ESPRITS CERVEAUX ET PROGRAMMES*

JOHN R. SEARLE
TRADUIT PAR ERIC DUYCKAERTS

Quelle signification psychologique et philosophique faut-il attacher aux récents efforts de simulation par ordinateur des capacités cognitives de l'homme ? Pour répondre à cette question, je trouve utile de distinguer ce que j'appellerai l'I.A. (Intelligence Artificielle) "forte" de l'I.A. "faible" ou "prudente". Pour celle-ci, la valeur principale de l'ordinateur dans l'étude de l'esprit est de nous offrir un outil très puissant. Par exemple, il nous permet de formuler et de tester des hypothèses d'une façon plus rigoureuse et précise. Pour l'I.A. forte, au contraire, l'ordinateur n'est pas simplement un outil pour l'étude de l'esprit ; bien plus, programmé de façon appropriée, il *est* réellement un esprit, au sens où l'on peut dire littéralement d'ordinateurs dotés de programmes corrects qu'ils *comprennent* et ont d'autres états cognitifs. En I.A. forte, comme l'ordinateur programmé a des états cognitifs, les programmes ne sont pas de simples outils destinés à tester des explications psychologiques : ils sont eux-mêmes les explications.

Je n'ai pas d'objections aux revendications de l'I.A. faible, du moins dans le cadre de cet article. Ma discussion sera dirigée ici vers les affirmations que j'ai définies comme relevant de l'I.A. forte, à savoir celles qui prétendent que l'ordinateur correctement programmé a des états cognitifs et que, dès lors,

les programmes expliquent la cognition humaine. Quand par la suite je me référerai à l'I.A. il s'agira de la version forte telle qu'elle s'exprime dans ces deux affirmations.

J'examinerai le travail de Roger Schank et de ses collègues (Schank et Abelson 1977), parce qu'il m'est plus familier que d'autres dont les prétentions sont similaires, et parce qu'il fournit un exemple très clair du type de travail que je souhaite prendre en considération. Cependant, rien de ce qui suit ne dépend du détail des programmes de Schank. Les mêmes arguments s'appliqueraient à SHDRLU de Winograd (Winograd 1973), ELIZA de Weisenbaum (Weisenbaum 1965) et, en fait, à toute simulation de phénomènes mentaux propres à l'homme, sur n'importe quelle machine de Turing

Très brièvement et en laissant de côté les divers détails, on peut décrire le programme de Schank comme suit : son but est de simuler l'aptitude humaine à comprendre des histoires. Une caractéristique de la capacité de compréhension d'histoires est, chez les êtres humains, qu'ils peuvent répondre à des questions sur l'histoire, même si l'information qu'ils donnent ne s'y trouvait pas explicitement posée. Ainsi par exemple, supposez que l'on vous soumette l'histoire suivante : "Un homme entre dans un restaurant et commande un hamburger. Quand celui-ci arrive, il est calciné. L'homme furieux, sort du restaurant en coup de vent, sans payer le hamburger ni laisser de pourboire". Si maintenant on vous demande : "L'homme a-t-il mangé le hamburger ?", on peut présumer que votre réponse sera : "Non, il ne l'a pas mangé". De même, si l'on vous soumet cette histoire : "Un homme entre dans un restaurant et commande un hamburger. Quand celui-ci arrive, l'homme en est très content et en quittant le restaurant, il laisse un large pourboire à la serveuse avant de payer la note". A la question : "L'homme a-t-il mangé le hamburger ?", votre réponse présumable sera : "Oui, l'homme a mangé le hamburger". De même, les machines de Schank peuvent répondre de cette façon à des questions sur des restaurants. Pour le faire, elles ont une "représentation" du type d'informations dont disposent les êtres humains quant aux restaurants; elles peuvent ainsi

répondre à des questions comme celles qui précèdent, pour le même genre d'histoires. Quant la machine reçoit l'histoire et, ensuite, la question, elle imprime des réponses du type de celles que nous attendrions d'un être humain face aux mêmes histoires. Les partisans de l'I.A. forte prétendent que dans cette séquence de questions et réponses, la machine n'a pas seulement simulé l'aptitude humaine, mais aussi (1) qu'on peut dire littéralement qu'elle *comprend* l'histoire et fournit les réponses aux questions, et (2) que ce que font machine et programme *explique* l'aptitude humaine à comprendre des histoires et à répondre à des questions sur celles-ci.

Ces deux prétentions ne me paraissent absolument pas soutenues par le travail de Schank, comme j'essayerai de le montrer dans ce qui suit. Bien sûr, je ne dis pas que Schank lui-même soit responsable de ces affirmations.

Une manière de tester n'importe quelle théorie de l'esprit est de se demander à quoi ressemblerait mon propre esprit s'il fonctionnait selon les principes qui, pour cette théorie, sont ceux du fonctionnement de tous les esprits. Appliquons ce test au programme de Schank avec la *Gedankenexperiment* que voici. Supposez que je sois enfermé dans une pièce et qu'on me donne une grosse liasse d'écriture chinoise. Supposez en outre (comme c'est effectivement le cas), que je ne connaisse rien au chinois, parlé ou écrit, et que je ne sois même pas sûr de pouvoir reconnaître l'écriture chinoise comme distincte, disons de l'écriture japonaise ou de gribouillis dépourvus de sens. Pour moi,l'écriture chinoise n'est qu'autant de gribouillis dépourvus de sens. Supposez maintenant qu'après cette première liasse, on me donne un second paquet d'inscription chinoises, en même temps qu'un ensemble de règles pour corréler l'une à l'autre. Les règles sont en anglais et je les comprends aussi bien que n'importe quel locuteur anglophone d'origine. Elles me permettent de corréler un ensemble de symboles formels avec un autre - "formel" ne signifiant ici rien d'autre que la possibilité d'identifier les symboles simplement à partir de leurs formes. Supposez encore que je reçoive une troisième liasse de symboles chinois

avec quelques instructions - en anglais - qui me permettent de corréler des éléments de cette troisième avec les deux premières. Ces règles m'instruisent de la manière de fournir certains symboles chinois d'une certaine forme en réaction ('réponse') à certaines formes de la troisième liasse. Les inconnus qui me donnent tous ces symboles appellent la première liasse un "script", la deuxième une "histoire" et la troisième des "questions". De plus, ils appellent les symboles que je leur fournis en réaction à la troisième liasse des "réponses" ('answer') aux questions. L'ensemble des règles en anglais que j'ai reçu, ils l'appellent le "programme". Ici, pour le plaisir de compliquer l'histoire, imaginez que ces gens me donnent aussi des histoires en anglais (que je comprends) et me posent sur elles des questions auxquelles je réponde en anglais. Supposez aussi qu'après un certain temps je devienne tellement habile à suivre les instructions de manipulation des symboles chinois et les programmeurs tellement forts dans la rédaction des programmes ,que d'un point de vue extérieur - c'est à dire du point de vue de quelqu'un extérieur à la pièce où je suis enfermé - mes réponses aux questions soient absolument impossibles à distinguer de celles de locuteurs d'origine chinoise. En se contentant de regarder mes réponses, personne ne pourrait dire que je ne parle pas un mot de chinois. Supposons également que mes réponses aux questions anglaises soient impossibles à distinguer de celles d'autres locuteurs anglophones d'origine ; elles le seraient sans aucun doute pour la simple raison que je suis un locuteur anglophone d'origine. Du point de vue extérieur - du point de vue de quelqu'un qui lit mes "réponses" - les réponses aux questions anglaises et chinoises sont également bonnes. Cependant, dans le cas du chinois, contrairement à celui de l'anglais, je produis les réponses en manipulant des symboles formels ininterprétés. Pour ce qui est du chinois, je me comporte simplement comme un ordinateur, j'exécute des opérations d'ordinateur ('computational') sur des éléments formellement spécifiés. Je ne suis qu'une instantiation du programme de l'ordinateur.

Les affirmations de l'I.A. forte sont donc que l'ordinateur

programmé comprend les histoires et que le programme explique en quelque façon l'entendement humain. Nous sommes maintenant à même d'examiner ces prétentions à la lumière de notre expérience imaginaire.

■ En ce qui concerne la première affirmation, il me parait assez évident que, dans mon exemple, je ne comprends pas un mot des histoires chinoises. J'ai des entrées et des sorties qui ne peuvent être distinguées de celles des locuteurs chinois d'origine et je puis être doté de tous les programmes formels qu'on voudra, je ne comprends toujours rien. Pour les mêmes raisons, l'ordinateur de Schank ne comprend rien à aucune histoire, qu'elle soit en chinois, en anglais ou en n'importe quoi ; en effet, dans le cas du chinois, l'ordinateur c'est moi et dans le cas où je ne suis pas l'ordinateur, celui-ci ne dispose de rien de plus que moi quand je ne comprends rien.

■ Pour la seconde affirmation (que le programme explique l'entendement humain), nous pouvons voir que l'ordinateur et son programme ne procurent pas de condition suffisante à l'entendement, puisqu'ils se contentent de fonctionner sans comprendre. Mais ceci pourrait-il quand même fournir une condition nécessaire ou une contribution significative à l'entendement ? L'une des affirmations émises par les tenants de l'I.A. forte est que lorsque je comprends une histoire en anglais je fais exactement la même chose - ou, peut-être, principalement la même chose - que ce que je faisais en manipulant les symboles chinois. Ce serait seulement une manipulation de symboles plus formelle qui distinguerait le cas du chinois (où je ne comprends pas) de celui de l'anglais (où je comprends). Je n'ai pas démontré la fausseté de cette affirmation, mais elle parait certainement incroyable au vu de mon exemple. Pour qu'elle devienne plausible, il faut supposer que nous puissions construire un programme qui ait les mêmes entrées et sorties que des locuteurs parlant leur langue d'origine et supposer, en outre, que ceux-ci soient redevables à un certain niveau d'une description en terme d'instantiation d'un programme. Sur la base de ces deux suppositions, on suppose que, même si le programme de

Schank n'est pas le fin mot de l'histoire en matière d'entendement, il peut être un morceau de cette histoire. Soit ; admettons qu'il y ait là une possibilité empirique. Mais jusqu'ici nous n'avons pas la moindre raison de croire à sa vérité puisque ce que suggère mon exemple (sans toutefois le démontrer) c'est que le programme de l'ordinateur est simplement dépourvu de toute pertinence en ce qui concerne ma compréhension des histoires. Dans le cas du chinois, je dispose de tout ce que l'I.A. peut me faire ingérer par voie de programme et je ne comprends rien ; dans le cas de l'anglais, je comprends tout et il n'y a pas la moindre raison, à ce point, de supposer que ma compréhension ait quelque chose à voir avec les programmes d'ordinateurs, c'est à dire avec les opérations d'ordinateurs ('computational') appliquées à des éléments spécifiés de façon purement formelle. Mon exemple suggère que ces opérations n'entretiennent pas de connexions intéressantes avec l'entendement. Elles ne constituent certainement pas des conditions suffisantes et nous n'avons pas la moindre raison de supposer qu'elles soient des conditions nécessaires ni même des contributions significatives quant à l'entendement. Il faut remarquer que la force de l'argument ne s'appuie pas simplement sur le fait que des machines différentes peuvent avoir les mêmes entrées et sorties, tout en opérant sur des principes formels différents - là n'est pas la question. Bien plutôt : quels que soient les principes formels que vous introduisiez dans l'ordinateur, ils ne suffiront pas à expliquer l'entendement puisqu'un homme sera capable de les suivre sans rien comprendre. Aucune espèce de raison ne nous a été offerte de supposer nécessaires, (voire seulement utiles) de tels principes ; en effet, rien ne nous permet de supposer que quand je comprends l'anglais j'opère selon un programme formel, quel qu'il soit.

Mais alors , qu'est-ce que j'ai de plus dans le cas de phrases anglaises et qui me fait défaut dans celui des phrases chinoises ? De toute évidence, la réponse est que je sais ce que signifie les premières, tandis que je n'ai pas la plus petite idée de la signification des secondes. En quoi cela consiste-t-il et, quelle qu'en soit la nature, pourquoi ne pourrions-nous en doter une

machine ? Je reviendrai plus loin sur cette question, mais je veux auparavant poursuivre mon exemple.

J'ai eu l'occasion de le présenter à plusieurs chercheurs en I.A. et il est intéressant de remarquer qu'ils ne semblent pas s'accorder sur la réplique adéquate à lui opposer. J'ai reçu une variété surprenante de réponses et, dans ce qui suit, je vais considérer les plus communes d'entre elles (en les spécifiant par leur origine géographique).

Mais d'abord, je voudrais barrer la route à quelques méprises courantes au sujet de "l'entendement" : un mot qui dans ce débat donne lieu à une profusion d'apostilles fantaisistes. Mes critiques visent des thèses comme celle-ci : il y a beaucoup de degrés d'entendement différents ; "comprendre" n'est pas un simple prédicat à deux places ; il y a même différentes sortes et différents niveaux d'entendement ; souvent la loi du tiers-exclu ne s'applique pas directement aux énoncés de forme "X comprend Y" ; dans de nombreux cas, la question de savoir si X comprend Y n'est pas une simple question de fait mais de décision ; etc.. Sur tous ces points, je dirais : bien sûr... mais ils n'ont rien à voir avec la question débattue. Il y a des cas où il est clair que "comprendre" s'applique littéralement et d'autres où il est clair qu'il ne s'applique pas : de tels cas sont tout ce dont j'ai besoin pour mon argumentation[1]. Je comprends des histoires en anglais, à un degré moindre des histoires en français, moindre encore en allemand. En chinois, pas du tout. Par ailleurs, ma voiture et ma machine à calculer ne comprennent rien : ce n'est pas leur affaire. Par métaphore et par analogie, nous attribuons souvent "comprendre" et d'autres prédicats cognitifs à des voitures, des calculatrices et autres produits de la technique mais de telles attributions ne prouvent rien. Nous disons "La porte *sait* quand elle doit s'ouvrir grâce à sa cellule photo-électrique, la calculatrice *sait comment* (*comprend comment* ou *est capable* de) faire l'addition et la soustraction, la thermostat *perçoit* des changements de température". La raison pour laquelle nous nous livrons à ces attributions est assez intéressante : elle se raporte au fait que dans les produits de la technique nous

prolongeons notre propre intentionnalité[2]. Nos outils sont des extensions de nos buts et dès lors, nous trouvons naturel de leur attribuer métaphoriquement de l'intentionalité, mais je crois que de tels exemples ne cassent rien, philosophiquement. Le sens dans lequel une porte automatique "comprend des instructions" de sa cellule photo-électrique n'est pas du tout celui dans lequel je comprends l'anglais. Si le sens dans lequel l'ordinateur programmé de Schank comprend des histoires est censé être le sens métaphorique de la porte qui comprend et non celui dans lequel je comprends l'anglais, le problème ne vaut pas la peine d'être discuté. Mais Newell et Simon écrivent que le type de cognition qu'ils revendiquent pour les ordinateurs est le même que pour les êtres humains. J'apprécie le caractère direct de cette affirmation ; tel est le type de proclamation que je prendrai en considération. Je veux montrer qu'au sens littéral l'ordinateur comprend ce que comprend la voiture et la calculatrice, à savoir : strictement rien. L'entendement de l'ordinateur n'est pas seulement partiel ou incomplet (comme pour l'allemand dans mon cas), il égale zéro.

Passons aux réponses :

- **La réponse systémique (Berkeley)**

"Bien qu'il soit vrai que la personne individuelle qui est enfermée dans la pièce ne comprenne pas l'histoire, elle n'est en fait qu'une partie d'un système global et celui-ci comprend vraiment l'histoire. La personne a sous les yeux un registre où les règles sont écrites, il dispose de beaucoup de papier de brouillon et de crayons pour faire ses calculs, il a des "banques de données" faites d'ensembles de symboles chinois. Aussi l'entendement n'est-il pas assigné au seul individu; au contraire, on l'assigne à tout le système dont il n'est qu'une partie".

Ma réponse à la théorie systémique est assez simple : faisons en sorte que les éléments du système soient internes à

l'individu. Il mémorise les règles du registre et les banques de données en symboles chinois ; il fait tous les calculs dans sa tête. Ainsi le système entier est incorporé à l'individu. Il n'est rien de tout le système qu'il n'intègre. Nous pouvons même nous débarasser de la pièce et imaginer qu'il travaille dehors. Cela revient au même : il ne comprend rien au chinois et, à fortiori, le système non plus, parce qu'il n'y a rien dans le système qui ne soit en lui. Si l'individu ne comprend pas, il n'y a aucune manière pour le système de comprendre car il n'est qu'une partie de l'individu.

En réalité, je me sens un peu gêné de donner même cette réponse à la théorie systémique qui me paraît trop invraisemblable dès le départ. L'idée est la suivante : alors qu'une personne ne comprend pas le chinois, d'une certaine façon, la *conjonction* de cette personne et de bouts de papiers pourrait le comprendre. Il m'est difficile d'imaginer que quelqu'un qui ne soit pas sous l'emprise d'une idéologie puisse trouver cette idée ne fût-ce que plausible. Pourtant, je crois que beaucoup de ceux qui sont engagés dans l'idéologie de l'I.A. forte pencheront finalement pour dire quelque chose qui y ressemble beaucoup. Aussi, poursuivons un peu plus loin. L'une des versions de cette opinion est celle-ci : dans l'exemple du système devenu interne, sans doute l'homme ne comprend-il pas le chinois à la façon d'un locuteur chinois d'origine (parce que, par exemple, il ignore que l'histoire renvoie à des restaurants, des hamburgers, etc.), pourtant "l'homme en tant que système de manipulation de symboles formels" *comprend réellement le chinois*. Il ne faudrait pas confondre deux sous-systèmes de cet homme : le premier est constitué du système de manipulation de symboles formels pour le chinois et le second est le sous-système pour l'anglais.

On a donc réellement deux sous-systèmes dans cet homme ; l'un comprend l'anglais, l'autre le chinois - "simplement, les deux systèmes n'ont pas grand-chose à voir l'un avec l'autre". Ici, je veux répliquer que non seulement ils n'ont pas grand-chose à voir l'un avec l'autre, mais qu'ils n'ont rien de semblable, même de très loin. Le sous-système qui comprend

l'anglais (en admettant que nous parlions un moment ce jargon "sous-systémique") sait que les histoires concernent les restaurants et la consommation de hamburgers, il sait qu'on lui pose des questions sur les restaurants et qu'il y répond du mieux qu'il peut en faisant diverses inférences à partir du contenu de l'histoire, etc... Par contre, le système chinois ne sait rien de tout ça. Alors que le sous-système anglais sait que les "hamburgers" renvoient à des hamburgers, le sous-système chinois sait seulement que "squiggle, squiggle" doit être suivi de "squoggle, squoggle"... Tout ce qu'il sait, c'est que divers symboles sont introduits d'un côté et manipulés selon des règles écrites en anglais, et que d'autres symboles sortent de l'autre côté. Toute la question de l'exemple de départ était l'argument selon lequel semblable manipulation de symboles ne pouvait suffire par elle-même à expliquer la compréhension du chinois dans un sens littéral ; en effet, l'homme pouvait arriver à écrire "squoggle, squoggle" à la suite de "squiggle, squiggle" sans rien comprendre au chinois. Cet argument n'est pas rencontré en postulant des sous-systèmes à l'intérieur de cet homme, car ceux-ci ne sont pas en meilleurs position que ne l'était l'homme dans le premier cas ; ils n'ont toujours rien qui ressemble à ce qu'a un locuteur anglais (ou son sous-système). En effet, dans le cas tel qu'on vient de le décrire, le sous-système chinois n'est qu'une partie du sous-système anglais, une partie qui entraîne la manipulation de symboles dépourvus de signification, selon des règles en anglais.

Demandons-nous ce qui est censé motiver la réponse systémique, en premier lieu ; autrement dit : quelles bases *indépendantes* doit-on supposer pour dire que l'agent doit être doté d'un sous-système interne qui, littéralement, comprend des histoires en chinois ? Pour autant que je puisse le voir, les seules bases sont que, dans mon exemple, j'ai les mêmes entrées et sorties que le locuteur chinois d'origine et un programme qui va des unes aux autres. Or le point visé par mes deux exemples a été d'essayer de montrer que cela ne pouvait suffire à expliquer la compréhension (au sens où je comprends des histoires en anglais) , parce qu'une personne et, de là, l'ensemble des systèmes qui concourent à constituer

une personne pourraient avoir la bonne combinaison d'entrée, de sortie et de programme sans comprendre quoi que ce soit, au sens littéral pertinent : celui où je comprends l'anglais. Le seul motif pour affirmer qu'il *doive* y avoir en moi un sous-système qui comprend le chinois, c'est que j'ai un programme et que je peux passer le test de Turing (je peux abuser des locuteurs chinois d'origine). Mais, précisément l'un des points en discussion est l'adéquation du test de Turing. L'exemple montre qu'il pourrait y avoir deux "systèmes" qui passent tous deux le test de Turing, mais dont un seul comprend. On ne répond pas à mon argument en disant que, puisqu'ils passent tous deux le test de Turing, l'un et l'autre doivent comprendre. Cette affirmation, en effet, ne rencontre pas l'argument selon lequel le système interne qui comprend l'anglais a, de loin, quelque chose en plus que le système de simple traitement du chinois. En bref, la réponse systémique fait une pétition de principe en affirmant sans argument que le système doit comprendre le chinois.

Qui plus est, la réponse systémique semble vouloir conduire à des conséquences indépendantes qui sont absurdes. S'il nous faut conclure qu'il doit y avoir cognition à partir du fait que j'ai une certaine sorte d'entrées et de sorties et un programme entre les deux, il semble que toutes sortes de systèmes non cognitifs vont s'avérer cognitifs. Par exemple, à un certain niveau de description, mon estomac fait du traitement de l'information et exemplifie ('instantiates') un certain nombre de programmes d'ordinateur ; pourtant, je ne pense pas que nous voulions soutenir qu'il a le moindre entendement (cf. Pylyshyn 1980). Mais si nous acceptons la réponse systémique, il est difficile d'éviter l'affirmation selon laquelle l'estomac, le coeur, le foie etc... sont tous des sous-systèmes qui comprennent, puisqu'on ne dispose pas de moyens de principe de distinguer le motif de l'affirmation d'entendement pour le sous-système chinois et pour l'estomac. En passant : on ne répondra pas à cette objection en disant que le système chinois a de l'information comme entrées et sorties, tandis que l'estomac n'a que de la nourriture et des produits de la nourriture. En effet, du point de vue de l'agent, de mon point de vue, il n'y a

pas plus d'information dans la nourriture que dans le chinois - celui-ci n'est qu'une suite de gribouillis sans signification. Il n'y a information dans le cas du chinois qu'aux yeux des programmeurs et des interprètes ; rien ne peut les empêcher de considérer les entrées et les sorties de mon estomac comme de l'information, s'ils le désirent.

Ce dernier point porte sur quelques problèmes indépendants de l'I.A. forte et il vaut la peine de digresser un moment pour l'expliquer. Si l'I.A. forte doit devenir une branche de la psychologie, elle doit devenir capable de distinguer les systèmes mentaux de ceux qui ne le sont pas. Elle doit être capable de distinguer les principes selon lesquels l'esprit fonctionne de ceux selon lesquels fonctionnent des systèmes non-mentaux. Autrement, elle ne nous offrira aucune explication de ce qui est spécifiquement mental dans le mental. De plus, la distinction mental/non-mental ne peut se faire seulement aux yeux de l'observateur, elle doit être intrinsèque aux systèmes : sinon, il dépendrait de n'importe quel observateur de traiter les gens comme non-mentaux et, par exemple, les ouragans comme mentaux, si ça lui plait. Pourtant, la littérature en I.A. brouille souvent la distinction d'une manière qui risque de s'avérer désastreuse à long terme pour ses prétentions à l'enquête cognitive. Mac Carthy, par exemple, écrit : "On peut dire de machines aussi simples que les thermostats qu'elles ont des croyances. Le fait d'avoir des croyances semble être une des caractéristiques de la plupart des machines susceptibles de résoudre des problèmes" (Mac Carthy 1979). Quiconque pense que l'I.A. forte a une chance comme théorie de l'esprit devrait peser les implications de cette remarque. On nous demande d'accepter comme une découverte de l'I.A. forte que le morceau de métal sur le mur, que nous utilisons pour régler la température, a des croyances, exactement au sens où nous-mêmes, nos épouses et nos enfants en ont ; de plus, que "la plupart" des autres machines dans la pièce - téléphone, enregistreur, calculatrice, interrupteur - ont aussi des croyances dans ce sens littéral. Le but de cet article n'est pas d'argumenter contre cette thèse de Mac Carthy, aussi me contenterai-je d'affirmer ce qui suit sans

argument. L'étude de l'esprit commence par des faits comme ceux-ci : les hommes ont des croyances, tandis que les téléphones et les calculatrices n'en ont pas. Si vous avez une théorie qui dément ces faits, vous avez produit un contre-exemple à la théorie de départ, qui est donc fausse. On a l'impression que les gens de l'I.A. qui écrivent le genre de choses citées pensent qu'ils peuvent s'épargner cette démarche parce qu'ils ne prennent pas ça vraiment au sérieux et ne pensent pas que quelqu'un le fera. Je propose, au moins pour un moment, de prendre ça au sérieux. Concentrez-vous une minute sur ce qui serait nécessaire pour établir que le morceau de métal, là sur le mur, a des croyances réelles, des croyances avec une direction d'ajustement, un contenu propositionnel et des conditions de satisfaction ; des croyances qui aient la possibilité d'être fortes ou faibles ; nerveuses, anxieuses ou paisibles, dogmatiques, rationnelles ou superstitieuses ; des fois aveugles ou des cogitations hésitantes; n'importe quelle croyance. Le thermostat n'est pas candidat. L'estomac, le foie, la calculatrice ou le téléphone non plus. Toutefois, puisque nous prenons cette idée au sérieux, remarquez que sa vérité serait fatale à la prétention de l'I.A. forte à être une science de l'esprit. Car à ce moment-là, l'esprit serait partout. Nous voulions savoir ce qui distingue l'esprit des thermostats et des foies ; si Mac Carthy avait raison, l'I.A. forte ne pourrait espérer nous le dire.

- **La réponse robotique (Yale)**

"Supposez que nos écrivions un programme différent de celui de Schank. Supposez que nous mettions un ordinateur dans un robot et que ce robot ne se contente pas de recevoir des symboles formels à l'entrée et d'en donner d'autres à la sortie, mais plutôt qu'il fasse quelque chose qui ressemble au plus près à percevoir, marcher, bouger, planter des clous, manger, boire - tout ce qu'on voudra. Par exemple, le robot aurait une caméra de télévision qui le rendrait capable de voir, des bras et des jambes pour "agir" et tout ceci serait contrôlé par son "cerveau" ordinateur. Un tel robot, à l'inverse de l'ordinateur

de Schank, aurait un entendement authentique et d'autres états mentaux".

La première chose à noter dans la réponse robotique, c'est qu'elle concède tacitement que la cognition n'est pas seulement une affaire de manipulaton de symboles formels. Cette réponse, en effet ajoute un ensemble de relations causales avec le monde extérieur (cf. Fodor 1980). Mais on opposera à la réponse robotique que l'addition de capacités "perceptuelles" et "motrices" n'ajoute rien au programme original de Schank en matière d'entendement, en particulier, ou d'intentionalité, en général. Pour le voir, remarquez que la même expérience imaginaire s'applique au cas du robot. Supposez que, plutôt que l'ordinateur dans le robot vous me mettiez dans la pièce et, comme pour l'exemple chinois de départ, vous me donniez des symboles chinois avec des instructions en anglais pour mettre certains symboles en correspondance avec d'autres et fournir des symboles chinois à la sortie. Supposez qu'à mon insu certains des symboles chinois que je reçois passent par une caméra de télévision attachée au robot et que d'autres symboles chinois que je sors servent à déclencher les moteurs qui font se mouvoir les bras et les jambes du robot. Il est important de souligner que tout ce que je fais, c'est de la manipulation de symboles formels : j'ignore tout des autres faits. Je reçois de "l'information" de l'appareil "perceptif" du robot et je donne des "instructions" à son appareil moteur sans connaître aucun de ces faits. Je suis l'homuncule du robot, mais à la différence de l'homuncule traditionnel, j'ignore ce qui se passe. Je ne comprends rien, à l'exception des règles de manipulation des symboles. Donc, dans cet exemple, je dirai que le robot n'a pas le moindre état intentionnel ; ses mouvements sont seulement un résultat de ses connexions électriques et de son programme. Bien plus, en mettant le programme en oeuvre ('instantiating'), je n'ai aucun état intentionnel d'un type pertinent. Tout ce que je fais se réduit à suivre des instructions formelles sur la manipulation de symboles formels.

- **La réponse du simulateur de cerveau (Berkeley et M.I.T.)**

"Supposez que nous concevions un programme qui ne représente pas notre information sur le monde comme dans le script de Schank, mais qui simule la séquence réelle des déclenchements neuronaux dans les synapses du cerveau d'un locuteur chinois d'origine, quand il comprend des histoires en chinois et y répond. La machine assimile des histoires et des questions en chinois comme entrées, elle simule sa structure formelle de cerveaux chinois réels dans le traitement de ces histoires et fournit des réponses chinoises comme sorties. Nous pouvons même imaginer que la machine ne travaille pas seulement à partir d'un seul programme séquentiel, mais avec tout un ensemble de programmes fonctionnant en parallèle d'une manière qu'on peut présumer être celle de cerveaux humains réels qui traitent du langage naturel. Dans un tel cas, il nous faudrait reconnaître que la machine a compris les histoires; si nous nous y refusions, ne faudrait-il pas dénier la compréhension des histoires au locuteur chinois d'origine ? Au niveau des synapses, qu'est ce qui serait, ou pourrait être, différent pour le programme d'ordinateur et le programme de cerveau chinois ?"

Avant de contrer cette réponse, je voudrais noter en passant qu'elle est bizarre, venant d'un partisan de l'intelligence artificielle (ou du fonctionnalisme, etc) : je pensais que toute l'idée de l'I.A. était qu'il n'était pas besoin de connaître le fonctionnement du cerveau pour connaître celui de l'esprit. L'hypothèse de base, du moins je le supposais, était l'existence d'un niveau d'opérations mentales consistant en processus de calculs ('computational') appliqués à des éléments formels ; ces opérations constitueraient l'essence du mental et pourraient se réaliser dans toutes sortes de processus cérébraux différents, de la même façon qu'un programme d'ordinateur peut être réalisé sur des hardwares différents. Selon les suppositions de l'I.A. forte, l'esprit est au cerveau comme le programme est au hardware et nous pouvons donc comprendre l'esprit sans faire de neurophysiologie. S'il nous fallait savoir comment fonctionne le cerveau pour faire de

l'I.A., on ne s'ocuperait pas d'I.A. Quoi qu'il en soit, il n'est pas suffisant de se rapprocher, même à ce point, des opérations cérébrales pour produire de l'entendement. Pour le voir, imaginez qu'à la place d'un unilingue occupé à mélanger des symboles dans une pièce, nous ayons quelqu'un qui opère sur un ensemble complexe de tuyaux d'eau munis de valves pour les connecter. Quand cet homme reçoit les symboles chinois, il consulte le programme écrit en anglais pour savoir quelles valves il doit ouvrir ou fermer. A chaque branchement d'eau correspond une synapse dans le cerveau chinois et tout le système est installé de manière à ce qu'en faisant tous les bons déclenchements (c'est-à-dire en ouvrant tous les bons robinets) les réponses chinoises jaillissent à la sortie de la tuyauterie.

Où est l'entendement dans un tel système ? Il prend du chinois comme entrée, il simule la structure formelle des synapses du cerveau chinois et il donne du chinois comme sortie. Mais l'homme ne comprend certainement pas le chinois et les tuyaux non plus. Si l'on est tenté d'adopter le point de vue que je tiens pour absurde selon lequel, d'une certaine façon, la *conjonction* de l'homme **et** des tuyaux comprend, qu'on se rappelle que l'homme peut en principe rendre interne la structure des tuyaux et faire tous les "déclenchements" dans son imagination. Le problème avec le simulateur de cerveau, c'est qu'il simule ce qui ne convient pas. Aussi longtemps qu'il ne simulera que la structure formelle des déclenchements neuronaux au niveau synaptique , il n'aura pas simulé ce qui compte avec le cerveau, à savoir ses propriétés causales, son aptitude à produire des états intentionnels. Que les propriétés formelles ne suffisent pas à expliquer les propriétés causales, c'est ce que montre l'exemple de la tuyauterie : toutes les propriétés formelles peuvent être détachées des propriétés causales pertinentes.

- **La réponse composite (Berkeley et Stanford)**

"Bien que chacune des trois réponses précédentes puisse ne

pas être convaincantes en elle-même pour réfuter le contre - exemple de la chambre chinoise, en les prenant toutes les trois ensemble elles le sont beaucoup plus et même, collectivement, elles sont décisives. Imaginez un robot avec un ordinateur en forme de cerveau logé dans la cavité cranienne ; imaginez que tout le comportement du robot soit impossible à distinguer du comportement humain et finalement pensez le tout comme un système unifié, pas seulement comme un ordinateur avec entrées et sorties. Il est sûr que dans un tel cas il nous faudra assigner de l'intentionalité au système".

J'admets que dans un tel cas nous trouverions rationnel et, en fait, irrésistible d'accepter l'hypothèse selon laquelle le robot a des états intentionnels, pour autant que nous n'en sachions rien de plus. En fait, à part l'apparence et le comportement, tous les autres éléments sont hors de propos. Si nous pouvions construire un robot dont le comportement serait impossible à distinguer d'une grande étendue de comportements humains, nous lui attribuerions l'intentionnalité en attendant une raison de ne pas le faire. Nous n'aurions pas besoin de savoir à l'avance que son cerveau -ordinateur est un analogue du cerveau humain.

Mais je ne vois vraiment pas en quoi ceci est d'aucun secours pour les prétentions de l'I.A. forte. Voici pourquoi : selon l'I.A. forte, l'instantiation d'un programme formel avec les bonnes entrées et sorties est une condition suffisante d'intentionnalité et, en fait, en est constitutive. Comme l'affirme Newell (1979), l'essence du mental est d'opérer sur un système de symboles physiques. Or dans cet exemple, les attributions d'intentionalité à l'égard du robot n'ont rien à voir avec des programmes formels. Elles sont simplement basées sur la supposition que si le robot a une apparence et un comportement suffisamment semblables aux nôtres, on pensera qu'il doit avoir des états mentaux semblables aux notres, qui causent son comportement et que celui-ci exprime ; on pensera qu'il doit être doté d'un mécanisme interne capable de produire de tels états mentaux. Si indépendamment nous savions comment rendre compte du comportement du robot

sans de telles suppositions, nous ne lui attribuerions pas d'intentionalité - spécialement si nous savions qu'il a un programme formel. Tel est précisément le point de ma réponse à l'objection 2.

Supposez que nous sachions que le comportement du robot est entièrement explicable par les faits suivants : un homme est placé à l'intérieur, il reçoit des symboles formels ininterprétés par l'intermédiaire des récepteurs sensoriels du robot et il envoie des symboles ininterprétés aux mécanismes moteurs, après manipulation des symboles en accord avec un groupe de règles. De plus, supposez que cet homme ignore tout de ces faits concernant le robot ; tout ce qu'il sait, c'est quelles opérations il doit exécuter sur des symboles dépourvus de signification. Dans un tel cas, nous considérerions le robot comme un mannequin mécanique ingénieux. L'hypothèse selon laquelle le mannequin aurait un esprit cesserait d'être garantie et nécessaire, car il n'y aurait plus de raison d'assigner de l'intentionalité au robot ou au système dont il fait partie (à l'exception, bien sûr, de l'intentionalité de l'homme qui manipule les symboles). Les manipulations de symboles ont lieu, les entrées et les sorties correspondent, mais l'unique site d'intentionalité est l'homme et celui-ci ne connait aucun des états intentionnels pertinents. Par exemple, il ne *voit* pas ce qui arrive aux yeux du robot, il n'a pas *l'intention* d'en faire se mouvoir les bras et il ne *comprend* aucune des remarques faites au robot ou émises par lui. Pour les raisons établies plus haut, il en va de même pour le système dont l'homme et le robot sont une partie.

Pour éclairer ce point, il suffit de contraster le cas avec ceux où il nous semble parfaitement naturel d'attribuer de l'intentionalité aux membres d'une autre espèce de primates comme les singes ou les chimpanzés, ou à des animaux domestiques comme les chiens. Nos raisons sont en gros au nombre de deux : le comportement de l'animal ne peut faire sens pour nous sans que nous lui attribuions de l'intentionnalité et nous pouvons constater que les bêtes sont faites d'un matériau semblable au nôtre - ça c'est un oeil, ça le

nez, voici sa peau, etc... A partir de la cohérence du comportement de l'animal et la supposition du même matériau causal qui lui serait sous-jacent, nous pensons, à la fois, que l'animal doit avoir des états mentaux qui sous-tendent son comportement et que ceux-ci doivent être produits par des mécanismes constitués à partir du même matériau que le nôtre. Nous ferions certainement les mêmes suppositions au sujet du robot tant que feraient défaut des raisons contraires, mais dès que nous saurions que son comportement est le résultat d'un programme formel et que les propriétés causales réelles de sa substance physique n'entrent pas en ligne de compte, nous abandonnerions la supposition d'intentionalité.

Deux autres ripostes à mon exemple sont venues fréquemment ; elles valent donc d'être discutées, mais sont complètement à côté de la question.

- **La réponse des autres esprits (Yale)**

"Comment savez-vous que d'autres personnes comprennent le chinois ou quoi que ce soit ? Seulement par leurs comportements. Quand l'ordinateur passe des tests comportementaux aussi bien que des personnes (en principe), si vous attribuez de la cognition à celles-ci vous devez en principe faire de même pour celui là ".

Cette objection ne mérite vraiment qu'une courte réponse. Le problème discuté ici n'est pas de savoir comment je sais que d'autres personnes ont des états cognitifs, mais plutôt ce que je leur attribue quand je leur attribue des états cognitifs. La pointe de mon argument est que ça ne peut se réduire à des processus de calcul ('computational') et leurs sorties, car les uns et les autres peuvent exister sans état cognitif. On ne répond pas à cet argument en feignant l'anesthésie. Les "sciences cognitives" présupposent la réalité et la connaissabilité du mental, de la même façon que les sciences physiques présupposent la réalité et la connaissabilité des objets physiques.

■ **La réponses des nombreuses demeures (Berkeley)**

"Toute votre argumentation présuppose que l'I.A. ne concerne que les ordinateurs analogiques et digitaux. Il se trouve que ce n'est là que l'état actuel de la technologie. Quels que soient ces processus causaux qui sont pour vous essentiels à l'intentionalité (à supposer que vous ayez raison), il se peut que nous devenions capables de construire des appareils dotés de ces procesus et nous aurons de l'intelligence artificielle. Aussi vos arguments ne sont-ils pas dirigés contre l'aptitude de l'intelligence artificielle à produire et expliquer la cognition".

Je n'ai vraiment aucune objection à cette réponse, sauf à dire qu'elle a pour effet de rendre trivial le projet de l'I.A. forte, en la redéfinissant comme tout ce qui produit et explique la cognition artificiellement. L'intérêt de la prétention originale de l'intelligence artificielle est qu'elle offrait une thèse précise et bien définie : les processus mentaux sont des processus de calcul ('computational') appliqués à des éléments définis. Mon problème était de mettre cette thèse au défi. Si l'on redéfinit la prétention de départ de telle sorte qu'elle ne comprend plus cette thèse, mes objections ne portent plus, car il n'y a plus d'hypothèses à tester sur lesquelles les appliquer.

Revenons maintenant à la question pour laquelle j'ai promis un essai de réponse. Etant admis que dans mon exemple je comprend l'anglais et pas le chinois, et que, pour cette raison, la machine ne comprend ni l'un ni l'autre, il doit bien y avoir en moi quelque-chose qui fait que je comprends l'anglais et il doit me manquer quelque-chose de correspondant pour comprendre le chinois. Quel que soit ce quelque-chose, pourquoi ne pourrions-nous pas le donner à une machine ?

En principe, je ne vois pas de raison qui nous empêche de donner à une machine la faculté de comprendre l'anglais ou le chinois, puisque dans un sens important nos corps et nos cerveaux sont précisément de telles machines. Mais je vois des arguments très forts pour dire que nous ne pourrions donner

une chose pareille à une machine dont les opérations sont seulement définies en termes de processus de calcul ('computational') appliqués à des éléments définis formellement ; c'est-à-dire dont les opérations sont définies comme l'instantiation d'un programme d'ordinateur. Ce n'est pas par ce que je suis l'instantiation d'un programme d'ordinateur que je suis capable de comprendre l'anglais et d'avoir d'autres formes d'intentionalité (je pense être l'instantiation d'un nombre indéfini de programmes d'ordinateurs). Pour autant que je sache, c'est parce que je suis une certaine forme d'organisme doté d'une certaine structure biologique (c'est-à-dire physico-chimique) et que cette structure, sous certaines conditions, a la capacité causale de produire de la perception, de l'action, de l'entendement, de l'apprentissage et d'autres phénomènes intentionnels. Un point de l'argument présent est que c'est seulement quelque-chose qui a ces capacités causales qui pourrait avoir cette intentionalité. Peut-être d'autres processus physico-chimiques pourraient-ils produire exactement ces effets-là ; peut-être, par exemple, les martiens ont-ils aussi de l'intentionalité, alors que leurs cerveaux sont taillés dans une autre étoffe. C'est là une question empirique, un peu comme celle de savoir si la photosynthèse peut être réalisée par quelque-chose dont la chimie diffère de celle de la chlorophylle.

Le point principal de l'argumentation est qu'aucun modèle purement formel ne suffira jamais par lui-même à rendre compte de l'intentionalité. En effet, les propriétés formelles, par elles-mêmes, ne sont pas constitutives d'intentionalité, elles n'ont pas de capacités causales, à l'exception du pouvoir (sous instantiation) d'amener l'étape suivante du formalisme quand la machine tourne. Toutes les autres propriétés causales dont sont dotées des réalisations particulières du modèle formel ne relèvent pas de lui, parce qu'il est toujours possible de mettre le même modèle formel dans une réalisation différente à laquelle ces propriétés causales feront clairement défaut. Même si par miracle les locuteurs chinois réalisaient exactement le programme de Schank, nous pourrions donner ce même programme à des anglophones, des tuyaux ou des

ordinateurs : ils ne comprendraient pas le chinois, nonobstant le programme.

Ce qui compte avec le cerveau, ce n'est pas l'ombre formelle portée par la séquence synaptique, mais bien plutôt les propriétés réelles des séquences. Tous les arguments que j'ai rencontrés en faveur de la version forte de l'intelligence artificielle s'appuient sur les dessins du contour des ombres portées par la cognition et l'affirmation que les ombres sont la chose même.

En guise de conclusion, j'essayerai d'établir quelques-uns des aspects philosophiques généraux qui sont implicites à mon argumentation. Pour plus de clarté, je tenterai de le faire sous forme de questions et réponses, en commencant par cette question éculée :

"Une machine pourrait-elle penser ?"

La réponse est, évidemment, oui. Nous sommes précisément de telles machines.

"Oui, mais un produit de la technique, une machine construite par l'homme pourrait-elle penser ?"

En supposant qu'il soit possible de produire artificiellement une machine dotée d'un système nerveux, de neurones avec dendrites et axones, et de tout le reste, qui nous ressemble suffisemment, la réponse semble évidemment devoir être encore positive. Si vous pouvez reproduire les causes avec exactitude, vous pouvez reproduire les effets. De plus, il pourrait en fait être possible de produire de la conscience, de l'intentionalité et tout le reste, en se servant de principes chimiques différents de ceux mis en oeuvre chez les êtres humains. C'est là, comme je l'ai dit, une question empirique.

"D'accord. Mais un ordinateur digital pourrait-il penser ?"

Si par "ordinateur digital" nous entendons, en tout et pour tout, quelque chose qui, à un certain niveau de description, peut se définir comme l'instantiation d'un programme d'ordinateur, encore une fois la réponse sera, bien sûr, oui. Car nous sommes les instantiations d'un nombre indéfini de programmes d'ordinateur et nous pensons.

"Mais quelque-chose pourrait-il penser, comprendre, etc... en vertu du seul fait d'être un ordinateur doté du bon programme ? Etre l'instantiation d'un programme (le bon, bien sûr) pourrait-il constituer, par soi-même, une condition suffisante de l'entendement ?"

Voilà, à mon avis, la bonne question, habituellement confondue avec une ou plusieurs de celles qui précèdent. Sa réponse est non.

"Pourquoi pas ?"

Parce que les manipulations de symboles formels n'ont par elles-mêmes aucune intentionalité ; elles sont assez bien dépourvues de signification ; elles ne sont même pas des manipulations de *symboles*, puisque ces symboles ne symbolisent rien. En jargon linguistique, ils n'ont qu'une syntaxe et pas de sémantique. L'intentionalité dont semble dotés les ordinateurs est seulement dans l' esprit de ceux qui les programment et les utilisent, de ceux qui font les entrées et interprètent les sorties.

L'exemple de la chambre chinoise visait à en faire la démonstration en montrant que dès qu'on introduit dans le système quelque-chose qui a de l'intentionalité (un homme) et qu'on le programme de manière formelle, il apparaît que le programme formel n'apporte pas d'intentionalité supplémentaire. Par exemple, il n'ajoute rien à l'aptitude d'un

homme à comprendre le chinois.

Précisément cette caractéristique de l'I.A. qui semblait si prometteuse - la distinction entre le programme et la réalisation - s'avère fatale à l'affirmation selon laquelle une simulation pourrait être une reproduction. La distinction entre le programme et sa réalisation dans le hardware semble parallèle à la distinction entre le niveau des opérations mentales et le niveau des opérations cérébrales. Si nous pouvions décrire le niveau des opérations mentales comme un programme formel, il semble que nous pourrions décrire ce qui est essentiel à l'esprit sans avoir recours ni à la psychologie introspective, ni à la neurophysiologie du cerveau. Mais l'équation : "l'esprit est au cerveau comme le programme est au hardware", est boîteuse en plusieurs points, parmi lesquels les trois suivants :

Premièrement, la distinction entre programme et réalisation a pour conséquence qu'un même programme pourrait avoir toutes sortes de réalisations folles qui n'auraient aucune forme d'intentionalité. Weizenbaum (1976, Ch. 2) par exemple, montre en détail comment construire un ordinateur à partir d'un rouleau de papier de toilette et une pile de petits cailloux. De même, le programme de compréhension du chinois peut être réalisé par une tuyauterie, un ensemble d'éoliennes ou un locuteur anglais unilingue, dont aucun n'acquiert par là la compréhension du chinois. Des cailloux, du vent, du papier de toilette et des tuyaux ne sont pas le bon matériau pour mettre en évidence l'intentionalité : c'est seulement quelque-chose de doté des mêmes facultés causales que le cerveau qui peut produire de l'intentionalité. Bien que le locuteur anglais soit le bon matériau pour l'intentionalité, on voit facilement qu'il n'en acquiert pas de supplémentaire en mémorisant le programme, puisque ceci ne lui apprend pas le chinois.

Deuxièmement, le programme est purement formel, alors que les états intentionnels ne sont pas formels de cette façon. Ils sont définis dans les termes de leur contenus, pas de leurs formes. Par exemple, croire qu'il pleut ne se définit pas par une certaine configuration formelle, mais par un certain

contenu mental avec des conditions de satisfaction, une direction d'ajustement (Searle, 1979) et des choses semblables. En fait, la croyance comme telle n'a pas même de configuration formelle au sens syntaxique, puisqu'une seule et même croyance peut s'exprimer dans un nombre indéfini d'expressions syntaxiques différentes de système linguistiques différents

Troisièmement, comme je l'ai mentionné plus haut, des états mentaux et des évènements sont littéralement un produit des opérations du cerveau, mais le programme n'est pas, dans ce sens, un produit de l'ordinateur.

> *"Bien. Si les programmes ne sont en aucune façon constitutifs des processus mentaux, pourquoi tellement de gens ont-ils cru le contraire ? Ceci au moins mérite une explication".*

Je ne connais pas réellement la réponse à cette question. L'idée selon laquelle les simulations par ordinateur pourraient être la chose même aurait dû d'emblée éveiller la suspicion, car l'ordinateur n'est absolument pas confiné à la simulation d'opérations mentales. Personne n'imagine que les simulations sur ordinateur d'un incendie catastrophique va réduire en cendre le voisinage, ni que la simulation d'une pluie de tempête nous laissera trempés. Pourquoi donc quelqu'un irait-il imaginer que la simulation sur ordinateur de l'entendement comprend quoi que ce soit ? On dit parfois qu'il serait terriblement difficile de faire ressentir de la douleur à un ordinateur ou de le faire tomber amoureux. Pourtant, amour et douleur ne sont ni plus ni moins difficiles que la cognition ou que n'importe quoi. Pour simuler, tout ce qu'il faut ce sont les bonnes entrées et sorties avec un programme entre les deux qui transforme les unes en les autres. C'est tout ce dont dispose l'ordinateur pour tout ce qu'il est appelé à faire. En confondant la simulation et la reproduction, on commet la même erreur, qu'il s'agisse de douleur, d'amour, de cognition, d'incendie ou de pluie de tempête.

Toutefois, il y a plusieurs raisons qui ont pu faire croire - ou font peut-être encore croire à beaucoup - que l'I.A. reproduisait en quelque sorte et, par là, expliquait les phénomènes mentaux. Je crois que nous ne réussirons pas à ôter ces illusions tant que nous n'aurons pas pleinement exposé les raisons qui leur ont donné naissance.

Premièrement, le point peut-être le plus important est une confusion sur la notion de "traitement de l'information" : en sciences cognitives, nombreux sont ceux qui croient que le cerveau humain, avec son esprit, fait quelque chose qu'on appelle "traitement de l'information" et que, de manière analogue, l'ordinateur avec son programme en fait aussi. Par contre, ni les incendies, ni les pluies de tempête ne font de "traitement de l'information". Ainsi, bien que l'ordinateur puisse simuler les traits formels de n'importe quel processus, il entretient une relation particulière avec l'esprit et le cerveau : en effet, l'ordinateur correctement programmé (dans l'idéal avec le même programme que le cerveau) traiterait l'information comme le cerveau et ce traitement constituerait l'essence du mental. Le problème avec cet argument est qu'il s'appuie sur l'ambiguité de la notion d'"information". L'ordinateur programmé ne fait pas de "traitement de l'information" au sens où les gens en font quand ils réfléchissent sur, disons, des problèmes d'arithmétique ou quand ils lisent et répondent à des questions sur des histoires. En fait, il se limite à manipuler des symboles formels. Le fait que pour le programmeur et l'interprète des sorties les symboles utilisés valent pour des objets du monde est totalement hors de portée pour l'ordinateur. Répétons-le : l'ordinateur a une syntaxe et pas de sémantique. Donc, si l'on tape sur le clavier de l'ordinateur "2 + 2 = ?", il sortira "4", mais sans avoir la moindre idée du fait que "4" signifie 4 ou signifie quoi que ce soit. Le problème n'est pas qu'il lui manque une information du second degré pour interpréter ses symboles du premier degré, mais bien que ses symboles du premier degré n'ont pas d'interprétation du tout tant qu'il s'agit de l'ordinateur. Tout ce que l'ordinateur aurait de plus, ce serait encore des symboles. L'introduction de la notion de

"traitement de l'information" produit donc un dilemme : soit nous construisons cette notion de manière à ce que l'intentionalité y soit impliquée, soit nous ne le faisons pas. Dans le premier cas, l'ordinateur ne fait pas de traitement de l'information mais seulement de la manipulation de symboles formels. Dans le second cas, l'ordinateur fait du traitement de l'information mais seulement au sens où les calculatrices, les machines à écrire, les estomacs, les pluies de tempêtes et les ouragans en font; à savoir, il existe un niveau de description où nous pouvons les envisager comme recevant de l'information d'un côté, la transformant et produisant de l'information à la sortie. Mais ici c'est à l'observateur qu'il incombe d'interpréter l'entrée et la sortie comme de l'information au sens ordinaire. On n'a pas établi de similarité entre l'ordinateur et le cerveau, qui soit basée sur une similarité dans le traitement de l'information.

Deuxièmement, on rencontre en I.A. de nombreux restes de behaviorisme ou d'opérationnalisme. Puisque les ordinateurs peuvent avoir des modèles d'entrée et de sortie semblables à ceux des être humains, on est tenté de postuler des états mentaux dans les ordinateurs comme dans les êtres humains. Mais une fois reconnue la possibilité conceptuelle et empirique de donner à un système des capacités humaines dans certains domaines, sans avoir la moindre intentionalité, on devrait arriver à vaincre cette tentation. Ma calculatrice de bureau a des capacités de calcul, mais pas d'intentionalité ; dans cet article, j'ai essayé de montrer qu'un système pourrait avoir des capacités d'entrées et de sorties qui reproduisent celles d'un locuteur chinois d'origine, sans pour autant comprendre le chinois - quelle que soit la manière dont on le programme. Le test de Turing est typique de cette tradition qui s'affirme behavioriste et opérationnaliste sans vergogne. Je crois qu'en rejettant totalement ces deux courants, les chercheurs en I.A. élimineraient beaucoup de la confusion qui règne entre simulation et reproduction.

Troisièmement, ce reste d'opérationnalisme se joint à un reste de dualisme d'une certaine forme ; en effet, l'I.A. forte ne peut

faire sens que sur base de la supposition dualiste qui veut que, là ou l'esprit est concerné, le cerveau n'a pas d'importance. En I.A. forte (et pour le fonctionnalisme aussi bien) ce sont les programmes qui comptent et ceux-ci sont indépendants de leurs réalisations dans des machines. En fait, tant qu'il s'agit d'I.A., le même programme pourrait être réalisé sur une machine électronique, une substance mentale cartésienne ou un esprit du monde hégelien. J'ai fait une découverte des plus surprenantes en discutant de ces problèmes : beaucoup de chercheurs de l'I.A. sont plutôt choqués par mon idée de dépendance entre les phénomènes mentaux réels chez l'homme et les propriétés physico-chimiques réelles des cerveaux humains réels. Pourtant, en y pensant une minute, je n'aurais pas dû être surpris; car si l'on n'adopte pas une forme de dualisme, le projet d'I.A. forte est voué à l'échec. Ce projet est de reproduire et d'expliquer le mental en confectionnant des programmes. Mais tant qu'on n'affirme pas l'indépendance non seulement conceptuelle mais aussi empirique de l'esprit par rapport à la matière, on ne peut mener à bien ce projet, car le programme est complètement indépendant de n'importe quelle réalisation. A moins de tenir l'esprit pour séparable du cerveau à la fois conceptuellement et empiriquement - version forte du dualisme - on ne peut espérer reproduire le mental en rédigeant et faisant tourner des programmes, puisque ceux-ci doivent être indépendants des cerveaux et de toute autre forme particulière d'instantiation. Si les opérations mentales consistent en opérations de calcul ('computational') sur des symboles formels, il s'ensuit qu'elles n'ont pas de connexion intéressante avec le cerveau. La seule connexion serait que le cerveau se trouve être l'une de ces sortes de machines en nombre indéfini qui peuvent instantier le programme. Cette forme de dualisme n'est pas la même que le cartésianisme traditionnel qui affirme l'existence de deux sortes de *substances*, mais elle est cartésienne en soutenant que ce qui est spécifiquement mental dans l'esprit n'a pas de connexion intrinsèque avec les propriétés effectives du cerveau. Ce dualisme sous-jacent est masqué par le fait que la littérature en I.A. fulmine abondamment contre le "dualisme" ; ces auteurs ne semblent pas s'apercevoir que leur position présuppose une

version forte de dualisme.

"Une machine pourrait-elle penser ?" A mes yeux, *seule* une machine peut penser. En fait, seules des machines d'un genre très spécial, à savoir les cerveaux et les machines douées des mêmes capacités causales que ceux-ci. Si l'I.A. forte nous en apprend si peu sur la pensée, c'est principalement parce qu'elle n'a rien à dire des machines. Selon sa propre définition, elle s'occupe de programmes et les programmes ne sont pas les machines. L'intentionalité est quelque chose d'autre. Quelle qu'elle soit, c'est un phénomène biologique et elle est soumise à un rapport de dépendance causale à l'égard de la biochimie spécifique de ses origines, tout autant que la lactation, la photosynthèse ou n'importe quel autre phénomène biologique. Personne n'irait imaginer qu'il est possible de produire du lait ou du sucre rien qu'en faisant tourner sur ordinateur une simulation des séquences formelles de la lactation ou de la photosynthèse. Pourtant, quand il s'agit de l'esprit, nombreux sont ceux qui veulent croire à un tel miracle, à cause d'un dualisme profond et persistant; ils pensent que l'esprit est affaire de processus formels et qu'il est indépendant de causes matérielles tout-à-fait spécifiques, semblables à celles dont dépendent le lait et le sucre.

Pour défendre un tel dualisme on se raccroche souvent à l'idée selon laquelle le cerveau est un ordinateur digital (d'ailleurs, les premiers ordinateurs étaient souvent appelés des "cerveaux électroniques"). Mais c'est sans espoir. A l'évidence le cerveau est un ordinateur digital. Puisque toute chose est un ordinateur digital, le cerveau aussi. Le problème est que la capacité causale du cerveau à produire de l'intentionalité ne peut consister en l'instantiation d'un programme d'ordinateur. En effet, pour tous les programmes qu'il vous plaira, il est possible de trouver quelque chose qui les instantie sans pour autant avoir aucun état mental. Quoi que fasse le cerveau pour produire de l'intentionalité, ça ne peut consister en l'instantiation d'un programme, car aucun programme n'est par lui-même condition suffisante d'intentionalité[3].

R E F E R E N C E S

Fodor,J.A. (1980) Methodogical solipsism considered as a research strategy in cognitive psychology. *The Behavioral and brain sciences, 3.*

Mac Carthy, J. (1979) Ascribing mental qualities to machines. In : *Philosophical perspectives in artificial intelligence*, ed. M. Ringle. Atlantic Highlands, N.J. Humanities Press.

Newell, A. (1979) *Physical symbols systems.* Lecture at the La Jolla Conference on Cognitive Science.

Pykyshyn, Z.W. (1980) Computation and cognition : issues in the foundations of cognitive sciences. *The Behavioral and brain sciences, 3.*

Schank, R.C. and Abelson, R.P. (1977) *Scripts, plans, goals and understanding.* Hilldale, N.J. : Lawrence Erlbaum Press

Searle, J.R. :
- (1979 a) Intentionality and the use of language. In : *Meaning and use* , ed. A. Margalit. Dordrecht : Reidel.
- (1979 b) The intentionality of intention and action . *Inquiry*, 22 : 253-80
- (1979 c) What is an international state ? *Mind* 88 : 74-92.

Weizenbaum, J. (1965) Eliza - a computer program for the study of natural language communication between man and machine, *Communication of the association for computing machinery* 9 : 36-45. Et (1976) *Computer power and human reason* . San Francisco : W.H. Freeman.

Winograd, T. (1973) A procedural model of language understanding. In : *Computer models of thought and language*, ed. R. Schank and R. Colby. San Francisco : Freeman.

* *Minds, brains and programs*, The behaviorial and Brain Sciences © Sept, 1980, Vol. 3, n° 3.

1. Signalons aussi que "comprendre" implique à la fois la possession d'états mentaux (intentionnels) et la vérité (la validité, le succès) de ces états. Pour les besoins de la discussion, nous nous limitons à la possession d'états mentaux.

2. L'intentionalité est par définition cette caractéristique de certains états mentaux, par laquelle ceux-ci sont dirigés vers (ou concernent) des objets ou des états de choses dans le monde. Ainsi, les croyances, les désirs et les intentions sont des états intentionnels ; des formes non orientées d'anxiété ou de dépression n'en sont pas.

3. Je suis redevable à l'égard de nombreuses personnes de discussions sur ces matières et de leurs tentatives de vaincre mon ignorance de l'intelligence artificielle. J'aimerais spécialement remercier Ned Block, Hubert Dreyfus, John Haugeland, Roger Schank, Terry Wilensky et Robert Winograd .

NOTE DU TRADUCTEUR

John R. Searle est professeur de philosophie à l'université de Berkeley. Ce sont, avant tout, ses travaux de philosophie du language qui l'ont rendu célèbre : Les Actes de langage, trad. H. Pauchard, Hermann, Paris, 1972 ; et Sens et expression, Minuit, Paris 1982. Cependant, Searle écrit : "A la base de mon approche des problèmes du langage, il y a l'hypothèse fondamentale que la philosophie du langage est une branche de la philosophie de l'esprit /'mind', les états mentaux/" (L'Intentionalité, p.9). Deux ouvrages récemment traduits offrent donc les lignes de force de cette théorie des états mentaux qui sert de fondement à sa philosophie du langage : L'Intentionalité, trad. de C. Pichevin, Minuit, Paris, 1985 et Du Cerveau au savoir, trad. de C. Chaleyssin, Hermann, Paris, 1985. On aura compris que c'est par ce développement second de son oeuvre que Searle croise la question de l'Intelligence Artificielle. Il y consacre le dernier chapitre de L'Intentionalité et elle traverse Du cerveau au savoir dans son ensemble.

Avant d'en arriver là, Searle aura fait mûrir son point de vue sur l'I.A. dans une série d'articles dont le premier est présenté ici en traduction : "Minds, Brains, and Programs", The Behavioral and Brain Sciences, vol. 3, 1980 ; "Intrinsic Intentionality", ibid. ; "Analytic philosophy and mental phenomena", Midwest Studies in Philosophy, vol.5, 1981 ; "The myth of the computer", New York Review of Books, vol. XXIX, n°7, 1982. Ces multiples approches sont bien dans la manière de Searle : "Il me paraît très utile de mettre des idées à l'épreuve, sous une forme préliminaire, à la fois pour les formuler et pour susciter commentaires et critiques. Ces articles sont comme les esquisses préparatoires d'un artiste en vue d'une toile de plus grandes dimensions" (L'Intentionalité, p.12). La traduction que nous présentons est donc bien un document : d'abord parce qu'il constitue la première esquisse du point de vue de Searle sur l'I.A., ensuite parce que son retentissement fut considérable - il a suscité 28 réponses de spécialistes lors de sa première publication et figure dans de nombreuses anthologies relatives à l'I.A. ou aux sciences cognitives.