In Vues de l'esprit, éd. par D. Hofstadter et D. Dennett.
InterEditions 1987.

JOHN SEARLE

Esprits, cerveaux et programmes

Quelle importance psychologique et philosophique devrions-nous attacher aux récents efforts de simulation des capacités cognitives humaines sur ordinateur? Avant de répondre à cette question, il me paraît utile de faire la distinction entre ce que j'appelerai l'IA (Intelligence Artificielle) « forte » et l'IA « faible », ou « prudente ». D'après l'IA faible, la principale valeur de l'ordinateur dans l'étude de l'esprit, c'est qu'il est pour nous un outil très puissant. Ainsi, il nous permet de formuler et de tester des hypothèses de façon plus rigoureuse et plus précise. D'après l'IA forte, en revanche, l'ordinateur n'est pas simplement un outil d'étude de l'esprit; l'ordinateur convenablement programmé est véritablement un esprit, en ce sens que des ordinateurs munis des bons programmes *comprennent* et ont d'autres états cognitifs. En IA forte, l'ordinateur programmé ayant des états cognitifs, les programmes ne sont pas simplement des outils nous permettant de tester des explications; ils sont, eux-mêmes, les explications.

Je n'ai rien à objecter aux thèses de l'IA faible, tout au moins dans les limites de cet article. Ce qui m'intéresse, ici, ce sont les affirmations de l'IA forte telles que je les ai définies, et plus particulièrement celle selon laquelle l'ordinateur spécialement programmé a littéralement des états cognitifs et ces programmes expliquent donc le processus cognitif humain. Dans la suite de cet article, quand je parlerai d'IA, j'entendrai donc sa version forte, décrite par ces deux affirmations.

Je traiterai des travaux de Roger Schank et de ses collègues à Yale

« Minds, Brains, and Programs » de John Searle, extrait de *The Behavioral and Brain Sciences,* vol. 3.

(Schank et Abelson, 1977), car je les connais mieux que tous les autres suivant la même orientation, et parce qu'ils sont un exemple très clair du type de recherche que j'ai envie d'examiner. Néanmoins, rien de ce qui suit ne repose sur les détails des programmes de Schank. Les mêmes arguments seraient vrais de SHRDLU, de Winograd (Winograd, 1973), d'ELIZA, de Weizenbaum (Weizenbaum, 1965), et de toute simulation des phénomènes mentaux humains sur une machine de Turing. [En ce qui concerne ces références, voir les « Autres lectures ».]

Sans entrer dans le détail, on peut décrire le programme de Schank comme suit : l'objectif du programme est de simuler la capacité des humains à comprendre des histoires. Une des caractéristiques des humains en matière de compréhension des histoires, c'est qu'ils peuvent répondre à des questions sur un récit même si les informations qu'il donne n'ont pas été explicitement exposées. Supposons, par exemple, que l'on vous raconte l'histoire suivante : « Un homme est entré dans un restaurant et a commandé un steak. Quand le steak lui a été servi, il était réduit à l'état de boulet de charbon, et le client furieux est parti sans payer. » Si l'on vous demandait « Le client a-t-il mangé le steak? » vous répondriez probablement « Non ». De même, si l'on vous raconte : « Un homme est entré dans un restaurant et a commandé un steak ; on lui a servi le steak et il en a été très content; quand il a quitté le restaurant, il a donné un gros pourboire à la serveuse avant de payer l'addition. » On vous demande alors : « A-t-il mangé le steak? »; vous répondrez probablement « Oui ». Eh bien, les machines de Schank peuvent fournir le même genre de réponses sur des histoires de restaurants. Pour cela, elles ont une « représentation » du type d'informations qu'ont les humains concernant les restaurants, ce qui leur permet, pour des histoires similaires, de répondre à des questions comme les précédentes. Quand l'histoire est fourme à la machine et qu'on lui pose la question, elle imprime des réponses du genre de celles que devraient donner des humains à qui l'on aurait raconté des histoires semblables. Les partisans de l'IA forte prétendent que dans cette séquence de questions-réponses, la machine ne se borne pas à simuler une capacité humaine, mais que l'on peut dire qu'elle *comprend* l'histoire et fournit les réponses aux questions, et que ce que font la machine et son programme *explique* la capacité humaine de comprendre l'histoire et de répondre à des questions la concernant.

Ces affirmations ne me semblent absolument pas prouvées par les résultats de Schank, et je vais essayer de le montrer dans ce qui suit. Je ne prétends évidemment pas que Schank lui-même soutiendrait ces affirmations.

Une façon de tester toute théorie de l'esprit consiste à se demander ce qui se passerait si mon esprit fonctionnait réellement selon les principes qui, d'après la théorie considérée, régissent tous les esprits. Appliquons ce test au programme de Schank avec l'expérience mentale (« Gedankenexperiment ») suivante. Supposons que je sois enfermé dans une chambre et que l'on me donne une masse de texte chinois. Supposons également (ce qui est le cas) que je ne connaisse pas le

chinois, ni écrit ni parlé, et que je ne sois même pas certain de pouvoir distinguer l'écriture chinoise de, par exemple, l'écriture japonaise ou de n'importe quels signes cabalistiques. Pour moi, l'écriture chinoise n'est justement constituée que de signes cabalistiques incompréhensibles. Supposons maintenant qu'après ce premier lot d'écrits, on me donne un deuxième lot de texte chinois avec un ensemble de règles de corrélation entre le second et le premier lot. Ces règles sont en anglais, et je les comprends aussi bien que n'importe quelle autre personne de langue maternelle anglaise. Elles me permettent de faire le lien entre deux ensembles de symboles formels, l'adjectif « formel » signifiant ici que je peux identifier les symboles uniquement par leurs formes. Supposons encore que l'on me donne un troisième lot de symboles chinois accompagné d'instructions, toujours en anglais, me permettant d'établir un rapport entre des éléments de ce troisième lot et les deux premiers lots, et que ces règles m'indiquent comment produire certains symboles chinois ayant certaines formes en échange de certaines formes qui m'ont été remises dans le troisième lot. Ce que je ne sais pas, c'est que ceux qui me donnent tous ces symboles appellent le premier lot une « écriture », le deuxième une « histoire » et le troisième des « questions ». Quant aux symboles que je leur rends en échange du troisième lot, ils les appellent des « réponses aux questions », l'ensemble de règles en anglais constituant le « programme ». Maintenant, pour compliquer un peu l'histoire, imaginez que ces personnes me donnent également des histoires en anglais, que je comprends donc, et qu'elles me posent ensuite des questions en anglais sur ces histoires, questions auxquelles je réponds en anglais. Imaginons aussi qu'au bout d'un moment, je suive les instructions de manipulation des symboles chinois avec une telle maestria, et que les programmeurs écrivent les programmes avec un tel brio que d'un point de vue externe — c'est-à-dire du point de vue de quelqu'un qui se trouverait hors de la pièce dans laquelle je suis enfermé — mes réponses aux questions soient absolument indiscernables de celles que donneraient des Chinois. Quelqu'un qui lirait mes réponses ne pourrait pas se rendre compte que je ne parle pas un traître mot de chinois. Imaginons encore que mes réponses aux questions en anglais soient, ce qui serait sans aucun doute le cas, indiscernables de celles d'autres personnes de langue maternelle anglaise, pour la simple raison que je suis de langue maternelle anglaise. D'un point de vue externe, c'est-à-dire du point de vue de quelqu'un qui lirait mes « réponses », les réponses aux questions chinoises et aux questions anglaises seraient tout aussi bonnes. Or, dans le cas du chinois, j'ai produit les réponses uniquement en manipulant des symboles formels non interprétés. En ce qui concerne le chinois, je me suis comporté comme un ordinateur : j'ai accompli des opérations de calcul sur des éléments formellement définis. Et je ne suis qu'un équivalent concret du programme.

Les prétentions de l'IA forte sont, comme nous l'avons vu, que l'ordinateur programmé comprend les histoires et que, d'une certaine façon, le programme explique la compréhension humaine. Nous sommes maintenant en mesure de les examiner à la lumière de notre expérience mentale.

1. En ce qui concerne la première prétention, il me semble parfaitement évident, dans l'exemple, que je ne comprends rien aux histoires chinoises. J'ai des entrées et des sorties identiques à celles d'une personne de langue maternelle chinoise, et je peux avoir n'importe quel programme formel, mais je ne comprends rien. Pour les mêmes raisons, l'ordinateur de Schank ne comprend rien aux histoires qu'on lui fournit, qu'elles soient en chinois, en anglais, ou autre, puisque dans le cas du chinois, l'ordinateur, c'est moi, et dans les cas où l'ordinateur n'est pas moi, il n'a rien de plus que moi quand je ne comprends rien.

2. Pour ce qui est de la deuxième prétention, selon laquelle le programme éclaire la compréhension humaine, nous pouvons voir que l'ordinateur et son programme ne fournissent pas assez d'éléments de compréhension, puisqu'ils fonctionnent tous les deux mais qu'il n'y a pas pour autant compréhension. Apportent-ils même un élément nécessaire ou quelque importante contribution à la compréhension ? Une des affirmations des partisans de l'IA forte est que quand je comprends une histoire en anglais, je fais exactement la même chose — et peut-être même plus — que quand je manipulais les symboles chinois. La manipulation de symboles est simplement plus formelle avec l'anglais, que je comprends, qu'avec le chinois, que je ne comprends pas. Je n'ai pas démontré que la deuxième prétention est fausse, mais elle paraîtrait certainement incroyable dans le cas de l'exemple. Le peu de plausibilité de cette thèse dérive de la supposition que nous pouvons écrire un programme qui aura les mêmes entrées et sorties que des locuteurs natifs, en présumant, de plus, qu'à un certain niveau, ces locuteurs fonctionnent également comme un programme. Sur la base de ces deux présomptions, nous devons penser que même si le programme de Schank n'est pas l'illustration totale de la compréhension, il l'est peut-être en partie. Je veux bien supposer que c'est une possibilité empirique, mais nul n'a jusqu'ici donné la moindre raison de le croire, puisque ce qui est suggéré par l'exemple — quoique certainement pas démontré — est que le programme d'ordinateur n'a simplement rien à voir avec ma compréhension de l'histoire. Dans le cas du chinois, j'ai tout ce que l'intelligence artificielle peut me fournir sous forme de programme, et je ne comprends rien; dans le cas de l'anglais, je comprends tout, et il n'y a jusqu'ici aucune raison de supposer que ma compréhension ait quoi que ce soit à voir avec des programmes d'ordinateur, c'est-à-dire avec des opérations de calcul sur des éléments définis de façon purement formelle. Tant que le programme est établi en termes d'opérations de calculs sur des éléments purement formels, notre exemple suggère que ceux-ci n'ont, en eux-mêmes, aucun rapport intéressant avec la compréhension. Ils ne sont certainement pas des conditions suffisantes, et il n'y a aucune raison de supposer qu'ils puissent être des conditions nécessaires, ni même qu'ils apportent une importante contribution à la compréhension. Remarquez bien que la force de l'argument ne réside pas simplement dans le fait que différentes machines peuvent avoir les mêmes entrées et les mêmes sorties tout en fonctionnant selon des principes formels différents; c'est tout à fait à côté du sujet. Il montre plutôt que, quels que

soient les principes purement formels que vous introduisez dans un ordinateur, ils ne suffiront pas pour qu'il y ait compréhension, puisqu'un humain peut suivre les principes formels sans rien comprendre. Rien n'a jamais permis de supposer que ces principes soient nécessaires, ni même utiles, puisque rien ne laisse penser que lorsque je comprends l'anglais, je mets en œuvre un quelconque programme formel.

Mais alors, qu'ai-je dans le cas des phrases anglaises que je n'ai pas dans celui des phrases chinoises? La réponse évidente est que je sais ce que signifient les premières, alors que je n'ai pas la moindre idée de ce que veulent dire les deuxièmes. Mais en quoi consiste cette connaissance, et pourquoi ne pourrait-on pas la donner à une machine, quelle qu'elle soit? Je reviendrai plus loin sur cette question, mais je voudrais d'abord poursuivre avec l'exemple.

J'ai eu l'occasion de le présenter à plusieurs chercheurs en intelligence artificielle et, chose intéressante, ils ne semblent pas d'accord sur la réponse à ma question. J'ai eu droit à une surprenante gamme de réponses, et je traiterai des plus courantes d'entre elles (classées par origines géographiques) dans ce qui suit.

Avant de poursuivre, je voudrais toutefois faire obstacle à certaines erreurs de compréhension concernant le terme « compréhension ». Dans bon nombre de ces réponses, on note un grand flottement sur le sens de ce mot. Mes critiques soulignent qu'il y a beaucoup de degrés différents de compréhension; que la « compréhension » n'est pas un simple prédicat bivalent; qu'il y a même différents types et différents niveaux de compréhension, et que bien souvent, la loi du tout l'un ou tout l'autre ne peut même pas s'appliquer directement à des énoncés du type « *x* comprend *y* »; que souvent, il faut décider si *x* comprend y, et non pas seulement se rendre à cette évidence; et ainsi de suite. Sur tous ces points, je réponds : bien sûr. Mais ils passent totalement à côté du problème. Il y a des exemples clairs auxquels le terme « compréhension » s'applique très bien, et des exemples clairs auxquels il ne s'applique pas ; et ces deux types d'exemples sont tout ce dont j'ai besoin aux fins de cette discussion*. Je comprends des histoires en anglais, à un moindre degré des histoires en français, à un degré encore moindre des histoires en allemand, mais en chinois, rien du tout. Ma voiture et ma calculatrice, elles, ne comprennent rien ; ce n'est pas leur rayon. Or, nous attribuons souvent la capacité de « comprendre » et d'autres capacités cognitives, par métaphore et analogie, à des voitures, des calculatrices et d'autres objets, mais cette attribution ne prouve rien. Nous disons : « La porte *sait* quand elle doit s'ouvrir grâce à sa cellule photoélectrique », « La calculatrice *sait comment (comprend comment, peut)* faire une addition et une soustraction, mais pas une division », et « Le thermostat *perçoit* des changements de température ». La raison de ces attributions est très intéressante, et elle est liée au fait que nous étendons notre intentionna-

* En outre, la « compréhension » implique à la fois la possession d'états mentaux (intentionnels) et la vérité (validité, réussite) de ces états. Nous ne sommes concernés, aux fins de cette discussion, que par la possession de ces états.

lité aux objets* : nos outils étant des extensions de nos desseins, nous trouvons tout naturel de leur attribuer métaphoriquement une intentionnalité; je pense toutefois que ces exemples sont sans valeur philosophique. Le sens selon lequel une porte automatique « comprend les instructions » provenant de sa cellule photoélectrique n'est pas du tout celui selon lequel je comprends l'anglais. Si les ordinateurs programmés de Schank sont censés comprendre des histoires comme la porte comprend qu'elle doit s'ouvrir, et non pas comme je comprends l'anglais, la discussion est alors dénuée d'intérêt. Mais Newell et Simon (1963) écrivent que le genre de connaissance qu'ont, selon eux, les ordinateurs, est exactement le même que celui des humains. J'aime l'aplomb de cette affirmation, et c'est le type d'affirmation dont je vais discuter. J'avancerai qu'au sens littéral, l'ordinateur programmé comprend autant que la voiture et la calculatrice, c'est-à-dire exactement rien. La compréhension de l'ordinateur n'est pas seulement (comme ma compréhension de l'allemand) partielle ou incomplète ; elle est nulle.

Venons-en maintenant aux réponses :

1. La réponse du système (Berkeley). « Il est vrai que l'individu enfermé dans la pièce ne comprend pas l'histoire, mais il n'est qu'une partie d'un système global, et c'est ce système qui comprend l'histoire. La personne a devant elle un grand cahier dans lequel sont écrites les règles, elle dispose d'une grande quantité de papier brouillon et de crayons pour faire des calculs, elle a des "banques de données" d'ensembles de symboles chinois. La compréhension n'est pas le fait de l'individu seul, mais du système global dont il fait partie. »

Ma réponse à la théorie du système est très simple : imaginons que l'individu intériorise tous ces éléments du système. Il apprend par cœur les règles inscrites dans le cahier et les banques de données de symboles chinois, et il fait tous les calculs de tête. Il inclut alors tout le système. Il n'y a rien, dans le système, qui lui soit extérieur. On peut même éliminer la pièce et supposer qu'il travaille dehors. Il ne comprend néanmoins toujours pas le chinois, et le système non plus, puisque tout ce qui constitue le système est en lui. S'il ne comprend pas, le système ne pourrait en aucune façon comprendre, puisqu'il est devenu partie intégrante de l'individu.

Je dois avouer que je me sens un peu embarrassé de donner cette réponse à la théorie du système, qui me paraît tellement peu plausible. L'idée qu'elle énonce est qu'alors qu'une personne ne comprend pas le chinois, la *conjonction* de cette personne et de bouts de papier pourrait le comprendre. Je vois mal comment quelqu'un qui n'est pas sous l'emprise d'une idéologie pourrait trouver cette idée vraisemblable. Je crois toutefois que bien des personnes séduites par l'idéologie de l'IA forte finiront par soutenir quelque chose de ce genre ; allons donc un peu plus

* L'intentionnalité est, par définition, cette caractéristique de certains états mentaux qui fait qu'ils sont dirigés vers des états et des situations du monde. Les convictions, les désirs et les intentions sont donc des états intentionnels ; certaines formes non dirigées d'anxiété et de dépression n'en sont pas.

loin. Selon une des versions de cette théorie, si l'homme de l'exemple du système intériorisé ne comprend pas le chinois au sens où un Chinois le comprend (parce que, par exemple, il ne sait pas que l'histoire parle de restaurants, de steaks, etc.), « l'homme en tant que système de manipulation de symboles formels », lui, *comprend vraiment le chinois.* Il ne faut pas confondre le sous-système de l'homme qu'est le système de manipulation de symboles formels pour le chinois avec le sous-système de traitement de l'anglais.

Il y a donc vraiment deux sous-systèmes dans cet homme ; l'un qui comprend le chinois, l'autre l'anglais; et « les deux systèmes n'ont simplement pas grand-chose à voir l'un avec l'autre ». Je répondrai que non seulement ils n'ont pas grand-chose à voir l'un avec l'autre, mais ils n'ont strictement rien de semblable. Le sous-système qui comprend l'anglais (admettons que nous adoptions un petit moment ce jargon des « sous-systèmes ») sait que les histoires concernent des restaurants et des steaks, il sait qu'on lui pose des questions sur les restaurants et qu'il répond de son mieux aux questions par inférences à partir du contenu de l'histoire, etc. Le système chinois, lui, ne sait rien de tout ça. Si le sous-système anglais sait que les « steaks » concernent des steaks, le sous-système chinois sait seulement que « gribouillis » est suivi de « gribouillas ». Tout ce qu'il sait, c'est que divers symboles formels sont introduits d'un côté et manipulés selon des règles écrites en anglais, et que d'autres symboles sortent de l'autre côté. L'idée de l'exemple d'origine était de montrer que ce genre de manipulation de symboles ne peut pas, seul, suffire à comprendre le chinois parce que l'homme concerné pourrait écrire « gribouillas » après « gribouillis » sans rien comprendre de cette langue. Et cet argument n'est pas renversé par celui des sous-systèmes, car ceux-ci ne sont pas plus malins que l'homme ne l'était au départ; ils sont toujours loin d'avoir les facultés de l'homme (ou du sous-système) parlant anglais. En fait, dans le cas décrit, le sous-système chinois n'est qu'une partie du sous-système anglais, une partie occupée à manipuler des symboles dépourvus de sens selon des règles écrites en anglais.

Pour commencer, demandons-nous ce qui est censé motiver la réponse du système, c'est-à-dire quelles raisons *indépendantes* permettraient de dire que l'acteur doit avoir en lui un sous-système qui comprend les histoires en chinois. Pour moi, la seule raison est que dans l'exemple donné, j'ai les mêmes entrées et les mêmes sorties que des gens de langue maternelle chinoise et un programme entre les deux. Mais les exemples visaient précisément à montrer que cela ne suffirait pas pour qu'il y ait compréhension au sens où je comprends des histoires en anglais, parce qu'une personne, et donc l'ensemble des systèmes constituant une personne, pourrait avoir la bonne combinaison d'entrées, de sorties et de programme, et ne pas comprendre pour autant au sens pertinent où je comprends l'anglais. Nous pourrions dire qu'il *doit* y avoir en moi un sous-système qui comprend le chinois uniquement parce que j'ai un programme et que je peux réussir le test de Turing; je peux tromper des Chinois. Mais l'adéquation du test de Turing est précisément un des points litigieux. L'exemple montre qu'il pourrait y avoir deux « systè-

mes », capables tous les deux de réussir le test de Turing, mais dont un seul comprendrait; et il ne serait pas valable de dire que puisqu'ils réussissent tous les deux le test de Turing, ils doivent tous les deux comprendre, car cela irait à l'encontre de la thèse selon laquelle le système qui, en moi, comprend l'anglais, est bien plus puissant que celui qui ne fait que traiter des symboles chinois. Bref, la réponse du système se borne à affirmer sans justification que le système doit comprendre le chinois.

De plus, cette réponse semble entraîner inévitablement des conséquences qui sont indépendamment absurdes. En effet, s'il faut conclure que je dois avoir des capacités cognitives parce que j'ai un certain type d'entrées et de sorties et un programme entre les deux, alors bon nombre de systèmes non cognitifs risquent de se voir également attribuer ces capacités. Par exemple, à un certain niveau de description, mon estomac traite des informations et exécute des programmes d'ordinateur, et pourtant, je suis sûr que nul n'accepterait de dire qu'il comprend (cf. Pylyshyn, 1980). En revanche, si l'on accepte la réponse du système, je vois mal comment on pourrait éviter de dire que l'estomac, le cœur, le foie, etc., sont tous des sous-systèmes doués de compréhension, puisqu'il n'y a pas, en théorie, de raison de dire que le sous-système chinois comprend sans dire que l'estomac comprend. Il ne serait par ailleurs pas acceptable d'objecter que le système chinois a des informations sous forme d'entrées et de sorties, alors que l'estomac ne reçoit et ne rejette que des aliments et des produits alimentaires, puisque du point de vue de l'acteur, c'est-à-dire de mon point de vue, il n'y a d'informations ni dans les aliments ni dans le chinois, le chinois n'étant qu'une suite de signes totalement dépourvus de sens. Il n'y a d'informations, dans le cas du chinois, qu'aux yeux des programmeurs et de ceux qui interprètent les sorties de l'ordinateur, et rien ne les empêche, s'ils le veulent, de traiter les entrées et sorties de mon système digestif comme des informations.

Ce dernier point touche quelques autres problèmes de l'IA forte et vaut bien par conséquent une petite digression. Si l'IA forte veut être une branche de la psychologie, elle doit être capable de faire la distinction entre les systèmes authentiquement mentaux et ceux qui ne le sont pas. Elle doit pouvoir distinguer les principes selon lesquels fonctionne l'esprit de ceux selon lesquels fonctionnent les systèmes non mentaux, sans quoi elle n'éclairera nullement notre lanterne quant à ce qu'il y a de spécifiquement mental dans le mental. Et la différence mental/non mental ne doit pas être perçue seulement par une personne extérieure ; elle doit être inhérente aux systèmes, sans quoi, n'importe quelle personne extérieure pourrait dire à sa guise que les humains sont non mentaux et, par exemple, que les ouragans sont des systèmes mentaux. Or, dans la littérature des spécialistes de l'IA, cette distinction est bien souvent gommée d'une façon qui, à long terme, pourrait ébranler très fortement la prétention que l'IA est une recherche cognitive. McCarthy, par exemple, écrit : « On peut dire de machines aussi simples que les thermostats qu'elles ont des convictions, et les convictions semblent être une caractéristique de la plupart des machines capables de résoudre des

problèmes » (McCarthy, 1979). Ceux qui croient que l'IA forte pourrait percer en tant que théorie de l'esprit devraient réfléchir aux implications de cette remarque. On nous demande d'accepter comme une découverte de l'IA forte que le morceau de métal fixé au mur que nous utilisons pour réguler la température a des convictions tout comme nous, nos épouses et nos enfants, nous en avons et, qui plus est, que « la plupart » des autres machines se trouvant dans la même pièce — le téléphone, le magnétophone, la calculatrice, l'interrupteur — ont aussi des convictions dans ce sens littéral. Le but de cet article n'étant pas de réfuter la thèse de McCarthy, je me contenterai d'exposer platement ce qui suit. L'étude de l'esprit est fondée sur des postulats disant, par exemple, que les humains ont des convictions, mais que les thermostats, les téléphones et les calculatrices n'en ont pas. Si vous trouvez une théorie le reniant, vous produisez un contre-exemple à la théorie, et la théorie est donc fausse. On a l'impression que les chercheurs en IA qui écrivent ce genre de choses peuvent aller contre ces postulats parce qu'ils ne prennent pas la question au sérieux et qu'ils pensent que personne ne le fera. Je propose, pour un petit moment au moins, que nous, nous la prenions au sérieux. Pensez très fort, pendant environ une minute, à ce qu'il faudrait pour établir que ce morceau de métal au mur a de véritables convictions, des convictions assorties d'une direction d'intention, d'un contenu propositionnel, et de conditions de satisfaction ; des convictions qui puissent être fermes ou indécises; des convictions timides, angoissées, ou sûres; des convictions dogmatiques, rationnelles, ou pleines de superstition; des croyances aveugles ou des cogitations hésitantes ; bref, toutes sortes de convictions. Avec le thermostat, ça ne colle pas. Pas plus qu'avec l'estomac, le foie, la calculatrice, ou le téléphone. Notons toutefois, puisque nous prenons cette idée au sérieux, que sa vérité serait fatale à la prétention de l'IA forte d'être une science de l'esprit. Car l'esprit serait partout. Ce que nous voulions savoir, c'est ce qui différencie l'esprit des thermostats et des foies. Et si McCarthy avait raison, ce n'est certainement pas l'IA forte qui nous apporterait la réponse.

2. La réponse du robot (Yale). « Supposons que nous écrivions un programme d'un type différent de celui de Schank. Supposons que nous placions un ordinateur à l'intérieur d'un robot, cet ordinateur ne se contentant pas de recevoir des symboles formels en entrée et de donner des symboles en sortie, mais actionnant véritablement le robot de telle sorte que celui-ci perçoive, marche, se promène, plante des clous, mange, boive, etc., ou tout au moins qu'il fasse quelque chose de ressemblant. Ce robot pourrait être équipé d'une caméra de télévision qui lui permettrait de voir, de bras et dejambes lui permettant d'"agir", le tout étant commandé par son "cerveau" ordinateur. Il serait, contrairement à l'ordinateur de Schank, vraiment capable de comprendre et aurait d'autres états mentaux. »

La première chose qui ressort de la réponse du robot, c'est qu'elle reconnaît tacitement que la cognition n'est pas seulement une affaire de

manipulation de symboles, puisque cette réponse ajoute tout un ensemble de relations causales avec le monde extérieur (cf. Fodor, 1980). La réponse à cet argument est néanmoins que l'addition de ces capacités « perceptuelles » et « motrices » n'apporte rien du point de vue de la compréhension en général, ou de l'intentionnalité en particulier, au programme d'origine de Schank. Pour vous en rendre compte, reprenez notre expérience mentale pour l'appliquer au robot. Imaginez qu'au lieu de l'ordinateur à l'intérieur du robot, vous me mettez dans une pièce et que, comme dans le cas d'origine du chinois, vous me donnez des quantités de symboles chinois et des quantités d'instructions en anglais selon lesquelles je dois associer des symboles chinois à des symboles chinois et renvoyer des symboles chinois à l'extérieur. Supposons qu'à mon insu, certains des symboles chinois que je reçois proviennent d'une caméra de télévision montée au robot et que certains des symboles chinois que je rends servent à actionner les moteurs internes du robot qui commandent les mouvements de ses jambes ou de ses bras. Il est important de bien se rappeler que je ne fais que manipuler des symboles formels; je ne sais rien du reste. Je reçois des « informations » de l'appareil de « perception » du robot et je donne des « instructions » à son appareil moteur en ignorant totalement ces deux faits. Je suis l'homoncule du robot, mais à la différence de l'homoncule traditionnel, je ne sais pas ce qui se passe. Je ne comprends rien d'autre que les règles de manipulation des symboles. Eh bien, dans ce cas, le robot n'a aucun état intentionnel; il agit uniquement en fonction de son câblage électrique et de son programme. Quant à moi, en tant qu'équivalent physique du programme, je n'ai aucun état intentionnel du type voulu. Je me borne à suivre des instructions formelles relatives à la manipulation de symboles formels.

3. La réponse du simulateur de cerveau (Berkeley et MIT). « Supposons que nous écrivions un programme qui ne représente pas des informations sur le monde, comme les informations des textes de Schank, mais qui simule la véritable séquence d'excitation des neurones, au niveau des synapses, d'une personne de langue maternelle chinoise quand celle-ci comprend des histoires en chinois et donne des réponses concernant ces histoires. La machine reçoit, en entrée, des histoires chinoises et des questions à leur sujet, simule la structure formelle de véritables cerveaux chinois traitant ces histoires, et donne, en sortie, des réponses en chinois. On peut imaginer que la machine fonctionne non pas avec un seul programme sériel, mais avec tout un lot de programmes tournant en parallèle, comme les cerveaux humains le font sans doute quand ils traitent le langage naturel. Dans un tel cas, nous devrions certainement dire qu'une telle machine comprendrait les histoires ; et si nous refusions de l'admettre, ne devrions-nous pas nier la compréhension des histoires également aux Chinois? Au niveau des synapses, qu'est-ce qui serait ou pourrait être différent entre le programme de l'ordinateur et le programme du cerveau chinois ? »

Avant de contrer cette réponse, je ferai une petite digression pour faire

remarquer qu'il est curieux, pour un partisan de l'intelligence artificielle (ou du fonctionnalisme, etc.) d'avancer un tel argument : je pensais que la grande idée de l'IA forte, c'était justement que l'on n'a pas besoin de savoir comment fonctionne le cerveau pour savoir comment l'esprit fonctionne. L'hypothèse fondamentale, à ce que je croyais, était qu'il y a un niveau d'opérations mentales fait de processus de calcul sur des éléments formels, et que ces processus constituent l'essence du mental et peuvent être réalisés de toutes sortes de façons par le cerveau, de même qu'un programme d'ordinateur peut tourner dans différents matériels d'ordinateurs : d'après les hypothèses de l'IA forte, l'esprit est au cerveau ce que le programme est au matériel, ce qui permet de comprendre l'esprit sans étudier la neurophysiologie. Si nous devions savoir comment fonctionne le cerveau pour faire de l'IA, nous n'aurions rien à faire de l'IA. De toute façon, ce n'est pas parce qu'on s'approchera d'aussi près du fonctionnement du cerveau que l'on produira forcément la compréhension. Pour bien saisir ce point, imaginez que notre homme monolingue, au lieu de mélanger des symboles dans une pièce, ait à contrôler un réseau de canalisations d'eau avec des vannes. Quand il reçoit des symboles chinois, il consulte le programme, rédigé en anglais, pour savoir quelles vannes il doit ouvrir et fermer. Chacun des raccords entre les canalisations correspond à un synapse du cerveau chinois, et le système est monté de telle sorte qu'une fois que tous les bons neurones se sont excités, c'est-à-dire une fois que tous les bons robinets ont été ouverts, les réponses chinoises sortent à l'extrémité de la série de tuyaux.

Où est la compréhension, dans ce système? Il reçoit le chinois en entrée, simule la structure formelle des synapses du cerveau chinois et donne du chinois en sortie. Or, l'homme ne comprend absolument pas le chinois, les canalisations non plus, et si vous êtes tenté d'adopter le point de vue, à mon avis absurde, qui consiste à dire que la conjonction de l'homme et des canalisations comprend, rappelez-vous qu'en théorie, l'homme peut intérioriser la structure formelle des tuyaux et produire toutes les « excitations neurales » dans son imagination. Le problème, avec le simulateur de cerveau, c'est qu'il se trompe d'objet cérébral à simuler. Tant qu'il ne simule que la structure formelle de la séquence d'excitations des neurones au niveau des synapses, il ne simulera pas l'important, c'est-à-dire les propriétés causales du cerveau, sa capacité de produire des états intentionnels. Et l'exemple des canalisations montre bien que les propriétés formelles ne suffisent pas pour avoir les propriétés causales : on peut très bien exciser toutes les propriétés formelles des propriétés causales neurobiologiques pertinentes.

4. La réponse de la combinaison (Berkeley et Stanford). « Si chacune des trois réponses précédentes ne suffit peut-être pas, à elle seule, pour réfuter le contre-exemple du chinois, les trois ensemble sont collectivement beaucoup plus convaincantes, voire implacables. Imaginez un robot dont la cavité crânienne serait occupée par un ordinateur en forme de cerveau, cet ordinateur contenant dans son programme tous les synapses d'un cerveau humain. Imaginez aussi que le comportement du

robot soit absolument identique à celui d'un être humain, et voyez l'ensemble comme un système unifié et non pas comme un ordinateur avec des entrées et des sorties. Dans un tel cas, on dirait certainement que le système est intentionnel. »

Je suis entièrement d'accord sur le fait que dans un tel cas, nous trouverions rationnel, et même irrésistible, de dire que le robot, tel qu'il a été décrit, a une intentionnalité. En fait, seuls comptent vraiment l'apparence et le comportement; les autres éléments de la combinaison sont totalement inutiles. Si l'on pouvait fabriquer un robot dont le comportement serait en grande partie indifférenciable de celui d'un humain, nous dirions, jusqu'à preuve du contraire, qu'il a une intentionnalité. Nous n'aurions pas besoin de savoir que son cerveau-ordinateur est formellement analogue à un cerveau humain.

Mais je ne vois vraiment pas en quoi cela soutient les thèses de l'IA forte, et voici pourquoi : d'après l'IA forte, la réalisation concrète d'un programme formel ayant les entrées et les sorties appropriées est une condition suffisante pour qu'il y ait intentionnalité, et constitue même l'intentionnalité. Comme dit Newell (1979), l'essence du mental est le fonctionnement d'un système physique de symboles. Mais quand nous attribuons l'intentionnalité au robot de cet exemple, ça n'a rien à voir avec les programmes formels. Nous partons simplement de l'hypothèse que si le robot nous ressemble et se comporte de façon suffisamment identique à nous, ça doit signifier, jusqu'à preuve du contraire, qu'il doit avoir des états mentaux comme les nôtres, qui causent son comportement et sont en même temps exprimés par celui-ci, et qu'il doit avoir un mécanisme interne capable de produire ces états mentaux. Si nous pouvions expliquer son comportement sans passer par ce genre d'hypothèse, nous ne lui attribuerions pas d'intentionnalité, surtout sachant qu'il contenait un programme formel. C'est justement le cœur de ma réponse à l'objection 2.

Supposons que nous sachions que le comportement du robot s'expliquait entièrement par le fait qu'un homme, à l'intérieur, recevait des symboles formels non interprétés transmis par les récepteurs sensoriels du robot et envoyait des symboles formels non interprétés à ses mécanismes moteurs, et que cet homme effectuait cette manipulation de symboles selon un ensemble de règles. Supposons, de plus, que l'homme ne sache rien du robot et se borne à savoir quelles opérations il doit exécuter sur quels symboles dépourvus de sens. Nous considérerions alors le robot comme une ingénieuse contrefaçon mécanique. L'hypothèse selon laquelle ce faux a un esprit deviendrait injustifiable et inutile, car il n'y aurait plus de raison de prêter l'intentionnalité au robot ou au système dont il fait partie (à l'exception de l'intentionnalité de l'homme lorsqu'il manipule les symboles). Les manipulations de symboles formels continuent, les entrées et les sorties sont correctes, mais le seul véritable centre d'intentionnalité est l'homme, qui ne sait rien des états intentionnels pertinents ; ainsi, il ne *voit* pas ce qui passe sous les yeux du robot, il ne *veut* pas faire bouger les bras du robot, et il ne *comprend* aucune des remarques faites par le robot ou à celui-ci. Et, pour les raisons déjà

exposées, le système dont l'homme et le robot font partie ne voit, ne veut et ne comprend rien de plus.

Pour mieux me suivre sur ce point, comparez cet exemple à d'autres exemples dans lesquels il nous paraît tout à fait naturel d'attribuer l'intentionnalité à des membres de certaines autres espèces de primates comme les singes et à des animaux domestiques comme les chiens. Les raisons pour lesquelles cela nous semble naturel sont essentiellement au nombre de deux : nous ne pouvons pas comprendre le comportement de l'animal sans supposer qu'il agit intentionnellement, et nous voyons que les animaux sont faits à peu près comme nous : ils ont des yeux, un nez, de la peau, etc. Par suite de la cohérence du comportement de l'animal et de l'hypothèse de l'existence du même matériau sous-jacent, nous supposons à la fois que le comportement de l'animal doit résulter d'états mentaux et que ces états mentaux doivent être produits par des mécanismes de même nature que les nôtres. Nous ferions certainement les mêmes suppositions au sujet d'un robot, mais dès que nous apprendrions que son comportement est le résultat d'un programme formel et que les propriétés causales de sa matière physique n'ont rien à voir avec ce comportement, nous abandonnerions l'hypothèse d'intentionnalité.

Deux autres réponses sont fréquemment opposées à mon exemple (et valent donc la peine d'être considérées), mais elles passent totalement à côté de la question.

5. La réponse des autres esprits (Yale). « Comment savez-vous que les autres gens comprennent le chinois, ou quoi que ce soit d'autre? Seulement d'après leur comportement. Or, l'ordinateur peut réussir les tests de comportement aussi bien qu'eux (en théorie) ; donc, si vous attribuez des facultés cognitives aux autres gens, vous devriez, en principe, en faire autant pour les ordinateurs. »

Cette objection ne mérite pas plus qu'une brève réponse. La question, ce n'est pas comment je sais que d'autres personnes ont des états cognitifs, mais plutôt ce que je leur attribue quand je pense qu'elles ont des états cognitifs. Le cœur de cet argument, c'est que les opérations de calcul et les sorties de celles-ci ne suffisent pas, à elles seules, parce que ces opérations et leurs sorties peuvent exister sans l'état cognitif. Et je n'accepte pas qu'on feigne d'ignorer cette réponse en faisant le mort intellectuellement. Dans le domaine des « sciences cognitives », on présuppose la réalité et l'identifiabilité du mental, tout comme en sciences physiques, on doit présupposer la réalité et l'identifiabilité des objets physiques.

6. La réponse des multiples maisons (Berkeley). « Votre argument présuppose que l'IA ne concerne que des ordinateurs analogiques et numériques. C'est effectivement là qu'en est la technologie. Quelles que soient les propriétés causales qui, d'après vous, sont essentielles à l'intentionnalité (en supposant que vous ayez raison sur ce point), nous finirons, un jour, par être en mesure de fabriquer des machines qui auront ces propriétés causales, et ce sera l'intelligence artificielle. Vos

arguments ne remettent donc absolument pas en question la capacité de l'intelligence artificielle à produire et expliquer la cognition. »

Je n'ai rien à objecter à cette réponse, si ce n'est qu'elle enlève tout sens au projet de l'IA forte en le redéfinissant corne tout ce qui produit et explique artificiellement la cognition. L'intérêt de la déclaration d'intention originelle de l'intelligence artificielle, c'est que c'était une thèse claire et nette : les processus mentaux sont des opérations de calcul sur des éléments formellement définis. Et moi, j'étais opposé à cette thèse. Mais si ce n'est plus la thèse de l'IA, mes objections ne sont plus valables, parce qu'elles ne concernent plus une hypothèse vérifiable.

Revenons-en maintenant à la question à laquelle j'ai promis d'essayer de répondre : étant donné que dans mon exemple d'origine, je comprends l'anglais mais pas le chinois, et considérant donc que la machine ne comprend ni l'anglais ni le chinois, il doit y avoir quelque chose en moi qui fait que je comprends l'anglais et, inversement, quelque chose de semblable qui me fait défaut et me rend incapable de comprendre le chinois. Mais pourquoi ne pourrions-nous pas donner ces quelques choses, quelles qu'elles soient, à une machine?

Je ne vois aucune raison de principe pour laquelle nous ne pourrions pas donner à une machine la capacité de comprendre l'anglais ou le chinois, puisque dans un sens important, nos corps, avec nos cerveaux, constituent précisément des machines de ce genre. Mais je vois des arguments militant fortement contre l'attribution de ces choses à une machine, tant que le fonctionnement de la machine sera défini uniquement en termes d'opérations de calcul sur des éléments formellement définis, c'est-à-dire tant que le fonctionnement de la machine sera défini comme la matérialisation d'un programme d'ordinateur. Ce n'est pas parce que je suis l'équivalent physique d'un programme d'ordinateur que je comprends l'anglais et que j'ai d'autres formes d'intentionnalité (je suis, je présume, l'équivalent physique d'un grand nombre de programmes), mais pour autant que nous le sachions, c'est parce que je suis un certain type d'organisme ayant une certaine structure biologique (c'est-à-dire chimique et physique), et que dans certaines conditions, cette structure est capable, selon des principes de causalité, de produire la perception, l'action, la compréhension, l'apprentissage, et d'autres phénomènes intentionnels. Et un des points importants de ce débat, c'est que seule une chose ayant ces capacités causales pourrait avoir cette intentionnalité. D'autres processus chimiques et physiques pourraient peut-être engendrer les mêmes effets; il se peut, par exemple, que les Martiens aient Pintentionnalité, mais que leurs cerveaux ne soient pas faits de la même matière. C'est là une question empirique, comme celle consistant à savoir si la photosynthèse peut être réalisée par une substance ayant une composition chimique différente de celle de la chlorophylle.

Le point principal du débat, c'est toutefois qu'aucun modèle purement formel ne sera jamais suffisant, à lui seul, pour qu'il y ait intentionnalité, car les propriétés formelles ne constituent pas, à elles seules,

l'intentionnalité, et n'ont aucune capacité causale à l'exception de celle, lorsqu'elles sont traduites sous une forme concrète, d'engendrer l'étape suivante du formalisme quand la machine fonctionne. Et toute autre propriété causale inhérente à des réalisations concrètes particulières du modèle formel est sans rapport avec le modèle formel parce que l'on peut toujours mettre le même modèle formel dans une réalisation dont ces propriétés causales seront de toute évidence absentes. Même si, par quelque miracle, des personnes de langue maternelle chinoise peuvent exécuter le programme de Schank, on peut introduire le même programme dans des personnes de langue anglaise, des canalisations d'eau ou des ordinateurs, qui ne comprennent pas le chinois, en dépit du programme.

Ce qui importe, en ce qui concerne les opérations du cerveau, ce n'est pas l'ombre formelle projetée par l'enchaînement de synapses, mais plutôt les propriétés de ces enchaînements. Tous les arguments en faveur de la version forte de l'intelligence artificielle que j'ai rencontrés s'obstinent à vouloir tracer un trait autour des ombres projetées par la cognition et à proclamer ensuite que ces ombres sont le cœur du problème.

Pour conclure, je voudrais essayer d'exposer quelques-unes des questions philosophiques générales implicites dans ce débat. Par souci de clarté, j'essaierai d'adopter la forme question-réponse. Je commencerai par l'inévitable question :

« Une machine pourrait-elle penser? »

La réponse est, de toute évidence, oui. Nous sommes précisément des machines pensantes.

« Oui, mais un objet fabriqué, une machine créée par l'homme, pourrait-elle penser? »

En supposant qu'il soit possible de produire artificiellement une machine ayant un système nerveux (c'est-à-dire des neurones avec des axones, des dendrites, et tout le reste) suffisamment semblable au nôtre, la réponse semble alors à nouveau clairement être oui. Si l'on peut reproduire exactement les causes, on peut reproduire les effets. Et il pourrait bien être possible de produire la conscience, l'intentionnalité, etc. en utilisant des types de principes chimiques différents de ceux des humains. C'est, comme je l'ai déjà dit, une question empirique.

« D'accord, mais est-ce qu'un ordinateur numérique pourrait penser? »

Si par « ordinateur numérique », nous désignons un objet quelconque comportant un niveau permettant de le décrire correctement comme une forme concrète de programme d'ordinateur, alors la réponse est encore oui, puisque nous sommes les formes concrètes d'un nombre indéterminé de programmes d'ordinateur et que nous pensons.

« Mais serait-il possible que quelque chose pense, comprenne, etc. pour la seule raison qu'il s'agit d'un ordinateur doté du bon type de programme? Le fait d'être l'équivalent concret d'un programme, d'un programme du type adéquat, bien sûr, pourrait-il suffire pour qu'il y ait compréhension? »

C'est là, à mon avis, la bonne question à poser, mais elle est généralement confondue avec une ou plusieurs des questions précédentes, et la réponse est non.

« Pourquoi? »

Parce que les manipulations de symboles formels n'ont aucune intentionnalité; elles n'ont aucun sens; en fait, ce ne sont même pas des manipulations de symboles, puisque les symboles ne symbolisent rien. Pour utiliser le jargon linguistique, ils ont une syntaxe, mais pas de sémantique. L'intentionnalité que semblent avoir les ordinateurs est exclusivement dans les esprits de ceux qui les programment, de ceux qui les utilisent, de ceux qui leur donnent des entrées et de ceux qui interprètent leurs sorties.

L'objet de l'exemple du chinois était justement d'essayer de le montrer en faisant apparaître que dès qu'on met dans le système quelque chose qui a une intentionnalité (un homme) et qu'on y introduit le programme formel, on se rend compte que le programme formel n'ajoute aucune intentionnalité. Il n'ajoute rien, par exemple, à la capacité d'un homme de comprendre le chinois.

Cette particularité de l'IA qui semblait si attirante — la distinction entre le programme et sa réalisation concrète — s'avère justement fatale à l'affirmation que la simulation pourrait être la reproduction. La distinction entre le programme et sa réalisation concrète, au niveau matériel, semble parallèle à la distinction entre le niveau des opérations mentales et celui des opérations cérébrales. Si nous pouvions décrire le niveau des opérations mentales comme un programme formel, alors il semble que nous pourrions décrire ce qu'il y a d'essentiel dans l'esprit sans recourir ni à la psychologie introspective ni à la neurophysiologie du cerveau. Mais l'équation « l'esprit est au cerveau ce qu'un programme est au matériel » a quelques points faibles, dont les trois suivants :

Tout d'abord, la distinction entre le programme et l'objet physique de réalisation a pour conséquence que le même programme pourrait avoir toutes sortes d'expressions concrètes abracadabrantes dénuées de toute intentionnalité. Weizenbaum (1976, Ch. 2), par exemple, décrit par le menu la construction d'un ordinateur à partir d'un rouleau de papier hygiénique et d'un tas de petites pierres. De même, le programme comprenant les histoires en chinois peut tourner dans un assemblage de canalisations d'eau, dans des moulins à vent, ou dans un anglophone monolingue, lesquels ne comprennent pas pour autant le chinois. Les pierres, le papier hygiénique, le vent et les canalisations peuvent difficilement avoir une intentionnalité — seules des choses possédant les mêmes propriétés causales que les cerveaux peuvent avoir une intentionnalité — et si l'anglophone a ce qu'il faut pour avoir une intentionnalité, vous vous rendez facilement compte qu'il ne gagne rien, du point de vue de l'intentionnalité, à apprendre le programme par cœur, puisque ce n'est pas parce qu'il le connaîtra par cœur qu'il apprendra le chinois.

Ensuite, le programme est purement formel, mais les états intentionnels ne sont pas formels de la même façon. Ils sont définis par leur contenu, et non pas par leur forme. La conviction qu'il pleut, par

exemple, n'est pas définie par un certain aspect formel, mais par un certain contenu mental assorti de conditions de satisfaction, d'une direction d'intention (cf. Searle, 1979), etc. En fait, la conviction, en tant que telle, n'a même pas d'aspect formel au sens syntaxique du terme, puisqu'une même conviction peut être désignée par un nombre indéfini d'expressions syntaxiques dans différents systèmes linguistiques.

Enfin, comme je l'ai déjà dit, les états mentaux et les événements sont un produit direct du fonctionnement du cerveau, alors que le programme n'est pas un produit de l'ordinateur.

« Eh bien, si les programmes ne sont pas équivalents à des processus mentaux, pourquoi tant de gens ont-ils cru le contraire? C'est un point qui mérite une explication. »

Je ne sais vraiment pas que répondre à cette question. L'idée que les simulations par ordinateur pourraient exister pour de vrai aurait toujours dû paraître suspecte, parce que l'ordinateur n'est pas limité à la simulation d'opérations mentales. Personne ne suppose que des simulations sur ordinateur d'un gigantesque incendie vont faire brûler le quartier, ni que la simulation d'un orage va nous laisser trempés jusqu'aux os. Alors, pourquoi diable supposerait-on qu'un ordinateur simulant la compréhension comprendrait quoi que ce soit? On entend parfois dire qu'il serait terriblement difficile de parvenir à ce que les ordinateurs souffrent ou tombent amoureux, mais l'amour et la douleur ne sont ni plus difficiles ni plus faciles à obtenir que la cognition ou quoi que ce soit d'autre. La simulation ne demande que les bonnes entrées, les bonnes sorties et, entre les deux, un programme qui transforme les unes en les autres. Et un ordinateur n'a que des entrées, des sorties et un programme. L'erreur, c'est de confondre la simulation et la reproduction, qu'il s'agisse de douleur, d'amour, de cognition, d'incendies ou d'orages.

Pourtant, plusieurs raisons expliquent que certains aient pu croire — et croient peut-être encore — que l'IA reproduit, et du même coup explique, des phénomènes mentaux, et je pense que l'on ne balaiera pas ces illusions tant que l'on n'aura pas clairement expliqué leur origine.

Tout d'abord, et c'est peut-être là le point le plus important, il y a une confusion au sujet de la notion de « traitement de l'information » : beaucoup de spécialistes des sciences cognitives pensent que le cerveau humain, avec son esprit, fait quelque chose qu'ils appellent du « traitement de l'information », de même que l'ordinateur, avec son programme, traite de l'information. Pourtant, les incendies et les orages, eux, n'effectuent aucun traitement d'information. Donc, bien que l'ordinateur puisse simuler les propriétés formelles de n'importe quel processus, il entretient un rapport particulier avec l'esprit et le cerveau, puisque quand il est correctement programmé, dans l'idéal avec un programme analogue au cerveau, le traitement de l'information réalisé est identique dans les deux cas; or, ce traitement de l'information est précisément l'essence du mental. Le problème, c'est que cet argument repose sur une ambiguïté de la notion d'« information ». Au sens où les gens « traitent l'information » quand ils réfléchissent, par exemple à des

problèmes d'arithmétique, ou quand ils lisent des questions sur des histoires et y répondent, l'ordinateur programmé n'effectue aucun « traitement de l'information ». Il se borne à manipuler des symboles formels. Le fait que le programmeur et la personne qui interprète les sorties de l'ordinateur utilisent les symboles pour représenter des objets du monde réel est totalement hors de portée de l'ordinateur. Je le répète, l'ordinateur a une syntaxe, mais pas de sémantique. Par conséquent, si vous tapez sur son clavier « 2 plus 2 égale? » il répondra « 4 ». Il n'a néanmoins pas la moindre idée que « 4 » signifie 4 ou quoi que ce soit d'autre. Ce n'est pas qu'il lui manque des informations de second ordre sur l'interprétation de ses symboles de premier ordre, mais plutôt que ses symboles de premier ordre n'ont pas d'interprétation en ce qui concerne l'ordinateur. Tout ce qu'il a, ce sont des symboles. L'introduction de la notion de « traitement de l'information » engendre donc un dilemme : nous interprétons l'expression « traitement de l'information » soit d'une façon qui implique l'intentionnalité, soit d'une façon qui ne l'implique pas. Dans le premier cas, l'ordinateur programmé ne fait pas de traitement de l'information, mais seulement de la manipulation de symboles formels. Dans le second, l'ordinateur réalise bien un traitement de l'information, mais seulement au sens où des calculatrices, des machines à écrire, des estomacs, des thermostats, des orages et des ouragans le font, au sens où ils ont un niveau de description auquel nous pouvons les définir comme recevant des informations, les transformant, et produisant des informations. Mais c'est alors aux observateurs extérieurs d'interpréter les entrées et les sorties comme des informations au sens habituel. Et aucune similitude n'est reconnue entre l'ordinateur et le cerveau du point de vue de leurs traitements de l'information.

Ensuite, il y a, dans l'IA, des restes de behaviorisme ou d'opérationnalisme. Parce que des ordinateurs spécialement programmés peuvent avoir des configurations d'entrées-sorties semblables à celles des êtres humains, nous sommes tentés de postuler qu'il existe, dans l'ordinateur, des états mentaux semblables à ceux des humains. Pourtant, une fois que nous nous apercevons qu'il est conceptuellement et empiriquement possible, dans un domaine donné, qu'un système ait des capacités humaines tout en étant dépourvu d'intentionnalité, nous devrions pouvoir surmonter cette tentation. Ma machine à calculer a des capacités de calcul, mais pas d'intentionnalité, et dans cet article, j'ai essayé de montrer qu'un système pourrait avoir des entrées et des sorties reproduisant celles d'une personne de langue maternelle chinoise sans pour autant comprendre le chinois, quel que soit son programme. Le test de Turing est typique de cette tradition, en ce qu'il est résolument behavioriste et opérationnaliste. Je suis convaincu que si les chercheurs en IA rejetaient totalement le behaviorisme et l'opérationnalisme, une grande partie de la confusion entre la simulation et la reproduction serait du même coup éliminée.

Troisièmement, à ce reste d'opérationnalisme s'ajoute un vestige de dualisme; en effet, l'IA forte est basée sur l'hypothèse que ce qui concerne l'esprit n'a rien à voir avec le cerveau. Dans l'IA forte (et le

fonctionnalisme), ce qui compte, ce sont les programmes, et les programmes sont indépendants des machines dans lesquelles ils tournent ; en ce qui concerne l'IA, le même programme pourrait être réalisé à travers une machine électronique, une substance mentale cartésienne ou un esprit universel hégélien. La découverte qui m'a le plus surpris, au cours de mes discussions sur ces questions, c'est qu'un grand nombre de chercheurs en IA sont très choqués par mon idée que les véritables phénomènes mentaux humains pourraient dépendre des véritables propriétés physico-chimiques des véritables cerveaux humains. Pourtant, en y réfléchissant une petite minute, on se rend compte que je n'aurais pas dû être surpris, car le projet de l'IA forte n'a aucune chance d'aboutir si l'on n'accepte pas une certaine forme de dualisme. Ce projet consiste à reproduire et expliquer le mental en concevant des programmes, mais si Pesprit n'est pas, conceptuellement, mais aussi empiriquement, indépendant du cerveau, ce projet ne pourra pas être mené à bonne fin, puisque le programme est totalement indépendant de toute réalisation concrète. Il est indispensable de croire que Pesprit est séparable du cerveau, tant conceptuellement qu'empiriquement — ce qui est une forme extrême de dualisme — pour espérer reproduire le mental en écrivant et en faisant tourner des programmes, puisque ceux-ci doivent être indépendants des cerveaux ou de toute autre forme particulière d'objet physique de réalisation. Si les opérations mentales sont des opérations de calcul sur des symboles formels, alors elles n'ont aucun lien intéressant avec le cerveau ; le seul serait que le cerveau est un de ces innombrables types de machines à travers lesquelles ce programme pourrait se réaliser. Cette forme de dualisme n'est pas la variété cartésienne traditionnelle, qui prétend qu'il y a deux sortes de *substances,* mais elle est néanmoins cartésienne en ce sens qu'elle insiste sur le fait que ce qu'il y a de spécifiquement mental dans Pesprit n'a aucun lien avec les véritables propriétés du cerveau. Ce dualisme sous-jacent nous est masqué par les fréquentes fulminations de la littérature de l'IA contre le « dualisme » ; mais les auteurs de ces écrits semblent ne pas s'apercevoir que leur position présuppose justement un fort dualisme.

« Une machine pourrait-elle penser? » Mon opinion est que *seules* des machines pourraient penser, mais uniquement de types très spéciaux, à savoir les cerveaux et les machines ayant les mêmes propriétés causales que les cerveaux. C'est la principale raison pour laquelle l'IA forte ne nous a pas appris grand-chose sur la pensée, puisqu'elle n'a rien à nous apprendre sur les machines. D'après sa propre définition, elle concerne les programmes, et les programmes ne sont pas des machines. Quoi que puisse être Pintentionnalité, c'est un phénomène biologique, et il y a autant de chances qu'elle soit causalement dépendante de la biochimie spécifique à ses origines que la lactation, la photosynthèse ou tout autre phénomène biologique. Personne ne supposerait que nous puissions produire du lait et du sucre en faisant une simulation par ordinateur des séquences formelles de la lactation et de la photosynthèse ; or, dès que Pesprit est enjeu, beaucoup de gens sont prêts à croire à un tel miracle, à cause d'un dualisme profondément ancré : ils supposent que Pesprit est

une affaire de processus formels et est indépendant de causes matérielles tout à fait spécifiques, tandis que le lait et le sucre ne le sont pas.

Pour défendre ce dualisme, ils avancent souvent que le cerveau est un ordinateur numérique (on appelait d'ailleurs souvent les premiers ordinateurs des « cerveaux électroniques »). Ça n'apporte malheureusement rien. Oui, bien sûr, le cerveau est un ordinateur numérique. Etant donné que tout et n'importe quoi est un ordinateur numérique, le cerveau l'est aussi. Le problème, c'est que la capacité causale du cerveau d'engendrer l'intentionnalité ne peut pas être sa capacité de concrétiser un programme d'ordinateur, puisqu'il est possible de trouver une forme concrète de réalisation pour n'importe quel programme sans pour autant lui donner des états mentaux. Quelle que soit la façon dont le cerveau produit l'intentionnalité, ce ne peut pas être en reproduisant concrètement un programme, puisque aucun programme n'est, à lui seul, suffisant pour engendrer l'intentionnalité*.