4

Les ordinateurs et l'intelligence

Alan Turing

TRADUIT DE L'ANGLAIS
PAR PATRICE BLANCHARD

1. Le jeu de l'imitation

Je propose de considérer la question : « Les machines peuvent-elles penser ? » Il faudrait commencer par définir le sens des termes « machine » et « penser ». Les définitions peuvent être conçues de manière à refléter, autant que possible, l'utilisation normale des mots, mais cette attitude est dangereuse. Si on doit trouver la signification des mots « machine » et « penser » en examinant comment ils sont communément utilisés, il est difficile d'échapper à la conclusion que la signification de la question « Les machines peuvent-elles penser ? » et la réponse à cette question doivent être recherchées dans une étude statistique telle que le sondage d'opinion. Mais cela est absurde. Au lieu de m'essayer à une telle définition, je remplacerai la question par une autre, qui lui est étroitement liée et qui est exprimée en des termes relativement non ambigus.

Le problème reformulé peut être décrit dans les termes d'un jeu que nous appellerons le « jeu de l'imitation ». Il se joue à trois : un homme (A), une femme (B) et un interrogateur (C) qui peut être de l'un ou l'autre sexe. L'interrogateur se trouve dans une pièce à part, séparé des deux autres. L'objet du jeu, pour l'interrogateur, est de déterminer lequel des deux est l'homme et lequel est la femme. Il les connaît sous les appellations X et Y et, à la fin du jeu, il doit déduire soit que « X est A et Y est B », soit que « X est B et Y est A ». L'interrogateur peut poser des questions à A et B de la manière suivante :

C : X peut-il ou peut-elle me dire, s'il vous plaît, quelle est la longueur de ses cheveux ?

A supposer à présent que X soit vraiment A, alors A doit répondre. La finalité du jeu pour A est d'essayer d'induire C en erreur. Sa réponse pourrait donc être :

A : « Mes cheveux sont coupés à la garçonne et les mèches les plus longues ont à peu près vingt centimètres de long. »

Pour que le ton de la voix ne puisse pas aider l'interrogateur, les réponses devraient être écrites ou, mieux, dactylographiées. L'installation idéale serait un téléimprimeur communiquant entre les deux pièces. A défaut, les questions et réponses peuvent être répétées par un intermédiaire. L'objet du jeu pour la joueuse (B) est d'aider l'interrogateur. La meilleure stratégie pour elle est probablement de donner des réponses vraies. Elle peut ajouter à ses réponses des choses telles que : « Je suis la femme, ne l'écoutez pas ! », mais cela ne servira à rien, car l'homme peut faire des remarques similaires.

Nous posons maintenant la question : « Qu'arrive-t-il si une machine prend la place de A dans le jeu ? L'interrogateur se trompera-t-il aussi souvent que lorsque le jeu se déroule entre un homme et une femme ? » Ces questions remplacent la question originale : « Les machines peuvent-elles penser ? »

2. Critique du nouveau problème

Au lieu de demander : « Quelle est la réponse à cette nouvelle forme de question ? », on pourrait tout aussi légitimement demander : « La nouvelle question vaut-elle la peine d'être examinée ? » Examinons cette dernière question sans autre forme de procès, coupant par là court à une régression infinie.

Le nouveau problème a l'avantage de tracer une ligne assez nette entre les capacités physiques et intellectuelles de l'homme. Aucun ingénieur ou chimiste ne prétend être capable de produire un matériau que rien ne distingue de la

peau humaine. Il est possible que cela puisse être fait un jour, mais, même en supposant que cette invention soit réalisée, nous jugerions sans intérêt de rendre une « machine pensante » plus humaine en l'habillant d'une telle chair artificielle. La forme sous laquelle nous avons posé le problème reflète ce fait à travers les conditions qui empêchent l'interrogateur de voir ou de toucher les autres participants, ou d'entendre leurs voix. On peut montrer d'autres avantages du critère proposé à travers un spécimen de questions et de réponses.

Ainsi :

C : Pouvez-vous, s'il vous plaît, m'écrire un sonnet au sujet du pont de la rivière Forth ?

A : Ne comptez pas sur moi pour ça. Je n'ai jamais réussi à écrire de la poésie.

C : Ajoutez 34 957 à 70 764.

(Un silence d'à peu près trente secondes, puis vient la réponse.)

A : 105 721[1].

C : Jouez-vous aux échecs ?

A : Oui.

C : J'ai mon roi en C8 et aucune autre pièce. Vous avez seulement votre roi en C6 et une tour en A1. C'est à vous de jouer, que jouez-vous ?

A (*après un silence de quinze secondes*) : Tour en A8, échec et mat.

La méthode des questions et réponses semble être adaptée pour introduire presque n'importe quel champ des capacités humaines que nous souhaitons inclure. Nous ne souhaitons pas pénaliser la machine pour son incapacité à briller dans des concours de beauté, et nous ne voulons pas pénaliser l'homme

1. Le texte anglais original, publié en octobre 1950 dans la revue *Mind*, donne un résultat différent (et erroné) : « 105 621 ». Le traducteur s'est donc permis de rectifier une « erreur » voulue, qui a pour but de dissimuler l'identité de la machine : l'erreur, comme chacun le sait, est humaine... (Note due à Jean Lessègue, que nous remercions.)

parce qu'il perd quand il court contre un avion. Les conditions de notre jeu rendent ces incapacités non pertinentes. S'ils le jugent souhaitable, les « témoins » peuvent se vanter autant qu'il leur plaît à propos de leurs charmes, de leur force ou de leur héroïsme. Mais l'interrogateur ne peut exiger une démonstration pratique.

Il est peut-être possible de critiquer le jeu sous le prétexte que la machine y est lourdement désavantagée. Si l'homme essayait de faire semblant d'être la machine, il est clair qu'il s'en sortirait fort mal. Il serait immédiatement trahi par sa lenteur et son inexactitude en arithmétique. Les machines ne peuvent-elles pas exécuter quelque chose qui relève d'une forme de « pensée », mais qui est très différent de ce qu'un homme fait ? Cette objection est très forte, mais nous pouvons au moins dire que, s'il est possible de construire une machine pour qu'elle joue le jeu de l'imitation de manière satisfaisante, nous n'avons pas besoin de nous occuper de cette objection.

On pourrait objecter que, lorsqu'elle joue au jeu de l'imitation, la meilleure stratégie pour la machine peut être autre que l'imitation du comportement humain. C'est possible, mais je pense qu'il est probable que cela n'ait pas beaucoup d'influence. De toute façon, nous n'avons pas l'intention d'étudier ici la théorie du jeu, et l'on présupposera que la meilleure stratégie est d'essayer de fournir des réponses qui seraient naturellement données par l'homme.

3. Les machines pouvant prendre part à ce jeu

La question que nous avons posée dans la section 1 ne sera totalement définie que lorsque nous aurons spécifié ce que le mot « machine » signifie.

Il est naturel que nous souhaitions permettre l'utilisation, dans nos machines, de n'importe quel type de technologie. Nous voulons aussi accepter la possibilité qu'un ingénieur, ou une équipe d'ingénieurs, puisse construire une machine qui fonctionne, mais dont les modalités de fonctionnement ne peuvent être décrites de manière satisfaisante par ses constructeurs, parce qu'ils ont appliqué une méthode en grande partie expérimentale.

Nous souhaitons enfin exclure de la catégorie des machines les hommes nés de la manière habituelle. Il est difficile d'élaborer des définitions qui satisfassent à ces trois conditions. On pourrait, par exemple, requérir que les ingénieurs soient tous du même sexe, mais cela ne serait pas vraiment satisfaisant, car il est probablement possible de créer un individu complet à partir d'une seule cellule (disons) de la peau d'un homme. Le faire serait un exploit de la technique biologique méritant les plus hauts éloges, mais nous ne serions pas tenté de le considérer comme un cas de « construction de machine pensante ». Ce qui nous pousse à abandonner l'idée d'accepter toutes les techniques. Nous sommes d'autant plus disposé à le faire que l'intérêt actuel pour les « machines pensantes » a été soulevé par un type particulier de machine, habituellement appelé « calculateur électronique », ou « ordinateur ». A la suite de cette suggestion, nous n'autoriserons que les ordinateurs à prendre part à notre jeu.

Cette restriction semble être, à première vue, très rigoureuse, j'essaierai de montrer qu'en réalité il n'en est rien. Pour ce faire, un bref exposé sur la nature et les propriétés des ordinateurs est nécessaire.

On peut dire aussi que cette identification des machines aux ordinateurs de même que mon critère de la « pensée » ne seront inadéquats que si (contrairement à ce que je crois) il se trouve que les ordinateurs sont incapables d'obtenir de bons résultats dans le jeu.

Il existe déjà un certain nombre d'ordinateurs en état de

marche, et l'on peut se demander : « Pourquoi ne pas tenter l'expérience tout de suite ? Il serait facile de satisfaire aux conditions du jeu. On pourrait utiliser un certain nombre d'interrogateurs et recueillir des données statistiques pour montrer la fréquence selon laquelle la bonne réponse est donnée. » Nous dirons, pour répondre brièvement, que nous ne nous demandons pas si tous les ordinateurs obtiendraient des résultats satisfaisants dans le jeu, ni si les ordinateurs actuellement disponibles obtiendraient ces résultats, mais s'il existe des ordinateurs imaginables qui les obtiendraient. Mais cela n'est qu'une réponse rapide. Nous examinerons plus loin cette question sous un jour différent.

4. Les ordinateurs

On peut expliquer l'idée qui est à l'origine des calculateurs numériques en disant que ces machines sont destinées à mener à bien toutes les opérations qu'un calculateur humain pourrait effectuer. Le calculateur humain est censé suivre des règles fixes ; il n'a pas l'autorisation de s'en éloigner si peu que ce soit. Nous pouvons supposer que ces règles lui sont fournies dans un livre, qui est modifié chaque fois qu'on lui donne un nouveau travail. Il dispose aussi d'une quantité illimitée de papier, sur lequel il fait ses calculs. Il peut encore faire ses multiplications et ses additions sur une « machine de bureau », mais cela n'a pas d'importance.

Si nous utilisons l'explication ci-dessus comme définition, le danger sera la circularité de l'argumentation. Nous l'évitons en donnant les grandes lignes des moyens par lesquels l'effet désiré est obtenu. On peut habituellement considérer qu'un ordinateur est composé de trois parties :

1) une mémoire ;

2) une unité d'exécution ;

3) une unité de contrôle.

La mémoire est une réserve d'informations et correspond au papier du calculateur humain, que ce soit le papier sur lequel il fait ses calculs ou celui sur lequel est imprimé son livre de règles. Dans la mesure où le calculateur humain fait des calculs dans sa tête, une partie de la mémoire correspondra à sa mémoire.

L'unité d'exécution est la partie qui effectue les différentes opérations individuelles qu'un calcul comporte. Ces opérations individuelles varieront d'une machine à l'autre. La machine peut habituellement exécuter des opérations relativement longues, telles que : « Multipliez 3 540 675 445 par 7 076 345 687 », mais dans certaines machines seules des opérations très simples, comme « Écrivez 0 », sont possibles.

Nous avons mentionné que le livre des règles fourni au calculateur est remplacé dans la machine par une partie de la mémoire. On l'appelle alors la *table des instructions*. C'est le rôle de l'unité de contrôle de vérifier que ces instructions sont correctement exécutées et dans le bon ordre. L'unité de contrôle est faite de telle manière que cela se produit nécessairement.

Les informations stockées dans la mémoire sont habituellement réparties en groupes de taille modérée. Dans une machine, par exemple, un groupe pourrait être constitué par dix chiffres décimaux. Des numéros sont attribués, de manière systématique, aux parties de la mémoire dans lesquelles les différents groupes d'informations sont enregistrés. Une instruction typique pourrait être :

« Ajoutez le nombre enregistré dans la position 6809 à celui qui est en 4302 et enregistrez le résultat dans cette dernière position. »

Il est inutile de dire que cela n'apparaîtra pas en français dans la machine, mais se trouvera probablement codé dans une forme telle que : 6 809 430 217. Ici, 17 indique laquelle

des différentes opérations possibles doit être exécutée à partir des deux nombres. Dans ce cas, l'opération est celle décrite ci-dessus, c'est-à-dire : « Ajoutez le nombre... » On remarquera que l'instruction a dix chiffres, et qu'elle forme ainsi un groupe d'informations facile à manipuler. L'unité de contrôle suivra normalement les instructions auxquelles elle doit obéir dans l'ordre des positions dans lesquelles elles sont enregistrées, mais on peut à l'occasion rencontrer une instruction comme :

« Obéissez maintenant à l'instruction enregistrée dans la position 5 606, et continuez à partir de là. »

Ou encore :

« Si la position 4 505 contient 0, obéissez ensuite à l'instruction enregistrée en 6 707, sinon continuez directement. »

Les instructions de ce dernier type sont très importantes, car elles rendent possible la répétition continue d'une suite d'opérations jusqu'à ce qu'une condition quelconque soit remplie, cependant que la machine obéit non pas à de nouvelles instructions à chaque répétition, mais aux mêmes.

Pour employer une analogie domestique, supposons que la mère de Tommy veuille qu'il passe chez le cordonnier tous les matins en allant à l'école pour voir si ses chaussures sont prêtes ; elle peut le lui redemander tous les matins, ou alors placer une fois pour toutes dans l'entrée une note qu'il verra à son départ pour l'école et qui lui dit de passer chez le cordonnier et aussi de détruire la note à son retour s'il ramène les chaussures.

Le lecteur doit accepter comme un fait établi que les ordinateurs peuvent être, et ont été, construits suivant les principes que nous avons décrits, et qu'ils peuvent en fait imiter de très près les actions d'un calculateur humain.

Le livre de règles dont nous avons dit que notre calculateur humain se servait était bien sûr une fiction commode. Les vrais calculateurs humains se rappellent en effet ce qu'ils ont à faire. Si l'on veut faire imiter par une machine les comportements du calculateur humain dans quelque opération complexe, on doit lui demander comment il fait, puis traduire la réponse sous la forme d'une table d'instructions, autrement dit : un *programme*. « Programmer une machine pour exécuter l'opération *A* » veut dire : mettre dans la machine la table d'instructions appropriée pour qu'elle exécute *A*.

Une variante intéressante de l'idée d'ordinateur est l'« ordinateur avec un élément de hasard ». Ces ordinateurs comportent des instructions qui incluent le jet d'un dé, ou tout autre procédé électronique équivalent. Une telle instruction peut par exemple être :

« Jetez le dé, et mettez le nombre obtenu dans la mémoire 1 000. » On décrit parfois une telle machine comme ayant un libre arbitre (bien que je n'utiliserais pas cette expression moi-même). Il n'est pas possible normalement de déterminer à partir de l'observation d'une machine si elle possède un élément de hasard, car un effet similaire peut être obtenu par des moyens tels que celui de faire dépendre les choix des décimales de π.

La plupart des ordinateurs existants ont seulement une mémoire fixe. Il n'y a pas de difficulté théorique à concevoir un ordinateur avec une mémoire illimitée. Bien entendu, seule une partie finie peut être utilisée à la fois. De même, on a pu seulement en construire une quantité finie, mais nous pouvons imaginer que l'on en rajoutera autant qu'il sera nécessaire. De tels ordinateurs ont un intérêt théorique particulier et on les appellera « ordinateurs à capacité infinie ».

L'idée d'ordinateur est ancienne. Charles Babbage, *Lucasian Professor of Mathematics* à Cambridge de 1828 à 1839, avait conçu une telle machine, appelée « machine analytique », mais elle ne fut jamais terminée. Bien que Babbage ait eu toutes les idées essentielles, sa machine ne représentait pas à l'époque un projet très intéressant. La vitesse qu'elle aurait pu atteindre aurait été nettement plus grande que celle d'un calculateur humain, mais quelque chose comme 100 fois plus

faible que celle de la machine de Manchester, qui est l'une des machines modernes les plus lentes. La mémorisation devait être purement mécanique, utilisant des rouages et des cartes.

Le fait que la machine analytique de Babbage ait dû être entièrement mécanique nous aidera à nous débarrasser d'une superstition. On attache souvent de l'importance au fait que les ordinateurs modernes sont électriques et que le système nerveux aussi est électrique. Puisque la machine de Babbage n'était pas électrique, et puisque tous les ordinateurs lui sont en un sens équivalents, nous voyons que l'utilisation de l'électricité ne peut guère avoir d'importance théorique. On trouve, bien sûr, habituellement l'électricité là où l'on a besoin de signaux rapides, ainsi il n'est pas surprenant que nous la trouvions dans les deux cas.

Dans le système nerveux, les phénomènes chimiques sont au moins aussi importants que les phénomènes électriques. Dans certains ordinateurs le système de mémorisation est principalement acoustique. On voit donc que l'utilisation de l'électricité n'est qu'une similarité très superficielle. Si nous souhaitons découvrir de telles similarités, nous devrions plutôt chercher des analogies mathématiques de fonction.

5. Universalité des ordinateurs

Les ordinateurs considérés dans la section ci-dessus peuvent être classés parmi les « machines à états discrets ». Ce sont des machines qui passent par bonds soudains d'un état parfaitement défini à un autre. Ces états sont suffisamment différents pour que toute possibilité de confusion entre eux soit négligeable. A strictement parler, il n'existe pas de telles machines. En réalité, tout bouge de manière continue. Mais il y a de nombreux types de machines qu'il vaut mieux *considérer* comme des machines à états discrets. Par exemple, si l'on considère les interrupteurs d'un éclairage, c'est une fiction commode de dire que chaque interrupteur doit être nettement ouvert ou nettement fermé. Il doit bien y avoir des positions intermédiaires, mais dans la plupart des cas nous pouvons l'oublier. Comme exemple d'une machine à états discrets, nous pourrions envisager une roue qui tourne d'un cran de 120° une fois par seconde, mais qui peut être arrêtée à l'aide d'un levier manipulé de l'extérieur ; de plus, une lampe s'allume dans l'une des positions de la roue. Cette machine pourrait, dans l'abstrait, être décrite comme suit : l'état interne de la machine (qui est décrit par la position de la roue) peut être q_1, q_2 ou q_3. Il y a un signal d'entrée i_0 ou i_1 (position du levier). L'état interne est déterminé à tout moment par le dernier état et le signal d'entrée, suivant le tableau ci-dessous :

		Dernier état		
		q_1	q_2	q_3
Entrée	i_0	q_2	q_3	q_1
	i_1	q_1	q_2	q_3

Les signaux de sortie, la seule indication externe visible de l'état interne (la lumière), sont décrits par le tableau :

État	q_1	q_2	q_3
Sortie	o_0	o_0	o_1

Cet exemple est typique des machines à états discrets. Elles peuvent être décrites par de telles tables, pourvu qu'elles aient seulement un nombre fini d'états possibles. Il apparaîtra que, à partir d'un état initial donné de la machine et de signaux d'entrée, il est toujours possible de prédire tous les états futurs. Cela nous rappelle les vues de Laplace selon lesquelles à partir de l'état complet de l'Univers à un moment donné, avec la description de la position et de la vitesse de toutes les particules, il serait possible de prédire tous les états futurs. La

prédiction que nous envisageons est, cependant, relativement plus effective que celle que Laplace considère. Le système de l'« Univers dans sa totalité » est tel que des erreurs absolument minimes dans les conditions initiales peuvent avoir un effet démesuré dans le futur. Le déplacement d'un seul électron d'un milliardième de centimètre à un moment donné peut faire qu'un homme sera tué par une avalanche un an plus tard, ou en réchappera. Une des propriétés essentielles des systèmes mécaniques que nous avons appelés « machines à états discrets » est que ce phénomène ne se produit pas. Même quand nous considérons des machines matériellement réelles au lieu de machines idéales, une connaissance raisonnablement exacte de l'état de la machine à un moment donné entraîne une connaissance exacte de son état à un moment ultérieur donné.

Comme nous l'avons mentionné, les ordinateurs sont classés parmi les machines à états discrets. Mais le nombre d'états dont une telle machine est capable est habituellement extrêmement grand. Par exemple, le nombre d'états pour la machine qui fonctionne maintenant à Manchester est à peu près de $2^{165\,000}$, c'est-à-dire à peu près $10^{50\,000}$. Comparez cela avec notre exemple de la roue décrite ci-dessus, qui avait trois états. Il n'est pas difficile de comprendre pourquoi le nombre d'états doit être si important. L'ordinateur comporte une mémoire correspondant au papier utilisé par un calculateur humain. Il doit être possible d'inscrire dans la mémoire chacune des combinaisons de symboles qui peuvent être écrites sur le papier. Pour la simplicité, supposons que seuls les chiffres de 0 à 9 sont utilisés comme symboles. Les différences d'écriture ne sont pas prises en compte. Supposons que le calculateur ait 100 feuilles de papier ayant chacune 50 lignes pouvant contenir chacune 30 chiffres. Alors le nombre d'états est de : $10^{100\times 50\times 30}$, c'est-à-dire $10^{150\,000}$. C'est à peu près le nombre d'états de trois machines de Manchester. Le logarithme de base 2 du nombre d'états est habituellement appelé « capacité de mémoire » de la machine. Ainsi, la

machine de Manchester a une capacité de mémoire d'à peu près 165 000, et la machine à roue de notre exemple d'à peu près 1,6. Si l'on met deux machines ensemble, il faut additionner leurs capacités pour obtenir la capacité de la machine ainsi obtenue. Cela rend possibles des affirmations comme : « La machine de Manchester contient 64 pistes magnétiques, chacune avec une capacité de 2 560, 8 tubes électroniques avec une capacité de 1 280. Diverses mémoires totalisant à peu près 300, cela fait un total de 174 880. »

Si l'on dispose de la table correspondant à une machine à états discrets, il est possible de prédire ce qu'elle fera. Il n'y a aucune raison pour que ce calcul ne puisse pas être exécuté au moyen d'un ordinateur. Pourvu qu'il puisse être exécuté suffisamment rapidement, l'ordinateur pourrait imiter ainsi le comportement de n'importe quelle machine à états discrets. Le jeu de l'imitation pourrait donc se jouer entre la machine en question (en tant que B), l'ordinateur qui l'imite (en tant que A) ; l'interrogateur serait incapable de les distinguer. L'ordinateur doit bien sûr avoir une capacité adéquate ainsi qu'une vitesse de travail suffisamment grande. De plus, il doit être re-programmé pour chaque nouvelle machine que nous désirons lui faire imiter.

On décrit cette propriété particulière des ordinateurs (qu'ils puissent imiter n'importe quelle machine discrète) en disant que ce sont des *machines universelles*. L'existence de machines possédant cette propriété entraîne la conséquence importante, en dehors de toute considération de vitesse, qu'il est inutile de concevoir différentes nouvelles machines pour réaliser différentes opérations de calcul. Elles peuvent être effectuées à l'aide d'un seul ordinateur, convenablement programmé pour chaque cas. On verra qu'en conséquence tous les ordinateurs sont en un sens équivalents.

Nous pouvons maintenant envisager de nouveau le problème soulevé à la fin de la section 3. Il a été suggéré à titre d'expérience que la question « Les machines peuvent-elles

penser ? » devrait être remplacée par : « Peut-on imaginer des ordinateurs qui fassent bonne figure dans le jeu de l'imitation ? » Si nous le souhaitons, nous pouvons rendre cette question superficiellement plus générale et demander : « Y a-t-il des machines à états discrets qui puissent y faire bonne figure ? » Mais, eu égard à la propriété d'universalité, nous voyons que chacune de ces deux questions est équivalente à celle-ci : « Fixons notre attention sur un ordinateur particulier O. Est-il vrai que, en modifiant cet ordinateur pour avoir une capacité de mémoire adéquate, en accroissant de manière satisfaisante sa vitesse de travail, et en lui fournissant un programme approprié, on peut faire jouer à O le rôle de A dans le jeu de l'imitation, le rôle de B étant tenu par un homme ? »

6. Vues contradictoires sur la question principale

Nous pouvons maintenant considérer que nous avons déblayé le terrain, et que nous sommes prêts à entrer dans le débat sur notre question « Les machines peuvent-elles penser ? » et sa variante citée à la fin du paragraphe précédent. Nous ne pouvons pas complètement abandonner la forme originale du problème, car les opinions différeront en ce qui concerne la validité de la substitution, et nous devons au moins être attentifs à ce qui peut être dit sur ce point.

Cela simplifiera les choses pour le lecteur si j'expose d'abord mes propres vues sur le sujet. Examinons, en premier lieu, la question sous sa forme la plus précise. Je crois que dans une cinquantaine d'années il sera possible de programmer des ordinateurs, avec une capacité de mémoire d'à peu près 10^9, pour les faire si bien jouer au jeu de l'imitation qu'un interrogateur moyen n'aura pas plus de 70 % de chances de procéder à l'identification exacte après cinq minutes d'interrogation.

Je crois que la question originale « Les machines peuvent-elles penser ? » a trop peu de sens pour mériter une discussion. Néanmoins, je crois qu'à la fin du siècle l'usage, les mots et l'éducation de l'opinion générale auront tant changé que l'on pourra parler de machines pensantes sans s'attendre à être contredit. Je crois de plus qu'il ne sert à rien de dissimuler ces croyances. L'idée populaire selon laquelle les savants avancent inexorablement d'un fait bien établi à un autre, sans être influencés par des hypothèses non vérifiées, est absolument fausse. Pourvu que nous sachions clairement quels sont les faits prouvés et quelles sont les hypothèses, aucun mal ne peut en résulter. Les hypothèses sont de grande importance puisqu'elles suggèrent d'utiles voies de recherches.

Je vais maintenant envisager les opinions opposées à la mienne.

1. L'objection théologique

Penser est une fonction de l'âme immortelle de l'homme. Dieu a donné une âme immortelle à tout homme ou femme, mais à aucun animal ni à aucune machine. En conséquence, ni l'animal ni la machine ne peuvent penser[a].

Je ne peux accepter en rien cette objection, mais j'essaierai d'y répondre en termes théologiques. Je trouverais l'argument plus convaincant si les animaux étaient classés avec les hommes, car il y a une plus grande différence, à mon avis, entre l'animé et l'inanimé qu'il n'y en a entre l'homme et les autres animaux. Le caractère arbitraire du point de vue ortho-

a. Il est possible que cette vue des choses soit hérétique. Saint Thomas d'Aquin (*Somme théologique* ; cité par Bertrand Russell, *A History of Western Philosophy*, New York, Simon et Schuster, 1945, p. 458) dit que Dieu ne peut pas faire qu'un homme n'ait pas d'âme. Mais il se peut que cela ne soit pas une restriction réelle de Ses pouvoirs mais seulement un résultat du fait que l'âme humaine est immortelle, et donc indestructible.

doxe devient plus clair si nous considérons comment il pourrait apparaître à un membre de quelque autre communauté religieuse. Comment les chrétiens considèrent-ils l'opinion musulmane : « Les femmes n'ont pas d'âme » ? Mais laissons cela de côté, et revenons à la discussion principale. Il m'apparaît que l'argument énoncé ci-dessus implique une sérieuse restriction de la toute-puissance de Dieu. Il est admis qu'il y a certaines choses qu'Il ne peut faire, comme de faire que 1 soit égal à 2, mais ne devrions-nous pas croire qu'Il a la liberté de donner une âme à un éléphant si cela lui semble convenable ? Nous pourrions nous attendre à ce qu'Il exerce seulement ce pouvoir en conjonction avec une mutation qui fournirait à l'éléphant un cerveau convenablement amélioré pour s'occuper des besoins de son âme. On peut imaginer un argument similaire pour le cas des machines, qui peut sembler différent parce qu'il est plus difficile à « avaler ». Mais cela signifie seulement que nous envisageons comme moins probable l'éventualité qu'Il considère que les circonstances sont favorables pour qu'Il leur donne une âme. On discutera les circonstances en question dans la suite de ce texte.

En essayant de construire de telles machines, nous ne devrions pas plus usurper irrévérencieusement Ses pouvoirs de créer des âmes que nous ne le faisons en engendrant des enfants : nous sommes plutôt, dans les deux cas, des instruments de Sa volonté, fournissant des demeures aux âmes qu'Il crée.

Cependant, cela est pure spéculation. Les arguments théologiques m'impressionnent peu, quel que soit l'objet qu'ils défendent. De tels arguments se sont souvent montrés peu satisfaisants dans le passé. Au temps de Galilée, on disait que les textes « Et le soleil s'arrêta [...] et ne se hâta pas de se cacher pendant toute une journée » (Jos X,13) et « Il posa les fondations de la Terre pour qu'elle ne bouge à aucun moment » (Ps CV,5) étaient une réfutation appropriée de la théorie copernicienne. Avec nos connaissances actuelles, un

tel argument paraît futile. Quand ces connaissances n'étaient pas établies, il faisait une impression tout à fait différente.

2. L'objection de l'autruche

« Le fait que les machines pensent aurait des conséquences trop terribles. Il vaut mieux croire et espérer qu'elles ne peuvent pas le faire. » L'argument est rarement exprimé aussi ouvertement que ci-dessus. Mais il affecte la plupart de ceux d'entre nous qui réfléchissent à ce sujet. Nous aimerions croire que l'homme est de quelque subtile façon supérieur au reste de la Création. Il serait encore mieux de pouvoir montrer qu'il est *nécessairement* supérieur, car alors il n'y aurait aucun risque qu'il perde sa position dominante. La popularité de l'argument théologique est clairement liée à ce sentiment. Il sera probablement plus fort parmi les intellectuels, puisqu'ils valorisent plus que les autres la capacité de penser comme base de leur croyance en la supériorité de l'homme. Je ne pense pas que cet argument soit suffisamment substantiel pour rendre nécessaire une réfutation. La consolation serait plus appropriée : peut-être devrait-on la chercher dans la métempsycose.

3. L'objection mathématique

Un certain nombre de résultats de la logique mathématique peuvent être utilisés pour montrer qu'il y a des limites aux pouvoirs des machines à états discrets. Le plus connu de ces résultats l'est sous le nom de *théorème de Gödel* et montre que dans tout système logique suffisamment puissant on peut formuler des affirmations qui ne peuvent ni être prouvées ni être réfutées à l'intérieur du système, à moins que le système lui-même ne soit inconsistant. Il existe d'autres résultats, similaires à certains égards, dus à Church, Kleene, Rosser et Turing. Le

dernier de ces résultats est le plus pratique à examiner, puisqu'il se réfère directement aux machines, alors que les autres ne peuvent être utilisés que comme des arguments comparativement indirects : par exemple, si l'on utilise le théorème de Gödel, il nous faut en plus nous donner des moyens de décrire les systèmes logiques en termes de machines, et les machines en termes de systèmes logiques. Le résultat en question se réfère à un type de machine qui est essentiellement un ordinateur à capacité infinie. Ce résultat établit qu'il y a certaines choses qu'une telle machine ne peut pas faire. Si elle est programmée pour répondre à des questions, comme dans le jeu de l'imitation, il y aura certaines questions auxquelles soit elle donnera une réponse fausse, soit elle ne donnera pas de réponse du tout, quel que soit le temps qui lui sera imparti pour répondre. Il se peut bien sûr qu'il y ait beaucoup de questions de ce genre, et des questions auxquelles une machine donnée ne saura pas répondre obtiendront peut-être une réponse satisfaisante de la part d'une autre. Nous supposons bien sûr pour le moment des questions appelant une réponse en « oui » ou en « non », plutôt que des questions telles que : « Que pensez-vous de Picasso ? » Nous savons que les machines doivent échouer dans des questions du type : « Considérez la machine spécifiée comme suit... Cette machine répondra-t-elle "oui" à n'importe quelle question ? » Les points de suspension doivent être remplacés par la description d'une machine de forme standard, qui pourrait ressembler à celle qui figure dans la section 5. Quand la machine décrite présente une relation évidente et comparativement simple avec la machine que l'on interroge, on peut montrer soit que la réponse est fausse, soit qu'elle n'apparaîtra jamais. Voici le résultat mathématique : on en déduit que cela prouve une incapacité des machines, qui ne se retrouve pas dans l'esprit humain.

Pour répondre brièvement à cet argument, il faut dire que, bien qu'il soit établi qu'il y a des limites à la puissance de n'importe quelle machine, il a seulement été affirmé, sans

aucune sorte de preuve, que de telles limites ne s'appliquaient pas à l'esprit humain. Mais je ne pense pas que nous puissions rejeter ce point de vue si légèrement. Chaque fois que l'on pose à l'une de ces machines la question cruciale appropriée et qu'elle donne une réponse définie, nous savons que cette réponse est forcément fausse, ce qui nous procure un certain sentiment de supériorité. Ce sentiment est-il illusoire ? Il est sans aucun doute tout à fait sincère, mais je ne pense pas qu'il faille y attacher trop d'importance. Nous donnons nous-mêmes trop souvent des réponses fausses à des questions pour que nous ayons le droit de nous réjouir d'une telle preuve de la faillibilité des machines. Nous ne pouvons de plus, en de telles occasions, ressentir notre supériorité que par rapport à la machine particulière sur laquelle nous avons remporté un triomphe insignifiant. Il est exclu de triompher simultanément de *toutes* les machines. En bref, il se peut qu'il y ait des hommes plus intelligents que n'importe quelle machine donnée, mais il se peut aussi qu'il y ait d'autres machines encore plus intelligentes, et ainsi de suite. Ceux qui tiennent à l'argument mathématique accepteraient pour la plupart volontiers, à mon avis, le jeu de l'imitation comme base de discussion. Ceux qui croient aux deux objections précédentes ne s'intéresseraient probablement à aucun critère.

4. L'argument issu de la conscience

Cet argument est très bien exprimé dans le discours Lister de 1949 du professeur Jefferson, dont j'extrais cette citation : « Nous ne pourrons pas accepter l'idée que la machine égale le cerveau jusqu'à ce qu'une machine puisse écrire un sonnet ou composer un concerto à partir de pensées ou d'émotions ressenties et non pas en choisissant des symboles au hasard, et non seulement l'écrire, mais savoir qu'elle l'a écrit. Aucun mécanisme ne pourrait ressentir (et non pas simplement produire

artificiellement un signal, ce qui relève d'un artifice facile) du plaisir quand il réussit, du chagrin quand ses lampes grillent ; il ne serait pas ému par la flatterie, malheureux de ses erreurs, charmé par le sexe, et ne se mettrait pas en colère ou ne se sentirait pas déprimé quand il ne peut pas obtenir ce qu'il veut. »

Cet argument revient à nier la validité de notre test. Selon ce point de vue extrême, la seule manière dont on pourrait s'assurer qu'une machine pense serait d'*être* la machine et de ressentir qu'on pense. On pourrait alors décrire ces sentiments au monde, mais bien sûr personne n'aurait de raisons d'en tenir compte. De même, suivant ce point de vue, la seule manière de savoir qu'un *homme* pense est d'être cet homme lui-même.

C'est en fait le point de vue solipsiste. Il se peut que ce soit la position la plus logique à tenir, mais cela rend difficile la communication des idées. A est enclin à croire que « A pense, mais B ne pense pas » ; de son côté, B croit que « B pense, mais pas A ». Au lieu de discuter continuellement ce point, on utilise habituellement la convention polie stipulant que tout le monde pense.

Je suis sûr que le professeur Jefferson ne souhaite pas adopter ce point de vue extrême et solipsiste. Il accepterait probablement volontiers le jeu de l'imitation comme test. Le jeu (en omettant le joueur B) est fréquemment utilisé en pratique sous le nom d'*examen oral* pour découvrir si quelqu'un comprend véritablement quelque chose ou a « appris comme un perroquet ». Imaginons une partie d'un tel examen :

L'examinateur : Dans le premier vers de votre sonnet qui dit : « Te comparerais-je à un jour d'été », est-ce que « un jour de printemps » serait aussi bien ou mieux ?

Le témoin : Cela ne rimerait pas.

L'examinateur : Et « un jour d'hiver » ? Cela rimerait très bien[2]...

2. En anglais, *summer* (été) et *winter* (hiver) ont les mêmes caractéristiques prosodiques. Alors que celles de *spring* (printemps) sont différentes *(NdT)*.

Le témoin : Oui, mais personne n'a envie d'être comparé à un jour d'hiver.

L'examinateur : Diriez-vous que M. Pickwick vous fait penser à Noël ?

Le témoin : D'une certaine manière, oui.

L'examinateur : Et pourtant Noël est un jour d'hiver, et je ne pense pas que la comparaison ennuierait M. Pickwick.

Le témoin : Je ne pense pas que vous soyez sérieux. Par « un jour d'hiver », on veut dire un jour d'hiver typique, plutôt qu'une journée spéciale comme Noël.

Et ainsi de suite.

Que dirait le professeur Jefferson si la machine à écrire des sonnets était capable de répondre ainsi à un examen ? Je ne sais pas s'il considérerait que la machine « produit simplement et artificiellement un signal » avec ces réponses, mais si les réponses étaient aussi satisfaisantes et fermes que dans le passage ci-dessus, je ne pense pas qu'il la décrirait comme un « artifice facile ». Cette expression a, je pense, pour but de recouvrir des dispositifs comme l'inclusion dans la machine de l'enregistrement de quelqu'un lisant un sonnet, qu'un système approprié mettrait en marche de temps en temps.

Donc, en bref, je pense que l'on pourrait persuader la plupart de ceux qui soutiennent l'argument issu de la conscience de l'abandonner plutôt que d'être contraints d'adopter la position solipsiste. Ils accepteront alors probablement volontiers notre test.

Je ne voudrais pas donner l'impression de penser qu'il n'y a pas de mystère relatif à la conscience. Il y a, par exemple, une sorte de paradoxe lié à toute tentative faite pour la localiser. Mais je ne crois pas que ces mystères doivent nécessairement être résolus avant que nous puissions répondre à la question qui nous intéresse ici.

5. Les arguments provenant de diverses incapacités

Ces arguments prennent la forme suivante : « Je vous concède que vous pouvez fabriquer des machines qui fassent tout ce que vous avez mentionné, mais vous ne serez jamais capable d'en fabriquer une qui fasse X. » On énumère à ce moment-là différents traits X. J'en présente une sélection :

Qu'elle soit gentille, débrouillarde, belle, amicale (p. 156-157), qu'elle ait de l'initiative, le sens de l'humour, qu'elle fasse la différence entre le bien et le mal, qu'elle fasse des erreurs (p. 157-158), qu'elle tombe amoureuse, qu'elle aime les fraises à la crème (p. 157), qu'elle rende quelqu'un amoureux d'elle, qu'elle apprenne à partir de son expérience (p. 166 *sq.*), qu'elle utilise les mots correctement, qu'elle soit l'objet de ses propres pensées (p. 159). (Certaines de ces incapacités sont examinées plus en détail, comme l'indiquent les numéros de pages.)

Aucune preuve n'est habituellement fournie pour soutenir ces affirmations. Je crois qu'elles sont surtout fondées sur le principe de l'induction scientifique. Un homme a vu des milliers de machines dans sa vie. De ce qu'il en voit, il tire un certain nombre de conclusions générales : elles sont laides ; chacune est conçue dans un but bien précis ; quand on leur demande un travail légèrement différent, elles sont incapables de le réaliser ; la variété de comportements de n'importe laquelle d'entre elles est très restreinte, etc. Il en conclut naturellement que ce sont des propriétés nécessaires des machines en général. Beaucoup de ces limites sont associées à la très faible capacité de mémoire de la plupart des machines. (Je suppose que l'idée de capacité de mémoire est étendue de manière à recouvrir des machines qui ne sont pas des machines à états discrets. La définition exacte n'a pas d'importance, puisque dans la présente discussion on ne revendique aucune exactitude

mathématique.) Il y a quelques années, alors qu'on avait peu entendu parler des ordinateurs, il était possible de faire disparaître une grande partie de l'incrédulité à leur égard en mentionnant leurs propriétés sans décrire leur réalisation. Cela était probablement dû à une application similaire du principe de l'induction scientifique. Les applications de ce principe sont bien sûr en grande partie inconscientes. Quand un enfant qui s'est brûlé craint le feu et montre qu'il le craint en l'évitant, je peux dire qu'il applique l'induction scientifique. (Je pourrais, bien sûr, aussi décrire son comportement de bien d'autres manières.) Les travaux et les coutumes de l'humanité ne semblent pas être un matériau très adapté à l'application de l'induction scientifique. On doit étudier une grande partie d'espace-temps si l'on veut obtenir des résultats fiables. Autrement, nous conclurions (comme le font la plupart des enfants français) que tout le monde parle français et qu'il est idiot d'apprendre l'anglais.

Il y a, cependant, des remarques particulières à faire concernant beaucoup des incapacités mentionnées. Le lecteur a pu être frappé par la futilité de l'incapacité à aimer les fraises à la crème. Il est possible qu'on puisse faire aimer ce plat délicieux à une machine, mais toute tentative pour le faire serait idiote. Ce qui importe, en ce qui concerne cette incapacité, est qu'elle contribue à d'autres incapacités : par exemple, on imagine mal comment un homme et une machine pourraient entretenir des liens d'amitié, comme ceux qui rapprochent des hommes blancs ou noirs.

Le fait d'affirmer que « la machine ne peut pas faire d'erreurs » semble curieux. On est tenté de répondre : « Est-elle pire pour cela ? » Mais adoptons une attitude plus sympathique et essayons de voir ce que cela veut dire. Je pense que cette critique peut être expliquée dans les termes du jeu de l'imitation. On affirme que l'interrogateur pourrait distinguer la machine de l'homme, simplement en lui posant un certain nombre de problèmes d'arithmétique. La machine

serait démasquée à cause de son exactitude implacable. La réplique est simple. La machine (programmée pour jouer le jeu) n'essaierait pas de donner les réponses *justes* aux problèmes d'arithmétique. Elle introduirait délibérément des erreurs d'une manière calculée pour dérouter l'interrogateur. Une erreur mécanique se révélerait probablement à cause d'une décision inopportune à propos du type d'erreur à commettre en arithmétique. Même cette interprétation de la critique n'est pas suffisamment sympathique. Mais la place nous manque pour y entrer plus avant. Il me semble que cette critique vient de la confusion entre deux types d'erreurs : nous pouvons les appeler « erreurs de fonctionnement » et « erreurs de conclusion ». Les erreurs de fonctionnement sont dues à quelque faute mécanique ou électrique qui fait que la machine ne se comporte pas comme elle le devrait. Dans les discussions philosophiques, on préfère ignorer la possibilité de telles erreurs ; on discute donc de « machines abstraites ». Ces machines abstraites sont des fictions mathématiques plutôt que des objets physiques. Elles sont par définition incapables d'erreurs de fonctionnement. En ce sens, nous pouvons effectivement dire que « les machines ne peuvent jamais faire d'erreurs ». Les erreurs de conclusion apparaissent seulement quand une signification est attribuée aux signaux de sortie de la machine. La machine peut, par exemple, imprimer des équations mathématiques, ou des phrases en anglais. Quand une proposition fausse se trouve imprimée, nous disons que la machine a commis une erreur de conclusion. Il n'y a évidemment absolument aucune raison de dire qu'une machine ne peut pas faire ce genre d'erreur. Elle pourrait ne rien faire d'autre qu'imprimer sans cesse « 0 = 1 ». Pour prendre un exemple moins pervers, elle pourrait disposer d'une méthode pour tirer des conclusions par induction scientifique. Nous pouvons nous attendre à ce qu'une telle méthode conduise occasionnellement à des résultats erronés.

Bien entendu, on peut répondre à l'affirmation qu'une machine ne saurait être l'objet de ses propres pensées que si l'on parvient à montrer que la machine *a* des pensées, et qu'elles *ont* des objets. Néanmoins, l'« objet des opérations d'une machine » semble bien avoir une signification, du moins pour les gens qui travaillent avec elle. Si, par exemple, la machine essayait de trouver une solution à l'équation $x^2 - 40x - 11 = 0$, on serait tenté de décrire l'équation comme une partie de l'objet de la machine à ce moment-là. Dans ce sens, une machine peut sans aucun doute être son propre objet. Elle peut être utilisée pour aider à la confection de ses propres programmes ou pour prévoir les effets de modifications de sa propre structure. En observant les résultats de son propre comportement, elle peut modifier ses propres programmes pour atteindre un but de manière plus efficace. Il s'agit là de possibilités du futur proche, plutôt que de rêves utopiques.

Souligner le fait qu'une machine ne peut pas avoir une grande diversité de comportements, c'est dire simplement qu'elle ne peut pas avoir une grande capacité de mémoire. Jusqu'à une période assez récente, une capacité de mémoire même de mille chiffres était très rare.

Les critiques que nous considérons ici sont souvent des formes déguisées de l'argument issu de la conscience. Habituellement, si l'on soutient qu'une machine *peut* vraiment faire l'une de ces choses, et si l'on décrit le type de méthode que la machine est susceptible d'utiliser, on ne fera pas une forte impression. La méthode (quelle qu'elle soit, car elle est forcément mécanique) est en effet estimée plutôt vile. A preuve, la remarque entre parenthèses dans la précédente citation de Jefferson.

6. L'objection de lady Lovelace

Les renseignements les plus détaillés que nous possédions sur la machine analytique de Babbage proviennent du mémoire de lady Lovelace. Elle y déclare : « La machine analytique n'a pas la prétention de *donner naissance* à quoi que ce soit. Elle peut effectuer *tout ce que nous savons lui ordonner* de faire » (les italiques sont de lady Lovelace). Cet extrait est cité par Hartree, qui ajoute : « Ceci n'implique pas qu'il ne soit pas possible de construire des machines électroniques qui "penseront par elles-mêmes" ou dans lesquelles, en termes biologiques, on pourrait inclure un réflexe conditionné qui servirait de base à un "apprentissage". Que cela soit en principe possible ou non est une question passionnante et stimulante, suggérée par certains développements récents. Mais il ne semble pas que les machines réalisées ou qui étaient à l'état de projet à cette époque aient eu cette propriété. »

Je suis entièrement d'accord avec Hartree à ce sujet. On remarquera qu'il n'affirme pas que les machines en question n'avaient pas cette propriété, mais plutôt que les preuves dont lady Lovelace disposait ne l'encourageaient pas à croire qu'elles avaient cette propriété. Il est fort possible que, en un sens, les machines en question l'aient eue. Car supposons qu'une quelconque machine à états discrets ait cette propriété. La machine analytique était un calculateur numérique universel, et, en conséquence, si sa capacité mémoire et sa vitesse étaient adéquates, on pourrait avec un programme adapté lui faire imiter la machine en question. Il est probable que cet argument ne vint pas à l'esprit de la comtesse, ni à celui de Babbage. De toute façon, ils n'étaient pas dans l'obligation d'avancer tout ce qu'il y avait à avancer.

On reconsidérera entièrement la question en examinant, plus loin, les machines à faculté d'apprentissage.

Une variante de l'objection de lady Lovelace affirme qu'une

machine ne peut « jamais rien faire de vraiment nouveau ». On peut y répondre pour l'instant avec le dicton : « Il n'y a rien de nouveau sous le soleil. » Qui peut être certain que le « travail original » qu'il a effectué n'était pas simplement la croissance de la semence plantée en lui par l'enseignement, ou la conséquence de principes généraux bien connus ? Une meilleure variante de l'objection affirme que la machine ne peut « jamais nous prendre par surprise ». Cette affirmation est un défi plus direct, et on peut y faire face plus franchement. Les machines me prennent très fréquemment par surprise. La raison principale en est que je ne fais pas de calculs suffisants pour décider de ce à quoi je peux m'attendre de leur part ou plutôt que, bien que je fasse des calculs, je les fais de manière rapide et bâclée, en prenant des risques. Je me dis peut-être : « Je suppose que le voltage ici devrait être le même que là : de toute façon, supposons qu'il en soit ainsi. » Naturellement, je me trompe souvent, et le résultat est surprenant, car au moment de l'expérience ces suppositions ont été oubliées. Ces suppositions justifient les remontrances qu'on pourrait me faire sur mes pratiques douteuses, mais ne jettent pas l'ombre d'un doute sur ma crédibilité quand je parle des surprises que je ressens.

Je ne m'attends pas à ce que cette réponse fasse taire les critiques. On m'objectera probablement que de telles surprises sont dues à quelque acte de création mentale de ma part, et ne sont pas à porter au crédit de la machine. Cela nous ramène à l'argument issu de la conscience et nous éloigne de l'idée de surprise. C'est une suite d'arguments que nous devons considérer comme close, mais il faut peut-être remarquer que le fait de trouver quelque chose surprenant requiert de toute façon un « acte de création mentale », que la surprise trouve son origine chez un homme, un livre, une machine ou quoi que ce soit d'autre.

Cette opinion selon laquelle les machines ne peuvent pas nous surprendre est due, à mon avis, à un sophisme dont les

philosophes et les mathématiciens sont tout particulièrement coutumiers. L'idée est que, dès qu'un fait se présente à l'esprit, toutes les conséquences de ce fait jaillissent simultanément avec lui dans l'esprit. C'est une hypothèse très utile dans de nombreuses circonstances, mais on oublie trop facilement qu'elle est fausse. Une conséquence naturelle est qu'on suppose qu'il n'y a aucun mérite à découvrir simplement les conséquences d'une information ou de principes généraux.

7. L'argument de la continuité dans le système nerveux

Le système nerveux n'est certainement pas une machine à états discrets. Une petite erreur dans l'information sur la taille d'une impulsion nerveuse affectant un neurone peut nous conduire à nous tromper grossièrement sur la taille de l'impulsion de sortie. On peut dire que, puisqu'il en est ainsi, il ne faut pas s'attendre à pouvoir imiter le comportement du système nerveux avec un système à états discrets.

Il est vrai qu'une machine à états discrets est forcément différente d'une machine continue. Mais, si nous acceptons les conditions du jeu de l'imitation, l'interrogateur ne pourra pas tirer avantage de cette différence. On peut rendre la situation plus claire en considérant une machine continue plus simple. Un analyseur différentiel conviendra très bien (un analyseur différentiel est un type de machine qui n'est pas à états discrets, et qu'on utilise pour certains types de calculs). Certains d'entre eux impriment leurs réponses et peuvent ainsi facilement prendre part au jeu. Il ne serait pas possible à un ordinateur digital de prédire exactement quelles réponses l'analyseur différentiel donnerait à un problème, mais il serait tout à fait capable de donner le genre de réponse adéquat. Par exemple, si on lui demandait de donner la valeur de π (en réalité, à peu près 3,1416), il serait raisonnable de choisir au hasard entre les valeurs : 3,12, 3,13, 3,14, 3,15, 3,16, avec des

probabilités disons de 0,05, 0,15, 0,55, 0,19, 0,06. Dans ces circonstances, il serait très difficile pour l'interrogateur de distinguer l'analyseur différentiel de l'ordinateur digital.

8. L'argument du comportement informalisable

Il n'est pas possible de produire un ensemble de règles qui ait la prétention de décrire ce qu'un homme devrait faire dans tout ensemble concevable de circonstances. On devrait, par exemple, établir une règle définissant qu'on doit s'arrêter quand on voit un feu rouge, et passer quand on voit un feu vert. Mais qu'arrive-t-il si par suite d'une erreur les deux apparaissent en même temps ? On peut peut-être décider qu'il est plus sûr de s'arrêter. Mais quelque autre difficulté peut bien se faire jour plus tard à cause de cette décision. Il paraît impossible d'élaborer des règles de conduite pour parer à toutes les éventualités, même à celles concernant les feux tricolores. Je partage entièrement ce point de vue.

A partir de là, on en déduit que nous ne pouvons pas être des machines. J'essaierai de reproduire l'argument, mais j'ai peur de ne pas être très juste à son égard. Il semble qu'il corresponde à peu près au syllogisme suivant : « Si chaque homme disposait d'un ensemble défini de règles de conduite d'après lesquelles il organiserait sa vie, il ne serait pas supérieur à la machine ; mais de telles règles n'existent pas ; ainsi, les hommes ne peuvent pas être des machines. » La non-distribution du moyen terme est manifeste. Je ne pense pas que l'argument soit jamais énoncé exactement dans ces termes, mais je crois néanmoins que c'est bien l'argument utilisé. Il se peut cependant qu'il y ait une certaine confusion entre les « règles de conduite » et les « lois du comportement » qui finisse d'obscurcir le problème. Par « règles de conduite » j'entends des préceptes tels que : « Arrêtez-vous quand vous voyez un feu rouge », sur lesquels on peut agir et dont on peut être

conscient. Par « lois de comportement » j'entends des lois naturelles comme celles qui s'appliquent au corps humain, par exemple : « Si vous le pincez, il criera. » Si nous substituons « les lois du comportement qui règlent sa vie » à « les lois de conduite d'après lesquelles il règle sa vie » dans l'argument cité, la non-distribution du moyen terme n'est plus un obstacle insurmontable. Car nous croyons non seulement que le fait d'être soumis à des lois de conduite implique que l'on soit une machine (bien que non nécessairement une machine à états discrets), mais que, réciproquement, le fait d'être une telle machine implique que l'on soit soumis à de telles lois. Cependant, nous ne pouvons pas nous convaincre de l'absence d'un ensemble complet de lois du comportement aussi facilement que nous l'avons fait pour l'ensemble complet des règles de conduite. La seule manière dont nous puissions découvrir de telles lois est l'observation scientifique, et nous ne connaissons aucune circonstance nous permettant de dire : « Nous avons assez cherché, de telles lois n'existent pas. »

Nous pouvons démontrer de manière plus convaincante qu'aucune affirmation de ce type ne serait justifiée. Supposons que nous puissions être sûrs de découvrir de telles lois si elles existaient. Alors, à partir d'une machine à états discrets donnée, il devrait certainement être possible de découvrir, par l'observation, assez d'éléments à son sujet pour prédire son comportement futur, et cela dans une période de temps raisonnable, disons mille ans. Mais il ne semble pas que ce soit le cas. J'ai introduit dans l'ordinateur de Manchester un petit programme utilisant seulement mille unités de stockage, par lequel la machine, lorsqu'on lui fournit un nombre de seize chiffres, répond par un autre nombre en deux secondes. Je défie quiconque d'en apprendre assez au sujet du programme à partir de ces réponses pour être capable de prédire la réponse pour des valeurs non encore utilisées.

9. L'argument de la perception extrasensorielle

Je pars du principe que le lecteur est familiarisé avec l'idée de la perception extrasensorielle et les quatre éléments qui en font partie, c'est-à-dire : la télépathie, la clairvoyance, la préconnaissance et la psychokinésie. Ces phénomènes troublants semblent remettre en cause toutes nos idées scientifiques habituelles. Comme nous aimerions les discréditer ! Malheureusement, l'évidence statistique, au moins pour la télépathie, est accablante. Il est très difficile de réorganiser ses idées pour y intégrer ces nouveaux faits. Une fois que nous les avons acceptés, ce n'est pas progresser beaucoup que de croire aux fantômes et aux spectres. L'idée que notre corps se déplace simplement suivant les lois connues de la physique, et suivant quelques autres qui n'ont pas encore été découvertes mais qui leur sont relativement similaires, serait la première à disparaître.

Cet argument est, à mon avis, très fort. On peut répondre que beaucoup de théories scientifiques semblent continuer à fonctionner dans la pratique malgré les conflits avec la perception extrasensorielle ; que l'on peut, en fait, très bien se débrouiller si on l'oublie. C'est d'un réconfort relatif, et l'on craint que la pensée ne soit justement le type de phénomène pour lequel la perception extrasensorielle est particulièrement adéquate.

Un argument plus spécifique, fondé sur la perception extrasensorielle, pourrait être rédigé en ces termes : « Jouons au jeu de l'imitation, en utilisant comme témoins un homme qui est un bon récepteur télépathique et un ordinateur digital. L'interrogateur peut poser des questions comme : "Quelle est la couleur de la carte que j'ai dans la main droite ?" L'homme, par télépathie ou clairvoyance, donne 130 fois la bonne réponse sur 400 cartes. La machine peut seulement deviner au hasard et peut-être obtenir 104 bonnes réponses. L'interrogateur peut ainsi l'identifier. »

Une possibilité intéressante apparaît ici. Supposons que

l'ordinateur renferme un générateur de nombres au hasard. Il est alors naturel de l'utiliser pour décider de la réponse à donner. Mais alors, le générateur de nombres au hasard sera sujet aux pouvoirs psychokinésiques de l'interrogateur. Cette psychokinésie fera peut-être que la machine devinera juste plus souvent que l'on ne s'y attend d'après le calcul des probabilités, et ainsi l'interrogateur ne pourra toujours pas l'identifier correctement. D'un autre côté, il se pourrait qu'il soit capable de deviner juste sans poser de questions, par clairvoyance. Avec la perception extrasensorielle, tout peut arriver.

Si la télépathie est admise, il sera nécessaire de renforcer notre test. La situation pourrait être considérée comme analogue à celle qui se produirait si l'interrogateur se parlait à lui-même et si l'un des participants écoutait avec l'oreille collée au mur. Le fait de placer les participants dans une « pièce à l'épreuve de la télépathie » satisferait toutes les exigences.

7. Les machines qui apprennent

Le lecteur aura compris que je n'ai pas d'argument positif très convaincant pour soutenir mon point de vue. Si j'en avais, je n'aurais pas pris tant de peine à montrer les erreurs des points de vue opposés au mien. Les preuves que j'ai, je vais maintenant les donner.

Revenons un moment sur l'objection de lady Lovelace, qui disait que la machine ne peut faire que ce qu'on lui dit de faire. On pourrait dire qu'un homme peut « injecter » une idée dans la machine, laquelle réagira jusqu'à un certain point, puis retournera à l'immobilité, comme une corde de piano frappée par un marteau. Un autre point de comparaison serait une pile atomique d'une masse inférieure à la masse critique : une idée injectée correspondra à un neutron entrant dans la pile, en provenance de l'extérieur. Tout neutron de ce type produira une certaine perturbation qui finira par cesser. Toutefois, si la masse de la pile est suffisamment accrue, la perturbation créée par l'entrée d'un tel neutron continuera probablement à s'accroître jusqu'à ce que toute la pile soit détruite. Existe-t-il un phénomène correspondant pour les esprits et en existe-t-il un pour les machines ? Il semble qu'il y en ait un pour l'esprit humain. La majorité des esprits humains paraissent « sous-critiques », c'est-à-dire semblent correspondre dans cette analogie aux piles à masse sous-critique. Une idée proposée à un tel esprit donnera lieu, en moyenne, à l'apparition de moins d'une idée en réponse. Une faible proportion est surcritique. Une idée proposée à un tel esprit pourra donner lieu à l'apparition de toute une « théorie » constituée d'idées secondaires, tertiaires ou encore plus éloignées. Les esprits des animaux semblent être absolument sous-critiques. En poursuivant cette analogie, nous nous demandons : « Peut-on rendre une machine surcritique ? »

L'analogie de la « peau de l'oignon » est aussi utile. En considérant les fonctions de l'esprit ou du cerveau, nous découvrons certaines opérations qui peuvent s'expliquer en termes purement mécaniques. Nous disons que cela ne correspond pas à l'esprit réel : c'est une espèce de peau que nous devons enlever si nous voulons trouver l'esprit réel. Mais, dans ce qui reste, nous rencontrons une autre peau à enlever, et ainsi de suite. En continuant de cette manière, arrivons-nous jamais à l'esprit « réel », ou arrivons-nous finalement à la peau qui ne contient rien ? Dans ce dernier cas, l'esprit serait entièrement mécanique (ce ne serait cependant pas une machine à états discrets, nous en avons discuté).

Ces deux derniers paragraphes ne prétendent pas être des arguments convaincants. On les décrirait mieux en disant que ce sont des « déclamations tendant à produire une croyance ».

Le seul élément vraiment satisfaisant qui puisse soutenir le point de vue exprimé au début de la section 6 nous sera fourni

par la réalisation, à la fin du siècle, de l'expérience décrite. Mais que pouvons-nous dire en attendant ? Quelle démarche devrions-nous entreprendre maintenant si l'expérience devait être couronnée de succès ?

Comme je l'ai expliqué, le problème est surtout un problème de programmation. Des progrès techniques devront aussi être réalisés, mais il semble improbable qu'ils ne puissent pas répondre aux exigences. Les estimations de la capacité de stockage du cerveau varient de 10^{10} à 10^{15} chiffres binaires. Je penche pour les valeurs les plus basses, et je crois que seule une très petite partie en est utilisée pour les types les plus élevés de pensée. La plus grande partie sert probablement à la conservation des impressions visuelles. Je serais surpris que plus de 10^9 soit nécessaire pour jouer de manière satisfaisante au jeu de l'imitation, du moins contre un aveugle (note : la capacité de l'*Encyclopaedia Britannica*, onzième édition, est de 2×10^9). Une capacité de stockage de 10^7 serait une possibilité très réalisable, même avec les techniques actuelles. Il n'est probablement pas nécessaire du tout d'accroître la vitesse opérationnelle des machines. Les pièces des machines modernes qui peuvent être considérées comme analogues aux cellules nerveuses fonctionnent à peu près mille fois plus vite que ces dernières. Cela devrait fournir une « marge de sécurité » couvrant les pertes de vitesse de multiples provenances. Le problème est alors de trouver comment programmer ces machines pour qu'elles jouent à notre jeu. Selon ma cadence actuelle de travail, je produis à peu près mille unités de programme par jour. En conséquence, une soixantaine de personnes travaillant assidûment pendant cinquante ans pourraient accomplir le travail, s'il n'y avait pas de perte. Une méthode plus expéditive serait la bienvenue.

En essayant d'imiter l'esprit humain adulte, il va nous falloir beaucoup réfléchir au processus qui l'a amené à l'état où il se trouve. Nous pouvons en signaler trois composantes :

a) l'état initial de l'esprit, disons à la naissance ;

b) l'éducation à laquelle il a été soumis ;

c) d'autres expériences, que l'on ne peut pas décrire comme éducatives, auxquelles il a été soumis.

Au lieu de produire un programme qui simule l'esprit de l'adulte, pourquoi ne pas essayer plutôt d'en produire un qui simule celui de l'enfant ? S'il était alors soumis à une éducation appropriée, on aboutirait au cerveau humain. Il est probable que le cerveau d'un enfant est une sorte de carnet acheté en papeterie : assez peu de mécanisme et beaucoup de feuilles blanches (mécanisme et écriture sont, de notre point de vue, presque synonymes). Notre espoir est qu'il y ait si peu de mécanisme dans le cerveau d'un enfant qu'il soit très facile de le programmer. Dans une première approximation, nous pouvons penser que la quantité de travail nécessaire à cette éducation serait à peu près la même que pour un enfant humain. Nous avons en conséquence divisé le problème en deux parties : le programme-enfant et le processus de l'éducation. Ces deux éléments restent très intimement liés. Nous ne pouvons pas nous attendre à découvrir dès la première tentative une bonne machine-enfant. Il faut tenter l'expérience de l'enseignement sur une telle machine et examiner comment elle apprend. On peut ensuite en essayer une autre et voir si c'est mieux ou moins bien. Il y a un lien évident entre ce processus et l'évolution, à travers les identités suivantes :

structure de la machine-enfant = matériel héréditaire ;
changement dans la machine-enfant = mutations ;
sélection naturelle = jugement de l'expérimentateur.

On peut cependant espérer que ce procédé sera plus expéditif que l'évolution. La survie du plus adapté est une méthode lente de mesure des avantages. L'expérimentateur, par l'exercice de son intelligence, devrait pouvoir l'accélérer. Le fait qu'il n'en soit pas réduit à des mutations aléatoires est également important. S'il sait trouver la cause d'une faiblesse, il est probablement en mesure d'imaginer le type de mutation qui l'améliorera.

Il ne sera pas possible d'appliquer exactement les mêmes procédés d'enseignement à la machine et à un enfant normal. Elle n'aura par exemple pas de jambes, et on ne pourra pas lui demander d'aller remplir le seau à charbon. Il est possible qu'elle n'ait pas d'yeux. Mais, même si ces manques étaient palliés au mieux par des techniques intelligentes, on ne pourrait l'envoyer à l'école sans que les autres élèves ne s'en moquent de manière excessive. Elle doit pourtant recevoir un certain enseignement. Il ne faut donc pas trop s'inquiéter à propos des jambes, des yeux, etc. L'exemple de M^lle^ Helen Keller montre que l'éducation est possible dès lors que la communication se produit dans les deux sens entre le maître et l'élève, quel que soit le moyen employé.

Nous associons normalement punitions et récompenses au processus de l'enseignement. On peut construire, ou programmer, des machines-enfants simples suivant ce genre de principe. Il faut construire la machine de manière que les événements qui précèdent immédiatement l'apparition d'un signal-punition aient peu de chances de se reproduire, alors qu'un signal-récompense doit accroître la probabilité de répétition de l'événement qui l'a provoqué. Ces définitions ne présupposent pas l'existence de sentiments de la part de la machine. J'ai fait quelques expériences avec une telle machine-enfant, et j'ai réussi à lui enseigner quelques petites choses, mais la méthode d'enseignement était trop peu orthodoxe pour que l'on considère que l'expérience a vraiment réussi.

L'utilisation de punitions et de récompenses peut, au mieux, faire partie du processus d'enseignement. En gros, si le maître n'a pas d'autre moyen de communiquer avec l'élève, la quantité d'informations qui peut lui parvenir ne dépasse pas le nombre total des récompenses et punitions utilisées. D'ici à ce qu'un enfant ait appris à répéter « Casabianca », il éprouverait sans doute quelques désagréments s'il ne pouvait découvrir le texte que par la méthode des « vingt questions », où chaque

« non » impliquerait qu'il reçoive un coup. Il est donc nécessaire d'avoir d'autres canaux de communication « non émotionnels ». Si ces derniers sont disponibles, il est possible d'enseigner à la machine, par un système de sanctions et de récompenses, à obéir aux ordres donnés dans un certain langage, par exemple un langage symbolique. Ces ordres doivent être transmis par des canaux « non émotionnels ». L'utilisation de ce langage diminuera de beaucoup le nombre de punitions et de récompenses requises.

Les opinions peuvent varier quant à la complexité la plus appropriée pour la machine-enfant. On pourrait essayer de la faire aussi simple que possible, conformément aux principes généraux. Ou bien on pourrait lui « intégrer[b] » un système complet d'inférences logiques. Dans ce dernier cas, la mémoire serait largement occupée par des définitions et des propositions. Les propositions pourraient avoir différents types de statuts, par exemple : des faits bien établis, des hypothèses, des théorèmes dont la preuve est mathématique, des affirmations provenant d'une autorité, des expressions ayant la forme logique d'une proposition mais sans valeur de croyance. La machine devrait être conçue de manière que, dès qu'un impératif est classé comme étant « bien établi », l'action appropriée ait automatiquement lieu. Pour illustrer cela, supposez que le maître dise à la machine : « Fais tes devoirs maintenant. » Cela peut avoir pour conséquence le classement parmi les faits bien établis de l'expression : « Le maître dit : "Fais tes devoirs maintenant." » Un autre fait identique pourrait être : « Tout ce que le maître dit est vrai. » En combinant les deux, on peut finalement arriver à ce que l'impératif « Fais tes devoirs maintenant » soit classé parmi les faits bien établis, et cela, d'après la conception de la machine, signifiera qu'elle commence réellement à faire ses devoirs, mais l'effet n'est

b. Ou plutôt « programmer », car notre machine-enfant sera programmée sur un ordinateur. Mais le système logique ne devra pas être appris.

pas très satisfaisant. Le processus d'inférence que la machine utilise n'a pas besoin d'être de nature à satisfaire les logiciens les plus exigeants. Il se pourrait par exemple qu'il n'y ait pas de hiérarchie de types. Mais cela ne signifie pas obligatoirement que des erreurs de types vont se produire, pas plus que nous ne sommes voués à tomber du haut de falaises non protégées. Des impératifs adéquats (ne faisant pas partie des règles *du* système, mais exprimés à l'*intérieur* du système) tels que : « N'utilisez pas une classe, à moins qu'elle ne soit une sous-classe de l'une de celles que le maître a mentionnées », peuvent avoir un effet similaire à : « Ne t'approche pas trop près du bord. »

Les impératifs auxquels une machine dépourvue de membres peut obéir sont forcément à caractère plutôt intellectuel, comme dans l'exemple donné ci-dessus (faire ses devoirs). Parmi ces impératifs, les plus importants seront ceux qui régleront l'ordre dans lequel les règles du système logique concerné devront être appliquées. Car, à chaque pas, lorsque l'on utilise un système logique, il y a un très grand nombre de progressions alternatives, chacune d'entre elles pouvant être utilisée, du moins en ce qui concerne l'obéissance aux règles du système logique. Ces choix font la différence entre un brillant ou un piètre raisonneur, mais non pas entre quelqu'un qui raisonne juste et quelqu'un qui raisonne faux. Des propositions conduisant à des impératifs de ce genre pourraient être : « Quand Socrate est mentionné, utilisez le syllogisme en Barbara », ou : « Si une méthode s'est avérée être plus rapide qu'une autre, n'utilisez pas la méthode la plus lente. » Certaines de ces propositions peuvent être « fournies par l'autorité », mais d'autres peuvent être produites par la machine elle-même, par exemple, par induction scientifique.

L'idée d'une machine qui apprend peut paraître paradoxale à certains lecteurs. Comment les règles d'opération de la machine peuvent-elles changer ? Elles devraient décrire complètement la manière dont la machine réagira, quels que soient

les changements qu'elle puisse subir. Les règles ne varient donc pas du tout dans le temps. C'est tout à fait vrai. L'explication du paradoxe est que les règles qui seront changées dans le processus d'apprentissage sont d'un type tout à fait modeste et ne revendiquent qu'une validité éphémère. Le lecteur peut mettre en parallèle la Constitution des États-Unis.

Une caractéristique importante de la machine qui apprend est que son maître ne saura souvent que très peu de choses sur ce qui se passe à l'intérieur, bien qu'il puisse dans une certaine mesure prévoir la conduite de son élève. Cela devrait plus particulièrement s'appliquer à l'éducation avancée de la machine qui proviendra de la machine-enfant résultant d'une conception (ou d'un programme) bien étudiée. Cela s'oppose clairement à la procédure normale d'utilisation d'une machine opérant des calculs : dans ce cas, l'objet est d'avoir une représentation mentale claire de la machine à tout moment du calcul. Cet objectif n'est atteint qu'à l'issue d'une lutte. L'opinion d'après laquelle « la machine peut faire seulement ce que nous savons lui ordonner de faire[c] » semble ici étrange. La plupart des programmes que nous pourrons introduire dans la machine auront pour résultat qu'elle fera quelque chose que nous ne pourrons pas du tout comprendre ou que nous considérerons comme un comportement totalement arbitraire. Le comportement intelligent consiste probablement à s'éloigner du comportement totalement discipliné que l'on utilise pour le calcul, mais pas trop, de façon que cela n'engendre pas un comportement arbitraire ou des boucles répétitives absurdes. Le fait de préparer notre machine à tenir son rôle dans le jeu de l'imitation par un processus d'enseignement et d'apprentissage aura un autre résultat important : la « faillibilité humaine » sera probablement négligée d'une manière assez naturelle, c'est-à-dire sans « entraînement » spécial. (Le

c. Comparez avec l'affirmation de lady Lovelace (p. 160 *sq*), qui ne contient pas le mot « seulement ».

lecteur devrait rapprocher cela du point de vue des pages 156 et 157.) Les processus qui sont appris ne produisent pas une certitude de résultats à 100 % ; si tel était le cas, ils ne pourraient pas être désappris.

Il est probablement sage d'inclure un élément de hasard dans une machine qui apprend (voir p. 143). Un élément de hasard est assez utile quand nous recherchons la solution de certains problèmes. Supposons, par exemple, que nous voulions trouver un nombre entre 50 et 200 qui soit égal au carré de la somme de ses chiffres. Nous pourrions commencer à 51 puis essayer 52 et continuer jusqu'à ce que nous trouvions un nombre qui convienne. Une autre solution serait de choisir des nombres au hasard jusqu'à ce que nous trouvions le bon. Cette méthode a pour avantage qu'il n'est pas nécessaire de garder une trace des valeurs qui ont été essayées ; son inconvénient est que l'on peut essayer deux fois la même, ce qui n'est pas très important s'il existe plusieurs solutions. L'inconvénient de la méthode systématique est qu'il peut y avoir un énorme bloc sans aucune solution à l'endroit où l'on opère d'abord. Le processus d'apprentissage peut être considéré comme la recherche d'une forme de comportement qui satisfera le maître (ou quelque autre critère). Puisqu'il y a probablement un grand nombre de solutions satisfaisantes, la méthode du hasard semble être meilleure que la méthode systématique. Il faut remarquer qu'elle est utilisée dans le processus analogue de l'évolution. Mais, là, la méthode systématique n'est pas possible. Comment pourrait-on garder une trace des différentes combinaisons génétiques qui ont été essayées, pour éviter de les essayer de nouveau ?

Nous pouvons espérer que les machines concurrenceront finalement l'homme dans tous les champs purement intellectuels. Mais par lesquels vaut-il mieux commencer ? Même cette décision est difficile à prendre. Beaucoup de gens pensent qu'une activité très abstraite comme le jeu d'échecs serait la meilleure. On peut aussi soutenir qu'il vaut mieux équiper

la machine avec les meilleurs organes sensoriels que l'on puisse acheter, puis lui apprendre à comprendre et à parler français. Ce processus pourrait se conformer à l'enseignement normal d'un enfant. On lui montrerait et nommerait des objets, etc. Encore une fois, je ne sais pas quelle est la bonne réponse, mais je pense qu'il faudrait essayer les deux voies.

Notre vision de l'avenir est limitée, mais du moins nous voyons qu'il nous reste bien des choses à faire.