aucune dissymétrie intrinsèque entre U et R, et l'on pourrait dire aussi bien que le réseau traite v dans le contexte de e. D'autre part, l'influence du contexte est peut-être ainsi figurée, elle n'est ni expliquée ni contrôlée. Enfin, une pareille conception du contexte tombe sous le coup des critiques adressées très tôt à l'Intelligence Artificielle classique[39] : elle limite préalablement le contexte à un nombre fini de situations identifiées — l'essence de la notion ne se trouve-t-elle pas ainsi niée ? On peut en discuter, mais le connexionnisme ne semble pas avoir sur ce point d'avantage décisif sur le cognitivisme.

Une proposition plus spécifique et plus originale a été faite par P. Smolensky[40] : la sémantique des unités est elle-même dépendante du contexte, dans le cas des représentations distribuées. Le café n'est pas tout à fait la même chose lorsqu'il remplit une tasse et lorsqu'il tache une chemise — ce sont des sous-populations légèrement différentes d'unités qui représentent « café » dans le contexte de la tasse et dans celui de la chemise. C'est pour le coup à un pilier du classicisme que l'on s'attaque ici : les primitives sémantiques sont classiquement indépendantes du contexte, et la sémantique est compositionnelle, ce qui signifie que les sens complexes sont exactement obtenus par des combinaisons appropriées de leurs composants, et elles-mêmes indépendantes du contexte. Il y aurait donc là une possibilité nouvelle — encore faut-il qu'elle reçoive un commencement de concrétisation.

b. Le connexionnisme perd le phénomène central

J. Fodor et Z. Pylyshyn[41] estiment que le connexionnisme est fondamentalement inadéquat comme théorie de la cognition, pour la raison

39. Cf. Hubert L. DREYFUS, *What Computers Can't Do. The Limits of Artificial Intelligence,* New York, Harper, 1972, 2d ed., 1979 ; trad. fr. par Rose-Marie VASSALLO-VILLANEAU, *Intelligence artificielle : mythes et limites,* Paris, Flammarion, 1984.

40. Reprenant une suggestion de Z. Pylyshyn ! *Cf. op. cit. supra* n. 23.

41. *Op. cit. supra* n. 3, et J. A. FODOR, *Psychosemantics. The Problem of Meaning in the Philosophy of Mind,* Cambridge, MA, MIT Press, 1987, Appendice : « Why there still has to be a language of thought ».

qu'il ne fait pas place aux représentations structurées, lesquelles sont seules susceptibles d'expliquer un aspect central de la cognition. Essayons de comprendre ce qu'ils veulent dire.

Voyons d'abord de quel aspect il s'agit. Le langage est notoirement *productif* et *systématique* : il permet d'engendrer une infinité de phrases à partir d'un nombre fini de mots et de règles ; et ses phrases se distribuent sur tout un espace logique, sans laisser d'interstice, au sens où dès qu'y figure quelque chose comme « Jean aime Marie », y figurent nécessairement aussi « Marie aime Jean », « Chacun aime Pierre », etc., et dès qu'y figure « La vache est brune et le cheval est gris », y figurent « La vache est brune » et « Le cheval est gris ». Maîtriser une langue, c'est notamment être en mesure de comprendre et d'émettre une infinité potentielle de phrases (productivité), et de ne pouvoir comprendre ni émettre « Jean aime Marie » sans pouvoir comprendre et émettre « Marie aime Jean », etc. (Maîtriser une langue, c'est donc, explique lumineusement Fodor, tout différent de connaître par cœur une liste de phrases telles que les fournissent les guides pour touristes étrangers.) Or, dit Fodor, ce qui est vrai du langage et de nos capacités linguistiques est vrai de nos pensées et de nos capacités cognitives. La raison ? Celle d'abord que le langage exprime la pensée — comprendre une phrase, c'est saisir la pensée que cette phrase exprime. Mais une réflexion directe sur les pensées dont nous sommes capables conduit à la même conclusion : si nous pouvons penser qu'il fait froid, nous pouvons nécessairement penser qu'il fait très froid, qu'il fait très très froid, etc. ; si nous pouvons penser que le chat est sur le tapis, nous devons pouvoir penser que le tapis est sur le chat ; si nous pouvons vouloir lever le pied gauche et le bras droit, nous devons pouvoir vouloir lever le pied gauche et vouloir lever le bras droit ; etc. (On remarquera en passant la tranquillité avec laquelle Fodor s'appuie sur l'introspection la plus élémentaire, sans même chercher une caution — il faut lui en savoir gré — dans d'hypothétiques expériences de psychologie.)

Voyons maintenant comment Fodor passe de ce constat à l'hypothèse du rôle fondamental des représentations structurées dans la cognition. Depuis Frege, nous rappelle-t-il, nous disposons d'une excellente explication de la productivité et de la systématicité du langage : les formes linguistiques sont obtenues par combinaison de formes primitives ; elles sont structurées en ce sens bien précis, et les processus linguistiques, à tous les niveaux, sont sensibles à la structure de ces représentations. D'un autre côté, on ne dispose d'aucune autre explication de ces phénomènes. La seule attitude rationnelle devant la manifestation des mêmes phénomènes dans un autre domaine, celui des pensées et des processus mentaux, est d'adopter le même genre d'explication. A défaut, donc, d'une

meilleure idée, on mettra au cœur des *états* mentaux des représentations structurées (au sens de la linguistique) — ce seront des formules dans un langage formel interne, le « langage de la pensée » ou « mentalais » ; et l'on attribuera aux *processus* mentaux la propriété essentielle d'être sensibles à la structure des représentations constitutives des états sur lesquels ils opèrent.

Dernière étape, la réfutation du connexionnisme. Les réseaux sont certes des machines à manipuler des représentations, nous disent Fodor et Pylyshyn. Mais leurs représentations ne sont pas structurées. Et à supposer qu'on puisse leur attribuer une structure, les manipulations ne pourraient y être sensibles, car elles reposent sur un principe mathématico-physique de minimisation d'écart, et sur un principe psychologique ou cognitif d'association selon des régularités statistiques — rien qui puisse, sinon accidentellement, assurer une sensibilité à la structure composite des formes matérielles (les populations d'unités activées) répondant à la structure syntaxique des entités représentées.

IV. — LES NIVEAUX

La raison évidente pour laquelle aucun des arguments exposés à l'instant ne saurait par principe être décisif n'est ni leur imprécision, ni leur source introspective ; elle réside dans la possibilité qu'a toute théorie — classique, connexionniste ou autre — de faire appel à une distinction de niveaux. Les classiques, mis en cause par les vertus « phénoménologiques » du connexionnisme, peuvent prendre appui sur l'idée qu'une véritable théorie des phénomènes que celui-ci ne fait somme toute que singer fera nécessairement appel à des entités et à des processus fondamentaux se situant à un niveau inférieur, niveau dont émergeront les propriétés phénoménales en question. Inversement, les connexionnistes, mis au défi par l'objection des représentations structurées, peuvent espérer produire celles-ci à un niveau supérieur, comme effet émergent des processus fondamentaux dont le connexionnisme énonce les lois.

L'appel aux distinctions de niveaux n'est pas l'apanage des théories de la cognition, bien au contraire. Mais celles-ci ont la particularité d'avoir lié leur sort dès l'origine à la postulation d'un niveau particulier, sur la nature duquel les doutes continuent de planer. Les entités qui peuplent ce niveau ne sont pas seulement inaccessibles à l'observation ; elles ne se rangent même pas clairement dans l'une des deux catégories d'inobservables distinguées, comme le rappelle souvent Dennett, par Reichen-

bach[42] : sont-ce des *illata,* c'est-à-dire des objets dont on postule qu'ils existent véritablement, à titre d'entités matérielles — comme des électrons ou des mésons ou des trous noirs ? ou sont-ce des *abstracta,* c'est-à-dire des entités purement théoriques, utiles, voire indispensables, pour ériger des théories portant sur certains objets matériels, mais ne faisant pas elles-mêmes partie de ces objets — comme des centres de gravité ou des trajectoires ?

Les entités dont les sciences cognitives parlent — l'information, les représentations, le contenu sémantique, le traitement de l'information... — demeurent profondément mystérieuses, ce qui n'empêche nullement bon nombre de programmes de recherche de progresser, mais qui rend difficile tout débat sur les fondements. Comme le dit Barwise[43], notre situation rappelle celle des hommes de l'âge du bronze, qui maniaient le bronze fort bien sans posséder ce que nous considérerions aujourd'hui comme une théorie acceptable du bronze.

Le statut ontologique et épistémologique du niveau informationnel, ou représentationnel, est pour les classiques d'obédience fodorienne un dogme — un dogme moderne sans doute : chacun reste libre de le rejeter. Mais, nous avertit charitablement Fodor, celui qui choisit de le faire s'exclut par là même de la science cognitive ; il fait de la biologie, de l'électronique, de la physique, ce qu'on voudra, il n'est plus dans la course, pas plus que ne demeurerait géologue celui qui prétendrait réfuter les théories régnantes en géologie en ne parlant que de la théorie quantique des champs. Ce n'est pas la possibilité d'une science cognitive que Fodor pose comme dogme ; c'est l'existence du niveau représentationnel comme condition nécessaire de cette possibilité. Et si le connexionnisme est à ses yeux la nouvelle Carthage à détruire, c'est qu'il prétend fonder la science cognitive sur un autre niveau, et en faire la science de ce niveau.

Mais quelle est la nature de la hiérarchie même qu'implique la référence à un niveau privilégié ? Voilà qui n'est pas clair non plus, et qui mériterait à soi seul de longs développements. Parle-t-on de niveaux de description ? Parle-t-on d'échelles de grandeur et de niveaux d'organisation, chacun étant peuplé d'agrégats d'entités peuplant le niveau immédiatement inférieur ? Parle-t-on de niveaux d'intégration, comme les neurophysiologistes ?

Les classiques parlent essentiellement de niveaux de description, et en

---

42. Hans REICHENBACH, *Experience and Prediction,* Chicago, University of Chicago Press, 1938, p. 211-212 ; cité *in* Daniel C. DENNETT, *The Intentional Stance,* Cambridge, MA, MIT Press, 1987, p. 53, trad. fr. à paraître, Paris, Gallimard.

43. « Information and Circumstance », *art. cit. supra* n. 37.

même temps, secondairement, de niveaux d'abstraction : le niveau privilégié est donc obtenu par une double opération, l'une, cruciale et originale à leurs yeux, de visée, l'autre, secondaire et banale, d'idéalisation. Acceptons sans la discuter ici la seconde, et décrivons la première. Les états cérébraux d'un organisme sont classés selon la relation d'équivalence « joue le même rôle que... », le rôle étant celui que tient un état mental vis-à-vis des autres états internes et des états perceptuels et moteurs. Mais bien entendu, c'est sous une certaine description que ces états peuvent être vus comme interagissant les uns sur les autres (sont-ce des molécules de bois, ou bien une latte de parquet posée de guingois qui provoquent ma chute ? ou bien encore ce qu'elles provoquent serait-il la modification de la trajectoire des atomes qui me composent ?). On ne peut donc que *postuler* que ces états sont notamment informationnels, ou comme on dit parfois, « sémantiquement évaluables » : le fonctionnalisme offre une solution au rapport corps-esprit, mais à un coût élevé ; en ce sens, il n'est aucunement réductionniste. Le niveau informationnel demeure infondé. Admettons donc qu'un état mental soit un état cérébral caractérisé sémantiquement ou informationnellement — ou plus exactement la classe d'équivalence des états cérébraux ayant le même rôle informationnel dans l'économie du système. Que seront alors les *processus* mentaux ? C'est ici que la métaphore de l'ordinateur joue un rôle crucial : l'équivalent des états mentaux sont les états internes de la machine de Turing (ces états sont eux-mêmes, et c'est heureux, des abstractions ; ce sont des classes d'équivalence dans un espace abstrait de configurations) ; or ces états se transforment les uns dans les autres sous l'effet des opérations de la machine, opérations parfaitement définies sur le plan abstrait, et dont la réalisation matérielle est maîtrisée théoriquement : il n'y a pas de mystère dans la façon dont un calculateur concret réalise ou incarne les opérations d'une machine de Turing. *Ergo* les opérations qui font passer d'un état mental à un autre sont identiques ou du moins semblables aux opérations d'une machine de Turing, et il n'y a pas de mystère *théorique* dans l'idée que les processus cérébraux réalisent ces opérations. (Il y a certes un mystère empirique, dont la solution est d'autant plus impatiemment attendue qu'elle devra révéler le pourquoi de la stabilité de ces processus vis-à-vis de la relation d'équivalence sur les états neurophysiologiques : il faut que deux états équivalents soient transformés en deux autres états équivalents, ce qui est une contrainte extrêmement forte.)

Comment concevoir alors le rapport entre le niveau qui est celui des classes d'équivalence d'états cérébraux informationnels et le niveau immédiatement inférieur ou sous-jacent ? La question est ambiguë, et c'est là une source de confusions. *Primo*, en tant que niveau de description, il a pour soubassement le niveau des états cérébraux eux-mêmes

(une classe d'états cérébraux renvoie, dans ce rapport, à n'importe lequel de ses éléments), mais bien entendu *sous une certaine description* ; cette description est celle que fournissent les neurosciences. Mais les neurosciences fournissent *plusieurs* descriptions, selon le niveau d'*organisation* auquel elles se placent. De laquelle s'agit-il ? Nécessairement de celle qui porte sur des entités de la « taille » de celles qui peuplent le niveau informationnel : si ce sont, par exemple, des assemblées de neurones d'un certain type qui sont porteuses de représentations atomiques variables, des colonnes de neurones d'un autre type qui sont porteuses de représentations logiques fixes, le niveau cherché est celui de ces assemblées et de ces colonnes. *Secundo,* ce niveau est à son tour en relation avec un niveau d'organisation immédiatement inférieur, qui pourrait être par exemple celui des neurones individuels caractérisés par le traitement de certains signaux qu'ils effectuent.

Finalement le rapport dont parlent les classiques, qui est celui de l' « implémentation » du niveau informationnel dans le « wetware », est purement conceptuel, vide de contenu empirique — on conçoit mal une preuve empirique de la non-existence d'un niveau d'intégration neurophysiologique correspondant au niveau informationnel, ou de la non-existence de processus neurophysiologiques réalisant les opérations postulées à ce dernier niveau, et stables pour la relation d'équivalence. Il n'y a, dans la doctrine classique, aucune théorie substantielle de l'émergence d'entités sémantiques à partir d'un niveau fondamental — contrairement à ce qui est parfois suggéré, le candidat avancé étant l'implémentation ou instantiation. Au contraire, l'existence de niveaux d'intégration du tissu nerveux menant par paliers du neurone individuel à un niveau « sous-jacent » au niveau informationnel est riche de contenu empirique ; en ce sens, il est vrai que les sciences cognitives peuvent avoir pour les neurosciences une valeur heuristique, alors qu'inversement celles-ci restent sans aucune influence sur une conception *classique* rigoriste des sciences cognitives. Lorsqu'on parle des contraintes qu'exerceraient les uns sur les autres les différents niveaux de description, et donc les différentes sciences du « cerveau/esprit », on va trop vite.

Le tableau classique se complique encore par le fait que le niveau informationnel lui-même peut se ramifier indéfiniment, on l'a vu, par le jeu de machines virtuelles nichées les unes dans les autres. Si bien que souvent, lorsqu'il est question de la relation classique entre niveaux, on parle d'une relation exacte d'implémentation — il vaudrait mieux dire de compilation — entre deux langages ou deux machines virtuelles. Si l'on veut par exemple, dans le cadre classique, faire droit à l'idée d'un niveau de micro-entités et de micro-processus, sous-jacent au niveau des phénomènes mentaux « visibles », par exemple accessibles à l'introspection, on

postulera un langage « machine » ou de niveau inférieur, considéré comme définissant les « véritables » opérations physiques de la machine, et un langage « évolué », « compilé » ou de niveau supérieur, dans lequel s'inscrivent, ou au niveau duquel émergent, les « macro »-phénomènes. Mais, ajoutera-t-on, il existe par définition une réduction sans résidu du niveau supérieur au niveau inférieur : les opérations « réelles » rendent *exactement* compte des phénomènes émergents ou virtuels commodément décrits dans le langage évolué.

C'est précisément à cette réduction sans résidu que veut échapper le connexionnisme. Le sous-titre même de la bible du courant PDP, « investigations de la *micro-structure* de la cognition », indique suffisamment son ambition de découvrir le niveau fondamental auquel les choses se passent réellement, c'est-à-dire auquel on peut donner une description systématique, exhaustive et exacte des *processus* en jeu dans la cognition. Or quel est ce niveau ? Il ne peut s'agir « du » niveau physique, c'est-à-dire d'un niveau de description physique ou biologique des états cérébraux. Il importe, en effet, de « rester dans la course », c'est-à-dire de parler d'entités sémantiquement évaluables, d'états et de processus informationnels. Sur ce point, les connexionnistes de tendance PDP, comme on l'a vu, acceptent la règle de Fodor : ils sont fonctionnalistes, et refusent de se laisser cantonner, comme les y invitent Fodor et Pylyshyn[44], au rôle d' « implémentateurs ». Ils refusent, en d'autres termes, d'être ceux qui montrent comment on peut réaliser les fonctions classiques dans un substrat (abstrait) — on parle dans ce contexte d' « architecture » — non classique.

Il leur faut donc postuler l'existence d'un niveau plus fondamental ou « micro » que le niveau « symbolique » des classiques, mais encore (dans l'image d'une descente de la cognition vers la matière, qui mène en même temps des phénomènes les plus complexes aux moins complexes) informationnel. Cependant, s'ils veulent que ce niveau soit effectivement différent du niveau symbolique (ce n'est pas le cas de tous les connexionnistes : beaucoup se contentent de proposer une théorie rivale au même niveau que les classiques), ils doivent éviter de le munir d'une sémantique classique. Ils postulent donc un « saut » *(shift)* sémantique[45] entre les macro-entités, les assemblées d'unités, et les micro-entités que sont les unités individuelles. Celles-ci, dans le paradigme des représentations distribuées que l'on a mentionné au § 2, se voient munies de particules de sens dont seul l'assemblage en cours de traitement peut fournir un véritable sens ou une représentation au sens classique. Ainsi, le niveau

44. « Connectionism as a Theory of Implementation » est le titre d'une section de leur *op. cit. supra* n. 3, p. 64-66.

45. Cf. P. SMOLENSKY, *art. cit. supra* n. 23.

fondamental serait-il « subsymbolique » ; il serait caractérisé par une « syntaxe » exacte, c'est-à-dire par des transformations exactement spécifiables dans le vocabulaire de la physique mathématique, et par une « proto-sémantique » floue. Au niveau supérieur d'organisation émergeraient des phénomènes affectant des assemblées d'unités munies d'une pleine sémantique, mais régies par une « syntaxe » floue : toute caractérisation des transformations les affectant serait approximative. La description au niveau supérieur ne se réduirait donc pas à la description au niveau inférieur, et le rapport entre les niveaux se rapprocherait plus de celui qui s'établit entre microphysique et physique newtonienne que du rapport de compilation entre langages informatiques.

Remarquons en passant que sous l'angle de la sémantique et de la « syntaxe » (toujours dans le sens, abusif, de système des lois de transition entre états), l'existence de sauts crée donc un espace logique à quatre positions, toutes occupées, présenté sans autre explication dans le tableau suivant :

Saut sémantique Saut « syntaxique »	NON	OUI
NON	Psychologie cognitive classique des processus « personnels » ; Intelligence Artificielle classique	Psychologie cognitive classique des processus « subpersonnels » ; p. ex. psycholinguistique, vision artificielle
OUI	Connexionnisme modéré de type « localiste » ; psychologie du sens commun *(folk psychology)*	Connexionnisme radical ; paradigme subsymbolique

Pour séduisante qu'elle soit par certains aspects, la position connexionniste radicale défendue par Smolensky (le « paradigme subsymbolique » occupant la case la plus « exotique » du tableau) demeure fragile. Car en refusant d'assimiler le niveau « subsymbolique » au niveau informationnel des classiques, elle se prive d'un solide ancrage dans l'intuition et l'introspection : que nos états mentaux soient sémantiquement évaluables, ou si l'on préfère intentionnels au sens de Brentano — qu'ils visent quelque chose qui leur est extérieur — est certes un grand mystère, mais porte en même temps la marque de l'évidence. Les entités subsymboliques seraient intentionnelles elles aussi, mais sans qu'on puisse spécifier dans le vocabulaire de la psychologie, même étendu, à quoi elles ren-

voient. Et en refusant, d'autre part, d'assigner au niveau subsymbolique, serait-ce au prix d'une schématisation, une place parmi les niveaux d'intégration ou d'organisation du tissu cérébral, ce connexionnisme-là se prive d'un solide ancrage dans la tradition de la modélisation physique. Dire, comme le fait Smolensky, que ce niveau est intermédiaire entre celui des classiques et celui des neurosciences n'est guère défendable — ce n'est sans doute pas même cohérent, si du moins l'on accepte l'analyse qui vient d'être proposée du niveau fondamental des classiques, dont Smolensky entend conserver la conception fonctionnaliste. Les connexionnistes qui interprètent leurs propres efforts comme des prolégomènes à une neurophysiologie théorique ou « computationnelle » (je préférerais, tout simplement, « mathématique ») se placent sur un terrain plus solide, même s'ils doivent faire face à de difficiles objections concernant la « plausibilité » neurophysiologique de leurs modèles.

V. — LES STRUCTURES, LE TEMPS ET LA SIGNIFICATION

Le reproche fondamental adressé par Fodor et Pylyshyn aux réseaux, qui est d'être constitutionnellement incapables de recevoir et de manipuler adéquatement des représentations structurées, n'a pas laissé les connexionnistes indifférents. Ils ont de fait proposé plusieurs procédés de représentations des structures dans les réseaux ; la première proposition, due à Geoffrey Hinton[46], précède même de plusieurs années l'interpellation par les défenseurs du classicisme !

S'il est impossible dans le cadre du présent article de passer en revue les solutions avancées, on peut indiquer certaines des conclusions qui se dégagent d'un tel examen. En premier lieu, il n'existe aucun système couvrant tous les aspects de la représentation et de la manipulation des structures. Il est même assez difficile de discerner, dès que l'on quitte le cadre classique, avec ses formules d'un langage formel, ce qu'ont en commun les procédés connexionnistes de représentations de relations (« Pierre est le père de Marie », « Marie est la sœur de Paul »...), ou de suites finies ($a_1 b_2$ c_3...), de parcours d'une liste, de représentations de classifications (les A sont des B ou des C, les B des X, des Y ou des Z, les C des U, des V ou des W, etc.), de réalisation de « systèmes de production » (au sens de Newell : systèmes modifiant graduellement une base de données par l'application de « règles de production » de la

46. « Implementing Semantic Networks in Parallel Hardware », *in* G. HINTON, J. ANDERSON, eds, *op. cit. supra* n. 5.

forme : si A(x) et B(x), alors C(x) — si, pour une valeur a, figurent dans la base à un instant donné les formules A(a) et B(a), on ajoute à la base, à l'instant suivant, C(a)). En second lieu, on peut distinguer trois types de solutions : celles qui n'affichent d'autre ambition que d' « implémenter » (en un sens qui demande à être précisé) des langages ou des traitements classiques ; celles qui consistent à mettre en évidence, par exemple par une analyse statistique de l'activité des unités au cours du traitement des différentes données, une hiérarchisation de ces unités reflétant la classification naturelle des données ; celles, enfin, qui consistent à inscrire dans le medium des unités des suites de symboles, sans perdre l'esprit anticlassique de la modélisation connexionniste[47].

De ces trois approches, la plus intéressante sur le plan théorique est évidemment la troisième. Mais les solutions auxquelles elle a conduit jusqu'à présent sont à la fois très compliquées et très en deçà de nos attentes car, d'une part, elles limitent *a priori* la longueur ou la complexité des représentations et, d'autre part, elles ne ménagent pas un libre recours à la récurrence. Quoi qu'il en soit, il est intéressant de repérer les différences par rapport à la solution classique. La première est que le réseau doit se subdiviser de manière permanente en sous-réseaux, chaque sous-réseau étant dédié à la représentation particulière d'un rôle ou place dans la structure[48]. La seconde est que pour manipuler les représentations, le réseau doit faire appel au temps d'une manière beaucoup plus fondamentale qu'un système classique. En effet, pour exécuter la transformation de XYZ en X'Y'Z', le réseau doit prendre à l'instant t la configuration XYZ, et à l'instant suivant t' la configuration X'Y'Z' : ne pouvant « dire » XYZ, comme le système classique, le réseau « est » en un sens, ou encore « mime » XYZ, donc pour « être » ou « mimer » X'Y'Z', il doit se transformer — quitte du reste à perdre la trace de son état antérieur et à oublier ainsi la provenance de son nouvel état. Le système classique, lui, dispose d'un tableau noir sur lequel il peut inscrire XYZ, et aussi, sans même l'effacer, X'Y'Z' : l'essence d'un tel système est la division intangible entre la partie variable, lieu des inscriptions, et la partie fixe, transformateur des inscriptions.

On peut donc se demander si le connexionnisme ne doit pas profiter

---

47. Trois exemples caractéristiques de ces approches sont respectivement : 1. David S. TOURETSKY, « BoltzCONS : Dynamic Structures in a Connectionist Network », CMU-CS-89-182 Technical Report, Carnegie Mellon University, August 1989 ; 2. Jeffrey ELMAN, « Representation and Structure in Connectionist Models », CRL Technical Report 8903, Center for Research in Language, University of California at San Diego, 1988 ; 3. P. SMOLENSKY, « Tensor Product Variable Binding and the Representation of Symbolic Structures in Connectionist Networks », *Artificial Intelligence,* sous presse.

48. Je simplifie : le schéma imaginé par Smolensky (cf. *supra* n. 47) est plus subtil ; il implique néanmoins une partition fixe du réseau en sous-réseaux spécialisés.

de sa différence pour opérer un changement radical de perspective, plutôt que de se contorsionner pour imiter le classicisme : s'il le fait trop fidèlement, il risque, en effet, de n'être qu'implémentation (ce qui peut comporter d'intéressantes retombées techniques : il y a là sans doute pour l'Intelligence Artificielle le moyen de surmonter certaines de ses faiblesses) ; s'il ne parvient que partiellement et maladroitement, c'est l'intérêt même de la tentative qui est en cause. Or, que suggère l'insistance des connexionnistes sur leur intérêt pour la « microstructure » de la cognition ? C'est la distinction simple dans son principe, mais si difficile à énoncer clairement dans le cas de la cognition, entre l'*organe* et ses *produits* (qui n'est autre, on peut le présumer, que la vénérable mais plus obscure distinction entre la structure et la fonction). Cette distinction est gommée dans la perspective classique, mais fait retour sous la forme du fantôme dans la machine — celui pour lequel les représentations, si commodément inscrites sur le tableau noir intérieur, représentent. L'intérieur reflète l'extérieur, le langage de la pensée (producteur, agent) mime le langage tout court (produit, agi), comme dans l'univers de la Renaissance — « univers de miroirs dans lequel tout se reflète dans tout » (selon l'expression de R. Gadoffre).

Pourquoi le connexionnisme ne tenterait-il pas de concevoir une machine dont le fantôme serait d'emblée exorcisé, au lieu d'attendre de l'être à la fin, combien hypothétique, des travaux, et qui aurait sa propre notion de ce qui représente quelque chose, de ce qui *pour elle* possède une signification ? Une machine que l'on doterait progressivement des capacités cognitives fondamentales ? La première de ces capacités serait la mémoire d'objets isolés — objectif qui, convenablement généralisé, peut être considéré comme celui de la première vague du connexionnisme contemporain. La seconde capacité fondamentale viserait, par un nouveau biais, le problème des structures ; il s'agit de la mémoire de suites enchaînées d'objets. Ces suites jouent, en effet, un rôle fondamental : toute la cognition, tout le comportement intelligent reposent sur la reconnaissance et la production de suites : suite de notes (air de musique), suite de phonèmes, de mots, de phrases, suite d'événements identiques (coups de cloche), d'où la numération, les suites de déductions, les suites de gestes... Pour obtenir, à partir de là, les représentations structurées des classiques, il ne manquerait que la capacité de typer certains objets.

C'est précisément à ce programme que s'est attelé, avec succès, un connexionniste physicien d'obédience ANN, Daniel Amit[49]. Souvenons-nous que les réseaux, étant munis de connexions multidirectionnelles

49. Cf. *op. cit. supra* n. 5, et « Neural Networks Counting Chimes », *Proc. Natl. Acad. Sc. USA,* vol. 85, 1988, p. 2141 sq.

(donc de *feedback*), sont ici vus non comme des machines entrée/sortie, mais comme des systèmes dynamiques doués d'une dynamique endogène, sujette à des perturbations provenant de l'environnement. Tout retour rapide à l'équilibre est, par définition, un événement cognitivement significatif pour le réseau — ce qui n'est pas arbitraire, dans la mesure où un tel événement est repérable par un certain type de neurone formel situé à l'extérieur du réseau. L'attracteur vers lequel le réseau vient de converger constitue le contenu de l'événement ; et ce contenu est à son tour la représentation de l'événement perturbateur initial. On pourra dire, revenant maintenant à l'idée de Hopfield[50], que le réseau a le souvenir *de* cet événement, souvenir dont le *contenu* est l'attracteur auquel il a conduit le réseau. Arrêtons-nous un bref instant sur cette proposition : quel saisissant contraste par rapport à la doctrine classique, et même par rapport au connexionnisme PDP ! Rien d'équivalent à ce critère de signification intrinsèque (de « signifiance ») ne semble se présenter de façon naturelle dans ces derniers. Tout stimulus ne doit-il pas être significatif, puisqu'il ne peut qu'ébranler un système qui, en l'absence de stimulus, est au repos ? On pourrait être tenté de restreindre la qualité de signifiance aux stimulus conduisant le système à l'équilibre — mais c'est là une propriété indécidable : on le sait dans le cas classique (c'est le célèbre problème de l'arrêt de Turing), et c'est encore vrai, pour de toutes autres raisons, dans le cas des réseaux PDP. Ce qui, en pratique, nous ramène au non-critère précédent. On pourrait aussi délimiter arbitrairement le domaine de la signifiance, ce qui reviendrait, quelle que soit la solution technique adoptée, à faire attribuer au système une valeur distinguée (« non significatif ») à certains stimuli. Mais qui ne voit que *pour le système* un stimulus ainsi traité serait bel et bien significatif ?

Venons-en à la seconde contribution d'Amit. Le problème initial, posé par Donald Hebb[51], était le suivant : quoique deux coups de cloche soient, en tant que stimuli, indiscernables, comment se fait-il que nous n'y réagissions pas de la même façon — que par exemple nous disions « 1 » après l'un, « 2 » après le suivant ? La solution d'Amit prend la forme d'un réseau capable de compter — dans les deux sens du terme : il sait compter « en l'air », c'est-à-dire réciter (un segment initial de) la suite des entiers, et il sait compter les éléments d'une suite d'événements identiques. Voici comment. Le réseau est muni de deux types de connexions, les synapses rapides, qui ont les caractéristiques habituelles,

50. Cf. *op. cit. supra* n. 22. C'est par commodité que j'attribue l'idée à cet auteur : il en partage la paternité avec plusieurs autres chercheurs (T. Kohonen, S. Grossberg, etc.), dont certains revendiquent même l'antériorité.

51. Donald O. HEBB, *Essay on Mind*, Hillsdale, NJ, Erlbaum, 1980.

et les synapses lentes, qui ne transmettent une quantité non négligeable d'influx qu'après un certain temps d'accumulation. La dynamique du réseau peut être vue, en première approximation, comme résultant de la superposition d'une dynamique rapide et d'une dynamique lente. Partant d'un état initial quelconque, le réseau atteint rapidement, soit spontanément (comptage « en l'air »), soit suite à un premier coup de cloche, une position d'équilibre provisoire (« quasi attracteur » de sa dynamique rapide). Au bout d'un certain intervalle de temps, dont l'ordre de grandeur est déterminé précisément (il est compris entre celui du retour à l'équilibre selon la dynamique rapide, et celui de la stabilisation « définitive »), la dynamique lente, aidée, dans le cas des coups de cloche, par un nouveau coup, déstabilise le système, qui regagne rapidement une nouvelle position d'équilibre provisoire, et ainsi de suite jusqu'à la stabilisation finale. Pendant ce temps, un neurone spécialisé détecte les phases successives de convergence vers un quasi-attracteur, et fait « tourner » un compteur. Il détecte également la stabilisation finale, ce qui permet au compteur de se remettre à zéro (pour éviter de compter « douze » au premier coup de minuit...).

Ce réseau n'est pas seulement doté de la capacité de compter des événements cognitivement significatifs ; il possède aussi une notion de temps intrinsèque, donnée par les stabilisations successives dans des quasi-attracteurs. Ce temps n'est pas celui de la pulsation des mises à jour des activités, temps uniforme dépourvu de signification pour le système, temps des transitions « inconscientes » entre deux états « conscients » ; ce n'est pas davantage le temps de l'horloge interne de l'ordinateur classique, également dépourvu de signification, incapable de distinguer, parmi les événements qui le scandent, ceux qui seraient significatifs. C'est vraiment le temps *pour* le système, et non pas seulement le temps *du* système. Voilà donc un concept — certes très particulier —, le temps, représenté *dans* et *pour* le système ; non par un symbole arbitraire, mais par une unité liée de façon très spécifique à l'ensemble du réseau (le neurone détectant les retours rapides au quasi-équilibre).

***

Le connexionnisme — peut-être serait-il préférable de parler des connexionnismes — ne fournissent pas, c'est bien clair, toutes les réponses ; encore n'avons-nous pas évoqué toutes leurs difficultés (notamment celles que pose le passage à des problèmes en vraie grandeur : seuls pourraient les résoudre des réseaux beaucoup plus grands que ceux que l'on utilise aujourd'hui ; or de tels réseaux sont parfaitement incontrôlables dans l'état actuel de nos connaissances — c'est la

version connexionniste de l'explosion combinatoire). Mais le plus difficile, pour le théoricien, est encore de déterminer les questions auxquelles ils apportent des réponses. C'est du reste un trait qu'ils partagent avec le classicisme (au sein duquel il faudrait aussi, mais c'est une autre histoire, distinguer plusieurs doctrines et un grand nombre de programmes de recherche). Chacune de ces approches apporte des solutions dont on a le plus grand mal à préciser la place exacte dans l'économie générale de la cognition. Le classicisme permet, semble-t-il, de comprendre comment des représentations élémentaires peuvent être combinées de manière effective pour permettre à un système ou agent cognitif de déployer des comportements complexes adaptés ou intelligents. Le connexionnisme PDP nous donne une conception étendue de la perception comme détection dans l'environnement de régularités statistiques d'ordre arbitrairement élevé. Le connexionnisme ANN nous aide à concevoir des systèmes auto-organisés, mus par une dynamique interne, capables de distinguer dans le flux des processus certains événements significatifs, munis de concepts intrinsèquement significatifs.

Tantôt ces théories semblent complémentaires — mais comment les articuler ? Tantôt elles se posent en concurrentes — mais comment trancher, et surtout comment, si l'on en choisit une, faire droit aux aspects que seules les autres semblent en mesure d'expliquer, de respecter, voire de seulement formuler ? Gageons que notre perplexité ne prendra pas fin de sitôt. Ayons du moins l'audace d'espérer qu'à force de suivre tantôt une voie, tantôt une autre, nous nous fassions petit à petit une idée plus riche et plus précise de ce singulier monument de la nature qu'est l'esprit juché sur le cerveau.

Daniel ANDLER,
Université Charles de Gaulle-Lille 3,
C.R.E.A., École polytechnique, Paris.