Harnad, S. (1990) The Symbol Grounding Problem. Physica D 42: 335-346.

LE PROBLÈME DE L'ANCRAGE DES SYMBOLES

Stevan Harnad
Department of Psychology
Princeton University
Princeton NJ 08544
harnad@cogsci.soton.ac.uk

ABSTRACT: Il y a eu récemment beaucoup de discussions sur la portée et les limites des modèles purement symboliques de l'esprit et sur le rôle propre du connexionnisme dans la modélisation cognitive. Cet article décrit le « problème de l'ancrage des symboles »: comment l'interprétation sémantique d'un système de symboles formel peut-elle être rendue intrinsèque au système, plutôt que simplement parasite sur les significations dans nos têtes? Comment les significations des jetons de symboles sans signification, manipulés uniquement sur la base de leurs formes (arbitraires), peuvent-elles être ancrées dans autre chose que d'autres symboles sans signification? Le problème est analogue à essayer d'apprendre le chinois à partir d'un dictionnaire chinois / chinois seul. Une solution candidate est esquissée: les représentations symboliques doivent être ancrées de bas en haut dans des représentations non symboliques de deux types: (1) «représentations iconiques», qui sont des analogues des projections sensorielles proximales d'objets et d'événements distaux, et (2) «représentations catégoriques ", qui sont des détecteurs de caractéristiques appris et innés qui sélectionnent les caractéristiques invariantes des catégories d'objets et d'événements à partir de leurs projections sensorielles. Les symboles élémentaires sont les noms de ces catégories d'objets et d'événements, attribués sur la base de leurs représentations catégoriques (non symboliques). Les "représentations symboliques" d'ordre supérieur (3), ancrées dans ces symboles élémentaires, consistent en des chaînes de symboles décrivant les relations d'appartenance à une catégorie (par exemple, "Un X est un Y qui est Z"). Le connexionnisme est un candidat naturel pour le mécanisme qui apprend les caractéristiques invariantes sous-jacentes aux représentations catégorielles, reliant ainsi les noms aux projections proximales des objets distaux qu'ils représentent. De cette manière, le connexionnisme peut être considéré comme une composante complémentaire dans un modèle hybride non symbolique / symbolique de l'esprit, plutôt que comme un rival de la modélisation purement symbolique. Un tel modèle hybride n'aurait cependant pas de «module» symbolique autonome; les fonctions symboliques émergeraient comme un système symbolique intrinsèquement «dédié» à la suite de l'ancrage ascendant des noms de catégories dans leurs représentations sensorielles. La manipulation des symboles serait régie non seulement par les formes arbitraires des jetons de symboles, mais par les formes non arbitraires des icônes et des invariants de catégorie dans lesquels ils sont ancrés.

KEYWORDS: symbol systems, connectionism, category learning, cognitive models, neural models

1. Modéliser l’esprit

1.1 Du béhaviorisme au cognitivisme

Pendant de nombreuses années, la seule approche empirique en psychologie a été le comportementalisme, ses seuls outils explicatifs des associations entrée / entrée et entrée / sortie (dans le cas du conditionnement classique; Turkkan 1989) et l'histoire des récompenses / punitions qui ont "façonné" le comportement (dans le cas du conditionnement opérant; Catania & Harnad 1988). En réaction contre la subjectivité de l'introspectionnisme en fauteuil, le behaviorisme avait déclaré qu'il était tout aussi illicite de théoriser ce qui se passait dans la tête de l'organisme pour générer son comportement que de théoriser ce qui se passait dans son esprit. Seuls les observables devaient faire l'objet de la psychologie; et, apparemment, on attendait d'eux qu'ils s'expliquent.

La psychologie est devenue davantage une science empirique lorsque, avec l'avènement progressif du cognitivisme (Miller 1956, Neisser 1967, Haugeland 1978), il est devenu acceptable de faire des inférences sur les processus inobservables sous-jacents au comportement. Malheureusement, le cognitivisme a également laissé entrer le mentalisme par la porte dérobée, car les processus internes hypothétiques se sont agrémentés d'interprétations subjectives. En fait, l'interprétabilité sémantique (signification), comme nous le verrons, était l'une des caractéristiques déterminantes du candidat le plus en vue en lice pour devenir le vocabulaire théorique du cognitivisme, le «langage de la pensée» (Fodor 1975), qui devint la vision dominante. en théorie cognitive depuis plusieurs décennies sous la forme du modèle «symbolique» de l'esprit: l'esprit est un système de symboles et la cognition est une manipulation de symboles. La possibilité de générer un comportement complexe grâce à la manipulation de symboles a été démontrée empiriquement par des succès dans le domaine de l'intelligence artificielle (IA).

1.2 Systèmes symboliques

Qu'est-ce qu'un système de symboles? À partir de Newell (1980) Pylyshyn (1984), Fodor (1987) et des travaux classiques de Von Neumann, Turing, Goedel, Church, etc. (voir Kleene 1969) sur les fondements du calcul, nous pouvons reconstruire la définition suivante:

Un système de symboles est:

1. un ensemble de «jetons physiques» arbitraires, des rayures sur le papier, des trous sur une bande, des événements sur un ordinateur numérique, etc.
2. manipulé sur la base de "règles explicites" qui sont
3. de même des jetons physiques et des chaînes de jetons. La manipulation de jeton de symbole régie par des règles est basée
4. uniquement sur la forme des jetons de symboles (et non sur leur "signification"), c'est-à-dire qu'il est purement syntaxique et se compose de
5. "Combiner avec régularité" et recombiner les jetons de symboles. Il y a
6. jetons de symboles atomiques primitifs et
7. chaînes de jetons de symboles composites. L'ensemble du système et toutes ses parties - les jetons atomiques, les jetons composites, les manipulations syntaxiques à la fois réelles et possibles et les règles - sont tous

8. "interprétable sémantiquement:" On peut attribuer systématiquement à la syntaxe une signification, par exemple comme représentant des objets, comme décrivant des états de choses).

Selon les partisans du modèle symbolique de l'esprit tels que Fodor (1980) et Pylyshyn (1980, 1984), les chaînes de symboles de ce type capturent ce que sont les phénomènes mentaux tels que les pensées et les croyances. Les symbolistes soulignent que le niveau symbolique (pour eux, le niveau mental) est un niveau fonctionnel naturel qui lui est propre, avec des régularités régulières qui sont indépendantes de leurs réalisations physiques spécifiques. Pour les symbolistes, cette indépendance de mise en œuvre est la différence critique entre les phénomènes cognitifs et les phénomènes physiques ordinaires et leurs explications respectives. Ce concept de niveau symbolique autonome est également conforme aux principes fondateurs généraux de la théorie du calcul et s'applique à tout le travail effectué dans l'IA symbolique, la branche de la science qui a jusqu'à présent été la plus réussie à générer (donc à expliquer) un comportement intelligent .

Les huit propriétés énumérées ci-dessus semblent être essentielles à cette définition du symbolique. [1] De nombreux phénomènes ont certaines propriétés, mais cela n'implique pas qu'ils soient symboliques dans ce sens explicite et technique. Il ne suffit pas, par exemple, qu'un phénomène soit interprétable comme régi par des règles, car à peu près tout peut être interprété comme régi par des règles. Un thermostat peut être interprété comme suivant la règle: Allumez la fournaise si la température descend en dessous de 70 degrés et l'éteignez si elle dépasse 70 degrés, mais cette règle n'est explicitement représentée nulle part dans le thermostat. Wittgenstein (1953) a souligné la différence entre les règles explicites et implicites: ce n'est pas la même chose de «suivre» une règle (explicitement) et simplement de se comporter «conformément à» une règle (implicitement). [2] La différence critique réside dans les critères de composition (7) et de systématique (8). La règle symbolique explicitement représentée fait partie d'un système formel, elle est décomposable (sauf si primitive), son application et sa manipulation sont purement formelles (syntaxique, dépendant de la forme), et l'ensemble du système doit être sémantiquement interprétable, pas seulement le bloc en question . Un bloc isolé («modulaire») ne peut pas être symbolique; être symbolique est une propriété systématique.

Ainsi, le simple fait qu'un comportement soit "interprétable" comme étant réglementaire ne signifie pas qu'il est réellement régi par une règle symbolique [3]. L'interprétabilité sémantique doit être couplée à une représentation explicite (2), une manipulabilité syntaxique (4) et une systématisation (8) pour être symbolique. Aucun de ces critères n'est arbitraire et, pour autant que je sache, si vous les affaiblissez, vous perdez l'emprise sur ce qui ressemble à une catégorie naturelle et vous coupez les liens avec la théorie formelle du calcul, laissant un sens de «symbolique "qui est simplement une métaphore inexpliquée (et diffère probablement d'un locuteur à l'autre). C'est donc seulement ce sens formel de «symbolique» et de «système de symboles» qui sera considéré dans cette discussion sur l'ancrage des systèmes de symboles.

1.3 Systèmes connexionnistes (réseaux de neurones artificiels)

Un premier rival du modèle symbolique de l'esprit est apparu (Rosenblatt 1962), a été vaincu par l'IA symbolique (Minsky & Papert 1969) et a récemment réapparu sous une forme plus forte qui rivalise actuellement avec l'IA pour être la théorie générale de la cognition et comportement (McClelland, Rumelhart et al.1986, Smolensky 1988). Décrite de manière variée comme «réseaux neuronaux», «traitement distribué parallèle» et «connexionnisme», cette approche a un programme multiple, qui comprend la fourniture d'une théorie de la fonction cérébrale. Maintenant, on peut dire beaucoup pour et contre l'étude indépendante des fonctions comportementale et cérébrale, mais dans cet article, on supposera que, avant tout, une théorie cognitive doit reposer sur ses propres mérites, qui dépendent de la façon dont elle explique notre comportement observable. capacité. Qu'il le fasse ou non d'une manière suffisamment cérébrale est une autre question, et une question en aval, au cours du développement de la théorie. On en sait très peu sur la structure du cerveau et ses fonctions «inférieures» (végétatives) à ce jour; et la nature de la fonction cérébrale «supérieure» est elle-même une question théorique. «Contraindre» une théorie cognitive à rendre compte du comportement de manière cérébrale est donc prématuré à deux égards: (1) il est loin d'être clair encore ce que signifie «cérébral», et (2) nous sommes loin d'avoir pris en compte une taille de vie. morceau de comportement encore, même sans contraintes supplémentaires. De plus, les principes formels qui sous-tendent le connexionnisme semblent reposer sur la structure associative et statistique des interactions causales dans certains systèmes dynamiques; un réseau de neurones n'est qu'une implémentation possible d'un tel système dynamique. [4]

Le connexionnisme ne sera donc considéré ici que comme une théorie cognitive. En tant que tel, il a récemment remis en question l'approche symbolique de la modélisation de l'esprit. Selon le connexionnisme, la cognition n'est pas une manipulation de symboles mais des modèles dynamiques d'activité dans un réseau multicouche de nœuds ou d'unités avec des interconnexions positives et négatives pondérées. Les modèles changent en fonction des contraintes du réseau interne régissant la manière dont les activations et les forces de connexion sont ajustées sur la base de nouvelles entrées (par exemple, la «règle delta» ou «rétropropagation» généralisée, McClelland, Rumelhart et al. 1986). Le résultat est un système qui apprend, reconnaît les modèles, résout les problèmes et peut même faire preuve de motricité.

1.4 Portée et limites des symboles et des réseaux neuronaux

Les capacités et les limites réelles de l'IA symbolique ou du connexionnisme sont loin d'être claires. Le premier semble meilleur pour les tâches formelles et langagières, le second pour les tâches sensorielles, motrices et d'apprentissage, mais il y a un chevauchement considérable et aucun des deux n'est allé bien au-delà du stade des tâches «jouets» vers une capacité comportementale grandeur nature. De plus, il y a eu un certain désaccord sur la question de savoir si le connexionnisme lui-même est symbolique ou non. Nous adopterons ici la position selon laquelle ce n'est pas le cas, car les réseaux connexionnistes ne remplissent pas plusieurs des critères pour être des systèmes de symboles, comme Fodor et Pylyshyn (1988) l'ont récemment soutenu. En particulier, bien que, comme tout le reste, leur comportement et leurs états internes puissent recevoir des interprétations sémantiques isolées, les réseaux neuronaux ne satisfont pas aux critères de compositeness (7) et de systématique (8) énumérés précédemment: les modèles d'interconnexions ne se décomposent pas, ne se combinent pas et ne recombiner selon une syntaxe formelle qui peut recevoir une interprétation sémantique systématique. [5] Au lieu de

cela, les réseaux neuronaux semblent faire ce qu'ils font de manière non symbolique. Selon Fodor & Pylyshyn, il s'agit d'une limitation sévère, car beaucoup de nos capacités comportementales semblent être symboliques, et donc l'hypothèse la plus naturelle sur les processus cognitifs sous-jacents qui les génèrent serait qu'ils doivent également être symboliques. Nos capacités linguistiques sont les principaux exemples ici, mais bon nombre des autres compétences que nous avons - raisonnement logique, mathématiques, jeu d'échecs, peut-être même nos capacités perceptives et motrices de plus haut niveau - semblent également être symboliques. Dans tous les cas, lorsque nous interprétons nos phrases, nos formules mathématiques et nos mouvements d'échecs (et peut-être certains de nos jugements perceptifs et stratégies motrices) comme ayant une signification ou un contenu systématique, nous savons de première main que c'est littéralement vrai, et pas seulement un figure de style. Le connexionnisme semble donc désavantagé pour tenter de modéliser ces capacités cognitives.

Pourtant, il n'est pas clair si le connexionnisme doit pour cette raison aspirer à être symbolique, car l'approche symbolique se révèle souffrir d'un handicap sévère, qui peut être responsable de l'étendue limitée de son succès à ce jour (en particulier dans la modélisation à l'échelle humaine capacités) ainsi que la nature inintéressante et ad hoc de la «connaissance» symbolique qu'elle attribue à «l'esprit» du système symbolique. Le handicap a été remarqué sous diverses formes depuis l'avènement de l'informatique; J'en ai surnommé une manifestation récente le «problème d'ancrage des symboles» (Harnad 1987b).

2. Le problème de l’ancrage des symboles

2.1 La chambre chinoise

Avant de définir le problème d'ancrage des symboles, je vais en donner deux exemples. Le premier vient du célèbre « Argument de la chambre chinoise » de Searle (1980), dans lequel le problème d'ancrage des symboles est appelé le problème de la signification intrinsèque (ou «intentionnalité»): Searle remet en question l'hypothèse fondamentale de l'IA symbolique selon laquelle un pour générer un comportement indiscernable de celui d'une personne, il faut avoir un esprit. Plus précisément, selon la théorie symbolique de l'esprit, si un ordinateur pouvait passer le test de Turing (Turing 1964) en chinois - c'est-à-dire s'il pouvait répondre à toutes les chaînes de symboles chinois qu'il reçoit en entrée avec des chaînes de symboles chinois indiscernables de les réponses qu'un vrai locuteur chinois ferait (même si nous continuons à tester pour une vie) - alors l'ordinateur comprendrait la signification des symboles chinois dans le même sens que je comprends la signification des symboles anglais.

La simple démonstration de Searle que cela ne peut pas être ainsi consiste à s'imaginer faire tout ce que fait l'ordinateur - recevoir les symboles d'entrée chinois, les manipuler uniquement sur la base de leur forme (conformément à (1) à (8) ci-dessus), et enfin renvoyer les symboles de sortie chinois. Il est évident que Searle (qui ne connaît pas le chinois) ne comprendrait pas le chinois dans ces conditions - donc l'ordinateur non plus. Les symboles et la manipulation des symboles, étant tous basés sur la forme plutôt que sur la signification, sont systématiquement interprétables comme ayant un sens - c'est après tout ce que c'est d'être un système de symboles, selon notre définition. Mais l'interprétation ne sera pas intrinsèque au système de symboles lui-même: elle sera parasite sur le fait que les symboles ont un sens pour nous, exactement de la même manière

que les significations des symboles dans un livre ne sont pas intrinsèques, mais dérivent du significations dans nos têtes. Par conséquent, si les significations des symboles dans un système de symboles sont extrinsèques, plutôt qu'intrinsèques comme les significations dans nos têtes, alors elles ne sont pas un modèle viable pour les significations dans nos têtes: la cognition ne peut pas être simplement une manipulation de symboles.

2.2 Le manège du dictionnaire chinois / chinois

Mon propre exemple du problème d'ancrage des symboles a deux versions, l'une difficile et l'autre, je pense, impossible. La version difficile est la suivante: supposons que vous deviez apprendre le chinois comme deuxième langue et que la seule source d’information dont vous disposiez était un dictionnaire chinois / chinois. Le voyage à travers le dictionnaire équivaudrait à un manège, passant sans cesse d'un symbole ou d'une chaîne de symboles sans signification (les definientes) à un autre (les definienda), sans jamais s'arrêter sur ce que quelque chose signifiait. [6]

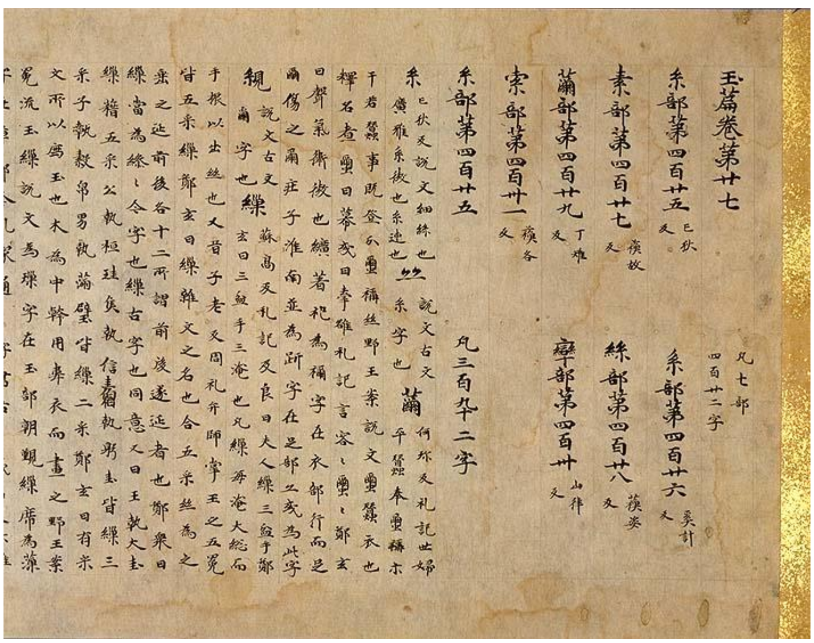

Figure 1: Dictionnaire Chinois/Chinois

La seule raison pour laquelle les cryptologues des langues anciennes et des codes secrets semblent être capables d'accomplir avec succès quelque chose de très semblable à cela est que leurs efforts sont ancrés dans une première langue et dans l'expérience et les connaissances du monde réel. [7] La deuxième variante du Dictionary-Go-Round, cependant, va bien au-delà des ressources concevables de la cryptologie: supposons que vous deviez apprendre le chinois comme première langue et que la seule source d'informations que vous ayez était un dictionnaire chinois / chinois! [8] Cela ressemble plus à la tâche réelle à laquelle est confronté un modèle purement symbolique de l'esprit: comment pouvez-vous jamais sortir du manège de symboles / symboles? Comment le symbole signifie-t-il être ancré dans autre chose que des symboles plus dénués de sens? [9] C'est le problème d'ancrage des symboles. [10]

2.3 Se connecter au monde

La réplique standard du symboliste (par exemple, Fodor 1980, 1985) est que la signification des symboles vient de la connexion du système de symboles au monde «de la bonne manière». Mais il semble évident que le problème de la connexion avec le monde de la bonne manière est pratiquement coextensif avec le problème de la cognition elle-même. Si chaque définien dans un dictionnaire chinois / chinois était en quelque sorte connecté au monde de la bonne manière, nous n'aurions guère besoin de la définienda! De nombreux symbolistes pensent que la cognition, en tant que manipulation de symboles, est un module fonctionnel autonome qui n'a besoin d'être raccordé qu'à des périphériques pour "voir" le monde des objets auquel ses symboles se réfèrent (ou, plutôt, auxquels ils peuvent être systématiquement interprété comme faisant référence). [11] Malheureusement, cela sous-estime radicalement la difficulté de repérer les objets, les événements et les états de choses dans le monde auxquels les symboles se réfèrent, c'est-à-dire qu'il banalise le problème d'ancrage des symboles.

C'est un candidat possible pour une solution à ce problème, confronté directement, qui va maintenant être esquissé: Ce qui sera proposé est un système hybride non symbolique / symbolique, un système «dédié», dans lequel les symboles élémentaires sont ancrés dans deux types de des représentations non symboliques qui sélectionnent, à partir de leurs projections sensorielles proximales, les catégories d'objets distaux auxquelles se réfèrent les symboles élémentaires. La plupart des composants dont est constitué le modèle (projections et transformations analogiques, discrétisation, détection d'invariance, connexionnisme, manipulation de symboles) ont également été proposés dans diverses configurations par d'autres, mais ils seront assemblés de manière spécifique bottom-up ici cela n'a pas, à ma connaissance, été suggéré précédemment, et c'est de cette configuration spécifique que dépend le succès potentiel du schéma d'ancrage.

Le tableau 1 résume les forces et les faiblesses relatives du connexionnisme et du symbolisme, les deux candidats rivaux actuels pour expliquer à lui seul toute la cognition. Leurs atouts respectifs seront mis à profit en coopération plutôt qu'en concurrence dans notre modèle hybride, remédiant ainsi également à certaines de leurs faiblesses respectives. Examinons maintenant de plus près les capacités comportementales qu'un tel modèle cognitif doit générer.

Tableau 1. Connectionisme Vs. Systèmes de symboles

Forces du connexionnisme:

(1) Fonction non symbolique: Tant qu'il n'aspire pas à être un système de symboles, un réseau connexionniste a l'avantage de ne pas être soumis au problème d'ancrage des symboles.

(2) Généralités: Le connexionnisme applique la même petite famille d'algorithmes à de nombreux problèmes, tandis que le symbolisme, étant une méthodologie plutôt qu'un algorithme, repose sur des règles symboliques sans fin spécifiques aux problèmes.

(3) "Neurosimilitude": L'architecture connexionniste ressemble plus à un cerveau qu'une machine de Turing ou un ordinateur numérique.

(4) Apprentissage des modèles: Les réseaux connexionnistes sont particulièrement adaptés à l'apprentissage de modèles à partir de données.

Faiblesses du connexionnisme:

(1) Fonction non symbolique: Les réseaux connexionnistes, parce qu'ils ne sont pas des systèmes de symboles, n'ont pas les propriétés sémantiques systématiques que de nombreux phénomènes cognitifs semblent avoir.

(2) Généralités: tous les problèmes ne se résument pas à l'apprentissage de modèles. Certaines tâches cognitives peuvent nécessiter des règles spécifiques au problème, une manipulation de symboles et un calcul standard.

(3) «Neurosimilitude»: la ressemblance cérébrale du connexionnisme peut être superficielle et peut (comme les modèles de jouets) camoufler des limitations de performances plus profondes.

Forces des systèmes de symboles:

(1) Fonction symbolique: Les symboles ont la puissance de calcul des machines de Turing et les propriétés systématiques d'une syntaxe formelle qui est sémantiquement interprétable.

(2) Généralités: Toutes les fonctions calculables (y compris toutes les fonctions cognitives) sont équivalentes à un état de calcul dans une machine de Turing.

(3) Succès pratiques: la capacité des systèmes de symboles à générer un comportement intelligent est démontrée par les succès de l'intelligence artificielle.

Faiblesses des systèmes de symboles :

(1) Fonction symbolique: Les systèmes de symboles sont sujets au problème d'ancrage des symboles.

(2) Généralité: le pouvoir de Turing est trop général. Les solutions aux nombreux problèmes de jouets de l'IA ne donnent pas lieu à des principes communs de cognition mais à une grande variété de stratégies symboliques ad hoc..

3. La capacité béhaviorale humaine

Depuis l'avènement du cognitivisme, les psychologues ont continué à collecter des données comportementales, bien que dans une large mesure les preuves pertinentes soient déjà disponibles: nous savons déjà ce que les êtres humains sont capables de faire. Ils peuvent (1) discriminer, (2) manipuler, [12] (3) identifier et (4) décrire les objets, les événements et les états de choses dans le monde dans lequel ils vivent, et ils peuvent aussi (5) "produire des descriptions" et (6) "répondre aux descriptions" de ces objets, événements et états de choses. Le fardeau de la théorie cognitive est maintenant d'expliquer comment les êtres humains (ou tout autre appareil) font tout cela. [13]

3.1 Discrimination et Identification

Examinons d'abord de plus près la discrimination et l'identification. Être capable de discriminer, c'est pouvoir juger si deux entrées sont identiques ou différentes et, si elles sont différentes, dans quelle mesure elles sont différentes. La discrimination est un jugement relatif, basé sur notre capacité à distinguer les choses et à discerner leur degré de similitude. Être capable d'identifier, c'est être capable d'attribuer une réponse unique (généralement arbitraire) - un «nom» - à une classe d'entrées, en les traitant toutes comme équivalentes ou invariantes à certains égards. L'identification est un jugement absolu, basé sur notre capacité à dire si une entrée donnée appartient ou non à une catégorie particulière.

Considérez le symbole «cheval». Nous sommes capables, en regardant différents chevaux (ou le même cheval dans des positions différentes, ou à des moments différents) de les distinguer et de juger lesquels d'entre eux se ressemblent le plus, voire à quel point ils se ressemblent. C'est de la discrimination. De plus, en regardant un cheval, nous pouvons l'appeler de manière fiable un cheval, plutôt que, disons, un mulet ou un âne (ou une girafe, ou une pierre). C'est de l'identification. Quelle sorte de représentation interne serait nécessaire pour générer ces deux types de performances?

3.2 Les representations iconiques et catégorielles

Selon le modèle proposé ici, notre capacité à discriminer les entrées dépend de la formation de «représentations iconiques» de celles-ci (Harnad 1987b). Ce sont des transformées analogiques internes des projections d'objets distaux sur nos surfaces sensorielles (Shepard & Cooper 1982). Dans le cas des chevaux (et de la vision), ils seraient des analogues des nombreuses formes que les chevaux jettent sur nos rétines. [14] Les jugements identiques / différents seraient basés sur la similitude ou la différence de ces représentations iconiques, et les jugements de similitude seraient basés sur leur degré de congruence. Aucun homoncule n'est impliqué ici; simplement un processus de superposition d'icônes et d'enregistrement de leur degré de disparité. Il n'y a pas non plus de problèmes de mémoire, puisque les entrées sont soit simultanément présentes soit disponibles en succession suffisamment rapide pour puiser dans leurs icônes sensorielles persistantes.

Nous avons donc besoin d'icônes de chevaux pour distinguer les chevaux. Mais qu'en est-il de les identifier? La discrimination est indépendante de l'identification. Je pourrais discriminer les choses sans savoir ce qu'elles étaient. L'icône me permettra-t-elle d'identifier les chevaux? Bien qu'il y ait des théoriciens qui croient que ce serait le cas (Paivio 1986), j'ai essayé de montrer pourquoi cela ne pouvait pas (Harnad 1982, 1987b). Dans un monde où il y avait des discontinuités naturelles audacieuses et facilement détectables entre toutes les catégories que nous aurions jamais à trier et identifier (ou choisir de) - un monde dans lequel les membres d'une catégorie ne pouvaient être confondus avec les membres une autre catégorie - les icônes peuvent être suffisantes pour l'identification. Mais dans notre monde sous-déterminé, avec son infinité de catégories potentielles confusables, les icônes sont inutiles pour l'identification parce qu'elles sont trop nombreuses et parce qu'elles se mélangent continuellement [15] les unes dans les autres, ce qui en fait un problème indépendant d'identifier lesquelles d'entre elles sont des icônes des membres de la catégorie et qui ne le sont pas! Les icônes des projections sensorielles sont trop peu sélectives. Pour l'identification, les icônes doivent être sélectivement réduites à ces "caractéristiques invariantes" de la projection sensorielle qui distingueront de manière fiable un

membre d'une catégorie de tout non-membre avec lequel il pourrait être confondu. Appelons la sortie de ce détecteur de caractéristiques spécifiques à une catégorie la "représentation catégorielle". Dans certains cas, ces représentations peuvent être innées, mais comme l'évolution pourrait difficilement anticiper toutes les catégories dont nous pourrions avoir besoin ou choisir d'identifier, la plupart de ces caractéristiques doivent être tirées de l'expérience. En particulier, notre représentation catégorique d'un cheval est probablement savante. (Je reporterai à la section 4 le problème de la façon dont les caractéristiques invariantes sous-jacentes à l'identification pourraient être apprises.)

Notez que les représentations iconiques et catégoriques ne sont pas symboliques. Les premiers sont des copies analogiques de la projection sensorielle, conservant fidèlement sa «forme»; ces dernières sont des icônes qui ont été sélectivement filtrées pour ne conserver que certaines des caractéristiques de la forme de la projection sensorielle: celles qui distinguent de manière fiable les membres des non-membres d'une catégorie. Mais les deux représentations sont toujours sensorielles et non symboliques. Il n'y a aucun problème quant à leur connexion aux objets qu'ils choisissent: c'est une connexion purement causale, basée sur la relation entre les objets distaux, les projections sensorielles proximales et les changements internes acquis qui résultent d'une histoire d'interactions comportementales avec eux. Il n'y a pas non plus de problème d'interprétation sémantique ni de justification de l'interprétation sémantique. Les représentations iconiques ne «signifient» pas plus les objets dont elles sont les projections que ne le fait l'image dans une caméra. Les icônes et les images de caméra peuvent bien sûr être interprétées comme signifiant ou représentant quelque chose, mais l'interprétation serait clairement dérivée plutôt qu'intrinsèque. [16]

3.3 Representations symboliques

Les représentations catégoriques ne peuvent pas encore être interprétées comme "signifiant" quoi que ce soit. Il est vrai qu'ils choisissent la classe d'objets qu'ils «nomment», mais les noms n'ont pas toutes les propriétés systématiques des symboles et des systèmes de symboles décrits précédemment. Ils ne sont qu'une taxonomie inerte. Pour la systématisation, il doit être possible de les combiner et de les recombiner régulièrement en propositions qui peuvent être interprétées sémantiquement. "Cheval" n'est pour l'instant qu'une réponse arbitraire qui est faite de manière fiable en présence d'une certaine catégorie d'objets. Il n'y a aucune justification pour l'interpréter de manière holophrastique comme signifiant «Ceci est un cheval [membre de la catégorie]» lorsqu'il est produit en présence d'un cheval, parce que les autres propriétés systématiques attendues de «ceci» et «a» et le tout-important "est" de prédication ne se manifeste pas par une simple taxonomisation passive. Que faudrait-il pour générer ces autres propriétés systématiques? Simplement que les noms ancrés dans la taxonomie des catégories soient regroupés dans des propositions concernant d'autres relations d'appartenance aux catégories. Par exemple:

(1) Supposons que le nom de «cheval» soit ancré dans des représentations iconiques et catégoriques, apprises par l'expérience, qui distinguent et identifient de manière fiable les chevaux sur la base de leurs projections sensorielles.

(2) Supposons que les "rayures" soient ancrées de la même manière.

Considérons maintenant que la catégorie suivante peut être constituée à partir de ces catégories élémentaires par une description symbolique de l'appartenance à une catégorie seule:

(3) "Zebra" = "cheval" et "rayures" [17]

Quelle est la représentation d'un zèbre? Il s'agit simplement de la chaîne de symboles «cheval et rayures». Mais parce que «cheval» et «rayures» sont ancrés dans leurs représentations iconiques et catégoriques respectives, «zèbre» hérite de l'ancrage, à travers sa représentation symbolique ancrée. En principe, quelqu'un qui n'avait jamais vu de zèbre (mais qui avait vu et appris à identifier les chevaux et les rayures) pouvait identifier un zèbre à sa première connaissance, armé de cette seule représentation symbolique (plus les représentations non symboliques - iconiques et catégoriques - des chevaux et rayures qui l'ancrent).

Une fois que l'on a l'ensemble ancré de symboles élémentaires fourni par une taxonomie de noms (et les représentations iconiques et catégoriques qui donnent du contenu aux noms et leur permettent de choisir les objets qu'ils identifient), le reste des chaînes de symboles d'un langage naturel peuvent être générés par la composition de symboles seule, [18] et ils hériteront tous de l'ancrage intrinsèque de l'ensemble élémentaire. [19] Par conséquent, la capacité de discriminer et de catégoriser (et ses représentations non symboliques sous-jacentes) a conduit naturellement à la capacité de décrire et de produire et de répondre à des descriptions à travers des représentations symboliques.

4. Un rôle complémentaire pour le connexionnisme

Le schéma d'ancrage des symboles que nous venons de décrire a une lacune importante: aucun mécanisme n'a été suggéré pour expliquer comment les représentations catégorielles essentielles pourraient être formées: comment le système hybride trouve-t-il les caractéristiques invariantes de la projection sensorielle qui permettent de catégoriser et d'identifier les objets correctement? [20] Le connexionnisme, avec sa capacité générale d'apprentissage de modèles, semble être un candidat naturel (bien qu'il puisse y en avoir d'autres): les icônes, associées à des commentaires indiquant leurs noms, pourraient être traitées par un réseau connexionniste qui apprend à identifier les icônes correctement à partir de l'échantillon d'alternatives confusables qu'il a rencontrées en ajustant dynamiquement les poids des caractéristiques et des combinaisons de caractéristiques qui sont associées de manière fiable aux noms d'une manière qui résout (provisoirement) la confusion, réduisant ainsi les icônes à l'invariant (confusion- résolution) des caractéristiques de la catégorie à laquelle elles sont affectées. En effet, la «connexion» entre les noms et les objets qui donnent lieu à leurs projections sensorielles et à leurs icônes serait assurée par des réseaux connexionnistes.

Ce rôle complémentaire circonscrit du connexionnisme dans un système hybride semble remédier aux faiblesses des deux concurrents actuels dans leurs tentatives de modéliser l'esprit de manière indépendante. Dans un modèle symbolique pur, la connexion cruciale entre les symboles et leurs référents fait défaut; un système de symboles autonome, bien qu'il se prête à une interprétation sémantique systématique, n'est pas ancré. Dans un modèle connexionniste pur, les noms sont connectés aux objets par des motifs invariants dans leurs projections sensorielles, appris par exposition et rétroaction, mais la propriété compositionnelle cruciale est absente; un

réseau de noms, bien qu'ancré, ne se prête pas encore à une interprétation sémantique systématique complète. Dans le système hybride proposé ici, il n'y a plus du tout de niveau symbolique autonome; à la place, il existe un système de symboles intrinsèquement dédié, ses symboles élémentaires (noms) connectés à des représentations non symboliques qui peuvent sélectionner les objets auxquels ils se réfèrent, via des réseaux connexionnistes qui extraient les caractéristiques invariantes de leurs projections sensorielles analogiques.

5. Conclusions

L'attente a souvent été exprimée que les approches «descendantes» (symboliques) de la modélisation de la cognition rencontreraient d'une manière ou d'une autre des approches «ascendantes» (sensorielles) quelque part entre les deux. Si les considérations d'ancrage dans cet article sont valables, alors cette attente est désespérément modulaire et il n'y a vraiment qu'une seule voie viable du sens aux symboles: de la base vers le haut. Un niveau symbolique flottant comme le niveau logiciel d'un ordinateur ne sera jamais atteint par cette route (ou vice versa) - et il n'est pas clair non plus pourquoi nous devrions même essayer d'atteindre un tel niveau, car il semble que pour y arriver revient simplement à déraciner nos symboles de leur signification intrinsèque (nous réduisant ainsi simplement à l'équivalent fonctionnel d'un ordinateur programmable).

Dans un système de symboles intrinsèquement dédié, il y a plus de contraintes sur les jetons de symboles que de simples contraintes syntaxiques. Les symboles sont manipulés non seulement sur la base de la forme arbitraire de leurs jetons, mais également sur la base de la "forme" résolument non arbitraire des représentations iconiques et catégoriques liées aux symboles élémentaires ancrés à partir desquels les symboles d'ordre supérieur sont composés . De ces deux types de contraintes, les iconiques / catégoriques sont primordiales. Je ne connais aucune analyse formelle de tels systèmes de symboles dédiés, [21] mais cela peut être dû au fait qu'ils sont uniques à la modélisation cognitive et robotique et que leurs propriétés dépendront des types spécifiques de capacités robotiques (c'est-à-dire comportementales) qu'ils sont conçus. exposer.

Il convient que les propriétés des systèmes de symboles dédiés se révèlent dépendre de considérations comportementales. Le schéma d'ancrage actuel est toujours dans l'esprit du behaviorisme en ce que les seuls tests proposés pour savoir si une interprétation sémantique portera le poids sémantique qui y est placé consistent en un test formel (répond-il aux huit critères pour être un système de symboles?) Et un test comportemental (peut-il discriminer, identifier et décrire tous les objets et états de choses auxquels se réfèrent ses symboles?). Si les deux tests sont réussis, alors l'interprétation sémantique de ses symboles est «fixée» par la capacité comportementale du système de symboles dédié, telle qu'elle s'exerce sur les objets et les états de choses du monde auxquels se réfèrent ses symboles; les significations des symboles ne sont donc pas seulement parasites sur les significations dans la tête de l'interprète, mais intrinsèques au système de symboles dédié lui-même. Ce n'est toujours pas une garantie que notre modèle a capturé une signification subjective, bien sûr. Mais si les capacités comportementales du système sont à la taille de la vie, elles sont aussi proches que nous pouvons espérer l'être.

Notes

1. Paul Kube (communication personnelle) a suggéré que (2) et (3) peuvent être trop forts, excluant certains types de machine de Turing et peut-être même conduisant à une régression infinie sur les niveaux d'explicitation et de systématisation.

2. Des considérations similaires s'appliquent au concept de «réalité psychologique» de Chomsky (1980) (c'est-à-dire si les règles chomskiennes sont réellement physiquement représentées dans le cerveau ou si elles «correspondent» simplement à nos régularités de performance, sans être ce qui les régit réellement). Une autre version de la distinction concerne les règles explicitement représentées par rapport aux contraintes physiques câblées (Stabler 1985). Dans chaque cas, une représentation explicite constituée d'éléments qui peuvent être recombinés de manière systématique serait symbolique alors qu'une contrainte physique implicite ne le serait pas, bien que les deux seraient sémantiquement «interprétables» comme une «règle» si elles étaient interprétées isolément plutôt que comme faisant partie de un système.

3. De manière analogue, le simple fait qu'un comportement soit interprétable comme intentionnel ou conscient ou significatif ne signifie pas qu'il est réellement intentionnel ou conscient. (Pour des arguments contraires, voir Dennett 1983).

4. Il n'est même pas encore clair qu'un «réseau neuronal» doit être implémenté en tant que réseau (c'est-à-dire un système parallèle d'unités interconnectées) pour faire ce qu'il peut faire; si les simulations symboliques de réseaux neuronaux ont la même capacité fonctionnelle que les réseaux neuronaux réels, alors un modèle connexionniste n'est qu'un type particulier de modèle symbolique, et le connexionnisme n'est qu'une famille spéciale d'algorithmes symboliques.

5. Il y a un malentendu sur ce point car il est souvent confondu avec un simple problème de mise en œuvre: les réseaux connexionnistes peuvent être simulés à l'aide de systèmes de symboles, et les systèmes de symboles peuvent être mis en œuvre en utilisant une architecture connexionniste, mais cela est indépendant de la question de savoir peut faire un système de symboles ou un réseau connexionniste, respectivement. Par analogie, le silicium peut être utilisé pour construire un ordinateur, et un ordinateur peut simuler les propriétés du silicium, mais les propriétés fonctionnelles du silicium ne sont pas celles du calcul, et les propriétés fonctionnelles du calcul ne sont pas celles du silicium.

6. L'IA symbolique regorge de symptômes du problème d'ancrage des symboles. Une manifestation bien connue (bien que mal diagnostiquée) de celui-ci est le soi-disant problème du «cadre» (McCarthy & Hayes 1969; Minsky 1974; NcDermott 1976; Pylyshyn 1987): C'est une expérience frustrante mais familière d'écriture «basée sur la connaissance». programmes qu'un système se comportant apparemment parfaitement intelligemment pendant un certain temps peut être déjoué par un cas inattendu qui démontre sa stupidité totale: un programme de "compréhension de la scène" décrira allègrement ce qui se passe dans une scène visuelle et répondra à des questions démontrant sa compréhension (qui a fait quoi, où, pourquoi?) et révèle soudainement qu'il ne «sait» pas que raccrocher le téléphone et quitter la pièce ne fait pas disparaître le téléphone, ou quelque chose du genre. (Il est important de noter que ce ne sont pas les types de lacunes et de lacunes dans les connaissances auxquelles les gens sont enclins; ils sont plutôt des hurleurs tels qu'ils jettent un doute sérieux sur le fait que le système a quelque chose comme la «connaissance».)

Le problème du "cadre" a été défini avec optimisme comme le problème de la spécification formelle ("cadrage") de ce qui varie et de ce qui reste constant dans un "domaine de connaissance" particulier, mais en réalité c'est le problème de remettre en question toutes les contingences du programmeur. pas anticipé en symbolisant la connaissance qu'il tente de symboliser. Ces contingences sont probablement illimitées, à des fins pratiques, parce que la «connaissance» purement symbolique n'est pas fondée. Le simple fait d'ajouter plus de contingences symboliques, c'est comme prendre quelques tours de plus dans le dictionnaire chinois / chinois. Il n'y a en réalité aucun motif en vue: simplement assez de manipulation de symboles «intelligente» pour endormir le programmeur en perdant de vue le fait que sa signification est juste parasite sur les significations qu'il projette sur lui à partir des significations ancrées dans sa propre tête. (J'ai appelé cet effet le "hall herméneutique des miroirs" [Harnad 1990];

c'est l'envers du problème d'ancrage des symboles). Pourtant, c'est le parasitisme, car le prochain "problème de cadre" qui se cache au coin de la rue est prêt à le confirmer. (Une forme similaire de surinterprétation s'est produite dans les expériences de «langage» des singes [Terrace 1979]. Peut-être que les singes et les ordinateurs devraient être formés à l'aide du code chinois, pour immuniser leurs expérimentateurs et programmeurs contre des surinterprétations fallacieuses. Mais depuis la réalité les tâches comportementales dans les deux domaines sont encore si triviales, il n'y a probablement aucun moyen d'empêcher leur déchiffrement. En fait, il semble y avoir une tendance irrésistible à surinterpréter les performances des tâches jouets elles-mêmes, en extrapolant de manière préventive et en "augmentant" conceptuellement la taille réelle sans aucune justification dans la pratique.)

7. Les cryptologues utilisent également des informations statistiques sur la fréquence des mots, des inférences sur ce qu'une culture ancienne ou un gouvernement ennemi est susceptible d'écrire, des algorithmes de décryptage, etc.

8. Il n'est bien entendu pas nécessaire de restreindre les ressources symboliques à un dictionnaire; la tâche serait tout aussi impossible si l'on avait accès à l'ensemble de la littérature chinoise, y compris tous ses programmes informatiques et tout ce qui peut être codifié sous forme de symboles.

9. Même les mathématiciens, qu'ils soient platoniciens ou formalistes, soulignent que la manipulation des symboles (le calcul) elle-même ne peut pas saisir la notion d'interprétation voulue des symboles (Penrose 1989). Le fait que les systèmes de symboles formels et leurs interprétations ne soient pas la même chose est donc évident indépendamment de la thèse de Church-Turing (Kleene 1969) ou des résultats de Goedel (Davis 1958, 1965), qui ont été avec zèle mal appliqués au problème de l'esprit. la modélisation (par exemple, par Lucas 1964) - pour laquelle ils sont largement hors de propos, à mon avis.

10. Notez que, à proprement parler, l'ancrage des symboles n'est un problème que pour la modélisation cognitive, pas pour l'IA en général. Si les systèmes de symboles réussissent à eux seuls à générer toutes les performances de la machine intelligente qui intéresse l'IA pure - par exemple, un dictionnaire automatisé - alors il n'y a aucune raison d'exiger que leurs symboles aient une signification intrinsèque. D'un autre côté, le fait que nos propres symboles aient une signification intrinsèque alors que ceux de l'ordinateur n'en ont pas, et le fait que nous puissions faire des choses que l'ordinateur ne peut pas faire jusqu'à présent, peut indiquer que même dans l'IA, il y a des gains de performance à faire. (en particulier en robotique et en vision industrielle) de l'effort aux systèmes de symboles au 11. Le point de vue homonculaire inhérent à cette croyance est tout à fait apparent, de même que l'effet de la «salle herméneutique des miroirs» (Harnad 1990).

12. Bien qu'elles soient sans aucun doute aussi importantes que les capacités perceptives, la motricité ne sera pas explicitement considérée ici. On suppose que les caractéristiques pertinentes de l'histoire sensorielle (par exemple, l'iconicité) se généraliseront à l'histoire motrice (par exemple, dans les analogues moteurs; Liberman 1982). De plus, de grandes parties de l'histoire motrice peuvent ne pas être cognitives, s'appuyant plutôt sur des schémas moteurs innés et un retour sensori-moteur. Le concept de Gibson (1979) «d'affordances» - les caractéristiques invariantes du stimulus qui sont détectées par les possibilités motrices qu'ils «offrent» - est également pertinent ici, bien que Gibson sous-estime les problèmes de traitement impliqués dans la recherche de tels invariants (Ullman 1980). Dans tous les cas, l'ancrage moteur et sensori-moteur sera sans doute aussi important que l'ancrage sensoriel sur lequel il s'agit ici.

13. Si un modèle candidat devait présenter toutes ces capacités comportementales, à la fois linguistiques (5-6) et robotiques (c'est-à-dire sensorimotrices), (1-3), il passerait le «test de Turing total» (Harnad 1989). Le test de Turing standard (Turing 1964) ne demande que la capacité de performance linguistique: symboles entrants et symboles sortants. Cela rend équivoque le statut, la portée et les limites de la manipulation de symboles purs, et donc soumis au problème d'ancrage des symboles. Un modèle qui pourrait passer le test de Turing total, cependant, serait ancré dans le monde.

14. De nombreux problèmes liés à la discrimination figure / fond, au lissage, à la constance de la taille, à la constance de la forme, à la stéréopsie, etc., rendent le problème de la discrimination beaucoup plus compliqué que ce qui est décrit ici, mais ils ne changent pas fait que les représentations iconiques sont un substrat candidat naturel pour notre capacité à discriminer.

15. Ailleurs (Harnad 1987a, b), j'ai essayé de montrer comment le phénomène de «perception catégorielle» pouvait générer des discontinuités internes là où il y a continuité externe. Il est prouvé que notre système perceptif est capable de segmenter un continuum, tel que le spectre des couleurs, en régions ou catégories relativement discrètes et délimitées. Les différences physiques de même ampleur sont plus discriminables à travers les frontières entre ces catégories qu'à l'intérieur d'elles. Cet effet de frontière, à la fois inné et appris, peut jouer un rôle important dans la représentation des catégories perceptives élémentaires à partir desquelles les catégories d'ordre supérieur sont construites.

16. D'un autre côté, la ressemblance sur laquelle repose la performance de la discrimination - le degré d'isomorphisme entre l'icône et la projection sensorielle, et entre la projection sensorielle et l'objet distal - semble être intrinsèque, plutôt qu'une simple question. d'interprétation. La ressemblance peut être objectivement caractérisée comme le degré d'inversibilité de la transformation physique d'objet en icône (Harnad 1987b).

17. La figure 1 est en fait l'entrée du dictionnaire chinois pour «zèbre», qui signifie «cheval rayé». Notez que le caractère pour "zèbre" se trouve être en fait le caractère pour "cheval" plus le caractère pour "rayé". Bien que les caractères chinois aient une structure iconique, ils fonctionnent comme des lexigrammes alphabétiques arbitraires au niveau de la syntaxe et de la sémantique.

18. Certains connecteurs et quantificateurs logiques standard sont également nécessaires, tels que non, et, tous, etc.

19. Notez qu'il n'est pas prétendu que "cheval", "rayures", etc. sont en fait des symboles élémentaires, avec un ancrage sensoriel direct; on prétend seulement que certains ensembles de symboles doivent être directement fondés. La plupart des représentations des catégories sensorielles sont sans aucun doute un hybride sensoriel / symbolique; et leurs caractéristiques peuvent changer par bootstrapping: "Horse" peut toujours être révisé, à la fois sensoriellement et symboliquement, même s'il était auparavant élémentaire. Kripke (1980) donne un bon exemple de la façon dont «l'or» pourrait être baptisé sur le métal jaune brillant en question, utilisé pour le commerce, la décoration et le discours, et alors nous pourrions découvrir «l'or du fou», qui ferait toutes les caractéristiques sensorielles que nous avait utilisé jusque-là insuffisant, nous obligeant à en trouver de nouveaux. Il souligne qu'il est même possible en principe que «l'or» ait été baptisé par inadvertance sur «l'or des fous»! Les aspects ontologiques de cette possibilité ne sont pas intéressants ici, mais les aspects épistémiques: nous pourrions amorcer avec succès de l'or réel même si chaque cas antérieur avait été de l'or du fou. "Or" serait toujours le mot juste pour ce que nous avions essayé de choisir depuis le début, et ses caractéristiques provisoires d'origine auraient toujours fourni une approximation suffisamment proche pour le fondre, même si des informations ultérieures venaient à tirer le sol de dessous il, pour ainsi dire.

20. Bien qu'il soit hors de la portée de ce document d'en discuter longuement, il faut mentionner que cette question a souvent été soulevée dans le passé, principalement au motif de "la disparition des intersections". On a prétendu que l'on ne peut pas trouver des caractéristiques invariantes dans la projection sensorielle parce qu'elles n'existent tout simplement pas: l'intersection de toutes les projections des membres d'une catégorie telle que «cheval» est vide. Les empiristes britanniques ont été critiqués pour penser autrement; par exemple, la discussion de Wittgenstein (1953) sur les «jeux» et les «ressemblances familiales» a été considérée comme ayant discrédité leur point de vue. Et les recherches actuelles sur la catégorisation humaine (Rosch & Lloyd 1978) ont été interprétées comme confirmant que les intersections disparaissent et que, par conséquent, les catégories ne sont pas représentées en termes de caractéristiques invariantes. Le problème de la disparition des intersections (avec l'argument «pauvreté du stimulus» de Chomsky [1980]) a même été cité par des penseurs comme Fodor (1985, 1987) comme justification du nativisme extrême. Le présent article est franchement empiriste. À mon avis, la raison pour laquelle les intersections n'ont pas été trouvées est que personne ne les a encore recherchées correctement. L'introspection n'est certainement pas la façon de regarder. Et les algorithmes généraux d'apprentissage de modèles tels que le connexionnisme sont relativement nouveaux; leur puissance inductive reste à tester. En outre, une distinction prudente n'a pas été faite entre les catégories sensorielles pures (qui, je le prétends, doivent avoir des invariants, sinon nous ne pourrions pas les identifier avec succès comme nous le faisons) et les catégories d'ordre supérieur qui sont fondées sur des catégories sensorielles; ces représentations abstraites peuvent être symboliques plutôt que sensorielles, et donc ne pas reposer directement sur des invariants sensoriels. Pour une discussion plus approfondie de ce problème, voir Harnad 1987b).

21. Bien que les mathématiciens étudient les propriétés formelles de systèmes de symboles non interprétés, toutes leurs motivations et intuitions proviennent clairement des interprétations voulues de ces systèmes (voir Penrose 1989). Peut-être que ceux-ci sont également ancrés dans les représentations iconiques et catégoriques dans leur tête.

Références

Catania, A. C. & Harnad, S. (eds.) (1988) The Selection of Behavior. The Operant Behaviorism of B. F. Skinner: Comments and Consequences. New York: Cambridge University Press.

Chomsky, N. (1980) Rules and representations. *Behavioral and Brain Sciences* 3: 1-61.

Davis, M. (1958) *Computability and unsolvability.* Manchester: McGraw-Hill.

Davis, M. (1965) *The undecidable.* New York: Raven.

Dennett, D. C. (1983) Intentional systems in cognitive ethology. Behavioral and Brain Sciences 6: 343 - 90.

Fodor, J. A. (1975) *The language of thought* New York: Thomas Y. Crowell

Fodor, J. A. (1980) Methodological solipsism considered as a research strategy in cognitive psychology. *Behavioral and Brain Sciences* 3: 63 - 109.

Fodor, J. A. (1985) Précis of "The Modularity of Mind." *Behavioral and Brain Sciences* 8: 1 - 42.

Fodor, J. A. (1987) Psychosemantics Cambridge MA: MIT/Bradford.

Fodor, J. A. & Pylyshyn, Z. W. (1988) Connectionism and cognitive architecture: A critical appraisal. Cognition 28: 3 - 71.

Gibson, J. J. (1979) *An ecological approach to visual perception.* Boston: Houghton Mifflin

Harnad, S. (1982) Metaphor and mental duality. In T. Simon & R. Scholes, R. (Eds.) *Language, mind and brain.* Hillsdale, N. J.: Lawrence Erlbaum Associates

Harnad, S. (1987a) Categorical perception: A critical overview. In S. Harnad (Ed.) *Categorical perception: The groundwork of Cognition.* New York: Cambridge University Press

Harnad, S. (1987b) Category induction and representation. In S. Harnad (Ed.) *Categorical perception: The groundwork of Cognition.* New York: Cambridge University Press

Harnad, S. (1989) Minds, Machines and Searle. *Journal of Theoretical and Experimental Artificial Intelligence* 1: 5-25.

Harnad, S. (1990) Computational Hermeneutics. *Social Epistemology* in press.

Haugeland, J. (1978) The nature and plausibility of cognitivism. *Behavioral and Brain Sciences* 1: 215-260.

Kleene, S. C. (1969) *Formalized recursive functionals and formalized realizability.* Providence, R.: American Mathematical Society.

Kripke, S.A. (1980) *Naming and Necessity.* Cambridge MA: Harvard University Press

Liberman, A. M. (1982) On the finding that speech is special. *American Psychologist* 37: 148-167.

Lucas, J. R. (1961) Minds, machines and G*"odel. Philosophy 36: 112-117.

McCarthy, J. & Hayes, P. (1969) Some philosophical problems from the standpoint of artificial intelligence. In: Meltzer B. & Michie, P. *Machine Intelligence* Volume 4. Edinburgh: Edinburgh University Press.

McDermott, D. (1976) Artificial intelligence meets natural stupidity. *SIGART Newsletter* 57: 4 - 9.

McClelland, J. L., Rumelhart, D. E., and the PDP Research Group (1986) *Parallel distributed processing: Explorations in the microstructure of cognition,* Volume 1. Cambridge MA: MIT/Bradford.

Miller, G. A. (1956) The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review* 63: 81 - 97.

Minsky, M. (1974) A framework for Representing knowledge. *MIT Lab Memo* # 306.

Minsky, M. & Papert, S. (1969) *Perceptrons: An introduction to computational geometry.* Cambridge MA: MIT Press (Reissued in an Expanded Edition, 1988).

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

Neisser, U. (1967) *Cognitive Psychology* NY: Appleton-Century-Crofts.

Paivio, A. (1986) *Mental representation: A dual coding approach.* New York: Oxford

Penrose, R. (1989) *The emperor's new mind.* Oxford: Oxford University Press
Pylyshyn, Z. W. (1980) Computation and cognition: Issues in the foundations of cognitive science. *Behavioral and Brain Sciences* 3: 111-169.
Pylyshyn, Z. W. (1984) *Computation and cognition.* Cambridge MA: MIT/Bradford
Pylyshyn, Z. W. (Ed.) (1987) *The robot's dilemma: The frame problem in artificial intelligence.* Norwood NJ: Ablex
Rosch, E. & Lloyd, B. B. (1978) *Cognition and categorization.* Hillsdale NJ: Erlbaum Associates
Rosenblatt, F. (1962) Principles of neurodynamics. NY: Spartan
Searle, J. R. (1980) Minds, brains and programs. *Behavioral and Brain Sciences* 3: 417-457.
Shepard, R. N. & Cooper, L. A. (1982) *Mental images and their transformations.* Cambridge: MIT Press/Bradford.
Smolensky, P. (1988) On the proper treatment of connectionism. *Behavioral and Brain Sciences* 11: 1 - 74.
Stabler, E. P. (1985) How are grammars represented? *Behavioral and Brain Sciences* 6: 391-421.
Terrace, H. (1979) Nim. NY: Random House.
Turkkan, J. (1989) Classical conditioning: The new hegemony. *Behavioral and Brain Sciences 12:* 121 - 79.
Turing, A. M. (1964) Computing machinery and intelligence. In: *Minds and machines,* A. R. Anderson (ed.), Engelwood Cliffs NJ: Prentice Hall.
Ullman, S. (1980) Against direct perception. *Behavioral and Brain Sciences* 3: 373 - 415.
Wittgenstein, L. (1953) *Philosophical investigations.* New York: Macmillan