Apports de l'Intelligence Artificielle à la Psychologie

Daniel Defays, Robert M. French & Jacques Sougné

1. Introduction

L'appellation “intelligence artificielle” (IA) est souvent considérée comme malheureuse : certains y voient une forme de provocation, un slogan voire une sorte d'incantation. Elle présente pourtant l'avantage d'encapsuler dans deux mots l'essence de la discipline. La gamme de fonctions à produire ou à reproduire (ou les comportements à simuler si l'optique est d'ordre plus psychologique) est décrite par le mot “intelligence” et les moyens d'y parvenir par l'adjectif “artificiel”. Qui dit “artificiel” fait nécessairement référence à une construction humaine. Il s'agit donc de construire des machines intelligentes. Cette formulation n'est pourtant pas dépourvue d'ambiguïté : de quel genre de machine parle-t-on ? Que couvre exactement la notion d'intelligence ? L'histoire de l'IA est émaillée de contributions qui ont cherché à répondre à ces questions.

L'objectif de ce chapitre est d'aider le lecteur à se faire une première idée de la nature des problèmes abordés en IA, des différentes tendances qui existent, des principaux modèles utilisés.

Les comportements mentaux que l'IA cherche à reproduire sont cependant fort diversifiés : résolution de problèmes, créativité, planification, reconnaissance de formes, apprentissage par généralisation, par explication , compréhension de la langue écrite ou parlée,… Dans un chapitre comme celui-ci, il n'était pas possible, ni souhaitable de tout couvrir. Nous avons fait des choix. Par exemple, malgré leur importance, les théories, modèles ou systèmes liés au langage ne sont guère abordés. De même, peu de références à la reconnaissance de formes, à la modélisation de la vision sont données. A la présentation chronologique classique de la discipline, nous avons préféré un exposé centré sur les contraintes que s'imposent les systèmes en matière de modélisation de la cognition. Les systèmes informatiques développés ne cherchent généralement qu'à simuler certaines caractéristiques d'une activité mentale donnée : un logiciel qui joue aux échecs ne connaît pas le stress, le niveau de vigilance est constant - le système n'est jamais distrait -, il perçoit toutes les pièces avec la même acuité; un logiciel qui raisonne par analogie a déjà extrait de sa mémoire les situations à mettre en correspondance; un réseau neuronal reçoit des informations qui ont déjà été décodées etc. Les systèmes mettent l'accent tantôt sur la capacité que possèdent nos cerveaux à manipuler des symboles, tantôt sur leur aptitude à résoudre des problèmes, ou à associer des idées, à apprendre, à faire des inférences, à agir sur le milieu; les architectures des systèmes informatiques correspondants sont bien entendu fort différentes.

Le chapitre débute par un bref rappel des origines de l'IA et de son intérêt pour les psychologues. Il se continue par une discussion sur la possibilité d'une mécanisation de l'intelligence. La question a suscité de nombreux débats philosophiques dont nous rendons compte sous le titre “les machines peuvent-elles penser ?".

Ensuite, nous présentons différentes contraintes que se sont imposés les modèles les plus courants de l'IA :

- l'activité mentale conçue comme une aptitude à manipuler des symboles (en explicitant l'hypothèse des systèmes physiques de symboles);
- l'activité mentale conçue comme une aptitude à résoudre des problèmes;
- l'activité mentale conçue comme une aptitude à raisonner logiquement.

Un système obéissant à ces trois contraintes est présenté de manière plus détaillée pour aider le lecteur à se faire une idée plus précise de la nature des formulations et des systèmes utilisés en IA. La pertinence de ces systèmes pour la psychologie et leurs limites sont évoquées.

Dans les paragraphes qui suivent, un nouvel ensemble de contraintes est passé en revue;

- l'activité mentale conçue comme une aptitude à associer des idées (et nous introduisons sous ce paragraphe les systèmes connexionnistes);
- l'activité mentale conçue comme une aptitude à apprendre;
- l'activité mentale conçue comme une aptitude à agir sur le milieu.

Les systèmes obéissant à ces contraintes sont généralement fort différents des précédents, ont en général été développés plus récemment. De nouveau, les principes de fonctionnement sous-jacents sont évoqués et des exemples élémentaires sont donnés.

Le lecteur intéressé par un exposé plus complet est invité à consulter les ouvrages donnés en référence.

1.1. Les origines de l'intelligence artificielle

A la fin du XVII^{ème} siècle apparaît déjà le premier aperçu moderne de la notion d'une machine "pensante". Blaise Pascal construit en 1642 une machine capable au moyen d'engrenages d'effectuer des additions et des soustractions. Quelques années plus tard, Gottfried Leibnitz améliore le mécanisme pour lui permettre de multiplier et de diviser. L'un de ses grands projets fut ensuite la création d'un *Characteristica Universalis*, un système de règles formelles destinées à résoudre tous les problèmes auxquels étaient confrontés les hommes. Un demi siècle plus tard, La Mettrie écrit *L'homme-machine*, aboutissement logique, voire inéluctable, de la pensée newtonienne : le comportement physique et mental de l'homme pourrait être entièrement décrit à l'aide de règles déterministes et, par conséquent, l'homme ne serait rien d'autre qu'une machine, d'une complexité extraordinaire, certes, mais une machine tout de même. Quant aux premiers pas vers la réalisation d'une véritable machine pensante, il a fallu attendre encore cent ans avant que Boole et de Morgan ne proposent un ensemble rigoureux de mécanismes logiques capables de servir de base théorique à cette machine. Charles Babbage tente de traduire ce système théorique en machine pratique, le célèbre *Moteur Analytique*. Les travaux de Frege et de Russell au début du XX^{ème} siècle continuent à renforcer cette réduction de la pensée à un système de propositions issues d'un petit ensemble d'axiomes, de symboles fondamentaux et de règles pour manipuler les symboles. Enfin, le

Tractatus de Ludwig Wittgenstein, écrit dans les années 20 est la clef de voûte philosophique de ce mouvement. Ensuite, Alan Turing décrit dès 1936 une machine ultra-simple, aujourd'hui appelée une machine de Turing, capable de réaliser les fonctions généralement acceptées comme étant suffisantes à la pensée. Mais, il faudra attendre l'avènement des ordinateurs vers les années 50 pour pouvoir réunir toutes ces idées sous une seule rubrique, celle de l'informatique et plus précisément, celle de l'intelligence artificielle.

L'IA a été définie de différentes manières. Dans ce chapitre nous adopterons la définition proposée par Shapiro (1992), à savoir, le domaine de la science et de l'ingénierie qui traite :

- de la compréhension, à l'aide de l'ordinateur, de ce qui est appelé couramment le comportement intelligent,
- et de la création de systèmes artificiels qui reproduisent ces comportements.

1.2. L'intérêt de l'IA pour les psychologues

Il y a plusieurs façons de considérer l'intelligence artificielle, la première est celle des ingénieurs. Elle vise à résoudre des problèmes et ce de manière optimale sans se préoccuper de la cognition humaine bien que ces chercheurs emploient souvent l'introspection pour développer leurs méthodes. La seconde voit l'intelligence artificielle comme partie des sciences cognitives ; ici on ne cherche plus à optimiser mais à imiter l'homme, c'est-à-dire à réussir là où les sujets réussissent et à échouer là où les sujets échouent. Au sein de cette seconde approche, Searle (1980) distingue deux paradigmes. Le premier, appelé "IA faible", considère la modélisation sur ordinateur comme un outil pour étudier l'esprit humain. Le second, "IA forte", estime possible la création au moyen de l'ordinateur d'un véritable esprit artificiel. Le programme dans ce cas constitue une explication de son fonctionnement.

Chacune de ces trois visions — l'IA comme moyen d'optimiser, comme outil et comme explication — a selon nous des vertus pour l'étude de l'esprit humain. Toutes trois permettent d'identifier des conditions suffisantes à la production de certains comportements intelligents.

1.2.1. L'IA comme optimisation

Les programmes d'optimisation cherchent à simuler des compétences intelligentes de manière indépendante de celles de l'homme. Ces tentatives débouchent parfois sur des impasses qui sont instructives. Elles amènent à nous poser la question : comment l'humain échappe-t-il à ce problème ? Cela nous conduit à remettre en cause certaines théories et à identifier de nouveaux problèmes, comme par exemple, le "frame problem" (voir 3.1.5). De même, les succès de cette intelligence artificielle optimisante poussent à distinguer la façon dont le programme fonctionne des stratégies humaines, comme par exemple, celles utilisées dans le jeux d'échecs.

1.2.2. L'IA comme outil ou méthode

Tout commence par la programmation de l'ordinateur de manière prescrite par une théorie. Déjà à ce stade, cette méthode apporte une contribution car une théorie essentiellement verbale peut receler des insuffisances ou de nombreux défauts qui resteront cachés jusqu'à ce que quelqu'un se mette à la programmer.

La seconde phase consiste à comparer les comportements de la machine à ceux des humains. On se concentre alors sur les écarts entre le modèle et les données humaines et sur les prédictions du modèle, à propos du comportement humain, qui ne sont pas vérifiables sur des données. En effet, quand une simulation fait face à une tâche, elle a le potentiel de réagir de façon originale. Ceci introduit le deuxième type de contribution de la méthode : les écarts indiquent comment le modèle peut être amélioré, les prédictions procurent des hypothèses à vérifier par des recherches expérimentales.

Lorsqu'un modèle a été construit et testé, la troisième phase consiste à réviser ce modèle ou à en construire un autre différent, de manière soit à le généraliser en l'appliquant à de nouveaux domaines, soit à mieux l'adapter aux données ou bien encore, à tester les nouvelles hypothèses formulées.

1.2.3. L'IA comme explication

Sans entrer dans le débat philosophique sur la capacité des ordinateurs d'expliquer l'esprit humain, notons que les apports les plus considérables de l'IA à la psychologie viennent de cet effort (citons le "General Problem Solver", voir 3.1.4., ou les travaux de McClelland et Rumelhart sur le "Parallel Distributed Processing" 3.2.1). Malgré le caractère partiel des explications données par ces modèles, leur apport est multiple. En plus des avantages présentés par les deux méthodes précédentes, elle offre les intérêts suivants.

1. Ces modèles recréent les compétences humaines de manière détaillée et débouchent sur de nouvelles théories, hypothèses et prédictions. Les théories de l'expertise, par exemple, ne sont apparues en psychologie, qu'à la suite des travaux de Newell et Simon. Le renouveau du courant connexionniste en psychologie est dû aux travaux de Rumelhart et McClelland (voir 3.2.1).
2. Ces modèles par leurs choix, s'imposent des contraintes. Par exemple, le connexionnisme modélise l'esprit à partir d'unités très simples interconnectées et évoluant sans système d'administration ou de contrôle. Une fois ces contraintes imposées, la réalisation des modèles débouche sur de nouveaux problèmes qui s'avèrent importants pour la psychologie. Citons par exemple le problème de l'affectation des variables (voir 3.2.1). Ces problèmes suscitent des solutions qui sont autant d'hypothèses sur la manière dont l'esprit humain procède.

2. Les machines peuvent-elles penser ?

Qu'entendons-nous par l'intelligence ? Comment la détecter ? Peut-on la définir rigoureusement ? Pour certains (Newell, Rosenbloom, Laird, 1989), elle se caractérise par un certain nombre de capacités.

1. Réagir de manière souple à l'environnement.
2. Exhiber un comportement rationnel.
3. Opérer en temps réel.
4. Opérer dans un environnement riche et complexe :
   percevoir, utiliser des connaissances, contrôler le système moteur.
5. Utiliser des symboles et des abstractions.
6. Utiliser des langages (naturels et artificiels).
7. Apprendre.
8. Acquérir de nouvelles compétences.
9. Vivre de manière autonome dans une communauté sociale.
10. Exhiber une conscience de soi.

Pour d'autres, de telles listes possèdent un caractère trop rigide; ils préfèrent une définition plus opérationnelle. La plus fine et célèbre expression de cette dernière approche est sans doute apparue dans l'article de Turing (1950) sur un test comportemental de l'intelligence que nous considérons ci-dessous.

2.1. Le test de Turing et la chambre chinoise

La question : "Les machines peuvent-elles penser ?" a été abordée en 1950 de manière originale par le mathématicien anglais Alan Turing. Auparavant toute discussion au sujet de l'intelligence mécanique s'appuyait sur divers critères prétendument nécessaires à l'intelligence, comme ceux cités ci-dessus. L'article de Turing a formulé cette question tout à fait autrement.

Son article est fondé sur la notion d'identité comportementale : tant qu'on ne constate pas de différence comportementale entre deux entités, on peut conclure que les deux appartiennent vraisemblablement à la même famille. Supposez que vous ne connaissiez qu'un seul individu de nationalité française. Cette personne parle français, vit à Paris, vote socialiste, considère Jerry Lewis comme un génie du cinéma, mange des baguettes à tous les repas, fume des Gauloises sans filtre, et porte un béret dans la rue. Ensuite, vous rencontrez quelqu'un de nationalité inconnue. Son comportement est cependant identique à celui du seul Français que vous connaissiez : il parle français, vit à Paris, vote socialiste, considère Jerry Lewis comme un génie du cinéma, mange des baguettes à tous les repas, fume des Gauloises sans filtre, et

porte un béret dans la rue. Il est probable que vous supposiez que ce deuxième individu est de nationalité française. Si, par contre, une personne ne remplit pas toutes ces conditions, vous hésiterez à vous prononcer sur sa nationalité. Il n'est d'ailleurs pas nécessaire de connaître la définition rigoureuse de "Français" pour arriver à cette conclusion. Ce faisant vous appliquez implicitement, comme Turing propose de le faire, le principe d'identité comportementale.

Le test envisagé, et qui aujourd'hui porte le nom de test de Turing, avait pour but explicite d'éviter une définition classique de l'intelligence par exemple, au moyen d'une liste de propriétés. Une machine qui pouvait réussir le test, un simple jeu d'imitation, serait intelligente. Il se joue comme suit.

Un interrogateur **I** cherche, à partir de questions, à identifier celui qui parmi ses deux interlocuteurs est un ordinateur. Pour ce faire, et pour ne pas défavoriser la machine, **I** est situé dans une salle différente et communique uniquement par télex avec ses partenaires. L'interlocuteur humain répond aussi honnêtement que possible aux questions posées; l'ordinateur, par contre, cherche à induire **I** en erreur en se faisant passer pour une personne. Si après une longue séance de questions, pouvant couvrir tous les sujets imaginables, **I** n'arrive toujours pas à identifier l'ordinateur, celui-ci, affirme Turing, a réussi le test et, par conséquent, doit être considéré comme intelligent.

L'article de Turing a, dès sa parution en 1950, suscité un très vif débat sur l'intelligence artificielle. On se demandait si machine pouvait réussir le test sans être réellement intelligente. Une réussite au test de Turing constituait-il une condition *suffisante* pour l'intelligence ? Pour démontrer la non-suffisance de cette condition, on imaginait l'existence d'une vaste base de données constituée de tous les faits, aussi évidents ou obscurs soient-ils, de notre univers humain, tels que la hauteur de la Tour Eiffel, la profondeur de la Seine sous le Pont Neuf, la date de la bataille de Marignan, la définition d'un triangle, le nombre de planètes du système solaire, le nombre de centimètres dans un pouce, la phrase de quatre notes de la 5ème symphonie de Beethoven, et ainsi de suite. On faisait toujours abstraction du temps de recherche nécessaire pour trouver une information donnée. La supposition était que la réponse, sous une forme ou une autre, à toute question possible se trouvait quelque part dans cette immense base de données. La machine serait donc capable de répondre convenablement à toutes les questions posées par l'interrogateur et par conséquent de réussir le test à l'aide de sa base de données et de ses algorithmes de recherche. On parvenait ainsi à la conclusion qu'une machine de ce type pourrait réussir le test de Turing sans pour autant posséder le moindre brin d'intelligence.

Ce même argument a été repris sous une autre forme trente ans plus tard dans un célèbre article écrit par le philosophe John Searle (1980). Il a imaginé une *Gedankenexperiment*, appelée "La chambre chinoise," où il met en relief la possibilité de réussir le test de Turing sans posséder de l'intelligence, la vraie, c'est-à-dire, la nôtre. Searle nous demande d'imaginer une chambre dans laquelle se trouve un homme qui n'a pas la moindre connaissance du chinois. Cette chambre est équipée d'une immense "bibliothèque" constituée d'énormes piles de

documents sur lesquels on a consigné des règles du type “SI vous recevez tel type de caractère chinois (ou chaîne de caractères) ALORS répondez par tel ou tel autre caractère chinois (ou chaîne de caractères).” La chambre comporte une ouverture par laquelle on peut échanger des questions et des réponses. A l’extérieur, un “interrogateur” chinois écrit des questions (uniquement en chinois) et les introduit dans la chambre. A l’intérieur, l’individu prend chaque question (dont il ignore complètement la signification) et commence à chercher une réponse parmi les piles de règles. Puisqu’il ne lit pas le chinois, il est obligé de comparer la forme de chaque caractère de la question aux caractères dans les piles de feuilles. Il passe de règle en règle jusqu’à ce qu’il parvienne à produire une suite de caractères qui constitue une réponse (mais dont il ignore complètement le contenu). Il la transmet, parfaitement rédigée en chinois, à l’interrogateur qui la lit. Si les règles ont été correctement définies, les réponses doivent être bien écrites et raisonnables et, par conséquent, cacher l’ignorance totale du chinois de leur auteur. Une conclusion à la Turing nous obligerait à inférer de manière erronée à partir des réponses une connaissance du chinois.

L’article de Searle, paru en 1980, a fait couler une quantité invraisemblable d’encre. Pour certains, l’argument démontre une proposition fort rassurante, à savoir, que l’intelligence humaine — et *a fortiori* la conscience — restera à jamais unique, au delà de toute simulation par ordinateur. D’autres n’y voient aucunement une démonstration de l’unicité théorique de l’intelligence humaine et ont démonté l’argumentation développée par Searle pour en montrer le caractère spécieux.

2.2. Difficultés avec l’argument de la chambre chinoise

Il existe de très nombreuses critiques de cet argument mais nous présenterons ici la plus convaincante à nos yeux.

La présence d’une personne au centre du dispositif imaginé par Searle n’est pas innocente, rien n’exige que la recherche automatique des éléments de réponses soit effectuée par un être humain. Lorsqu’il parle d’une personne dans une chambre, nous imaginons tous une vraie personne en chair et en os dans une vraie chambre. Quand il nous parle d’ouverture pour échanger les questions et les réponses, nous pensons à une ouverture réelle dans notre chambre imaginaire, et ainsi de suite. Il est difficile de ne pas s’identifier à la personne dans la chambre, recevant des questions composées de chaînes de caractères chinois, cherchant des éléments de réponses dans d’énormes piles de papiers, dessinant péniblement des caractères chinois censés constituer une réponse. Et on se dit, “Effectivement, je vois comment quelqu’un — en l’occurrence *moi* — pourrait répondre à une question en chinois de manière appropriée sans connaître le chinois.”

L’argument de Searle paraît convainquant. Il résiste pourtant mal à quelques déformations mineures. Commençons par une modification des dimensions des objets dans cette *Gedankenexperiment*. En principe, ces changements ne devraient pas influer sur l’essence

de l'argument de Searle. La taille de l'homme dans la chambre va être réduite à l'échelle du micron. Appelons-le le "démon de Haugeland" en honneur de John Haugeland (1980), qui fut le premier à formuler cette réponse à Searle. Ce démon sera doté d'une rapidité extraordinaire. Maintenant, transformons la chambre et réduisons son volume à environ 1300cc., celui du cerveau humain. Et les piles de règles ? Le cerveau humain est constitué de neurones dont le fonctionnement peut être schématisé par une règle : *SI il y a une accumulation suffisante d'activité dendritique, ALORS produire un déclenchement axonal.* Les piles de règles sont donc remplacées par des neurones. Lorsqu'une question est posée, le démon de Haugeland s'active fébrilement. Tout comme l'homme (taille normale) dans la chambre chinoise (taille normale) doit passer de règle en règle pour trouver sa réponse, son homologue micronal doit sauter de neurone en neurone pour les "chatouiller". De cette manière, le démon parvient à simuler le fonctionnement neuronal habituel d'un Chinois qui aurait lu la question. La réponse émise par "la petite chambre chinoise" (c'est-à-dire, une tête), tout comme celle qui est venue de la grande chambre chinoise de Searle, sera parfaitement sensée. Mais maintenant nos intuitions sont complètement différentes. Le démon ne comprend certainement pas le chinois — pas plus que son homologue dans la grande chambre chinoise — mais *le système comprenant le démon et les règles-neurones*, lui par contre comprend bel et bien le chinois, car c'est n'est rien d'autre que le cerveau d'un chinois. (Voir Hofstadter et Dennett, 1981, pour un exposé détaillé des arguments et contre-arguments de la chambre chinoise.)

2.3. D'autres problèmes : La surdifficulté du test de Turing

L'énorme difficulté qu'aurait une machine à réussir le test de Turing (French, 1990) a également été mentionnée. L'article de Turing se divise en deux thèses, la première théorique, l'autre pragmatique. La thèse théorique affirme qu'une machine qui réussit le test de Turing est intelligente. Cette thèse-là est depuis presque 50 ans est au coeur de presque toutes les discussions sur le test. Mais si l'on pouvait démontrer que sur le plan pratique il est impossible de réaliser une machine capable de réussir le test, la discussion autour de la thèse théorique deviendrait beaucoup moins intéressante. Sa contrepartie pragmatique dont on entend moins parler peut se formuler comme suit : il sera un jour possible de construire une machine qui réussisse le test de Turing.

Montrons la difficulté de cette entreprise; *une machine qui n'a pas grandi, vécu et évolué dans notre monde ne pourrait jamais réussir le test de Turing.*

Il ne faut pas oublier que *n'importe quelle question* peut être posée par l'interrogateur (à condition qu'elle puisse être tapée à la machine). Cherchons par exemple, plutôt qu'à tester des connaissances explicites telles que le nom de la capitale du Sénégal ou une description de la manière d'attacher ses lacets, à sonder le niveau *sous-cognitif* de l'intelligence. Qu'entendons-nous par une question sous-cognitive ? Il s'agit de questions dont la réponse fait appel non pas à des connaissances explicites, mais à des connaissances acquises de manière non explicite

à travers notre interaction avec le monde. Si l'on vous demande, par exemple, de répondre le plus rapidement possible et sans souci d'exactitude à la question suivante :

"Que boivent les vaches ?"

Vous allez sans doute donner la réponse "Lait" — qui vient spontanément à l'esprit. — et non pas la bonne réponse qui est "Eau". Cette dernière réponse, donnée plus lentement, est dérivée de connaissances explicites sur la soif, sur la manière dont les mammifères la satisfont, sur le fait qu'une vache est un mammifère, etc. Les questions sous-cognitives veulent provoquer des réponses spontanées, automatiques. Voici un petit échantillon de questions sous-cognitives :

"Evaluez sur une échelle de 0 (complètement impossible) à 10 (tout à fait possible) la plausibilité des associations suivantes :

Choupette comme le nom d'une nouvelle fabrique d'ordinateurs;
Choupette comme le nom d'un aspirateur-jouet pour enfants;
Bubulle comme le nom d'un mannequin;
Bubulle comme le nom d'un "nounours" d'enfant;
Un *cornet de glace vanille* comme un *médicament*;
Un *stylo* comme une *arme;*
Un *sac à main* comme une *arme*;
Une *pile de feuilles mortes* comme une *cachette*;
Etc."

Les réponses à ces questions ne se trouveront pas dans un livre ou dans une quelconque base de données, mais sont directement tirées de nos expériences avec le monde. Aucun dictionnaire, par exemple, ne contient les mots inventés "Choupette" et "Bubulle." Pourtant nous sommes tous capables de dire que "Bubulle" paraît un mauvais nom pour un mannequin, tout en étant un nom tout à fait plausible pour un petit ours en peluche. De même, l'association pile de feuilles mortes et cachette est construite à travers notre vécu. Mais comment programmer tout cela *explicitement* dans une machine ? Sûrement pas en explicitant toutes les associations possibles et tous les poids de ces associations pour tous les objets, situations, effets, sentiments, endroits, idées, combinaisons de mots et de néologismes, expressions, etc. Nous humains acquérons tout cela sans difficulté grâce à notre contact quotidien avec le monde. Comment serait-il possible de connaître toutes ces associations à l'avance afin de les programmer explicitement en machine ? La difficulté, voire l'impossibilité, d'une programmation de ces connaissances sous-cognitives est la clef de la méthode que l'interrogateur peut utiliser pour dévoiler infailliblement la machine dans le test de Turing.

Le test peut donc paraître trop difficile pour servir de critère d'intelligence. Imaginez

donc une machine identique à nous à une seule différence près : ses yeux sont attachés aux genoux. Ceci ne devrait pas avoir un impact sur son intelligence. Par contre, ses réponses à des questions sous-cognitives permettraient facilement de la repérer. Son réseau d'associations serait différent de celui des humains et ces différences seraient mises en évidence. Par exemple, ses associations avec une chute en vélo, avec un coup de pied au genou lors d'un match de football, avec un pansement autour du genou, et même avec le simple port d'un pantalon long, etc. — autrement dit, des événements sans grande importance chez les humains — la mettrait dans des situations pour le moins incommodes, voire dangereuses. Cela entraînerait inéluctablement des différences qui seraient détectées par des questions sous-cognitives. Cette machine tout à fait intelligente serait donc écartée par le test de Turing. Il est probable en fin de compte que la seule chose capable de passer le test soit un être humain.

Serait-il possible d'éviter cette difficulté en interdisant les questions sous-cognitives ? Non, car la frontière est trop floue. Imaginez, par exemple, un jugement sur un poème, une plaisanterie, une publicité, ou le ton d'une lettre. Le sous-cognitif semble fort proche ! Le dissocier de la cognition classique, de l'intelligence est illusoire.

3. Présentation des différents types de modèles

3.1. Un premier ensemble de contraintes

3.1.1. L'activité mentale conçue comme une aptitude à manipuler des symboles

Introduction

Lorsqu'on regarde autour de soi, on peut sans problème donner un nom à chacun des objets situés dans notre champ visuel. On voit, par exemple, une porte, un stylo, un ordinateur, des pièces de monnaie, un mur, une fenêtre, une chaussette, etc. De la même manière, toutes les actions qu'on perçoit portent un nom : l'enfant *court*, on *sort* la poubelle, on *se lave* les mains, etc. Nos états d'âme sont qualifiés, labélisés : on est *nostalgique*, on *croit* en quelque chose, on *espère*… La mise en rapport d'une sensation et d'une catégorie symbolique paraît du reste à beaucoup un aspect fondamental de la perception.

Une petite expérience suffira à démontrer la nature "grammaticale" de la pensée. Choisissez n'importe quel objet dans la pièce où vous vous trouvez actuellement et portez votre réflexion sur cet objet, sans parler. Maintenant, écoutez-vous. Vous "entendrez" sans doute des phrases, des mots, etc., qui donnent l'impression d'une pensée constituée de phrases, structurées comme celle que l'on verbalise ou que l'on communique à d'autres. Quoi de plus logique ? Le monde est organisé en objets et actions qui portent des noms. Il peut être décrit

de manière apparemment satisfaisante à l'aide de phrases qui enchaînent ces mots à partir de règles grammaticales précises. Quand on écoute nos propres pensées, on "entend" des phrases parlées. On en déduit qu'il existe un "langage de la pensée" (Fodor, 1975) qui ressemble très fort à la langue que l'on parle.

Cette capacité à manipuler des symboles qui renvoient à des objets, des actions, des états d'âme, des catégories du monde qui nous entoure est vite apparue comme une propriété essentielle des systèmes intelligents. Certains ont même fait un pas supplémentaire : pour reproduire la pensée sur un ordinateur, il suffit essentiellement de lui inculquer cette aptitude à manipuler les mots que nous utilisons pour décrire le monde et toutes les règles qui gèrent l'usage de ces mots. L'hypothèse affirmant que les symboles que les systèmes artificiels intelligents doivent manipuler correspondent à des objets de, et à des actions sur, notre environnement est souvent appelée l'hypothèse déclarative. Elle a pris une forme légèrement plus générale dans l'hypothèse des systèmes physiques de symboles proposée en 1975 par Newell et Simon (1976) et présentée dans un paragraphe suivant. L'IA lui doit de nombreux développements théoriques qui ont culminé dans des systèmes de type SOAR, ACT*. Boden (M. Boden, 1991b) parle également du paradigme symbolique pour évoquer cette famille de modèles. Comme il est à la base des premiers systèmes développés en IA, certains parlent aussi d'IA traditionnelle, ou de BOVIA (BOnne Vieille Intelligence Artificielle) ou plutôt de GOFAI puisque l'abréviation est d'origine anglophone (Good Old-Fashioned Artificial Intelligence) (Haugeland, 1980).

L'hypothèse des systèmes physiques de symboles

Comme nous venons de le signaler, la conception de l'intelligence comme la capacité à manipuler des symboles a trouvé son expression la plus forte dans l'hypothèse des systèmes physiques de symboles proposée en 1975 lors d'une conférence organisée par l'Association for Computing Machinery par Newell et Simon, deux figures importantes de la discipline (Newell et Simon, 1981). Cette formulation a eu et a toujours une grande influence sur la manière dont l'IA perçoit ses fondements .

"Un système physique de symboles consiste en un ensemble d'entités, appelées symboles. Ces symboles sont des structures physiques qui peuvent apparaître comme des composantes d'un autre type d'entités appelées expressions (ou structures symboliques). Une structure symbolique est donc composée de symboles particuliers physiquement reliés (ils peuvent être adjacents, par exemple)... De plus, le système contient également un ensemble de procédures qui transforment les expressions en d'autres expressions : création, modification, reproduction et destruction. Il se présente comme une machine qui produit au cours du temps une collection dynamique de structures symboliques".

Pour pouvoir être à la base de conduites intelligentes de tels systèmes doivent posséder des propriétés supplémentaires, à savoir :

1. un symbole doit pouvoir désigner n'importe quelle expression;

2. toute procédure que la machine est capable d'exécuter doit pouvoir être désignée par une expression;
3. le nombre d'expressions n'est pas limité;
4. elles peuvent générer ou modifier n'importe quelle autre expression;
5. elles sont stables.

Si de telles hypothèses sont satisfaites, Newell et Simon affirment - il s'agit bien entendu d'une hypothèse - que *ces systèmes possèdent les caractéristiques nécessaires et suffisantes pour réaliser des actions intelligentes***.**

L'hypothèse des systèmes physiques de symboles est bien entendu de nature empirique. Elle repose essentiellement sur deux constatations. Le caractère suffisant de l'hypothèse paraît justifié par les résultats déjà acquis en IA au cours de sa brève existence; la manipulation d'entités symboliques a permis de simuler plus ou moins bien de nombreuses conduites réputées intelligentes, de la reconnaissance de certaines formes, au raisonnement logique, en passant par l'apprentissage, le raisonnement analogique, la compréhension de langues naturelles. Le caractère nécessaire de l'hypothèse viendrait de la psychologie cognitive; l'étude des comportements intelligents a mis en évidence l'utilisation par l'homme de mécanismes qui s'apparentent à ceux des systèmes physiques de symboles.

Essentielle est également la capacité d'interprétation d'expressions symboliques : le système doit pouvoir exécuter la procédure désignée par une expression. Ceci confère donc aux symboles un rôle de déclenchement d'actions et pas uniquement un rôle de dénotation.

Une autre manière de bien comprendre la notion de système physique de symboles est de la sonder à partir de ses différentes phases de développement telles que Newell et Simon les présentent. L'hypothèse s'enracine dans des travaux de formalisation de la logique de Frege, Whitehead et Russell. Ceux-ci démontrent la possibilité de se débarrasser en quelque sorte des problèmes de signification en se limitant à garantir la cohérence des manipulations de symboles. Plus tard ces manipulations sont automatisées au moyen d'ordinateurs capables de considérer les programmes comme des données c'est-à-dire comme des objets susceptibles de transformation. L'étape ultime dans l'avènement des systèmes physiques de symboles coïncide avec le développement de langages de programmation de type LISP. Qu'ont-ils de particulier ? Leur capacité de désignation : les listes renvoient à d'autres listes, sont des pointeurs vers d'autres structures. Ceci éclaire le concept de systèmes physiques de symboles. Il s'agit de machines capables d'effectuer des calculs sur des expressions structurées interdépendantes, qui peuvent s'interpréter. S'il est facile de comprendre que l'hypothèse de Newell et Simon ait heurté ceux qui ne croient pas à l'intelligence artificielle, il est permis de s'interroger sur les débats qu'elle a suscité et continue de susciter au sein de la communauté des chercheurs de cette discipline. La formulation paraît en effet excessivement générale et paraît couvrir la plupart des systèmes construits jusqu'à présent. Le rôle central attribué aux symboles dans la plupart des systèmes - ils constituent les atomes de la pensée - pose cependant problème à certains. Comme déjà signalé, les symboles appartiennent au langage commun, ils font référence aux concepts

que la pensée manipule quotidiennement. La formulation de Newell et Simon paraît ramener l'intelligence à la maîtrise d'un langage adéquat pour organiser et manipuler ces entités conscientes et atomiques.

3.1.2. L'activité mentale conçue comme une aptitude à résoudre des problèmes

La plupart des systèmes de la BOVIA ont abordé des tâches de résolution de problèmes. L'intelligence est conçue comme la capacité à trouver des solutions en enchaînant des opérations mentales. Dans cette approche, l'accent est mis sur notre capacité à concevoir la résolution d'un problème comme la recherche d'une solution dans un espace approprié, en procédant par essais et erreurs ou en utilisant des méthodes particulières.

La notion de problème est difficile à cerner simplement. Newell, Shaw et Simon (1963) définissent comme problématique pour un individu toute situation dans laquelle il ne voit pas immédiatement la manière d'obtenir un résultat voulu. Les exemples les plus typiques sont certainement les jeux abondamment étudiés en IA. Leur intérêt est multiple : les problèmes se posent en termes clairs et simples, ils permettent, à partir d'une structure initiale simple, d'aborder des structures beaucoup plus complexes, des méthodes ayant des domaines de validité plus larges. Un jeu copieusement cité dans les manuels pour illustrer l'approche “résolution de problème” est le *taquin*. Il s'agit d'une plaquette carrée composée de petits pavés plats numérotés que l'on peut faire glisser. L'objectif est de les disposer dans un ordre défini a priori. La figure 1, ci-dessous illustre une configuration initiale possible et une configuration cible du taquin.

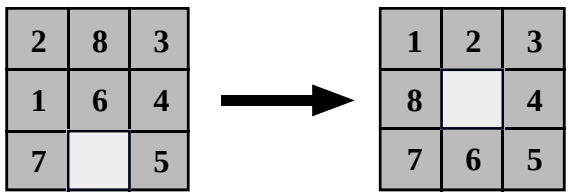

Figure 1 : le jeu de taquin

Ce type de problème est bien caractéristique d'une classe importante de problèmes étudiés en IA. Les situations initiales et finales sont clairement définies; il est possible au moyen d'un certain nombre d'opérations de changer de configurations. Pour passer d'un état à l'autre on utilise ce qu'on appelle un opérateur. Dans l'exemple du *taquin* il existe quatre opérateurs correspondant aux 4 types de mouvements possibles, en principe, à partir d'une configuration donnée. La case sans pavé peut généralement, en effet, en bougeant un autre pavé, se retrouver à droite, à gauche, au-dessus ou en-dessous de son emplacement initial. Trouver une solution revient à trouver une séquence d'opérateurs qui permettent de passer de la configuration initiale à la configuration finale.

Une autre manière souvent rencontrée de résoudre les problèmes est la méthode de réduction. Elle consiste essentiellement à décomposer le problème initial en sous-problèmes plus faciles à résoudre. Ceci peut se faire naturellement si la résolution comporte des étapes stratégiques facilement identifiables. Si, à partir de Liège, je veux aller à l'université d'Oxford, je planifie chaque étape en commençant par les plus importantes. Je dois m'occuper en premier lieu du trajet de Bruxelles jusqu'à Londres. Je décide de prendre l'avion. Ensuite, je détermine comment arriver à Bruxelles. Ce problème résolu (je prends le train), il me reste le trajet "domicile-gare de Liège". Et ainsi de suite. Chaque étape, aussi petite soit-elle, consiste en la *résolution d'un sous-problème.* Cette méthode est à la base d'un des logiciels les plus connus en IA appelé GPS (General Problem Solver). Il fut développé dans les années 60 par Newell, Shaw et Simon (1963; voir aussi Newell et Simon, 1963). Sa renommée est liée à sa capacité de résoudre différents types de problèmes (jeux simples, problèmes de logique, d'intégration (symbolique) de fonctions) et à son architecture dans laquelle les éléments de connaissance nécessaires à la résolution (opérateurs, objectif, importance des différences entre la cible et l'état initial, ...) sont clairement séparés de la connaissance générale dont dispose le système sur la manière de résoudre les problèmes.

Les opérations menées dans le cadre d'une résolution de problèmes nécessitent généralement une analyse logique. Il convient en effet de répondre à des questions du type "Cet état est-il l'état que l'on veut obtenir ?", "Est-il possible d'effectuer tel mouvement à partir de tel type de configuration ?", "Tel type de sous-problème a-t-il une solution ?". La résolution de problèmes paraît donc fortement dépendante de notre capacité à raisonner logiquement, d'où la contrainte suivante (voir également la deuxième des caractéristiques des systèmes intelligents proposées par Newell et Simon et mentionnées dans un paragraphe précédent).

3.1.3. L'activité mentale conçue comme une aptitude à raisonner logiquement

La reconnaissance du rôle de la logique dans la pensée est bien entendu largement antérieure aux tentatives d'automatisation des activités mentales. Certains y ont vu l'essence même de l'intelligence. Il n'est donc pas surprenant que les premiers systèmes artificiels lui aient réservé une place de choix. Bien que certains logiciens réfutent toute adéquation des logiques au raisonnement humain, l'idée que la pensée humaine suit une logique particulière a longtemps dominé la psychologie (Inhelder et Piaget, 1955). Cette idée est encore prônée par certains (Rips, 1994) tout en étant décriée par d'autres au nom des erreurs systématiques et des effets contextuels (Cheng et Holyoak, 1985, Johnson-Laird, 1983, Cosmides, 1989). Du côté de l'intelligence artificielle, certains voient encore la logique comme la clé de voûte de l'esprit humain. SNePS (Shapiro, 1979) en est un bon exemple. SNePS est un réseau sémantique fonctionnant selon une logique de la pertinence.

Le premier algorithme capable de prouver des théorèmes est l'algorithme de Wang (Wang, 1963). Celui-ci est capable de rechercher si des expressions constituent des théorèmes (on dit prouver des théorèmes). Prenons un exemple : le programme doit prouver l'expression

suivante ~(p ∨ q) ⊃ (~p ∨ ~q). Nous conviendrons, pour faciliter la compréhension, que l'expression formelle ci-dessus peut-être interprétée comme suit : “Nier que les pommes soient vertes ou bleues implique qu’elles ne soient pas vertes ou qu’elles ne soient pas bleues”. Le tableau 1 montre les étapes du raisonnement suivi par le programme pour résoudre ce problème.

Liste des formules vraies	Liste des formules fausses
(() (((NON (OU BLEUES VERTES))) (() (() (() ((BLEUES) T	((⊃ (NON (OU BLEUES VERTES)) (OU (NON BLEUES) (NON VERTES))))) ((OU (NON BLEUES) (NON VERTES)))) ((OU BLEUES VERTES) (OU (NON BLEUES) (NON VERTES)))) (BLEUES VERTES (OU (NON BLEUES) (NON VERTES)))) ((NON BLEUES) (NON VERTES) BLEUES VERTES)) ((NON VERTES) (BLEUES VERTES))

Tableau 1 : Un résultat de l’exécution de l’algorithme de Wang (1963).

L’algorithme commence par considérer que la conclusion à prouver est fausse et que les prémisses sont vraies. Si après une série de transformations, on arrive à une contradiction, alors le théorème est prouvable. L’algorithme comporte deux listes: une liste de formules vraies et une liste de formules fausses. En voici les étapes :

1. Si une formule (par exemple `BLEUES ou (NON VERTES)`) est dans les deux listes, alors retourner vrai; on est arrivé à une contradiction (à l’avant dernière ligne du tableau 1, `BLEUES` se trouve dans la liste des formules vraies et dans la liste des formules fausses).
2. Si les deux listes contiennent uniquement des formules atomiques (`BLEUES` ou `VERTES` dans l’exemple), alors retourner faux.
3. Enlever une formule non atomique (par exemple `(⊃ (NON (OU BLEUES VERTES))` entre la première et la deuxième ligne du tableau 1) d’une des listes et appeler l’algorithme de manière récursive. Cela sera fait de manière différente pour chaque opérateur (Et, Ou, Non, ⊃, ≡) et pourra être effectué de deux manières différentes (dans le cas de ≡ par exemple) auquel cas si les deux retournent vrai, l'algorithme retourne vrai, sinon il retourne faux.

Les systèmes à règles de production

Notre aptitude à enchaîner logiquement des règles, mises en évidence lorsqu'on demande par exemple à des experts d'expliquer la manière dont ils opèrent, a débouché sur une architecture de systèmes largement rencontrée en IA : les systèmes à règles de production. Ils constituent le coeur des systèmes dits experts qui ont connu un certain succès commercial. Au centre du dispositif, la règle de production :

“si telle condition est satisfaite alors exécuter telle action”.

Certains y voient un des mécanismes fondamentaux sous-jacent à toutes nos activités intellectuelles dites supérieures (résolution de problèmes, déduction, induction, ...). Le présupposé est ici que notre activité cognitive passe par une sélection, une utilisation et un enchaînement de règles stockées dans une grande base de données. Même si l’on n’est pas

toujours conscient de ces règles, elles existent et sont à la base de notre activité mentale. Dans cette approche les règles manipulent des éléments symboliques du type de ceux présentés précédemment : des objets, des actions, des états d'âme, des catégories, ... Il ne s'agit pas de la règle au sens informatique du terme qui, elle, se trouve au coeur de tous les langages de programmation et, par conséquent, de tous les systèmes évoqués dans ce chapitre.

3.1.4. Des systèmes obéissant à ces contraintes

La BOVIA

La conjugaison des contraintes reprises ci-dessus (contrainte *symbolique*, l'intelligence conçue comme une manière *rationnelle* de *résoudre des problèmes*) a donné naissance aux premiers systèmes de l'IA et a permis d'obtenir des résultats spectaculaires lors des quarante dernières années. Mêmes si les systèmes construits ne sont pas à la hauteur des ambitions initiales, certaines percées de la BOVIA méritent d'être soulignées. Donner une liste des principaux systèmes n'est ni possible, ni vraiment utile. Mentionnons simplement à titre illustratif que l'IA s'est attaquée à des domaines aussi différents que :

- la traduction automatique (SYSTRAN, Eurotra ...),
- les jeux (Dames : Samuel; Echecs : DEEP-BLUE de IBM, Backgammon BKG de Berliner, 1980,...),
- la démonstration de théorèmes mathématiques (Logic Theorist, Gelernter's Geometry Theorem-proving Machine, EURISKO de Lenat, 1983),
- la manipulation d'expressions formelles (MACSYMA),
- la vision (Algorithme de Waltz, 1975, le programme SEE de Guzman, ...),
- la reconnaissance d'écritures manuscrites (Structural Feature Extraction Method de Baird, 1988,…),
- la reconnaissance de certaines formes géométriques (ACRONYM Brooks, 1981,...)
- la reconnaissance à l'audition (HEARSAY,...),
- la compréhension du langage naturel (SHRDLU, BORIS, XCALIBUR de Carbonell et al., 1985,...),
- la création d'oeuvres d' "art" (AARON de Cohen, 1988,...),
- la découverte scientifique (BACON, AM, ...),
- l'apprentissage analytique par explication (LEX de Mitchell (1981),
- l'apprentissage basé sur les cas (ICARUS de Langley et al., 1992)
- l'apprentissage par induction de règles (Automatic Classification of Celestial Objects de Fayad et al, 1995....),
- la résolution de problèmes en général (GPS, SOAR, ACT*, ...).

Le lecteur intéressé trouvera un examen systématique et complet de ces réalisations dans l'"Encyclopedia of Artificial Intelligence" Shapiro (1992).

Ces systèmes ont en général des structures assez différentes, souvent très dépendantes

des tâches que l'on simule. Newell, Rosenbloom et Laird (1989) ont cependant introduit une notion d'architecture cognitive qui permet de donner une vision intégrée des différentes fonctions et des différentes structures présentes dans la plupart des systèmes artificiels intelligents du type BOVIA. Ce concept permet d'attirer l'attention sur des mécanismes fondamentaux comme le contrôle (c'est-à-dire sur la logique et la mécanique de déclenchement et d'enchaînement des différentes opérations mentales), sur des traitements élémentaires (la manière d'apparier des variables, les mécanismes de rappel de certaines informations, d'interprétation d'instructions, etc.), sur les notions de tractabilité (voir plus loin), et sur les effets de contexte (frame problem en anglais) généralement peu couverts dans les modèles psychologiques classiques. Ce que doit pouvoir réaliser un système intelligent selon Newell et al - résumé en 10 caractéristiques élémentaires - a déjà été introduit. Les éléments constitutifs des systèmes symboliques doivent couvrir cinq types de fonctionnalités différentes selon ces mêmes auteurs (Newell et al, 1989) :

1. Mémorisation: capacité de stocker des valeurs de symboles.
2. Symbolisation: existence de structures qui permettent d'accéder à d'autres symboles.
3. Traitement: opérations qui permettent de transformer des structures symboliques.
4. Interprétation: transformations de structures symboliques en actions en exécutant des opérations.
5. Interaction avec le monde extérieur: capacité de communiquer en temps réel à travers des interfaces.

Le système SOAR

Cette architecture peut être illustrée par le programme SOAR, développé par Laird, Newell, et Rosenbloom (1987). SOAR est le fils spirituel d'un des systèmes les plus connus de la BOVIA, le General Problem Solver. Tout comme son "ancêtre", il cherche à modéliser une grande gamme de conduites mentales conçues comme des problèmes à résoudre. Il revendique une certaine plausibilité psychologique et sa conception s'enracine dans des observations de comportements de sujets lors de certaines tâches intellectuelles. De plus, il fait appel à des modes classiques de représentation des connaissances (espace des états et règles de production). Ajoutons qu'il encapsule également des mécanismes d'apprentissage. Ces caractéristiques lui donnent un caractère paradigmatique de la BOVIA qui justifie sa présentation dans le paragraphe qui suit.

SOAR permet la réalisation de systèmes susceptibles d'effectuer des tâches routinières comme des tâches complexes de résolution de problèmes, d'utiliser toutes les connaissances nécessaires à la réalisation de ces tâches et d'apprendre avec l'expérience. Il présente les principales caractéristiques suivantes.

Résolution de problèmes par des recherches dans des espaces d'états

Les tâches sont représentées comme des problèmes à résoudre. Un problème se caractérise par

un but à atteindre, un état initial du système et un ensemble d'opérateurs à utiliser. Résoudre le problème revient à cheminer d'un état à un autre (imaginez par exemple que vous passez d'une configuration d'un damier à une autre en déplaçant des pions) en appliquant des opérateurs jusqu'à ce que la solution soit atteinte.

Représentation uniforme des connaissances par des règles

Dans SOAR, toutes les connaissances spécifiques à une tâche sont exprimées sous forme de règles: si telle condition est satisfaite, alors exécuter telle action. Ce mode de représentation est très courant dans la BOVIA. Dans SOAR, les règles servent essentiellement à piloter les déplacements dans les différents espaces d'états. Elles ne représentent pas uniquement les connaissances qu'un expert formulerait spontanément sous forme de règles mais également des opérations élémentaires nécessaires à la conduite des tâches. Lorsque le programme joue au *taquin* par exemple, il peut utiliser des règles du type:

> *Si l'espace est 'taquin', alors créer une préférence 'acceptable' pour un état nouvellement créé et enrichir le nouvel état avec des liaisons qui permutent les pavés de l'état courant qui sont affectées par l'opérateur appliqué.*

> *Si l'espace est 'taquin' et s'il existe une préférence 'acceptable' pour un nouvel état, alors copier de l'ancien état toutes les liaisons qui ne sont pas affectées par l'opérateur appliqué.*

Il n'est pas nécessaire de comprendre à quoi correspondent les opérations prescrites par ces règles pour voir le rôle d'intendant qu'elles peuvent jouer dans le système et leur niveau de sophistication.

Contrôle exercé par des préférences générées par des règles

Les deux règles données en exemple ci-dessus font appel à la notion de préférences. Contrairement aux deux premières caractéristiques qui se rencontrent fréquemment dans les programmes de la BOVIA (espace des états et règles), l'utilisation de préférences lorsqu'il y a des conflits de règles (différentes opérations sont possibles, laquelle choisir ?) est tout à fait typique de SOAR. Ceci permet en qualifiant certains éléments - "à rejeter", "acceptable", "le meilleur" - de fournir des éléments au système pour l'aider à prendre des décisions. Le contrôle est donc piloté par des connaissances sur ce qui est désirable et ce qui l'est moins, générées par des règles.

Mécanisme unique d'apprentissage par "chunking"

SOAR possède également un mécanisme unique d'apprentissage appelé "chunking". Ceci lui permet de tirer parti, en créant une nouvelle règle, de la manière dont il a réussi à sortir des impasses face auxquelles il s'est trouvé. Pour ce faire, durant la résolution d'un sous-

problème, il garde une trace des règles activées, il identifie les éléments, préexistants au sous-problème, qui ont permis d'activer ces règles, et à partir de ces connaissances en fabrique une nouvelle. Ce mécanisme lui permet d'enrichir progressivement sa base d'informations, d'éviter les errements antérieurs et d'améliorer ses performances au cours du temps.
La figure 2, ci-dessous, représente l'organisation des différentes composantes de SOAR.

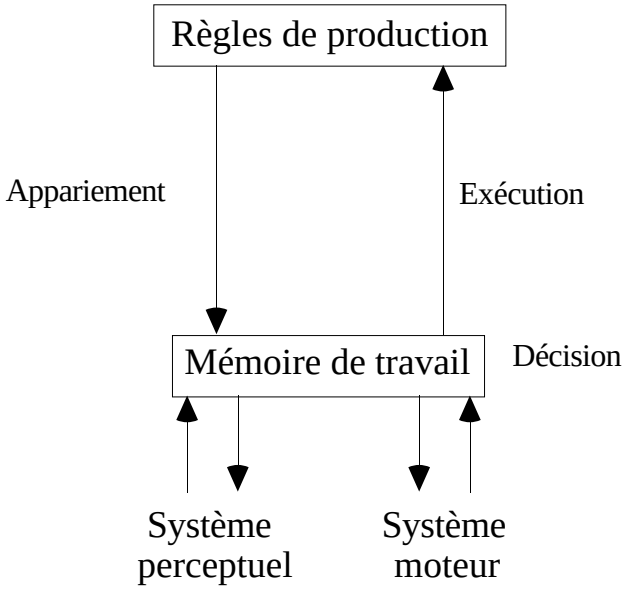

Figure 2 : organisation des différentes composantes de SOAR

SOAR fonctionne suivant quelques principes simples. Le problème à résoudre est représenté en mémoire de travail par un certain nombre de caractéristiques : le but, l'état actuel etc. Voici un exemple du type de représentation utilisée par SOAR.

(contexte G1 §espace P1 §état S1 §opérateur non défini)
(but G1 §objectif D1)
(objectif D1 §liaison DB1 §liaison DB2 ...)
(espace P1 §nom taquin)
(état S1§liaison B1 B2 B3 ...)
(liaison B1 §cellule C1 §pavé T1 ...)
(préférence §objet O1 §role opérateur §valeur acceptable §espace P1 §état S1)

Un cycle d'exécution commence par une phase d'élaboration : toutes les règles de production applicables (c'est-à-dire dont les parties condition sont satisfaites) sont activées. La phase de décision permet, lorsque l'élaboration est terminée, d'examiner la pile des problèmes en cours et d'identifier celui pour lequel une action est possible (application d'un opérateur, par exemple). Pour ce faire, il se sert des préférences. Si aucune action ne s'avère possible, SOAR est dans une impasse et crée dans la mémoire de travail un nouveau sous-problème. Sa résolution ultérieure donne lieu à la création de nouvelles règles.

SOAR a permis d'automatiser des tâches de nature très différentes.

Le système couvre de nombreux domaines permettant d'automatiser des :

- Petites tâches logiques, comme le problème des missionnaires et des cannibales, le taquin, la tour d'Hanoi, et des résolutions de problèmes dans les micromondes familiers à la BOVIA comme le monde de blocs de différentes formes, tailles et couleurs utilisé au MIT.
- Tâches de routine, comme la résolution de syllogismes, l'unification d'expressions formelles en logique, l'extrapolation de séquences numériques.
- Tâches exigeant des connaissances, comme la recherche de configurations d'ordinateurs répondant à certaines caractéristiques, diagnostic (comme le logiciel bien connu Mycin) de certaines maladies infectieuses particulières comme les méningites.
- Tâches diverses qui ont été étudiées par la BOVIA au moyen d'autres programmes comme la formation de concepts, l'analyse syntaxique, la résolution de théorèmes.

Pour effectuer ces tâches, des connaissances spécifiques doivent bien entendu être formalisées et encodées de manière appropriées. SOAR est capable d'utiliser également des méthodes de résolution dites faibles (exploration systématique et aveugle de l'espace des problèmes par exemple) qui ont un caractère universel et qui ne sont généralement mises en oeuvre que lorsque les connaissances spécifiques se révèlent insuffisantes.

3.1.5. Des problèmes mis en évidence et leur pertinence en psychologie

L'utilisation de ce type de modèle a fait apparaître de nombreux problèmes jusqu'alors ignorés ou peu traités; nous en reprendrons ici trois que nous avons déjà évoqués. Le premier est lié à la notion de contrôle; la manière précise dont les différentes opérations mentales peuvent ou doivent s'enchaîner a été étudiée avec beaucoup de précision en IA et a sûrement contribué à mettre en évidence des modèles, des alternatives qui n'apparaissaient pas de manière aussi explicite auparavant. Le deuxième concerne la complexité des espaces à parcourir pour résoudre les problèmes, l'explosion combinatoire des états à considérer. Ceci débouche sur la notion de tractabilité pour caractériser la complexité des algorithmes de recherche à mettre en oeuvre. Enfin, la nécessité d'un recours continuel au contexte, aux connaissances acquises pour appréhender de nouvelles informations ou simplement pour mettre à jour les anciennes a également été soulignée en IA. Ce problème est généralement appelé en anglais le *frame problem* [GPS, SOAR, A*, etc.].

Le contrôle

Vous êtes en voiture et vous prenez comme d'habitude la route qui vous mène au travail alors que cette fois vous vous rendez autre part..., vous vous retrouvez dans votre salle de bains en train de vous brosser les dents alors que vous y alliez pour vous donner un coup de

peigne..., dans une conférence internationale où la plupart des participants parlent anglais, vous vous entretenez en anglais avec un collègue francophone. Ces quelques exemples constituent ce qu'on appelle des erreurs de routine, elles sont dues à une certaine forme d'automatisation du comportement et sont liées à la logique et la mécanique d'enchaînement des opérations.

Cette mécanique a été soigneusement étudiée en IA et relève du contrôle. Qu'est ce qui vient après ? Comment vérifier que telle opération est applicable ? Lorsque différentes opérations conflictuelles sont envisageables y a t-il un mécanisme d'arbitrage ? De quelle nature est-il ? Est-il possible de considérer différentes options simultanément et si la réponse est oui, comment et dans quelles circonstances ? Comment faire marche arrière lorsqu'on est dans une impasse ? Pourquoi ne se retrouve-t-on pas toujours dans les mêmes impasses ? Si on recommence à travailler sur le même problème enchaînera-t-on les opérations de manière identique ? Voici quelques questions que l'IA a dû affronter et auxquelles différentes réponses ont été apportées. Des modèles ou plutôt des mécanismes ont été proposés: gestion de piles d'opérations dans les méthodes de résolution de problèmes dites faibles, chaînage avant ou arrière de règles de production, techniques sophistiquées d'appariement pour identifier les règles qui peuvent être activées, méthodes multiples de résolution de conflits lorsque plusieurs règles peuvent être simultanément activées (utilisation de préférences, par exemple dans SOAR), enchaînement probabiliste d'opérations auxquelles on a associé des niveaux de priorité qui définissent leur tendance a être appliquées, examen en parallèle de différentes démarches, modulation du niveau de vigilance du système et de son niveau “d'indéterminisme” au moyen d'un concept de température etc.

De nombreuses théories psychologiques paraissent avoir simplifié si pas dénaturé le problème du contrôle. Il est, soit renvoyé à un mécanisme non décrit, soit ramené à un enchaînement plus ou moins aléatoire, soit encapsulé dans une boîte noire, un homonculus tout puissant au coeur du système. La théorie de la mémoire de Baddeley (1986) est un exemple de ce problème. Elle postule une mémoire de travail composée de deux systèmes coordonnés par un administrateur central : une boucle articulatoire qui contient une information phonologique stockée sous forme sérielle et temporelle comme une bande magnétique et un bloc notes visuo-spatial qui, comme son nom l'indique, contient des informations visuelles et spatiales. Si vous tentez de simuler cette théorie sur ordinateur, la programmation de l'administrateur central pose problème. Son rôle de chef d'orchestre est trop peu explicite, de plus il paraît pourvu de capacités qui vont bien au-delà de la simple coordination.

La tractabilité

Qu’est-ce que l’intractabilité ? Pour mieux comprendre cette difficulté, une des premières auxquelles l’intelligence artificielle a été confrontée, considérons une technique utilisée par le programme d’analogie ACME (Holyoak et Thagard, 1989) pour trouver le meilleur ensemble de correspondances entre deux situations. Imaginons, par exemple, un parallèle entre le système solaire et l’atome de Rutherford. On commence par donner au

programme un découpage en attributs de chaque situation (d'un côté, soleil, planètes, ... de l'autre, neutrons, protons, électrons, noyaux, etc.). On crée ensuite un réseau dont chaque noeud consiste en une correspondance entre un des attributs de la première situation et un de la deuxième. Si deux situations plus complexes sont décrites au moyen de, disons, 30 attributs chacune, 900 correspondances sont théoriquement possibles, et donc un réseau de 900 noeuds. Sans rentrer dans le détail du fonctionnement de ACME, il suffit de dire que ce programme passera en revue les 900 correspondances afin de déterminer celles qui feraient la meilleure analogie entre les deux situations.

Mais parfois plusieurs attributs d'une situation peuvent correspondre à un seul attribut d'une autre. Si l'on voulait mettre en correspondance le soleil avec le noyau d'un atome, il serait raisonnable de mettre en correspondance non seulement chacun des neutrons et des protons du noyau avec le soleil, mais aussi l'ensemble des neutrons et protons à l'unique soleil de notre système solaire. Autrement dit, dans la logique d'ACME, lorsqu'on crée notre réseau, on doit mettre non seulement chaque attribut de la première situation en correspondance avec chaque attribut de la deuxième situation, mais aussi *chaque regroupement théoriquement possible* d'attributs de la première situation avec *chaque regroupement théoriquement possible* d'attributs de la deuxième situation. Du coup, il y a une véritable explosion — appelée *explosion combinatoire* — du nombre de correspondances à analyser. Maintenant, le programme aura besoin de traiter non pas 900 correspondances, mais plus de 1.000.000.000.000.000.000 (!) et ceci simplement pour trouver une bonne analogie entre deux situations dont les descriptions se limitaient à 30 attributs chacune. Même une machine capable d'analyser 250 millions de correspondances par seconde (la vitesse maximale de Deep Blue, le meilleur programme d'échecs à l'heure actuelle) mettrait environ 125 ans pour analyser toutes ces correspondances. Imaginons maintenant la difficulté de ce programme devant deux situations réelles ayant chacune non pas douze attributs mais des centaines, voire des milliers. Aucune machine, aussi puissante soit-elle, ne pourrait traiter de telles situations à la manière d'ACME. Voici donc un grand défi de l'IA : éviter le problème de l'explosion combinatoire.

Un problème est dit "traitable" si sa solution peut être découverte en un nombre fini et raisonnable d'étapes. Les algorithmes et, plus particulièrement, les problèmes sont évalués en fonction de leur degré de complexité. Celui-ci se mesure en terme de ressources de calcul utilisées (temps et mémoire) pour résoudre un problème. L'ordre de complexité décrit comment le temps et l'espace mémoire nécessaire à l'accomplissement d'une tâche augmentent en fonction de la quantité de données à traiter.

Sur base des algorithmes connus, on classe les problèmes en deux groupes (voir Garey et Johnson, 1979) :

- les problèmes traitables, qui peuvent se résoudre au maximum en un temps qui augmente comme un polynôme du nombre de données à traiter;
- les problèmes non traitables pour lesquels aucun algorithme pouvant trouver une solution en un temps polynomial n'est connu.

Ce problème est interpellant pour la psychologie et invalide certaines théories. Prenons par exemple les théories du raisonnement humain.

La théorie des modèles mentaux de Johnson-Laird (1983) explique le raisonnement humain par un processus de création de modèles suivi d'une évaluation par une recherche de contre-exemples dans la base de connaissances. Comment cette recherche va-t-elle éviter l'explosion combinatoire lorsque la taille de la base de connaissance augmente ? Les êtres humains ne semblent pourtant pas handicapés dans leurs raisonnements lorsque leur base de connaissance s'enrichit. Le mécanisme de recherche de contre-exemples n'est pas spécifié dans la théorie. C'est sur ce point que cette théorie est limitée. Pour être convainquante, une théorie doit prouver que les phénomènes qu'elle explique restent “tractables”. Le lecteur intéressé trouvera une excellente discussion dans Oaksford et Chater (1995).

Le frame problem

Les modèles utilisés dans la BOVIA utilise souvent des représentations et des modes de fonctionnement tirés de la logique. Leur mise en oeuvre nécessite souvent des processus de recherche guettés par l'explosion combinatoire comme nous venons de le montrer. Une deuxième difficulté, appelée le “frame problem”, est liée à la lourdeur des mises à jour des bases de connaissances dans une telle approche. Tout nouvel état de ces bases requiert en effet a priori une réévaluation de l'entièreté des faits connus en vue de déterminer ce qui peut être encore considéré comme exact. L’adjonction d’une information par exemple modifie la base de connaissances et ainsi est soumise au “frame problem.”

Imaginez que tous les oiseaux que vous ayiez vus volent. Il est logique de supposer que dans votre “base de connaissances” vous avez inféré une proposition du type, “Tous les oiseaux volent.” Un jour un ami vous parle des autruches, des oiseaux qui ne volent pas. Ceci vous oblige à corriger votre base de connaissances. Vous inférez qu'un oiseau vole sauf si c'est une autruche. Vous venez d'effectuer un raisonnement non monotone. L’ajout d’un nouveau fait vous a obligé à annuler des inférences précédentes. Remarquez que cette suppression requiert également la réévaluation complète des faits connus puisque certains sont peut-être dérivés de “tous les oiseaux volent.” Quand une base de connaissances est composée de milliers de faits et que l’expérience vous confronte à des milliers d’expériences nouvelles, imaginez l’ampleur de la tâche. Ici encore la pertinence pour la psychologie est réelle.

Une théorie cognitive devrait pouvoir expliquer comment, à partir d’une base de connaissances riche, toute nouvelle donnée peut être incorporée sans générer une réévaluation de l'entièreté de la base de connaissances. Une réévaluation complète est intraitable au vu de l'importance de la base de connaissances d'un être humain. Cette question est rarement abordée en psychologie, mais apparaîtra immanquablement lorsque l'on désire modéliser un raisonnement non monotone à partir d'une base de connaissances symbolique sur ordinateur.

3.1.6. Les limites de la BOVIA

La BOVIA a suscité beaucoup d'espoirs et beaucoup de désillusions. Différentes critiques ont été formulées à l'égard des systèmes dont nous venons de parler, elles peuvent se grouper en cinq classes que nous allons brièvement présenter. Alors que pour certains chercheurs, l'IA est définie comme le domaine de construction de systèmes physiques de symboles et rejettent toute autre vue, d'autres se montrent critiques envers la BOVIA. Ils posent différentes questions. L'intelligence est-elle dans l'esprit du programmeur ou dans le programme ? Le symbole est-il l'unité élémentaire de la cognition ? Le symbole est-il nécessaire pour exhiber les conduites intelligentes ? Comment accéder à la signification en manipulant des symboles ? Les performances des systèmes de la BOVIA sont-ils en rapport avec celles des êtres humains ?

L'argument de Popper

Popper écrit "Donner une définition précise d'une opération est déjà un schéma de programme... L'affirmation selon laquelle un tel programme peut être mis en oeuvre par une machine est alors presque tautologique... Le problème est de savoir si toutes les opérations de la pensée peuvent être décrites... dans un langage précis pouvant être traduit en algorithmes." (Voir, par exemple, le développement de cet argument par Ladrière dans l'Encyclopedia Universalis, 1991.) Autrement dit, le chercheur introduit dans son programme tout ce qui est nécessaire à l'accomplissement d'une tâche, et ce faisant la définit implicitement. Le problème est de savoir si toutes les tâches sont ainsi programmables.

Le rêve booléen

L'utilisation du mot symbole dans la formulation de l'hypothèse de Newell et Simon (1976) a créé certains malentendus et a soulevé de nombreuses objections. L'acception commune du mot fait en effet référence à une capacité de représentation, d'association à autre chose. Et les systèmes qui se revendiquent de l'hypothèse des systèmes physiques de symboles utilisent généralement cette acception. Néanmoins, dans la formulation originale de Newell et Simon, les symboles ont un rôle essentiellement instrumental qui va peut-être au-delà de la signification classique de ce concept. Le rôle central attribué aux symboles - ils constituent les atomes du système - pose cependant problème à certains. Comme déjà signalé, les symboles appartiennent au langage commun, ils font référence aux concepts que la pensée manipule quotidiennement. La formulation de Newell et Simon paraît ramener l'intelligence à la maîtrise d'un langage adéquat pour organiser et manipuler ces entités conscientes et atomiques. Cette approche que l'on a qualifiée de "paradigme symbolique" est opposée au "paradigme sous-symbolique" qui conçoit les symboles comme des émergences, des épiphénomènes d'activités plus élémentaires (Boden, 1991b). Les unités de base dans cette approche ne correspondent plus nécessairement à nos mots, à nos expressions, à nos phrases. Ils ne représentent plus les

objets de nos pensées mais plutôt des micro-caractéristiques dont la présence massive provoque l'émergence de ce qui est verbalisé. L'intelligence apparaît ici comme une propriété statistique du système. Les tenants de la théorie émergente accusent les “traditionalistes” de s'être débarrassés du cerveau trop vite. Il est un peu court de le concevoir comme une merveilleuse machine à calculer sur les symboles, disent-ils et puis de se pencher sur d'autres machines, artificielles cette fois ! De manière plus articulée, disons qu'ils remettent en cause à la fois les caractères suffisant et nécessaire de l'hypothèse des systèmes physiques de symboles.

Le peu de succès de la BOVIA dans la modélisation du sens commun, les difficultés à modéliser des choses aussi élémentaires que la reconnaissance en une fraction de seconde d'un style particulier, d'une scène familière, l'incapacité de s'accommoder de bruits, de perturbations dans les données sont souvent mis en exergue pour souligner les insuffisances de l'approche symbolique.

De même, il ne paraît pas évident comme l'affirment Newell et Simon (ref) que la pensée humaine puisse se ramener à du calcul symbolique. Pour certaines opérations mentales, les hommes paraissent fonctionner sur le mode des programmes informatiques en enchaînant des conduites mentales, en manipulant des connaissances. En résolvant certains problèmes, on paraît appliquer, par exemple, de manière séquentielle des opérations à des configurations, à des états d'un système. Mais dans la plupart des situations l'homme opère sous un autre mode, plus rapide, inconscient, moins régi par des règles. La rapidité de la reconnaissance des objets d’une scène ne porte pas à croire en la nécessité d’un enchaînement ou d’une manipulation de symboles. Toute perception de la similarité entre une scène actuelle et une autre préalablement rencontrée relève de l’analogie (voir le paragraphe sur les réseaux glissants et la représentation des concepts). Les premières étapes de la pensée analogique ressemble plus à un glissement non rationnel qu’à un processus régi par des règles. Simon (1980, cité dans Hofstadter, 1988, p. 664) disait, “Tout ce qui présente un intérêt dans le domaine de la cognition se produit au-delà du seuil des 100 millisecondes - le temps qu'il vous faut pour reconnaître votre mère.” Hofstadter (1988) répondait, “Tout ce qui présente un intérêt dans le domaine de la cognition se produit en deçà du seuil des 100 millisecondes- le temps qu'il vous faut pour reconnaître votre mère”. Une partie de l'activité mentale ne se formaliserait donc pas nécessairement en termes de manipulation de symboles.

L’impérialisme des représentations

Le traitement d’expressions symboliques est-il nécessaire pour réaliser des actions intelligentes comme affirmé dans l'hypothèse des systèmes physiques de symboles ? La relation entre senseurs et effecteurs passe-t-elle toujours par des représentations ? Certains proposent de répondre “non” à ces questions et de mettre l'accent sur notre aptitude à évoluer dans un environnement complexe, à percevoir les signaux qu'il nous envoie et à y répondre de manière appropriée. Pour reprendre notre logique de présentation, ils privilégient un nouveau type de contrainte que nous avons baptisée “agir sur le milieu”. Elle fait l'objet d'un paragraphe

ultérieur (Section 3.2.3).

Comment sortir de la syntaxe ?

"Comment peut-on en manipulant uniquement des expressions symboliques avoir accès à des significations ?" demandent également certains. Et cette interrogation remet en cause toute approche syntaxique de l'intelligence. S'il est vrai qu'il est difficile de comprendre comment un système uniquement capable de traiter des chaînes de “0” et de “1” peut exhiber une intelligence comparable à la nôtre, il est également troublant de remarquer qu'un cerveau biologique qui opère également des transformations sur des signaux peut lui penser. Le cerveau transforme les signaux chimiques et physiques grâce à un processus biologique. L’ordinateur transforme des signaux physiques par des processus physiques. Pour certains, comme Searle (1980), la nature biologique des processus de traitement syntaxique du cerveau serait capitale.

Rigidité de l’approche symbolique

La BOVIA a clairement montré ses limites dans certaines tâches où d'autres systèmes se sont montrés plus performants : production de généralisations, dégradation progressive des performances, accès aux informations à partir de descriptions partielles. Alors que les systèmes de la BOVIA se montrent peu efficaces dans ces domaines, l’homme lui y excelle. De même, les systèmes symboliques ont une capacité rationnelle exacerbée alors que l'être humain est victime de biais systématiques (voir le chapitre sur les processus intellectuels).

3.2. Des contraintes alternatives ou complémentaires

3.2.1. L'activité mentale conçue comme une aptitude à associer des idées

Les origines du connexionnisme

Le connexionnisme intègre le courant associationiste et la Gestalt. Les premières idées associationistes remontent peut-être à Aristote dans son "De memoria et Riminescentia”. Pour Aristote, les unités élémentaires sont les images sensitives, il les appelle mémoires. Les associations et les liens entre mémoires servent de base à la pensée, il les appelle réminiscences. Les associations d’Aristote s’opèrent grâce à la succession, la similarité, l’opposition et le voisinage. Ceci est très proche de l’idée des réseaux sémantiques (voir le paragraphe sur les réseaux glissants et la représentation des concepts). Ajoutons que pour Aristote la cause des mémoires et réminiscences est physiologique.

L'associationnisme considère le cerveau comme un central téléphonique d’où part une multitude de câbles. Le rôle de ce central est de coupler les “câbles” venant de la perception aux câbles allant vers les zones motrices, le rôle de l’apprentissage étant de fixer ces couplages. Les cellules du système sensoriel acquièrent leurs connections avec les cellules du système moteur. Trois problèmes surgissent de cette hypothèse :

- le cerveau ne devrait être actif que lorsque cela est nécessaire (or on sait que l'activité du cerveau n'est jamais arrêtée que par la mort),
- toute lésion dans le cerveau devrait avoir des effets très spécifiques sur un souvenir ou une habileté (ce qui n'est pas le cas),
- un objet ne pourrait pas être reconnu, une porte par exemple, quelle que soit l'angle sous lequel il est perçu (problème de l'équipotentialité).

La première tentative de rompre avec cette idée est celle de William James (1890). Pour cet auteur, l'association est une fonction du cortex régi par un principe élémentaire. Lorsque deux processus cérébraux sont actifs simultanément ou se succèdent immédiatement, l'activité de l'un tendra à activer l'autre. Ceci annonce l'apprentissage Hebbien (Hebb, 1949). De plus, James préfigure l'idée de distribution et d'inhibition dans la règle de sommation : "la quantité d'activation en un point du cerveau est la somme de tous le points déchargeant sur lui". L'activation est proportionnelle au nombre de fois où les deux points ont été activés simultanément, à l'intensité de l'excitation et à l'absence de points rivaux.

A l'associationnisme et son hypothèse du central téléphonique, s'oppose la Gestalt. La Gestalt décrit la forme comme l'organisation des parties et cette organisation est différente de la somme des parties (c'est ce qui la différenciait de l'associationnisme). Pour la théorie du champ, le cortex est constitué de tellement de cellules qu'il peut être considéré comme un médium statistique homogène. Le contrôle des centres moteurs dépend de la distribution des excitations sensorielles et des rapports entre ces excitations et non du lieu ou de l'action de cellules spécifiques. La représentation n'est ni dans les connexions, ni dans les localisations, mais dans les formes variables d'excitations affectant n'importe quelle cellule cérébrale. Le cerveau agit comme un tout. Les traces mnésiques sont dynamiques et non structurelles (voir Lashley, 1950).

Hebb (1949) propose une théorie tirant parti des idées de la Gestalt tout en y incluant un mécanisme d'apprentissage. Tout d'abord, Hebb est entièrement opposé au behaviorisme pour lequel l'explication psychologique réside dans la relation entre stimuli et réponse observée. Le behaviorisme rejette tout recours à la physiologie à des fins d'explication psychologique sous prétexte que la physiologie avait peu apporté à la psychologie. La position de Hebb est que tant que la neurophysiologie ne peut expliquer les transmissions corticales et que la psychologie ne peut étudier la pensée, on aura des difficultés à expliquer des données expérimentales et cliniques. Ces deux études sont rejetées par le behaviorisme. Pour Hebb l'objet de la psychologie est d'étudier ce qui se passe entre le stimulus et la réponse ou entre les voies afférentes et les voies efférentes c'est à dire le processus central de contrôle.

Hebb rejette aussi bien la vision associationniste que la théorie du champ en vertu du fait qu'aucune de ces théories n'est capable d'expliquer les délais variables entre stimulus et réponse. Mais, en même temps, il va intégrer les idées de localisation de l'associationnisme, les propriétés de distribution et de dynamisme de la théorie du champ tout en incluant un mécanisme d'apprentissage auquel on se réfère encore à l'heure actuel sous le nom

d'apprentissage Hebbien. Il s'énonce comme suit : quand un axone d'une cellule A excite une cellule B et que de manière répétée et persistante il prend part à son déclenchement, un processus de croissance ou un changement métabolique survient dans l'une ou les deux cellules de telle façon que l'efficacité de A, en tant que cellule provoquant la décharge de B, est augmentée.

Hebb postule la notion de "cell assemblies" (assemblée de cellules), structures comprenant des cellules pouvant être anatomiquement éloignées. Ces ensembles sont capables d'agir de manière brève comme un système clos et d'activer d'autres systèmes. Une série de ces activations successives de différents systèmes constitue une "phase séquence", le processus de pensée. Les cellules assemblées dans un système ont des propriétés temporelles précises. La perturbation temporaire de ces propriétés conduit aux dysfonctionnements émotionnels, alors que la perturbation chronique mène aux névroses ou aux psychoses. Une fois active une "cell assembly" se maintient pendant un certain temps par un mécanisme de réverbération pour constituer une image ou une idée. L'équipotentialité, le fait que l'on considère un objet vu de points de vue différents comme étant le même objet, est obtenue grâce à des voies alternatives permettant à la même assemblée de cellules d'être active.

L'oeuvre de Hebb (1949) est centrale dans le connexionnisme actuel. Bien que les premiers modèles n'intégraient que les propriétés de localisation et de distribution, les modèles actuels tels que les réseaux récurrents et les réseaux à assignation par synchronies temporelles intègrent les propriétés dynamiques.

Comme on a vu, une des critiques de la BOVIA est fondée sur la non suffisance de l'unité élémentaire, le symbole. Le connexionnisme va utiliser des unités encore plus élémentaires, dépourvues de toute sémantique et va s'inspirer du système nerveux. Si les premières réalisations de ce type de système connexionniste se sont montrées réductionnistes, les recherches à l'heure actuelle se rapprochent de plus en plus du fonctionnement réel du système nerveux. L'idée centrale du connexionnisme est que les activités mentales sont le résultat de l'activité parallèle d'unités élémentaires interconnectées. Les unités dont le réseau est composé (ou noeuds) reçoivent de l'activation ou de l'inhibition d'un grand nombre d'autres unités et en envoient à beaucoup d'autres. Ce qui sépare les réseaux connexionnistes des systèmes symboliques sont l'utilisation d'unités élémentaires plus simples que le symbole, le parallélisme (un grand nombre d'unités sont simultanément actives), et la distribution (une image, un symbole, ... sont représentés par un ensemble important d'unités).

Le perceptron

Une des premières tentatives de réalisation de réseaux de neurones artificiels sur ordinateur revient à Rosenblatt (1958). Dès la fin des années 50, il en proposait une, appelée le perceptron. Ce dispositif tentait de reproduire notre capacité

- à apprendre à associer des réponses adéquates aux stimuli auxquels nous sommes confrontés
- et à pouvoir fonctionner en présence d'informations dégradées.

Le premier but et l'architecture de ce perceptron (voir ci-dessous) pourraient amener le lecteur à associer ces premiers pas du connexionnisme au behaviorisme. Si le perceptron partage avec le behaviorisme une association étroite entre stimulus (= entrée du perceptron) et réponse (= sortie), il s'en distingue néanmoins par sa référence à la neurophysiologie et son aspect distribué qui le rapproche de la Gestalt (Lashley, 1950), de l'oeuvre de Hebb (1949) et, en fin de compte, des travaux de James (1890).

La conception du perceptron s'inspire explicitement du neurone. Il s'agit d'un dispositif qui, dans sa forme la plus simple, fonctionne comme suit. Une unité de perceptron est équivalente à un noeud de Mc Culloch et Pitts (1943), elle reçoit des excitations venant de différentes sources (des unités d'entrée), et en fait une somme pondérée. Sur la figure 3, l'entrée I d'un noeud est :

$$I = w_1 x_1 + w_2 x_2 \qquad (1)$$

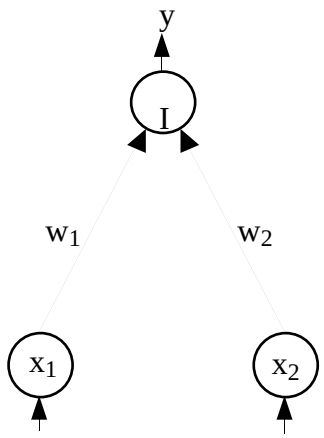

Figure 3 : le perceptron

Ensuite ce noeud répond positivement ou négativement suivant que cette somme est supérieure ou inférieure à un seuil donné T. La figure 4 représente graphiquement cette fonction de sortie dont la valeur est déterminée comme suit :

$$y = \begin{pmatrix} +1, & si \;\; I \geq T \\ -1, & si \;\; I < T \end{pmatrix} \qquad (2)$$

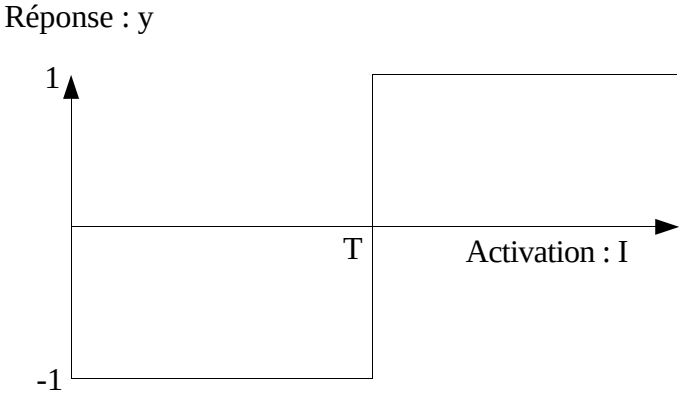

Figure 4 : la fonction de sortie d'un noeud de perceptron

Rosenblatt (1958) propose un algorithme permettant d'entraîner un tel réseau de neurones en ajustant les poids w_i. Il crée alors le premier réseau de neurones apprenant. La fonction d'apprentissage permet de modifier le poids des liens entre unité d'entrée et de sortie comme suit :

$$w_{nouveau} = w_{ancien} + ß \; y \; x \tag{3}$$

où $ß = \begin{pmatrix} +1, \; si \; la \; réponse \; est \; correcte \\ -1, \; si \; la \; réponse \; est \; erronée \end{pmatrix}$

et y est la réponse du perceptron.

De manière plus concrète, on peut imaginer, par exemple, que les différentes excitations afférentes sont liées à l'intensité lumineuse avec laquelle certains points de la “rétine” sont éclairés à partir d'un objet donné. La réponse du perceptron doit être positive ou négative suivant que l'objet est d'un certain type ou non.

Donnons un exemple de mode de fonctionnement possible. Le perceptron qui nous intéresse ici doit pouvoir reconnaître une lettre “X” à un endroit donné de la “rétine”. La figure 5 représente notre perceptron.

Chaque unité d'entrée reçoit une activation x_i - il y en a 25 dans l'exemple donné. Supposons que ces entrées correspondent à certains points situés sur la rétine. Si un point est noir, x_i prendra la valeur 1, si un point est blanc, il prendra la valeur -1. Les activations x_i vont être transmises à une unité de sortie via un lien de poids w_i. La valeur d'excitation I de l'unité de sortie sera la somme du produit des activations x_i et des poids des liens w_i. Si cette somme dépasse le seuil T, l'unité de sortie répondra positivement.

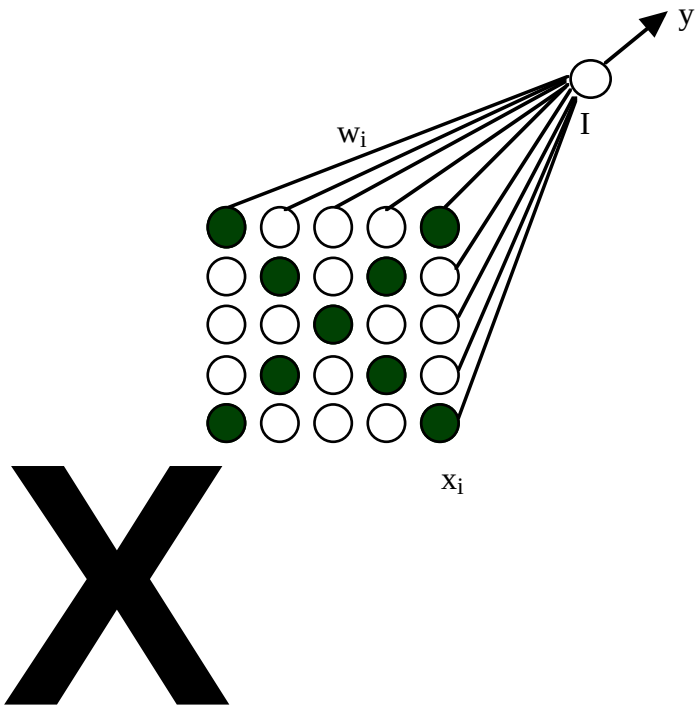

Figure 5 : reconnaissance de la lettre X par un perceptron

En cours d'apprentissage les poids des connexions vont être modifiés. Prenons un exemple, l'unité de sortie a répondu négativement alors que le contraire était désiré (ß = -1). L'algorithme d'apprentissage va modifier une connexion dont la valeur est, par exemple -0.2 (w_{ancien} = -0.2), connexion liant un point d'entrée noir (dont la valeur d'activation est x=1) à l'unité de sortie dont la réponse est -1 (y = -1) comme suit :

$w_{nouveau} = -0.2 + (-1 * -1 * 1) = 0.8$

Pour une connexion de même poids reliant cette fois un point blanc à l'unité de sortie, la valeur serait devenue -1.2. Après quelques essais le perceptron est sensible au stimulus et parvient à le catégoriser de manière efficace.

Dans ce type d'architecture, la *reconnaissance* d'objets et l'apprentissage des associations entre objets sont primordiaux. L*'association d'événements* (et d'objets), plutôt que la résolution de problèmes, est considérée comme un mécanisme de base de la cognition.

L'espoir secret était d'être capable, en interconnectant un nombre suffisant d'éléments simples de ce type, de reproduire les performances du cerveau. On a réussi à montrer que les perceptrons étaient capables d'apprendre à reconnaître des formes en modifiant progressivement les poids des liens. Il était donc permis d'imaginer des systèmes qui se configurent spontanément en fonction des tâches à réaliser, les rétroactions du milieu guidant le système dans ses apprentissages. Malheureusement, une étude mathématique des perceptrons a mis en évidence leurs limites, du moins dans leur version la plus simple (Minsky et Papert, 1969). Le célèbre problème d'XOR - il n'existe aucune combinaison de w_1, w_2, et seuil T (en général, T = 0) tel qu'un perceptron puisse produire la fonction suivante:

En entrée	En sortie
0 0	1
0 1	0
1 0	0
1 1	1

a ruiné les prétentions initiales.

Pour répondre aux critiques de Minsky et Papert, le perceptron multi-couche (PMC) a été développé.

Les PMCs

Le perceptron multi-couches (PMC) diffère du perceptron essentiellement par l'ajout d'une troisième couche de noeuds entre la couche d'entrée et la couche de sortie (Figure 6), appelée la couche cachée. De plus, l'utilisation d'une fonction sigmoïdale (Figure 7) au lieu

d'une simple fonction de seuil pour déterminer l'activation de sortie de chaque noeud a permis le développement d'un nouvel algorithme de changement des poids dans ces réseaux, un algorithme qui s'appelle la rétro-propagation ("backpropagation" en anglais). Actuellement ce type de réseau et ses variantes dominent très largement les recherches en connexionnisme. Il connaît un champ d'application important, allant de la reconnaissance des visages à la conduite automatique des voitures.

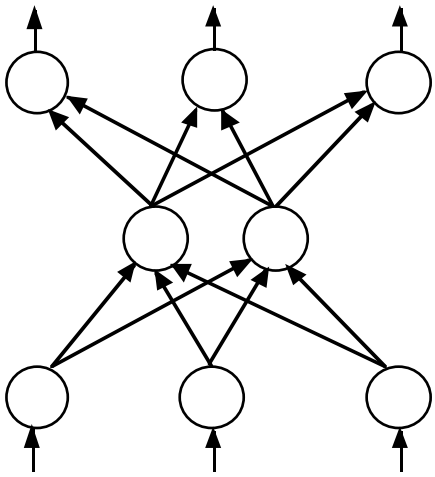

Figure 6 : le perceptron multi-couches

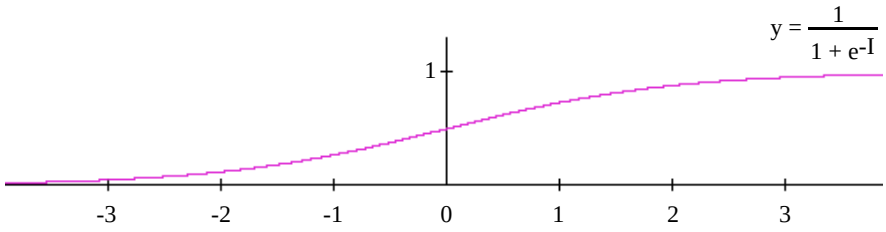

Figure 7 : La fonction sigmoïdale

Un exemple célèbre d'un PMC : NETtalk

L'un des premiers grands succès de la retropropagation fut le programme NETtalk, écrit par Sejnowski et Rosenberg (1987), visant à prononcer les mots écrits. Ce programme comprend 29 unités en entrée, chacune correspondant à une lettre d'un alphabet étendu et 55 unités en sortie, correspondant à 55 phonèmes prononcés par un synthétiseur. Le but du réseau est d'associer une séquence de lettres à une suite de phonèmes en tenant compte du contexte dans lequel chacune des lettres se trouvent. Par exemple, pour prononcer correctement la lettre "s", il faut tenir compte des lettres qui l'entourent. Au début d'un mot, comme dans le mot "sentir", le "s" ne se prononce pas de la même manière que dans le mot "peser," et encore moins dans le mot "erreurs." NETtalk est capable de faire ces distinctions.

Ce réseau commence par apprendre à prononcer un ensemble des mille mots les plus couramment utilisés dans la langue anglaise. La première phase passe par des balbutiements

qui se transforment peu à peu en une prononciation approximative pour parvenir à un taux de 91% de prononciation correcte sur l'ensemble des mille mots appris. Pour tester ses capacités de généralisation, le réseau doit prononcer mille nouveaux mots. Il réussit dans 80% des cas. Cette capacité de généraliser s'accroît en fonction de l'importance de l'ensemble d'apprentissage.

Ce qui frappa la communauté scientifique de l'époque était que ce réseau parvenait à prononcer et à généraliser correctement sans connaissances explicites des règles de prononciation.

Les réseaux récurrents

Ces réseaux de type rétropropagation associent uniquement des données en entrée à des données en sortie. En entrée, il pourrait par exemple y avoir un visage, en sortie un nom; ou en entrée, un mot, en sortie, la prononciation du mot, etc. Mais ces réseaux ne sont pas capables de prendre en compte le contexte dans lequel se situent les données. L'apprentissage de séquences d'événements où un même événement suit tantôt un événement E1, tantôt un événement E2 peut, ainsi, poser des difficultés. Par exemple, les réseaux rétropropagation peuvent apprendre sans difficulté la séquence :

A-B-C-D-E-F-G-H-I-J-K (1)

en apprenant les associations A-B, B-C, C-D, D-E, . . ., J-K. Le réseau reproduit la séquence en recevant un "A" en entrée, produisant puis récupérant un "B" en sortie, qui est ensuite envoyé en entrée pour produire un "C", etc. (Ce type de système porte le nom de " boite à musique ". Par contre, une séquence du type:

A-B-C-D-E-C-F-G-H-I-J-K (2)

serait impossible à apprendre car la lettre "C" est d'abord suivie d'un "D", et puis d'un "F". La solution requiert la prise en compte du contexte du premier "C" (précédé d'un "B") et de le différencier du deuxième "C" (précédé d'un "E").

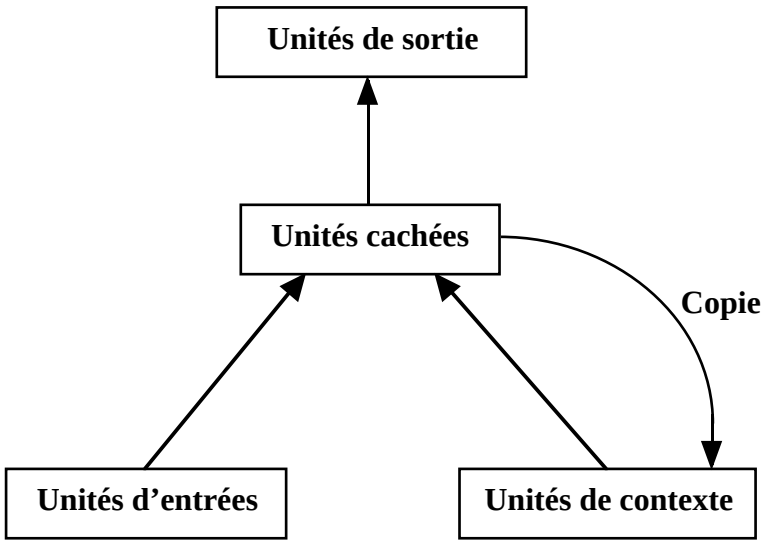

Figure 8 : le reseau récurrent de type Elman

Elman (1990) propose un réseau qui injecte en entrée une copie des activations de la couche cachée provenant de l'étape précédente (figure 8). De cette manière, le réseau apprend le contexte de la donnée qu'il est en train de traiter. Ce type de réseau est capable d'apprendre la séquence (2). Ainsi, lorsque " C " est présenté au réseau, si une copie de la couche cachée générée par " B " est également présentée, le réseau répondra " D ". Par contre, s'il s'agit d'une copie de la couche cachée générée par " E ", le réseau répondra " F ".

Le problème de l'affectation des variables

Comment représenter dans un réseau connexioniste la simple description : "Une rose rouge sur une table verte", sachant que les concepts "rouge", "rose", "verte", et "table" ont une représentation dans le système .? Comment associer "rouge" à "rose" et "verte" à "table" tout en évitant d'associer "rouge" à "table" ou "verte" à "rose" ? Les systèmes connexionistes des années 80 n'en étaient pas capables sans donner lieu à une confusion. Ce problème est connu sous le nom de "problème de l'affectation" (binding problem). La première solution venant à l'esprit est d'augmenter d'une part le poids des liens entre la représentation de "rose" et la représentation de "rouge" et d'autre part le poids des liens entre la représentation de "table" et la représentation de "verte". Mais les noeuds participant à cette représentation doivent, le cas échéant, participer utilement à d'autres. Certains de ces noeuds devraient, par exemple, être réutilisés pour encoder "La rose jaune sur la table rouge". Une assignation dynamique paraît donc nécessaire. Ce problème a été décrit par Feldman (1982).

D'autres problèmes découlent du problème de l'assignation. McCarthy (1988) soulignait que la plupart des systèmes connexionistes de cette époque étaient incapables de représenter des prédicats ayant plus d'un argument. Comment utiliser le prédicat ACHETER (acheteur, objet-acheté) dans différents contextes ? Lorsque le système représente : "Marie a acheté un disque" il doit, d'une part éviter de confondre le sujet "Marie" et le complément "disque" et d'autre part, si un nouveau fait apparaît tel que : "Sacha a acheté un bonbon" il doit clairement les distinguer.

Le second problème découlant du problème de l'assignation dynamique est le problème de la systématicité. Dans un article provoquant, Fodor et Pylyshyn (1988) remettent en cause la capacité des modèles connexionistes de rendre compte de la cognition en vertu de leur incapacité à exhiber ce qu'il appellent "la systématicité" qualité que possède les systèmes classique de la BOVIA. La systématicité est mal définie chez ces auteurs. Ils affirment que les capacités cognitives humaines sont systématiquement reliées et apparaissent par ensembles. Ils donnent une série d'exemples faisant référence au problème de l'assignation dynamique. Par exemple, on ne trouve pas de gens capables de penser que "Jean aime Hélène" et incapables de penser que "Hélène aime Jean". De même, nul n'est capable d'inférer "Jean est allé au magasin" à partir de "Jean, Suzanne, Marie et Sally sont allés au magasin" et incapable d'inférer "Jean est allé au magasin" à partir de "Jean, Suzanne et Marie sont allés au magasin".

Donc la systématicité pourrait être approximativement définie comme suit : la capacité de penser P(a, b) est liée à la capacité de penser P (b, a), la capacité de penser A à partir de (A & B) est liée à la capacité de penser A à partir de (A & B & C),... Mais, objecte Hadley (1996) ou s'arrête cette capacité ? La capacité de penser (A⊃B) est elle liée à la capacité de penser (˜B⊃˜A) ? De toute manière Fodor et Pylyshyn (1988) ne se posent pas la question de la réalité empirique de ce qu'ils avancent. Ingram (1985), par exemple, montre que l'enfant commençant à parler n'exhibe pas beaucoup de systématicité. Dans le domaine du raisonnement, les adultes ne se montrent guère systématiques.

L'article de Fodor et Pylyshyn a suscité un effort considérable au sein du monde connexioniste. Plusieurs solutions ont été proposées : l'utilisation du calcul vectoriel "Tensor Product" de Smolensky (1990), l'utilisation du degré d'activation des noeuds : "les signatures" de Lange et Dyer (1989), "CONSYDERR" de Sun (1992) et les "pattern similarity association" de Barnden et Srinivas (1991). La solution la plus intéressante, parce qu'elle s'appuie sur des données neurobiologiques est "l'assignation par synchronie temporelle" (Shastri et Ajjanagadde, 1993; Henderson, 1994; Hummel et Holyoak, 1996 et Sougné, 1996).

Dans ces systèmes, les noeuds, comme les neurones, accumulent de l'activation avant de se déclencher. Les noeuds représentant les objets à associer vont se déclencher de manière simultanée, tandis que les noeuds représentant des objets à distinguer se déclencheront de manière déphasée. Pour représenter "La rose rouge sur la table verte", il faut deux synchronies asynchrones. La première synchronie associe "rose" à "rouge" et la seconde "table" à "verte". Les deux ensembles se distinguent par leur succession. De nombreuses études neurobiologiques montrent que ce mécanisme est compatible avec celui permettant au cerveau de résoudre le problème de l'assignation (voir Singer, 1995). Ce système permet d'affecter les variables à leur contenu. L'utilisation d'un prédicat P(a, b) peut servir quelque soit le contenu des variables "a" et "b", "a" doit simplement se synchroniser avec son contenu et "b" avec le sien. La systématicité est alors résolue.

Les réseaux glissants et la représentation des concepts

Les modèles qui conçoivent l'activité mentale comme une capacité à associer des idées ont placé des réseaux au centre de leurs architectures. Si les modèles connexionnistes sont les plus connus et les plus répandus, d'autres approches ont été suggérées.

Hofstadter (1984) a, par exemple, proposé d'utiliser certains types de réseaux pour représenter les concepts. Ses travaux s'enracinent dans une analyse fine de phénomènes aussi différents que les lapsus, la pensée analogique, l'humour, la traduction, la découverte scientifique, la créativité, les jeux de chiffres. Tous paraissent exiger des capacités de manipulation, transformation, déformation, combinaison, transposition, éclatement et réarrangement d'éléments mentaux que Hofstadter estime centraux dans l'activité intellectuelle. Il utilise souvent un exemple simple pour illustrer le rôle joué par cette fluidité mentale dans notre pensée: la notion de "première dame". Ce concept est à première vue aisé à définir : pour

la plupart des gens, il s'agit de l'épouse du chef d'État, Madame Chirac pour les français, Madame Clinton pour les américains. Sa transposition à l'Allemagne, par contre, complique la tâche. Parle-t-on de Madame Herzog, épouse du président ou de Madame Kohl, épouse du chancelier ? Et que penser des déformations à lui imposer si on s'intéresse au Royaume-Uni. Le prince Philippe – pourtant du sexe masculin -- est-il un candidat crédible ? Très clairement, le contexte exige d'adapter notre définition initiale. L'épouse du chef d'État devrait sûrement être transformée en “conjoint ou conjointe du personnage le plus important du pays”. Et voici notre notion, a priori simple, sous la pression du contexte, qui commence à se déformer. Et l'exportation de la première dame dans des contextes plus exotiques exigera de poursuivre cette gymnastique intellectuelle. Comment définir la première dame d'un club qui n'a plus de président, d'une république où l'épouse d'un ancien président pourrait avoir reçu ce titre à vie etc. Ces opérations mentales constituent l'ingrédient fondamental de nos activités intellectuelles, dit Hofstadter et ceci l'amène à postuler un type d'architecture cognitive originale.

Premièrement, les relations entre la cognition et la perception méritent d'être réexaminées. Certaines conduites intellectuelles de haut niveau peuvent se représenter comme des activités perceptives. L'intelligence est ainsi conçue comme le produit de milliers de processus qui fonctionnent en parallèle, de manière essentiellement inconsciente. Ils s'exécutent de manière quasi autonome, partiellement en parallèle et de manière aléatoire (voir COPYCAT, Mitchell, 1993; TABLETOP, French, 1995; NUMBO, Defays, 1988).

Deuxièmement, les représentations mentales doivent évoluer sous pression du contexte, d'éléments extérieurs, comme illustré dans l'exemple de la “ première dame ”. Ceci exige des modes de description et d'encodage qui se prêtent aux glissements, aux rapprochements d'idées, aux associations. Pour ce faire, une structure complexe de réseau est proposée. De nouveau un exemple peut faciliter la compréhension des mécanismes et des structures. Le programme COPYCAT (Mitchell, 1993; Hofstadter et Mitchell, 1991), qui a permis d'introduire et de tester la notion de réseau glissant, étudie comment produire des variations mentales sur un thème donné. Le domaine choisi est constitué de lettres. Par exemple, si “abc se transforme en abd”, que devient “pqrs” si vous lui faites “la même chose” ? La plupart des gens interrogés répondent “pqrt”. Ils ont représenté la transformation initiale comme un remplacement de la dernière lettre de la chaîne “abc” par son successeur dans l'alphabet. Mais d'autres réponses sont bien entendu possibles. Pourquoi pas “abd”, puisque la règle initiale pourrait être “remplacer la première chaîne de caractères par abd”? La gymnastique mentale nécessaire exige de percevoir la première transformation de différentes manières: le “c” est tantôt perçu comme la troisième lettre de la chaîne : tantôt comme la dernière, comme le prédécesseur de “d” dans l'alphabet, comme la troisième lettre de l'alphabet, la chaîne “abc” est décrite comme une suite croissante, composée de lettres différentes et successives, de longueur 3, partiellement répétée dans “abd” etc.

Hofstadter propose de représenter l'univers conceptuel sous-jacent par un réseau dont les noeuds symbolisent des concepts impliqués (les différentes lettres, les notions de “premier”,

“dernier”, “successeur”, “prédécesseur”, “droit”, “gauche”, “groupe”, etc.) et dont les arcs représentent les relations que ces concepts entretiennent entre eux (“opposés”, “semblables”, “proches”, etc.). Remarquez que ces relations sont elles-mêmes des concepts. Voici un extrait de ce réseau (Figure 9).

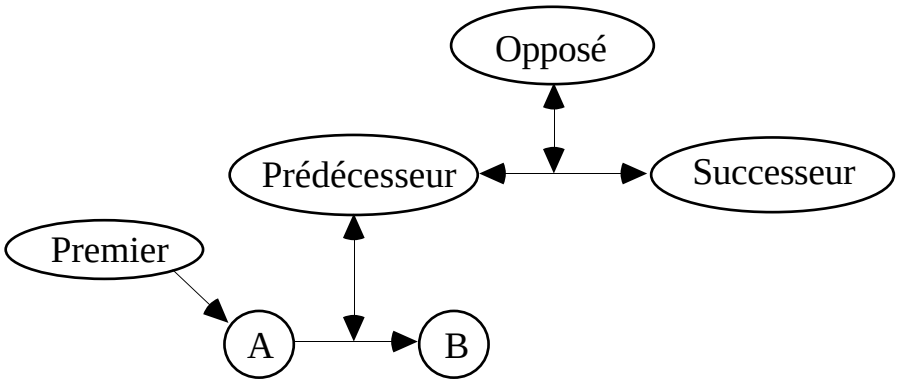

Figure 9 : extrait de réseau glissant

Comme dans les modèles connexionnistes, les noeuds sont caractérisés par une activation qui se propage au cours du temps le long des liens. Ceux-ci véhiculent plus ou moins d'activation suivant leur poids qui est lui-même fonction du concept qui leur est associé. Ainsi dans l'exemple ci-dessus, si le noeud “opposé/symétrique” est très actif, “successeur” et “prédécesseur” échangent beaucoup d'activation. Leur distance conceptuelle est faible, une notion se transforme aisément en son opposé, l'esprit glisse d'un point de vue à l'autre, d'où le nom de réseau glissant qui qualifie ce type de réseau.

Le réseau glissant est souvent comparé à un réseau sémantique, une autre structure fréquemment utilisée en IA pour représenter certaines connaissances; les noeuds représentent des objets, des concepts, des situations, voire des images, des propositions et les arcs, les relations qu'ils entretiennent entre eux. Celles-ci sont en général plus limitées que dans les réseaux glissants : liens de type “est-un” (Jumbo “est-un” programme) ou de type “possède” ou “fait-partie” (Dumbo “possède” des défenses). Dans un réseau sémantique, il n'y a pas de diffusion d'activation. Ils capturent la nature des liens qui unissent les noeuds, permettent de mettre en oeuvre des mécanismes d'héritage de propriétés (un objet spécifique hérite des caractéristiques d'un objet générique).

Ces structures, avant d'être utilisées en IA, ont été développées (Quillian, Loftus et Collins (ref),...) pour modéliser la mémoire.

3.2.1. L'activité mentale conçue comme une aptitude à apprendre

L'aptitude à apprendre est aussi rapidement apparue comme une caractéristique fondamentale de l'intelligence. La machine ne peut faire que ce qu'on lui commande d'exécuter

alors que l'esprit s'enrichit continuellement, à travers ses interactions avec le milieu, de nouvelles représentations qui lui permettent d'aborder de nouvelles tâches ou de mieux exécuter d'anciennes. Comment reproduire cette aptitude ? Différentes approches ont été proposées. Elles sont directement liées aux architectures cognitives utilisées pour modéliser les activités mentales.

Modification du poids des connexions dans des réseaux

Dans les premiers réseaux neuronaux (Rosenblatt, 1958), l'apprentissage était conçu comme une adaptation progressive des poids associés aux liens du réseau en fonction de l'erreur observée. Lorsque deux noeuds voisins sont simultanément actifs, le poids de leur connexion a tendance à augmenter. Cette loi fut suggérée comme mentionné précédemment (voir le paragraphe sur les origines du connexionnisme) par Hebb (1949) et mise en oeuvre de différentes manières sur machine.

Généralement, les réseaux démarrent avec une configuration plus ou moins aléatoire et se modifient de manière incrémentale. Certains apprentissages sont supervisés : ceci permet à la machine de modifier sa structure interne en fonction de la différence entre ce qui est produit et ce qui devrait être produit (voir le paragraphe sur le perceptron). Une de ces méthodes, largement utilisée, s'appelle la rétropropagation. Les algorithmes utilisés s'inspirent de méthodes d'optimisation développées en mathématiques. D'autres apprentissages ne nécessitent pas une intervention extérieure, ils sont dits non supervisés: ils consistent à renforcer les liens entre unités simultanément activées ou à pénaliser (en diminuant le poids) les associations entre unités dont le niveau d'activation est fort différent. Cette forme d'apprentissage, quelquefois appelée apprentissage compétitif permet de catégoriser en regroupant dans des classes des entités jugées similaires par le système. Elle s'inspire implicitement de techniques statistiques de classification automatique.

Ajustement d'une fonction d'évaluation

Samuel (1959) a présenté un programme capable non seulement de jouer aux dames, mais également de s'améliorer progressivement. Le principe de fonctionnement était simple. Jouer aux dames était conçu comme trouver la bonne séquence d'opérations à effectuer (de coups à jouer) pour battre son adversaire. A chaque coup, le nombre d'options à prendre en compte étant considérable, le système utilisait certaines heuristiques pour faire son choix : maximiser son nombre de pions, minimiser celui de l'adversaire, contrôler le centre du damier etc. Plus précisément, il associait à chaque configuration du jeu une valeur calculée comme une somme pondérée du type

$$w_1 s_1 + w_2 s_2 + \ldots + w_n s_n \qquad (4)$$

où les s_1, s_2, ..., sn sont des mesures relatives à des caractéristiques de la configuration et les w_i des facteurs de pondération. Il choisissait la configuration qui maximisait cette somme pondérée. Progressivement, il modifiait les w_i en augmentant ceux correspondant à des caractéristiques liées à des résultats favorables et en diminuant les autres. Le programme commençait généralement les parties de manière faible et peu conventionnelle mais les terminait remarquablement.

Apprentissage par généralisation d'exemples, discrimination, explication

Winston (1975) a développé un programme capable d'apprendre des nouveaux concepts à partir d'exemples et de contre-exemples: on lui “montrait” des exemples d'arches constituées de 2 piliers verticaux supportant un plateau horizontal et des constructions similaires qui n'étaient pas des arches (car les deux pieds verticaux se touchaient, par exemple). A partir d'une description initiale que le système était capable de construire en analysant un spécimen du concept à représenter, des affinements successifs étaient apportés jusqu'à ce qu'une description capable de caractériser tous les exemples du concept et d'exclure tous les contre-exemples soit construite. Plus tard, à côté de cet apprentissage par généralisation, d'autres modes ont été testés:

- apprentissage par discrimination au cours duquel le système se focalise non plus sur ce qui rapproche les objets d'une même catégorie, mais sur ce qui différencie ceux qui appartiennent à des classes différentes (Langley, 1987);
- apprentissage par explication où le programme cherche à partir de données théoriques à expliquer en quoi un prototype présenté constitue un membre de la catégorie à assimiler (Mitchell et al, 1986);
- apprentissage par agrégation (chunking) où il crée de nouvelles règles censées simplifier et généraliser des procédures plus complexes précédemment découvertes (voir SOAR présenté dans un paragraphe précédent).

Algorithmes génétiques

Des procédures plus aveugles de création de nouvelles connaissances par des techniques s'inspirant de la génétique ont également été proposées (Holland, 1975). Pour être appliquées, elles nécessitent une codification particulière des informations que manipule le système : généralement des représentations par des chaînes de 0 et de 1, de longueur fixe sont utilisées. Ces chaînes sont couplées, coupées en deux et “croisées” en échangeant une de leurs parties. Les nouvelles chaînes ainsi constituées sont alors testées et celles qui sont adaptées sont conservées (Mitchell, 1996).

La pertinence de ces travaux pour la psychologie

Le paragraphe ci-dessus ne donne qu'un aperçu fort incomplet de l'ensemble des techniques qui ont été proposées pour amener les programmes à innover, à s'affranchir de leurs

créateurs en quelque sorte, à les rendre plus semblables à ce que l'on observe chez les hommes. Certains algorithmes s'inspirent clairement de théories existantes (algorithmes génétiques ou apprentissage hebbien dans les réseaux connexionnistes) et les transposent en IA. Ils en montrent les qualités : caractère universel, simplicité de mise en oeuvre, caractère optimal dans certaines situations, et les faiblesses : risques d'oublis catastrophiques dans les réseaux neuronaux, longueur de la période d'apprentissage, caractère quelquefois sub-optimal des algorithmes génétiques. D'autres algorithmes simulent des modes d'apprentissage de plus haut niveau et donnent des formes opérationnelles à des concepts quelquefois déjà connus des psychologues. Ils permettent d'aborder de manière précise des questions fondamentales. Quand apprend-t-on ? Comment évite-t-on les généralisations sans intérêt ? Quel type de représentation se prête à tel type d'apprentissage ? Ils montrent comment et quand un système peut réagir à des pressions venant de l'environnement en adaptant des éléments de connaissance existants ou en en créant de nouveaux. Quelques faits importants établis par ces simulations ont été mis évidence : l'importance de la séquence des exemples et contre-exemples présentés dans une généralisation, la capacité ou l'incapacité de certaines représentations à se modifier, le lien entre les représentations, la résolution de problèmes et les techniques d'apprentissage, l'importance des informations dont on dispose initialement. Ces travaux ont également permis de montrer les limites d'une description générique des concepts et de proposer des modes de représentation alternatifs à partir de prototypes, d'exemples judicieusement choisis (ce courant de pensée a débouché sur ce qu'on appelle le “case base reasoning”).

3.2.3. L'activité mentale conçue comme une aptitude à agir sur le milieu

Il existe également un nouveau courant de pensée en IA qui s'interroge sur le statut des représentations dans les comportements intelligents. Celles-ci sont générées par le système perceptif; elles décrivent explicitement des entités (objets, propriétés, concepts, désirs, ...) du monde réel et tout ce qui n'est pas explicite paraît ne pas exister. Il a fallu des milliards d'années pour passer de la cellule aux mammifères, les premiers primates apparaissaient il y a 120 millions d'années, l'homme il y a seulement 2,5 millions d'années, et l'écriture il y a 5000 ans. Ceci suggère que la capacité à résoudre des problèmes, à raisonner, à acquérir de l'expertise apparaissent lorsque le système peut évoluer et réagir de manière appropriée dans son environnement. Les auteurs qui appartiennent à ce nouveau courant (Brooks, 1991) dénoncent le caractère “ impérialiste ” des représentations telles qu'elles sont conçues dans les systèmes relevant du paradigme symbolique. Celles-ci doivent s'enraciner dans le monde physique. Pour être intelligents, ces systèmes doivent être directement reliés à la réalité par des senseurs et des effecteurs. Et ils construisent, pour illustrer leur position, des robots, non plus basés sur la manipulation de symboles mais capables de comportements élémentaires (évitement d'obstacles, atteinte de cibles, recherche de sources de lumières, etc.), stimulés par des informations sensorielles. Pour ce courant, ces comportements complexes ne peuvent pas être appréhendés sans avoir modélisé au préalable ces fonctions élémentaires. C'est leur interaction,

la manière dont elles coexistent et coopèrent qui crée l'intelligence.

4. Conclusions

Le progrès scientifique dépend très souvent des développements croisés des théories et de la technologie. Les premières proposent un arsenal conceptuel et des langages. La seconde, des techniques d'investigation, d'exploration, d'expérimentation. Les interrogations théoriques appellent les expériences, les faits observés nécessitent le recours à de nouveaux concepts. L'astronomie et la physique nous en donnent des exemples nombreux : les études fouillées des orbites des planètes, les lois de Kepler, de Newton, l'invention du télescope, du microscope, les raffinements théoriques qui en résultent illustrent à merveille ce jeu subtil entre les appareils théoriques et expérimentaux.

L'introduction de l'informatique dans le domaine de la psychologie, il y a peine quarante ans, a peut-être apporté un élément nouveau dans cette interaction. A la fois outil et langage, elle servait deux maîtres à la fois. Machine à calculer prodigieuse pour les expérimentateurs, les psychométriciens, l'ordinateur permettait un traitement statistique rapide des données, une grande souplesse et une nouvelle précision dans la conception des expériences. Certains chercheurs, tel que John McCarthy, Marvin Minsky, John Holland, Allen Newell, et Herbert Simon, ont cependant très vite réalisé que ces avantages ne s'arrêtaient pas là. L'informatique offrait des possibilités de modélisation jusqu'alors inconnues. Ils ont ainsi développé et testé des théories de la cognition, de la mémoire, de l'apprentissage, du comportement grâce aux nouveaux langages mis à disposition. Des concepts vagues de traitement interne, représentation, contrôle acquéraient une réalité nouvelle. Leurs programmes donnaient la possibilité de dépasser les formulations littéraires ou mathématiques et proposaient des descriptions rigoureuses de processus dynamiques.

Dans ce chapitre, nous avons passé en revue certains des ces modèles, des architectures symboliques aux réseaux connexionnistes. Nous avons présenté non seulement les contributions mais aussi les limites de ces tentatives. Ces dernières ont permis de mieux apprécier les capacités extraordinaires de nos propres mécanismes cognitifs. Qui aurait pu imaginer les énormes difficultés associées à la simple reconnaissance d'un objet, à la compréhension d'une phrase, la saisie d'une analogie?

L'introduction de l'informatique en psychologie et les développements observés ces 40 dernières années ne constituent probablement que les premiers balbutiements d'une nouvelle manière de conceptualiser les phénomènes psychologiques. Ils ouvrent des nouvelles portes, livrent à l'investigation scientifique des domaines peu modélisés et suscitent beaucoup d'enthousiasme. Mais, faire la part de la mode et de la tendance de fond dans les travaux actuels n'est assurément pas facile. Trop souvent, la conviction parait tenir lieu de preuve. Des espoirs ont été déçus, des promesses non tenues. Et pourtant, comme nous venons de le mentionner, l'histoire de la science est irrémédiablement liée à celle des technologies.

L'apparition de l'ordinateur ne peut avoir qu'un impact durable sur la recherche en psychologie. La forme que prendront les travaux du futur est encore imprécise; de la logique aux réseaux neuronaux en passant par les systèmes experts et les algorithmes génétiques, l'évolution se caractérise par une grande inventivité qui rend toute prédiction hasardeuse. Le rôle de la simulation y parait prépondérant. L'étude du mental ne pourra qu'en bénéficier.

5. Références

Baddeley, A. D. (1986). *Working Memory*. Oxford: Oxford University Press.

Baird, H. S. (1988). Feature Identification for hybrid structural/statistical pattern classification. *Computer vision, graphics and image processing, 42*, 318-333.

Barnden, J., & Srinivas, K. (1991). Encoding techniques for complex information structures in connectionist systems. *Connection Science, 3,* 269-315.

Berliner, H. J. (1980). Computer backgammon. *Scientific American, 249,* 64-72.

Boden, M. (1991a). *The Philosophy of Artificial Intelligence*. Oxford: Oxford University Press.

Boden, M. (1991b). Horses of a differnt color? In W. Ramsey, S. P. Stich, D. E. Rumelhart (Eds.) *Philosophy and Connectionist Theory* . Hillsadale: Lawrence Erlbaum Associates.

Braine, M. D. S. (1978). On the relation between the natural logic of reasoning and standard logic. *Psychological Review, 85,* 1-21.

Brooks, R. (1981). Symbolic reasoning among three-dimensional models and two-dimensional images. *Artificial intelligence, 17,* 285-348.

Brooks, R. (1991). Elephants don't play chess. In P. Maes (Ed.) *Designing Autonomious Agents* . Cambridge MA.: MIT Press.

Carbonell, J. G., Boggs, W. M., Mauldin, M. L., & Anick, P. G. (1985). The EXCALIBUR project, A natural language interface to expert systems and data bases. In *Proceedings of the eighth IJCAI.* San Mateo: Morgan-Kaufman.

Cheng, P. W., & Holyoak, K. J. (1985). Pragmatic Reasoning Schemas. *Cognitive Psychology, 17,* 391-416.

Cohen, H. (1988). How to Draw Three People in a Botanical Garden. *Proceedings of the Senventeen National Conference on AI.*

Cosmides, L. (1989). The Logic of Social Exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition, 31,* 187-276.

Defays, D. (1988). *L'esprit en friche*. Liège: Mardaga.

Elman, J. L. (1990). Finding Structure in Time. *Cognitive Science 14*, 179-211.

Feldman, J. A. (1982). Dynamic Connections in Neural Networks. *Biological Cybernetics,46,* 27-39.

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

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

Fodor, J. A., & McLaughlin, B. P. (1990). Connectionism and the problem of systematicity: Why Smolensky solution doesn't work. *Cognition, 35,* 183-204.

French, R. M. (1990) Subcognition and the Limits of the Turing Test. *Mind, 99,* 53-65.

French, R. M. (1995). *The Subtlety of Sameness: A Theory and Computer Model of Analogy-Making.* Cambridge, MA: The MIT Press.

Garey, M. R., & Johnson, D. S. (1979). *Computers and Intractability.* New York: Freeman.

Hadley, R. F. (1996). Connectionism, Systematicity and Nomic Necessity. *Proceedings of the Eighteen conference of the Cognitive Science Society.* Mahwah,NJ: Lawrence Erbaum Ass.

Haugeland, J. (1980). Commentary on Searle "Minds, Brains, and Programs". *The Behavioral and Brain Sciences, 4,* 432-433.

Hebb, D. O. (1949). *The Organization of Behavior.* New York: Wiley.

Henderson, J. (1994). *Description Based Parsing in a Connectionist Network.* PhD thesis, University of Pennsylvania, Philadelphia, PA. Technical Report MS-CIS-94-46.

Hofstadter, D. R. (1984). The Copycat Project: An Experiment in Nondeterminism and Creative Analogies. AI Memo 755, Artificial Intelligence Laboratory, MIT.
Hofstadter, D. R. (1988). *Ma Thémagie*. Paris: InterEditions.
Hofstadter, D. R., & Dennett. D. C. (1981), *The Mind's I*. New York: Basic Books. Traduction française: *Vues de l'Esprit* Paris: InterEditions (1987).
Hofstadter, D. R., & Mitchell, M. (1991) The Copycat Project: A model of mental fluidity and analogy-making. In K. Holyoak, & J. Barnden (Eds.) *Advances in connectionist and neural computation theory, Vol. 2: Analogical connections.* New York: Ablex.
Holland, J. H. (1975) *Adaptation in Natural and Artificial Systems*. Ann Arbor, MI: The University of Michigan Press.
Holyoak, K. J., & Thagard, P. (1989). Analogical mapping by constraint satisfaction. *Cognitive Science 13,* 295-355.
Hummel, J. E., & Holyoak, K. J. (1996). LISA: A Computational Model of Analogical Inference and Schema Induction. *Proceedings of the Eighteen conference of the Cognitive Science Society.* Mahwah,NJ: Lawrence Erbaum Ass.
Ingram, D. (1985). The psychological reality of children's grammars and its relation to grammatical theory. *Lingua, 66,* 79-103.
Inhelder, B., & Piaget, J. (1955). *De la logique de l'enfant à la logique de l'adolescent.* Paris: PUF.
James, W. (1890). *Psychology (Briefer course).* New York: Holt.
Johnson-Laird, P. N. (1983). *Mental Models: Towards a cognitive science of language, inference, and conciousness*. Cambridge: Cambridge University Press.
Laird, J. E., Newell. A., & Rosenbloom, P. S. (1987). SOAR: An architecture for general intelligence. *Artificial Intelligence, 33,* 1-64.
Lange, T., & Dyer, M. (1989). High-level inferencing in a connectionist network. *Connection Science, 1,* 181-217.
Langley, P. (1987). A general Theory of Discrimination Learning. In Klahr, D., Langley, P. et Neches, R. (Eds.) *Production systems models of learning and development* . Cambridge MA: MIT Press.
Langley, P., Thompson, K., Iba, W., Gennari, J. H., & Allen, J. A. (1992) An Integrated Cognitive Architecture for Autonomous Agents. In W. Van De Velde (Ed.) *Representation and Learning in Autonomous Agents.* Amsterdam: North Holland.
Lashley, K. (1950). In search of the engram. In *Society of Experimental Biology Symposium, 4: Psychological Mechanisms in Animal Behavior.* Cambridge: Cambridge University Press.
Lenat, D.B. (1983). EURISKO: a program that learns new heuristics and domain concepts. *Artificial Intelligence, 21,* 61-98.
Mc Carthy, J. (1988). Epistemological Challenges for Connectionism. *Behavioral and Brain Science. 11,* 44.
Mc Culloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics, 5,* 115-133.
McClelland, J. L., & Rumelhart, D. E. (1986). *Parallel Distributed Processing*. Cambridge, MA: The MIT Press.
Minsky, M. L., & Papert, S. (1969). *Perceptrons*. Cambridge, MA: The MIT Press.
Mitchell ,T.M., Keller,R.M., et Kedar-Cabelli,S.T. (1986), Explanation-based generalization: an unifying view. *Machine Learning 1,* 47-80.
Mitchell, M. (1993). *Analogy-Making as Perception*. Cambridge, MA: The MIT Press.
Mitchell, M. (1996). *An Introduction to Genetic Algorithms*. Cambridge, MA: The MIT Press.
Newell, A., & Simon, H. A. (1963). GPS: A Program that Simulates Human Thought. In E. A. Feigenbaum & J. A. Feldman (Eds.) *Computers and Thought*. New York: McGraw-Hill.
Newell, A. & Simon, H. A. (1972). *Human Problem Solving*. Englewood-Cliffs, NJ: Prentice-Hall.
Newell, A, & Simon, H. A. (1976). Computer Science as Empirical Enquiry: Symbols and Search. *Communications of the ACM, 19,* 113-126.
Newell, A., Rosenbloom, P. S., & Laird, J. E. (1989). Symbolic architectures for cognition. In M. I. Posner (Ed.) *Foundations of Cognitive Science*. Cambridge, MA: The MIT Press.

Newell, A., Shaw, J. C., & Simon, H. A. (1963). Empirical Explorations with the Logic Theory Machine: A Case Study in Heuristics. In E. A. Feigenbaum, & J. A. Feldman (Eds.) *Computers and Thought*. New York: McGraw-Hill.

Oakford, M., & Chater, N. (1995). Theories of Reasoning and the Computational Explanation of Everyday Inference. *Thinking and Reasoning, 1,* 121-152.

Rips, L. J. (1994). *The Psychology of Proof: Deductive Reasoning in Human Thinking.* Cambridge, MA.: MIT Press.

Rosenblatt, F. (1958). The Perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review, 65,* 386-408.

Samuel, A. (1959) Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development,* 3, ...

Searle, J., R. (1980). Minds Brain, and Programs. *Behavioral and Brain Science. 3,* 417-424

Sejnowski, T. J., & Rosenberg, C. R. (1987). Parallels networks that learn to pronounce English text. *Complex Systems, 1,* 145-168.

Shapiro, S. (1979). The SNePS semantic network processing system. In N. Findler (Ed.) *Associative Networks: Representation and Use of Knowledge by Computers.* New York: Academic Press.

Shapiro, S. (1992). *Encyclopedia of Artificial Intelligence, 2nd Edition.* New York: Wiley.

Shastri, L., & Ajjanagadde, V. (1993). From Simple Associations to Systematic Reasoning: A connectionist representation of rules, variables and dynamic bindings using temporal synchrony. *Behavioral and Brain Sciences, 16,* 417-494.

Singer, W. (1995). Synchronization of neuronal responses as a putative binding mechanism. In M. A. Arbib (Ed.) *The Handbook of Brain Theory and Neural Networks.* Cambridge MA.: MIT Press.

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

Sougne, J. (1996). A Connectionist Model of Reflective Reasoning Using Temporal Properties of Node Firing. *Proceedings of the Eighteen conference of the Cognitive Science Society.* Mahwah,NJ: Lawrence Erbaum Ass.

Sun, R. (1992). On variable binding in connectionist networks. *Connection Science, 4,* 93-124.

Turing, A. (1950), Computing machinery and intelligence. *Mind, 50,* 433-60.

Waltz, D. (1975). Understanding line drawings of scenes with shadows. In P. H. Winston (Ed.) *The psychology of computer vision.* New-York: McGraw-Hill.

Wang, H. (1963) Toward Mechanical Mathematics. In K. M. Sayre, & F. J. Crosson (Eds.) *The Modeling of Mind: Computers and Intelligence.* New York: Simon and Schuster.

Winston, P. H. (1975). Learning Structural Descriptions from Examples. In P. H. Winston (Ed.) *The Psychology of Computer Vision*. New York: McGraw-Hill.