Machines informatiques et intelligence

A. M. Turing

1 Le jeu de l'imitation

Je propose de considérer la question, "Les machines peuvent-elles penser ?". On devrait commencer par définir les termes "machine" et "pensée." Les définitions devraient être choisies de manière à refléter aussi bien que possible l'usage courant de ces mots, mais cette attitude est dangereuse. Si les significations des mots "machine" et "pensée" doivent être utilisées de la manière dont elles le sont habituellement, il est difficile d'échapper à la conclusion que le sens de la question "Les machines peuvent-elles penser ?" et la réponse à cette question doivent être recherchés de façon statistique, comme par sondage. Mais cela est absurde. Plutôt que de tenter une telle définition, je remplacerai la question par une autre, qui lui est intimement reliée et qui s'exprime en termes relativement non-ambigus.

La nouvelle forme du problème peut être décrite en termes d'un jeu que nous appelons 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 reste dans une pièce et n'est pas vu par les deux autres. L'objectif du jeu pour l'interrogateur est de déterminer qui est l'homme et qui est la femme des deux autres. Il les connaît par leur étiquette (X et Y), et à la fin du jeu, il dit soit "X est A et Y est B" soit "X est B et Y est A." L'interrogateur a le droit de poser des questions à A et B comme :

C : X peut-il ou elle me dire la longueur de ses cheveux ?

Maintenant supposons que X est vraiment A, alors A doit répondre. L'objectif de A pour ce jeu est d'essayer de faire que C se trompe dans son identification. Sa réponse pourrait donc être :

"J'ai les cheveux attachés, et ils sont longs de 20 cm."

Pour que les hauteurs des voix ne puissent pas aider l'interrogateur, les réponses seront écrites ou mieux, tapées à la machine. Le meilleur dispositif consiste à avoir un téléscripteur de communication entre les deux pièces. Sinon, les questions et les réponses

peuvent être répétées par un intermédiaire. L'objectif du jeu pour le troisième joueur (B) est d'aider l'interrogateur. La meilleure stratégie pour cette personne est probablement de donner les vraies réponses. Elle peut ajouter des choses comme "Je suis la femme, ne l'écoutez pas !" à ses réponses, mais ça ne servira à rien parce que l'homme pourra faire des remarques similaires.

Maintenant posons la question, "qu'arrivera-t-il si une machine prend la place de A dans ce jeu ?". L'interrogateur(-machine) se trompera-t-il aussi souvent que lorsque le jeu est joué par des hommes et des femmes ? Ces questions remplacent notre question originale, "Les machines peuvent-elles penser ?"

2 Critique du nouveau problème

De la même façon qu'on peut se demander "Quelle est la réponse à la question sous sa nouvelle forme", on peut aussi se demander "Cette nouvelle question est-elle digne de réflexion ?". On réfléchira à cette dernière question sans tergiverser davantage, coupant là une régression infinie.

Le nouveau problème présente l'avantage de dessiner une frontière assez nette entre les capacités physiques et intellectuelles d'un homme. Aucun ingénieur ou chimiste ne se targue de pouvoir produire un matériau qui ne soit pas distinguable de la peau humaine. Il est possible qu'un jour cela soit réalisé, mais même en supposant que cette invention ait un jour été faite, nous pouvons sentir combien il y a peu en commun entre le fait d'essayer de rendre une "machine pensante" plus humaine en la recouvrant de cette peau artificielle. La forme dans laquelle nous avons spécifié le problème montre que les conditions empêchent l'interrogateur de voir ou toucher les autres personnages, ou d'entendre leurs voix. D'autres avantages du critère proposé peuvent se voir dans des questions et réponses specimen. Par exemple :

Q : S'il vous plaît, écrivez-moi un sonnet avec comme sujet le quatrième pont.

A : Ne comptez pas sur moi. Je ne pourrai jamais écrire de poèmes.

Q : Ajoutez 34957 à 70764.

A : (Pause d'environ 30 secondes et réponse donnée ensuite) 105721.

Q : Jouez-vous aux échecs ?

A : Oui.

Q : J'ai K en K1, et pas d'autres pièces. Vous avez K en K6 et R en R1. C'est votre tour. Que jouez-vous ?

A : (Après une pause de 15 secondes) R-R8 mat.

La méthode des questions et réponses semble adaptée pour introduire presque tous les champs d'application que nous souhaiterions inclure. Nous ne voulons pas pénaliser la machine pour son incapacité à briller dans des concours de beauté, ni pénaliser un humain parce qu'il perd une course contre un avion. Les conditions de notre jeu rendent ces incompétences non pertinentes. Les "témoins" peuvent se vanter, s'il pensent que c'est judicieux, autant qu'ils le souhaitent sur leurs charmes, leur force, ou leur héroïsme, mais l'interrogateur ne peut pas leur demander de démonstrations pratiques.

Le jeu peut peut-être être critiqué sur la base que les chances sont trop désavantageuses contre la machine. Si l'homme voulait faire semblant d'être une machine, il se montrerait vraisemblablement très médiocre. Il serait mis en échec du premier coup par sa lenteur et ses erreurs en arithmétique. Les machines ne sont-elles pas capables de faire quelque-chose qu'on a l'habitude de nommer *penser* mais qui est très différent de ce qu'un humain fait ? Cette objection est une objection très forte, mais au moins, nous pouvons dire que si, néanmoins, une machine peut être construite pour jouer de façon satisfaisante au jeu de l'imitation, il ne faudrait pas être troublé par cette objection.

On doit souligner qu'en jouant au "jeu de l'imitation", la meilleure stratégie pour la machine peut possiblement être quelque-chose de différent de l'imitation du comportement humain. C'est possible, mais je pense qu'il est peu probable que cela ait un grand effet. Dans tous les cas, on n'a pas d'intention ici de faire des recherches en théorie des jeux, et on supposera que la meilleure stratégie est d'essayer de fournir des réponses qui seraient naturellement données par un humain.

3 Machines concernées par le jeu

La question que nous avons posée en 1 ne sera pas bien définie tant que nous n'aurons pas précisé ce que signifie le mot "machine." Il est naturel de vouloir autoriser toutes sortes de techniques d'ingénierie dans nos machines. Nous pouvons aussi souhaiter la possibilité que des ingénieurs puissent construire une machine qui fonctionne, mais qui ne peut être considérée comme satisfaisante car ses constructeurs ont utilisé une méthode qui est très expérimentale. Finalement, nous souhaitons exclure des machines les humains nés de la façon (biologique) habituelle. Il est difficile de cadrer les définitions de manière à satisfaire les trois conditions. On peut par exemple insister pour que l'équipe

d'ingénieurs soient tous du même sexe, mais cela ne serait pas vraiment satisfaisant, car il est possible de construire un individu complet à partir d'une seule cellule de peau (par exemple) d'un humain. Réaliser cela serait un exploit de technique biologique méritant les plus grandes louanges, mais nous ne saurions considérer cela comme un cas de "construction d'une machine pensante". Cela nous enjoint à abandonner la nécessité que tout type de technique puisse être permis. Nous sommes d'autant plus prêts à cela étant donné que l'intérêt actuel pour les "machines pensantes" a été motivé par une sorte particulière de machine, habituellement appelées "ordinateurs électroniques" ou "ordinateurs digitaux". Selon cette suggestion, nous n'autorisons que les ordinateurs digitaux à prendre part à notre jeu.

Cette restriction apparaît au premier abord comme étant très drastique. Je vais essayer de montrer qu'elle ne l'est pas en réalité. Faire cela nécessite une légère prise en compte de la nature et des propriétés de ces ordinateurs.

On peut aussi dire que cette identification des machines aux ordinateurs digitaux, selon notre critère qualifiant la "pensée" ne sera pas satisfaisant si (contrairement à mon sentiment), il s'avère que les ordinateurs ne se montrent pas bons dans le jeu.

Il y a déjà un certain nombre d'ordinateurs digitaux capable de travailler, et on peut se demander "Pourquoi ne pas tenter cette expérience ? Il serait facile de satisfaire les conditions du jeu. Un certain nombre d'interrogateurs pourraient être utilisés, et des statistiques pourraient être calculées qui compteraient le nombre de fois où l'identification aurait été correcte". La réponse rapide à cela est que nous ne nous demandons pas si tous les ordinateurs digitaux seraient capables de jouer à ce jeu, ni si les ordinateurs actuels pourraient le faire, mais si on peut imaginer des ordinateurs capables de le faire. Mais c'est seulement une réponse rapide. Nous verrons cette question sous un autre angle ultérieurement.

4 Ordinateurs digitaux

L'idée derrière les ordinateurs digitaux peut être expliquée en disant que ces machines sont destinées à prendre en charge toutes les opérations qui pourraient être effectuées par des calculateurs humains. Le calculateur humain est supposé suivre des règles fixées ; il n'a pas autorité à dévier d'elles dans le moindre détail. On peut supposer que ces règles sont fournies dans un livre, qui est modifié à chaque fois qu'une nouvelle tâche est à réaliser. Il a aussi une quantité de papier illimitée sur laquelle il peut faire ses calculs. Il peut aussi faire ses multiplications et additions sur une "calculatrice de bureau", mais cela n'a pas d'importance.

Si nous utilisons l'explication ci-dessus comme une définition, nous pouvons risquer

d'être confronté à un argument circulaire. On évite cela en donnant un aperçu des moyens par lesquels l'effet désiré peut être obtenu. Un ordinateur digital peut habituellement être vu comme constitué de 3 parties :

(i) la mémoire.

(ii) l'unité d'exécution.

(iii) le contrôle.

La mémoire stocke l'information, et correspond au papier du calculateur humain, que ce soit le papier sur lequel il fait ses calculs ou celui sur lequel son livre de règles est imprimé. Puisque l'humain peut effectuer une partie de ses calculs de tête, une partie de la mémoire de la machine correspond à cette mémoire du calculateur humain.

L'unité d'exécution est la partie qui effectue les opérations individuelles impliquées dans un calcul. Ce que sont ces opérations variera d'une machine à l'autre. Habituellement des opérations assez longues peuvent être effectuées comme "Multiplier 3540675445 par 7076345687" mais pour certaines machines, seules des opérations très simples comme "Ecris 0" sont envisageables.

Nous avons mentionné que le "livre de règles" fourni à l'ordinateur est stocké dans la machine dans une partie de sa mémoire. On appelle cette partie de la mémoire la "table d'instructions". C'est la tâche du contrôle de voir que ces instructions sont exécutées correctement et dans le bon ordre. Le contrôle vérifie que ces contraintes sont respectées.

L'information en mémoire est habituellement découpée en paquets de taille modérément petite. Dans une machine, par exemple, un paquet peut consister en dix unités décimales. Des nombres sont assignés aux parties de la mémoire dans lesquelles les différents paquets d'information sont stockés, de manière systématique. Une instruction typique peut dire :

"Ajoute le nombre stocké à la position 6809 à celui en 4302 et met le résultat obtenu dans cette dernière unité de mémoire utilisée."

Il est inutile de dire que cela ne sera pas stocké en machine en anglais courant. Ca a plus de chances d'être codé dans une forme comme 6809430217. Ici 17 dit quelle opération doit être faite sur les deux nombres. Dans ce cas, l'opération est celle décrite plus haut, i.e. "Ajouter les nombres...". On notera que l'instruction prend 10 chiffres et constitue ainsi un paquet d'information, de façon très pratique. Le contrôle prendra normalement les instructions devant être effectuées dans l'ordre dans lequel elles ont été stockées, mais occasionnellement, une instruction comme "N'exécute pas l'instruction à

la position 5606, et continue à partir de là" peut être rencontrée, ou à nouveau "Si à la position 4505, il y a un 0, exécute ensuite l'instruction stockée en 6707, sinon continue séquentiellement".

Les instructions de ces dernières sortes sont très importantes parce qu'elles rendent possible le remplacement d'une séquence d'instructions par une autre plusieurs fois de suite jusqu'à ce qu'une certaine condition soit remplie, et ce faisant, d'exécuter non pas les nouvelles instructions à chaque répétition, mais la même instruction plusieurs fois successives. Pour prendre une analogie domestique, supposons que Maman veuille que Tommy passe chez le cordonnier chaque matin lorsqu'il va à l'école pour demander si ses chaussures sont prêtes, elle peut le lui rappeler chaque matin. Alternativement, elle peut coller un papier une fois pour toutes dans le hall et il le verra quand il part à l'école et cela lui rappellera de demander pour les chaussures, et le post-it sera détruit quand Tommy reviendra avec les chaussures réparées.

Le lecteur doit accepter cela comme un fait que les ordinateurs digitaux peuvent être construits, et par exemple ont été construits, selon les principes que nous avons décrits, et qu'ils peuvent quasiment simuler les actions d'un calculateur humain.

Le livre de règles qu'utilise le calculateur humain dont nous avons parlé est bien sûr une fiction pratique. Les véritables calculateurs humains se rappellent vraiment ce qu'ils ont à faire. Si l'on veut qu'une machine simule le comportement d'un calculateur humain pour des tâches complexes, on doit lui demander comment il fait et ensuite traduire sa réponse en utilisant une table d'instructions. Construire des tables d'instructions est habituellement appelé "programmer". "Programmer une machine pour qu'elle fasse l'opération A" signifie mettre l'instruction appropriée dans la machine de manière à ce qu'elle exécute A.

Une variante intéressante à l'idée d'ordinateur digital est celle d'"ordinateur digital contenant un composant aléatoire". Ces ordinateurs ont des instructions impliquant un lancer de dé ou un processus électronique équivalent ; une telle instruction par exemple peut être "lancer le dé et mettre le nombre résultant dans la case mémoire 1000". Parfois une telle machine est décrite comme possédant un libre-arbitre (même si je n'utiliserai pas cette expression moi-même). Il n'est normalement pas possible de déterminer, en observant la machine, si elle contient un composant basé sur l'aléa, car un effet similaire peut être produit par les composants non aléatoires en rendant les choix dépendant des chiffres des formes décimales des nombres en jeu.

La plupart des ordinateurs digitaux actuels ont seulement une mémoire limitée. Il n'y a pas de difficulté théorique dans l'idée d'un ordinateur avec une mémoire illimitée. Bien sûr que seule une partie limitée de la mémoire peut être utilisée à tout instant,

puisque seulement une quantité finie de mémoire a pu être fabriquée, mais on peut imaginer qu'on en rajoutera de plus en plus lorsque ça sera nécessité. De tels ordinateurs ont un intérêt théorique spécifique et nous les appellerons ordinateurs de capacité infinie.

L'idée d'ordinateur digital est une idée ancienne. Charles Babbage, Professeur Lucasien de Mathématique à Cambridge de 1828 à 1839, avait conçu une telle machine, appelée le moteur analytique, mais il ne l'a jamais terminé. Bien que Babbage ait eu les idées principales, sa machine n'était pas à l'époque un projet très attractif. La vitesse qui aurait été atteinte par cette machine aurait définitivement été plus rapide que celle d'un calculateur humain mais quelque-chose comme 100 fois plus lente que la machine de Manchester, elle-même l'une des plus lentes des machines modernes. La mémoire devait être purement mécanique, et utiliser des roues et des cartes.

Le fait que le moteur analytique de Babbage soit complètement mécanique nous aidera à nous débarrasser d'une superstition. On attache souvent de l'importance au fait que les ordinateurs digitaux modernes sont électriques, et que le système nerveux est également électrique. Puisque la machine de Babbage n'était pas électrique, et puisque tous les ordinateurs digitaux sont en quelque sorte équivalents, nous voyons que cette utilisation de l'électricité ne peut pas avoir d'importance théorique. Bien sûr, l'électricité intervient lorsqu'on doit traiter des signaux rapidement, et ce n'est donc pas surprenant que nous la trouvions dans ces deux sortes de connexions. Dans le système nerveux, les phénomènes chimiques sont au moins aussi importants que les phénomènes électriques. Dans certains ordinateurs, le système de mémoire est principalement acoustique. Le fait d'utiliser l'électricité est ainsi vu comme une simple similarité superficielle. Si nous souhaitons trouver de telles similarités, nous devrions plutôt chercher des analogies mathématiques au niveau des fonctions.

5 L'universalité des ordinateurs digitaux

Les ordinateurs digitaux considérés dans la section précédente peuvent être classés parmi les "machines à états discrets". Ce sont des machines qui bougent par sauts ou clics d'un état parfaitement défini à un autre. Ces états sont suffisamment différents pour que la possibilité de confusion entre eux soit ignorée. De façon stricte, il n'y a pas de sous-machines. Tout va continûement dans la réalité. Mais il y a de nombreuses sortes de machines pour lesquelles il est profitable de les penser comme des machines à états discrets. Par exemple, pour les états d'un système lumineux, c'est une fiction pratique que d'imaginer le système soit comme complètement allumé, soit comme complètement éteint. Il doit y avoir des positions intermédiaires, mais dans de nombreux contextes, nous pouvons les oublier. Comme exemple de machine à états discrets, nous pouvons considérer une roue qui tourne 120 fois par seconde, mais qui peut être arrêtée par un levier de l'extérieur ; supposons en plus qu'une lampe doive être allumée lorsque la roue

est dans une certaine position. Cette machine pourrait être décrite abstraitement 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 en entrée i_0 ou i_1 (position du levier). L'état interne à tout moment est déterminé par l'état final et l'état d'entrée selon la table

		Etat final		
		q_1	q_2	q_3
Signal en entrée	i_0	q_2	q_3	q_1
	i_1	q_2	q_3	q_1

Les signaux de sortie, les seules indications visibles de l'extérieur de l'état interne (les lumières) sont décrites par la table

Etat interne	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 en supposant qu'elles n'ont qu'un nombre fini d'états possibles.

Il semble que selon un état initial donné de la machine et selon les signaux d'entrée, il soit toujours possible de prédire tous les états futurs. Cela rappelle le point de vue de Laplace selon lequel si l'on connaît l'état complet de l'univers à un instant temporel donné, comme décrit par les positions et vitesses de toutes les particules, on devrait pouvoir prédire tous ses états futurs. La prédiction que nous sommes en train de considérer ici, pourtant, est, cependant, bien plus proche de la prédictabilité que celle considérée par Laplace. Le système de l'"univers comme un tout" est tel que des erreurs plutôt petites dans les conditions initiales peuvent avoir un gros effet plus tard. Le déplacement d'un seul électron d'un billionième de centimètre à un moment peut faire toute la différence entre un homme tué par une avalanche un an après, ou bien en réchappant. C'est une propriété essentielle des systèmes mécaniques que nous avons appelées "machines à états discrets" que ce phénomène ne se produise pas. Même quand nous considérons de vraies machines physiques plutôt que des machines idéelles, une connaissance raisonnable de l'état à un moment donné amène une connaissance raisonnable de l'état quelques étapes plus tard.

Comme nous l'avons mentionné, les ordinateurs digitaux font partie de la classe des machines à états discrets. Mais le nombre d'états d'une telle machine peut être vraiment très grand. Par exemple, le nombre d'états pour une machine de Manchester est environ de $2\ 165,00$, i.e., environ $10 \times 50\ 000$. Comparons cela à notre exemple de la roue à clics décrite précédemment, qui a trois états. Il n'est pas difficile de voir pourquoi le nombre d'états pourrait être aussi grand. L'ordinateur contient une mémoire correspondant au papier utilisé par un calculateur humain. Il doit être possible d'écrire dans

la mémoire n'importe quelle combinaison de symboles qui pourrait avoir été écrite sur le papier. Pour des raisons de simplicité, supposons que l'on n'utilise comme symboles que les chiffres de 0 à 9. Les variations d'écriture sont ignorées. Supposons que l'ordinateur puisse avoir 100 feuilles de papier, chacune contenant 50 lignes de 30 chiffres chacune. Alors le nombre d'états des trois machines mises ensemble est $10\ 100 \times 50 \times 30$ i.e., 10 150 000. C'est à peu près le nombre d'états de trois machines de Manchester mises ensemble. Le logarithme en base 2 de ce nombre d'états est habituellement appelé la "capacité mémoire" de la machine. Ainsi, la machine de Manchester a une capacité mémoire d'environ 165 000 et la machine à roues de notre exemple a une mémoire de taille 1.6. Si deux machines sont mises ensemble, leurs capacités doivent être ajoutées pour obtenir la capacité de la machine résultante. Cela amène à la possibilité d'assertion comme "La machine de Manchester contient 64 barrettes magnétiques chacune d'une capacité de 2560, huit tubes électroniques avec une capacité de 1280. Des petits ajouts accessoires de mémoires s'élèvent à 300, ce qui fait un total de 174 380."

La table correspondant à la machine à états discrets étant donnée, il est possible de prédire ce qu'elle fera. Il n'y a pas de raison que ce calcul ne puisse pas être effectué par un ordinateur digital. Si l'on suppose qu'il peut l'effectuer assez rapidement, l'ordinateur digital devrait pouvoir simuler le comportement d'une machine à états discrets quelconque. On pourrait alors imaginer jouer au jeu de l'imitation avec la machine en question (comme B) et l'ordinateur digital simulateur (comme A) et l'interrogateur ne pourrait pas les distinguer l'un de l'autre. Bien sûr, l'ordinateur digital doit avoir une capacité mémoire adéquate et doit être en mesure d'exécuter ses instructions suffisamment rapidement. De plus, il doit être programmé à nouveau à chaque fois qu'on souhaite le faire simuler une nouvelle machine.

Cette propriété spéciale des ordinateurs digitaux, le fait qu'ils puissent simuler n'importe quelle machine à états discrets, est décrit en les désignant par l'expression "machines universelles". L'existence de machines avec cette propriété a comme conséquence importante que, les considérations de vitesse étant mises à part, il n'est pas nécessaire de concevoir de nouvelles machines variées pour réaliser des processus de calcul variés. Cela peut aussi être fait avec un ordinateur digital, chacun adéquatement programmé pour chaque cas. On verra qu'une conséquence de cela est que tous les ordinateurs digitaux sont équivalents dans un certain sens.

Nous pouvons maintenant considérer à nouveau le point mis en avant à la fin du §3. On avait suggéré conjecturalement que la question "Les machines peuvent-elles penser" soit remplacée par "Les ordinateurs digitaux imaginables pourraient-ils se débrouiller dans le jeu de l'imitation ?". Si nous le souhaitons, nous pouvons rendre cela plus général de façon artificielle et demander "Y a-t-il des machines à états discrets qui seraient capables de faire cela ?". Mais en regardant la propriété universelle, nous voyons que

l'une ou l'autre de ces questions est équivalente à celle-ci : "Fixons notre attention sur un ordinateur digital C. Est-il vrai qu'en modifiant cet ordinateur pour avoir une mémoire adéquate, et en augmentant adéquatement sa rapidité d'exécution, et en lui fournissant un programme adéquat, C peut être fabriqué de façon à jouer de façon satisfaisante la partie A du jeu de l'imitation, la partie B étant tenue par un humain ?".

6 Des vues différentes de la même question

Nous pouvons considérer maintenant que le socle a été clarifié et nous sommes prêts à débattre de notre question "Les machines peuvent-elles penser ?" et de la variante qui a été citée à la fin de la section précédente. Nous ne pouvons pas abandonner la forme originale du problème, car les opinions diffèreront sur le caractère approprié de la substitution des problèmes et nous devons au moins entendre ce qui doit être dit au sujet de cette relation.

Cela simplifiera les choses pour le lecteur si j'explique d'abord mes propres convictions sur le sujet. Considérons d'abord la forme la plus précise de la question. Je crois que dans cinquante ans environ, il sera possible de programmer des ordinateurs d'une capacité d'envison 109, pour les faire jouer au jeu de l'imitation de façon à ce qu'un interrogateur n'ait plus que 70 pour cent de chances de faire l'identification correcte après cinq minutes de questions/réponses. La question originale, "Les machines peuvent-elles penser ?", je la crois trop privée de sens pour mériter une discussion. Pourtant, je crois qu'à la fin du siècle, l'usage des mots et l'opinion éduquée en général aura tellement changé qu'on sera capable de parler de pensée des machines sans s'attendre à être contredit. Je crois même qu'on ne sert aucun but utile en cachant de telles convictions. L'idée populaire que les scientifiques avancent inexorablement d'un fait établi à un fait établi, en n'étant jamais influencés par aucune conjecture améliorée, est légèrement erronée. A partir du moment où l'on sait clairement ce qui constitue les faits et ce qui constitue les conjectures, il n'en résulte aucun préjudice. Les conjectures sont d'une grande importance puisqu'elles suggèrent les orientations utiles de la recherche.

Je vais maintenant considérer les opinions opposées à la mienne.

(1) L'objection théologique

Penser est une fonction de l'âme immortelle humaine. Dieu a donné une âme immortelle à chaque homme et à chaque femme, mais non pas aux animaux ou aux machines. De ce fait, aucun animal et aucune machine ne peut penser.

Je ne peux accepter aucun élément de l'argument ci-dessus, mais vais essayer d'y répondre en termes théologiques. Je trouverais l'argument plus convaincant si les animaux

étaient classés du côté des humains, parce qu'à mon sens, la différence entre l'inanimé et l'animé est plus grande que celle entre l'humain et les autres animaux. Le caractère arbitraire d'un tel point de vue orthodoxe devient clair si nous considérons la manière dont cet argument peut être perçu par un membre d'une autre communauté religieuse. Comment les Chrétiens regardent-ils l'opinion musulmane selon laquelle les femmes n'ont pas d'âme ? Mais laissons ce point de côté et retournons à l'argument principal. Il me semble que l'argument cité ci-dessus entraîne une sérieuse restriction à l'omnipotence du Tout-Puissant. Il est admis qu'il y a certaines choses qu'Il ne peut pas faire comme par exemple faire que un soit égal à deux, mais ne devrions-nous pas croire qu'Il a la liberté de donner une âme à un éléphant s'Il trouve cela approprié ? Nous pourrions nous attendre à ce qu'Il puisse exercer son pouvoir conjointement à une mutation qui pourvoierait l'éléphant d'un cerveau adéquatement amélioré pour gérer des besoins de ce type. Un argument du même type peut être utilisé dans le cas des machines. Il peut sembler différent parce qu'il est plus difficile à "avaler". Mais cela signifie seulement que nous pensons qu'il serait moins vraisemblable qu'Il considère les circonstances appropriées pour leur conférer une âme. Les circonstances en question sont discutées dans le reste de cet article. En essayant de créer de telles machines, nous ne serions pas irrespectueux en usurpant Son pouvoir de créer des âmes, de même que nous ne le sommes pas quand nous procréons et avons des enfants : nous sommes plutôt, dans les deux cas, des instruments de Sa volonté pour créer les réceptacles des âmes qu'Il crée.

Pourtant, ceci, c'est de la pure spéculation. Je ne suis pas très impressionné par les arguments théologiques, quel que soit ce qu'ils sont destinés à expliquer. De tels arguments se sont souvent avérés insatisfaisants par le passé. A l'époque de Galilée, il avait été dit que les textes, "Et le soleil dura alors... et ne descendit pas pendant une journée complète" (Josué 10 :13) et "Il a posé les fondations de la Terre, de manière à ce qu'elle ne bouge jamais" (Psaume 104) étaient une réfutation de la théorie de Copernic. Avec nos connaissances actuelles, un tel argument semble futile. Quand cette connaissance n'était pas encore acquise, cela a fait une impression plutôt différente.

(2) L'objection "la tête dans le sable"

"Les conséquences du fait que des machines pensent seraient trop horribles. Espérons et croyons qu'elles ne pourront jamais le faire."

Cet argument est rarement exprimé de manière si claire qu'il ne l'est ci-dessus. Mais il affecte beaucoup d'entre nous qui le pensent complètement. Nous aimons penser que l'Homme est en quelque sorte supérieur au reste de la création. Il est mieux qu'il puisse être vu comme nécessairement supérieur, car alors il n'y a pas de danger qu'il perde sa position dominante. La popularité de l'argument théologique est clairement reliée à ce sentiment. Il est vraisemblable qu'un tel avis soit très partagé par les intellectuels, parce

qu'ils considèrent le pouvoir de la pensée comme plus important que ne le font d'autres personnes, et ils sont donc plus enclins à baser leurs convictions sur la supériorité de l'Homme concernant cette possibilité.

Je ne crois pas que cet argument soit suffisamment substantiel pour nécessiter une réfutation. La consolation serait plus appropriée : peut-être qu'elle pourrait être recherchée dans la métempsychose.

(3) L'objection mathématique

Il y a un certain nombre de résultats de logique mathématique qui peuvent être utilisés pour montrer qu'il y a des limitations à ce que peuvent les machines à états discrets. Le plus connu de ces résultats est le théorème de Gödel (1931) qui montre que dans tout système logique suffisamment puissant, des assertions peuvent être formulées qui ne peuvent ni être prouvées ni être réfutées dans ce système, à moins que le système lui-même dans son ensemble ne soit démontré comme étant inconsistant. Il y a d'autres résultats, similaires à celui-là en quelque sorte, dus à Church (1936), Kleene (1935), Rosser, et Turing (1937). Le dernier résultat est le plus pratique à considérer, puisqu'il fait directement référence aux machines, alors que les autres peuvent seulement être utilisés dans un argument indirect : par exemple, si le théorème de Gödel doit être utilisé, on a besoin d'avoir en plus des moyens de décrire les systèmes logiques par rapport aux machines, et les machines par rapport aux systèmes logiques. Le résultat en question [1] fait référence à un type de machine qui consiste essentiellement en un ordinateur digital avec une capacité infinie. Il établit qu'une machine ne peut effectuer certaines tâches. Si une telle machine doit donner des réponses à des questions comme dans le jeu de l'imitation, il y aura des questions auxquelles soit elle donnera une réponse fausse, soit elle échouera à donner une quelconque réponse quel que soit le temps qui lui est alloué pour y répondre. Il peut, bien sûr, exister de nombreuses telles questions, et les questions auxquelles une certaine machine ne pourra pas répondre pourront cependant recevoir une réponse satisfaisante de la part d'une autre machine. Nous supposons là que les questions sont de type fermé, i.e. elles attendent comme réponse appropriée une réponse "Oui" ou "Non", plutôt que des questions ouvertes comme "Que pensez-vous de Picasso ?". Les questions dont nous savons que les machines doivent échouer sont de ce type. "Considérons les machines spécifiées comme suit... Cette machine répondra-t-elle "oui" à toute question ?". Les points de suspension doivent être remplacés par une description d'une machine de forme standard, qui pourrait être du type de celles qui ont été envisagées au §5. Quand la machine décrite partage une certaine relation comparativement simple avec la machine que l'on interroge, on peut montrer que la réponse est soit fausse soit ne vient jamais. C'est cela le résultat mathématique : on soutient qu'il prouve l'incapacité des machines qui n'ont pas un intellect humain.

1. *ndt* : le résultat le plus pratique

La réponse courte à cet argument est que même s'il est établi qu'il y a des limitations à la possibilité de penser pour une machine particulière, il a seulement été affirmé, sans aucune sorte de preuve, que de telles limitations ne s'appliquent pas à l'intellect humain. Mais je ne pense pas qu'un tel point de vue doive être rejeté si clairement. A chaque fois que l'on pose à une telle machine une question critique appropriée, et qu'elle fournit une réponse, et que nous savons que cette réponse est fausse, cela nous donne un certain sentiment de supériorité. Ce sentiment est-il illusoire ? C'est sans doute vrai, mais je ne pense pas que trop d'importance doive être accordée à cela. Nous donnons nous-mêmes trop souvent des mauvaises réponses à des questions pour être satisfaits que cela soit utilisé comme évidence de la faillibilité des machines. De plus, notre supériorité peut seulement se ressentir dans certaines occasions précises, en relation avec une machine spécifique contre laquelle nous avons enregistré un petit triomphe. Il ne saurait être question de triompher simultanément de toutes les machines. En bref, de ce fait, il se pourrait qu'il existe un humain qui soit plus intelligent que n'importe quelle machine, mais alors à nouveau, il y aura des machines plus intelligentes que lui, et etc.

Ceux qui croient en cet argument mathématique devraient, je pense, accepter la plupart du temps le jeu de l'imitation comme base de discussion. Ceux qui sont convaincus des deux objections précédentes ne seraient probablement intéressés par aucun critère quel qu'il soit.

(4) L'argument de la conscience

Cet argument est très bien exprimé par le Professeur Jefferson dans son discours d'obtention de la médaille Lister, dont je cite : "Pas avant qu'une machine n'ait écrit un sonnet ou composé une symphonie à cause d'émotions et pensées ressenties, et pas par le hasard de concordance de symboles, nous ne pourrons être d'accord sur le fait qu'une machine égale un cerveau humain, c'est-à-dire que non seulement cette machine écrit mais qu'en plus, elle sait ce qu'elle a écrit. Aucun mécanisme ne pourrait ressentir de plaisir lorsqu'il réussit (et pas seulement des signaux artificiels, des stratagèmes faciles), ne pourrait ressentir de difficulté quand ses vannes fusionnent, être réconforté par des flatteries, ou rendu misérable par ses erreurs, charmé par le sexe, en colère ou déprimé parce qu'il n'arrive pas obtenir ce qu'il veut.".

Cet argument semble être un déni de la validité de notre test. Selon une forme extrême de ce point de vue selon lequel le seul moyen d'être sûr qu'une machine pense est d'être une machine et de se sentir penser. On pourrait alors décrire ce qu'une machine ressent au monde, mais bien sûr personne n'aurait la possibilité de donner son avis. Selon ce point de vue également, le seul moyen de savoir si un homme pense est d'être cet homme particulier. C'est en fait un point de vue solipsiste. Il peut être le point de vue le plus

logique mais il rend la communication des idées difficile. A est susceptible de croire “A pense mais B ne pense pas” tandis que B croit “B pense mais pas A.”. Plutôt que de continuer à tergiverser éternellement sur ce point, il est habituel d'avoir la convention polie que tout le monde pense.

Je suis sûr que le Professeur Jefferson ne souhaite pas adopter le point de vue extrême et solipsiste. Il serait vraisemblablement prêt à accepter le jeu de l'imitation comme test. Le jeu (sans le joueur B) est fréquemment utilisé en pratique sous le nom de *viva voce* pour découvrir si quelqu'un comprend effectivement quelque-chose ou bien “répète cette chose comme un perroquet”. Ecoutons un tel extrait du jeu *viva voce* :

Interrogateur : Dans la première ligne de votre sonnet “Je te comparerai à une journée d'été”, est-ce qu'“un jour de printemps” serait aussi bien ou mieux ?

Témoin : Ca n'irait pas.

Interrogateur : Et “un jour d'hiver”, ça sonnerait bien.

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

Interrogateur : Diriez-vous que M. Pickwick vous rappelle Noël ?

Témoin : D'une certaine manière, oui.

Interrogateur : Bien, Noël est un jour d'hiver, et je ne pense que M. Pickwick soit gêné par la comparaison.

Témoin : Je ne pense pas que vous ayez raison. Par un jour d'hiver, on entend un jour d'hiver basique, plutôt qu'un jour aussi spécial que le jour de Noël.

Et le tout à l'avenant... Que dirait le Professeur Jefferson si la machine à écrire des sonnets était capable de répondre ainsi au jeu de *viva voce* ? Je ne sais pas s'il considèrerait la machine comme “fournissant plutôt artificiellement” ces réponses, mais si les réponses étaient aussi satisfaisantes et soutenues que dans l'extrait ci-dessus, je ne pense pas qu'il la décrirait comme “un stratagème facile”. Cette phrase est destinée, je pense, à couvrir des mécanismes tels que l'inclusion dans la machine d'un enregistrement de quelqu'un lisant un sonnet, avec la possibilité appropriée de la mettre en marche de temps en temps.

En résumé donc, je pense que la plupart de ceux qui sont en faveur de l'argument de la conscience devraient être persuadés d'abandonner un tel point de vue plutôt que d'être forcés à prendre une position solipsiste. Ils souhaiteraient alors probablement accepter

notre test.

Je ne veux pas donner l'impression que je pense qu'il n'y a aucun mystère à propos de la conscience. Il y a, par exemple, quelque-chose de paradoxal lié au fait de souhaiter la localiser. Mais je ne pense pas que ces mystères aient nécessairement besoin d'être résolus avant que nous puissions répondre à la question à laquelle nous nous intéressons dans le présent article.

(5) Des arguments d'impossibilités diverses

Ces arguments sont de la forme "Je vous accorde le fait que vous puissiez fabriquer des machines qui font toutes les choses que vous avez mentionnées mais vous ne serez jamais capable d'en fabriquer une qui puisse faire X.". De nombreuses possibilités sont suggérées pour X dont je fournis une sélection :

Etre gentil, ingénieux, beau, amical, prendre des initiatives, avoir le sens de l'humour, dire ce qui est vrai ou pas, faire des erreurs, tomber amoureux, aimer les fraises à la chantilly, rendre quelqu'un amoureux de soi, apprendre de l'expérience, utiliser les mots à bon escient, être le sujet de ses propres pensées, présenter autant de variété dans son comportement qu'un être humain, faire quelque-chose de vraiment nouveau.

Aucun étayage pour soutenir ces assertions n'est en général fourni. Je crois que ces arguments sont la plupart du temps fondés sur le principe de l'induction scientifique. Un humain a vu des milliers de machines dans sa vie. De ce qu'il a vu d'elles, il tire un certain nombre de conclusions. Elles sont laides, chacune d'elle est conçue pour atteindre un objectif spécifique, si on a un autre objectif que celui-là, elles ne nous sont d'aucune utilité, la variété de comportement de toutes ces machines est très faible, etc., etc. Naturellement, il conclut que ces propriétés sont des propriétés nécessaires des machines en général. Beaucoup de telles limitations sont associées à la très faible capacité mémoire de la plupart des machines (je suppose que l'idée de capacité mémoire est étendue de façon à couvrir les machines autres que celles à états discrets. La définition exacte n'importe pas puisqu'aucune précision mathématique n'est revendiquée dans la présente discussion). Il y a quelques années, quand on avait encore peu entendu parlé des ordinateurs digitaux, il était possible d'éluder une telle incrédulité, si l'on mentionnait leurs propriétés sans décrire leur construction. Cela était sûrement dû à une application similaire du principe d'induction scientifique. Ces applications du principe sont bien sûr largement inconscientes. Quand un enfant qui s'est brûlé craint le feu et montre qu'il le craint en l'évitant, on pourrait dire qu'il applique une induction scientifique (je pourrais décrire son comportement de nombreuses autres façons). Les travaux et les habitudes des humains ne semblent pas être un matériau adéquat auquel appliquer l'induction scientifique. Une grande partie de l'espace-temps doit être étudié, si on souhaite ob-

tenir des résultats fiables. Sinon nous pourrions (comme le font la plupart des enfants anglais) décider que tout le monde parle anglais, et qu'il est idiot d'apprendre le français.

Il y a, cependant, des remarques particulières à faire à propos d'un certain nombre d'incapacités (des machines) qui ont été mentionnées. L'impossibilité d'aimer les fraises à la chantilly pourrait avoir semblé futile au lecteur. On pourrait fabriquer une machine pour apprécier ce met délicieux, mais toute tentative de faire cela semblerait débile. Ce qui est important dans cette incapacité, c'est qu'elle contribue à d'autres incapacités, e.g. à la difficulté qu'il y ait la même sorte d'amitié entre un homme et une machine qu'entre un homme et un autre.

L'argument "Les machines ne peuvent pas se tromper." semble curieux. On est tenté de rétorquer "Sont-elles pires de ce fait ?". Mais adoptons une attitude plus sympathique, et essayons de voir ce que l'on cherche réellement à dire par là. Je pense que cette critique peut s'expliquer selon le jeu de l'imitation. L'argument dit que l'interrogateur pourrait distinguer la machine de l'humain simplement en lui posant un certain nombre de problèmes d'arithmétique. La machine serait démasquée à cause de sa piètre compétence à les résoudre. La réponse à cela est simple. La machine (programmée pour jouer au jeu) n'essaierait pas de donner des réponses correctes aux problèmes arithmétiques. Elle introduirait délibérément des erreurs de manière à tromper l'interrogateur. Une erreur mécanique lui montrerait probablement à travers une décision inadéquate quelle sorte d'erreur faire dans un problème arithmétique. Même cette interprétation de la critique n'est pas suffisamment sympathique. Mais nous ne pouvons perdre de la place à entrer plus avant dans les détails. Il me semble que cette critique dépend d'une confusion entre deux sortes d'erreurs. Nous pourrions les appeler les "erreurs de fonctionnement" et les "erreurs de raisonnement". Les erreurs de fonctionnement sont dues à des défauts électriques ou mécaniques qui empêchent la machine de se comporter comme elle a été programmée à le faire. Dans les discussions philosophiques, on aime ignorer ces possibilités d'erreurs ; ce faisant, on discute de "machines abstraites". Ces machines abstraites sont des idées mathématiques plutôt que des objets physiques. Par définition, elles sont incapables d'erreurs de fonctionnement. C'est en ce sens qu'on peut dire que "les machines ne font jamais d'erreurs". Les erreurs de raisonnement peuvent quant à elles seulement advenir quand un sens est attaché aux signaux en sortie de la machine. La machine pourrait, par exemple, écrire des équations mathématiques, ou des phrases en anglais. Quand une phrase fausse est tapée, nous disons que la machine a fait une erreur de raisonnement. Il n'y a bien sûr aucune raison de dire qu'une machine ne peut jamais faire ce genre d'erreur. Elle pourrait ne rien faire et écrire sans fin "$0 = 1$". Pour prendre un exemple moins pervers, elle pourrait avoir une méthode pour trouver ses conclusions par induction scientifique. Nous pouvons nous attendre à ce qu'une telle méthode puisse parfois amener à des résultats erronés.

On peut répondre à l'argument selon lequel une machine ne peut être le sujet de sa propre pensée seulement si l'on peut montrer qu'une machine pense à certains sujets. Néanmoins, "Le sujet des opérations d'une machine" semble ne rien vouloir dire, au moins pour les personnes qui s'intéressent à un tel sujet. Si, par exemple, la machine essaye de trouver une solution de l'équation $x^2 - 40x - 11 = 0$, on peut être tenté de décrire cette équation comme faisant partie du sujet de la pensée de la machine à ce moment-là. Dans ce sens-là, une machine peut sans aucun doute être le sujet de sa propre pensée. Elle peut être utilisée pour mettre à jour ses propres programmes, ou pour prédire les effets des altérations de sa propre structure. En observant les résultats de son propre comportement, elle peut modifier ses propres programmes pour atteindre certains buts plus efficacement. Ce sont des possibilités de l'avenir proche, plutôt que des rêves utopiques.

La critique selon laquelle une machine ne peut pas avoir un comportement aussi diversifié est juste une manière de dire qu'elle ne peut pas avoir beaucoup de capacité mémoire. Jusqu'à assez récemment, une capacité mémoire d'un millier de digits était très rare.

Les critiques que nous considérons ici sont souvent des formes déguisées de l'argument concernant la conscience. Habituellement, si on maintient qu'une machine peut faire l'une de ces choses, et si l'on décrit le genre de méthode que la machine devrait utiliser pour ce faire, cela ne fera pas plus qu'une impression. On pense qu'une telle méthode (quelle qu'elle soit, car elle peut être mécanique) est vraiment plutôt basique. Comparer cela à ce qui est entre parenthèses dans l'assertion de Jefferson vue précédemment.

(6) L'objection de Lady Lovelace

L'information en notre possession la plus détaillée concernant le moteur analytique de Babbage est un mémoire d'Ada Lovelace (1842). Dans celui-ci, elle écrit "Le moteur analytique ne prétend pas *inventer* quoi que ce soit. Il peut faire *tout ce dont nous savons comment le lui ordonner*." (caractères en italique selon l'écrit original d'Ada Lovelace). Cette assertion est citée par Hartree (1949) qui ajoute : "Cela n'implique pas qu'il ne puisse être possible de construire un équipement électronique qui "penserait par lui-même", ou dans lequel, en termes biologiques, quelqu'un pourrait incorporer un réflexe conditionné, qui servirait de base à un "apprentissage". Que cela soit ou pas possible en principe est une question stimulante et excitante, suggérée par certains développements récents. Mais il ne semble pas que les machines construites ou projetées de l'être ont cette propriété.".

Je suis en profond accord avec Hartree sur cela. Il sera noté qu'il ne dit pas que les machines en question n'avaient pas cette propriété, mais plutôt que la conviction de Lady Lovelace ne l'encourageait pas à penser que les machines avaient cette propriété.

Il est assez possible que les machines en question avaient cette propriété dans un certain sens. Car supposons qu'une machine à états discrets ait cette propriété. Le moteur analytique était un ordinateur digital universel, et donc, si sa capacité mémoire et sa vitesse étaient adéquates, il aurait pu par programmation simuler la machine en question. Il est probable que cet argument n'ait pas été trouvé par la Comtesse ou par Babbage. Dans tous les cas, ils n'avaient aucune obligation de dire tout ce qui aurait pu être dit.

L'ensemble de cette question sera considéré à nouveau dans le paragraphe concernant les machines apprenantes.

Une variante de l'objection de Lady Lovelace affirme qu'une machine ne pourra "jamais faire quelque-chose de vraiment nouveau". On peut parer à cet argument par la litanie "Rien de nouveau sous le soleil". Qui peut être certain qu'un "travail original" qu'il a effectué n'était pas simplement la récolte d'une semaille qui a été plantée en lui lorsqu'il a reçu son enseignement, ou l'effet de principes généraux bien connus. Une meilleure variante de l'objection dit qu'une machine ne peut jamais "nous prendre par surprise". Cet argument est un défi plus direct et on peut le contrer directement. Les machines m'ont surpris très souvent. Cela est largement dû au fait que je ne fais pas suffisamment de calcul pour décider de ce que je dois attendre d'elles, ou plutôt parce que, même si je fais un tel calcul, je le fais à toute vitesse, d'une façon négligeante, en prenant des risques. Peut-être que je me dis à moi-même "Je suppose que le voltage ici doit être le même que là : bon, supposons que c'est le cas.". Bien sûr, j'ai souvent tort, et le résultat m'est une surprise car quand l'expérimentation arrive à son terme, j'ai oublié les suppositions que j'ai faites. Admettre cela me laisse ouvert pour écouter des conférences au sujet de ces mauvaises manières, mais ne laisse aucun doute sur ma crédibilité quand je témoigne de ces surprises que je rencontre.

Je ne m'attends pas à ce que cette réponse fasse se taire les points de vue critiques sur mon point de vue. On dira probablement que les surprises sont dues à une action mentale créative de ma part, et ne sauraient être attribuées à la machine. Cela nous ramène à l'argument de la conscience, et loin de l'idée de surprise. C'est une ligne d'argumentation que nous devons considérer comme fermée, mais il est peut-être pire de remarquer que l'appréciation de quelque-chose comme étant surprenant nécessite autant d'"action mentale créative", que l'événement surprenant ait été provoqué par un humain, un livre, une machine ou quoi que ce soit d'autre.

Le point de vue selon lequel les machines ne peuvent pas créer de surprises est dû, je pense, à une erreur que les philosophes et les mathématiciens font particulièrement souvent. C'est la supposition selon laquelle dès qu'un fait est présenté à un esprit, toutes les conséquences de ce fait germent dans l'esprit simultanément à ce fait. Cette supposition est très utile dans de nombreuses circonsctances, mais on oublie trop souvent qu'elle est

fausse. Une conséquence naturelle du fait de faire de la sorte est que l'on suppose alors qu'il n'y pas de vertu à traiter les conséquences des données et des principes généraux.

(7) L'argument au sujet de la continuité du système nerveux

Le système nerveux n'est certainement pas une machine à états discrets. Une petite erreur concernant l'information à propos de la taille de l'impulsion nerveuse en entrée d'un neurone peut engendrer une très grande différence sur l'impulsion en sortie. On peut arguer que, cela étant, on ne peut s'attendre à être capable de simuler le système nerveux par un système à états discrets.

Il est vrai qu'une machine à états discrets doit être différente d'une machine continue. Mais si nous nous plaçons dans les conditions du jeu de l'imitation, l'interrogateur ne pourra pas tirer avantage de cette différence. La situation peut être clarifiée si nous considérons une machine continue sonore la plus simple qui soit. Un analyseur différentiel sera parfait (un analyseur différentiel est une certaine sorte de machine qui n'est pas du type à états discrets et qui est utilisée pour certaines sortes de calculs). Certains d'entre eux fournissent leurs réponses sous une forme typée, et sont ainsi capables d'être utilisés dans le jeu d'imitation. Il ne serait pas possible qu'un ordinateur digital prédise exactement quelles réponses l'analyseur digital donnerait à un problème, mais il pourrait être capable de donner la bonne sorte de réponse. Par exemple, si on lui demande de donner la valeur de π (environ 3.1416), il serait raisonnable de choisir au hasard parmi les valeurs 3.12, 3.13, 3.14, 3.15, 3.16 avec les probabilités de 0.05, 0.15, 0.55, 0.19, 0.06 (disons). Dans ces circonstances, il serait très difficile pour l'interrogateur de distinguer l'analyseur digital de l'ordinateur digital.

(8) L'argument provenant du caractère informel du comportement

Il n'est pas possible de produire un ensemble de règles destinées à décrire ce qu'un homme devrait faire dans tout ensemble concevable de circonstances. On pourrait par exemple avoir une règle qui est qu'on doit s'arrêter quand on voit un feu rouge, et avancer quand on voit un feu vert mais que faire dans le cas où ils apparaissent tous les deux simultanément ? On peut peut-être décider qu'il est plus prudent de s'arrêter. Mais d'autres difficultés peuvent alors découler plus tard de cette prise de décision. Essayer de fournir des règles de conduite couvrant toutes les éventualités, même celles provenant des feux de la circulation, semble impossible. Avec tout ça, je suis d'accord.

A cause de ces raisons, on dit que nous ne pouvons pas être des machines. Je vais essayer de reproduire l'argument, mais je crains de ne pas lui rendre justice. C'est un peu quelque chose comme ça : "si tout homme avait un ensemble défini de règles de conduite par lesquelles il pouvait régenter sa vie, il ne serait rien de plus qu'une machine. Mais

de telles règles n'existent pas, et donc les hommes ne peuvent pas être des machines.". Le déséquilibre est flagrant. Je ne pense pas que l'argument soit jamais exprimé en ces termes, mais je crois que c'est cependant ce genre d'argument qui est utilisé. La confusion certaine entre les "règles de conduite" et les "lois de comportement" obscurcit le problème. Par "règles de conduite", je veux parler de préceptes tels que "Arrêtez-vous si vous voyez des feux rouges", que l'on peut faire, et dont on peut être conscient. Par "règles de comportement", je veux parler des lois de la nature telles qu'elles s'appliquent à un corps humain comme "si vous le pincez, il va crier". Si vous substituez "lois de comportement qui régentent sa vie" à "lois de conduite par lesquelles il régente sa vie" dans l'argument cité, le juste équilibre devient atteignable. Car nous croyons qu'il est non seulement vrai que voir son comportement commandé par des règles de comportement implique d'être une sorte de machine (même si cela ne signifie pas nécessairement être une machine à états discrets), mais qu'inversement, être une machine implique d'être régenté par de telles lois de comportement. Pourtant, nous ne pouvons pas nous convaincre aussi simplement de l'absence de lois pour tous les comportements possibles, ou de règles de conduite dans toutes les circonstances possibles. La seule manière que nous connaissions pour trouver de telles lois est l'observation scientifique, et nous ne connaissons sûrement aucun contexte dans lequel nous pourrions dire "Nous avons assez cherché. Il n'y a pas de telles lois.".

Nous pouvons démontrer de façon plus forte encore qu'une telle assertion serait injustifiée. Car supposons que nous soyons certains de trouver de telles lois si elles existaient. Alors étant donnée une machine à états discrets, il serait certainement possible de découvrir en l'observant suffisamment comment prédire son comportement futur et cela dans un temps raisonnable, disons mille ans. Mais cela ne semble pas être le cas. J'ai écrit un petit programme sur la machine de Manchester qui n'utilise que 1000 unités mémoire, et ce programme est tel que si l'on fournit à la machine un nombre à seize chiffres en renvoie un autre en deux secondes. Je défie quiconque d'en apprendre suffisamment sur ce programme en étudiant les réponses qu'il fournit, et d'être capable de prédire la réponse qu'il renverrait pour n'importe quelle valeur non déjà testée.

(9) L'argument de la perception extra-sensorielle

Je suppose que le lecteur est familier avec l'idée de perception extra-sensorielle (ESP), et qu'il connaît la signification des quatre termes suivants : la télépathie, la clairvoyance, la prescience et la psychokinèse. Ces phénomènes troublants semblent mettre en défaut 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 réordonner les idées de quelqu'un pour que ces faits cadrent dans sa pensée. Une fois qu'on les a admis, ça n'a pas l'air d'être une grosse étape de plus que de croire aux fantômes. L'idée que nos corps bougent simplement selon les lois connues

de la physique, et de quelques autres idées non encore découvertes mais à peu près similaires, va être la première idée vers laquelle se diriger.

Cet argument est selon moi un argument assez fort. On peut dire en réponse à cela que de nombreuses théories scientifiques semblent rester applicables en pratique, malgré leur opposition aux idées de perceptions extra-sensorielles ; c'est en fait ce que l'on obtient très simplement, quand on oublie celles-ci. C'est plutôt confortable, et on craint que la pensée soit la seule sorte de phénomène où les ESP puissent être particulièrement pertinents.

Un argument plus spécifique basé sur les ESP pourrait être le suivant : "Jouons au jeu de l'imitation, en utilisant comme témoins un homme qui est un bon récepteur télépathe, et un ordinateur digital. L'interrogateur peut poser des questions comme "A quelle suite la carte dans ma main droite appartient-elle ?". L'homme par clairvoyance télépathique donnera la bonne réponse pour 130 cartes sur 400. La machine peut seulement deviner au hasard et elle obtiendra peut-être 104 bonnes réponses, et alors l'interrogateur fera la bonne identification.". Il y a une possibilité intéressante qui s'ouvre ici. Supposez que l'ordinateur digital contienne un générateur de nombres aléatoires. Alors il sera naturel qu'il utilise ce dispositif pour décider de la réponse à fournir. Mais alors le générateur de nombres aléatoires pourra être manipulé par les pouvoirs psychokinétiques de l'interrogateur. Peut-être que ces pouvoirs psychokinétiques pourraient permettre à la machine de répondre juste plus souvent que ce qui est attendu selon un calcul de probabilités, de telle façon que l'interrogateur ne soit plus capable de faire la bonne identification. D'un autre côté, il pourrait être capable de deviner correctement sans poser aucune question, par sa clairvoyance. Avec l'ESP, tout peut arriver.

Si on admet la télépathie, il sera nécessaire de resserrer nos tests. La situation devrait être regardée comme analogue à celle qui arriverait si l'interrogateur se parlait à lui-même et si l'un des compétiteurs écoutait avec son oreille collée au mur. Mettre les compétiteurs dans une "pièce pour preuve télépathique" satisferait toutes les contraintes.

7 Machines apprenantes

Le lecteur aura compris que je n'ai aucun argument convaincant de nature positive pour appuyer mon point de vue. Si j'en avais, je n'aurais pas fait tant d'efforts pour montrer la fausseté des points de vue contraires. Je vais exposer la conviction qui est la mienne.

Retournons un instant à l'objection de Lady Lovelace, qui exprime que la machine ne peut que faire ce qu'on lui dit de faire. On pourrait dire qu'un humain "injecte" une idée dans la machine, et qu'elle répondra en quelque sorte, et puis retournera dans le calme, comme une corde de piano frappée par un marteau. Une autre image similaire serait

celle d'une pile atomique avec moins d'énergie qu'une certaine dose critique : une idée qu'on lui injecte correspond là à un neutron entrant dans la pile de l'extérieur. Un tel neutron provoquera une perturbation certaine qui éventuellement n'aura aucun effet. Si, cependant, la taille de la pile est suffisamment augmentée, la perturbation causée par un électron entrant va continuer à augmenter jusqu'à ce que la pile entière soit détruite. Y a-t-il un phénomène correspondant pour les esprits, y en a-t-il un pour les machines ? Il semble y en avoir un pour l'esprit humain. La majorité d'entre eux semblent être "sous-critiques", i.e. correspondre à cette analogie des piles de taille sous-critique. Une idée présentée à un tel esprit va en moyenne donner naissance à une idée en retour. Une petite proportion d'entre eux sont super-critiques. Une idée présentée à un tel esprit donnera naissance à une "théorie" complète consistant en une idée secondaire, une troisième et d'autres plus éloignées. Les esprits des animaux semblent être définitivement sous-critiques. En adhérant à cette analogie, on se demande "Une machine peut-elle être rendue super-critique ?".

L'analogie de la "peau d'oignon" est aussi utile. En considérant les fonctions de l'esprit ou du cerveau, nous trouvons certaines opérations que nous pouvons expliquer en termes purement mécaniques. Nous disons que cela ne correspond pas à l'esprit véritable : c'est une sorte de peau que nous devons arracher si nous voulons trouver le véritable esprit. Mais là, nous trouvons une nouvelle peau, que nous devons arracher aussi, et etc. En procédant de cette manière, nous n'arrivons jamais au "véritable" esprit, ou bien est-ce que nous finissons par arriver à une peau qui ne contient rien ? Dans ce dernier cas, tout l'esprit est mécanique (ce ne sera pas pour autant une machine à états discrets. Nous avons déjà discuté de cela.).

Ces deux derniers paragraphes ne prétendent pas être des arguments convaincants. Ils devraient plutôt être décrits comme des "récitations tendant à produire une croyance".

Le seul appui réellement satisfaisant qui peut être donné au point de vue exprimé au début du §6, sera d'attendre la fin du siècle et puis de réaliser l'expérience décrite. Mais que pouvons-nous dire en attendant ? Par quelles étapes devrions-nous passer aujourd'hui si l'expérience devait s'avérer une réussite ?

Comme je l'ai expliqué, le problème est principalement un problème de programmation. Des avancées dans l'ingénierie devront avoir être effectives également, mais il ne semble pas qu'elles ne puissent pas permettre ce que l'expérience requiert. Les estimations de la capacité mémoire du cerveau varie de 10^{10} à 10^{15} unités binaires (bits). Je penche pour les valeurs les plus basses et je crois que seule une petite fraction est utilisée pour les types de pensées les plus hautes. La plupart de ces unités sont probablement utilisées pour retenir les impressions visuelles, je ne serais pas surpris si plus de 10^9 était nécessitées pour jouer de manière satisfaisante au jeu de l'imitation, à n'importe quel niveau contre

un homme aveugle (Note : la capacité de l'Encyclopaedia Britannica, $11^{\text{ème}}$ édition, est de 2×10^9). Une capacité mémoire de 10^7 serait une possibilité praticable même par les techniques actuelles. Il ne serait probablement pas nécessaire du tout d'augmenter les vitesses de traitement des opérations des machines. Des parties des machines modernes qui peuvent être regardées comme analogues aux cellules nerveuses travaillent environ mille fois plus vite que ces dernières. Cela fournirait une "marge de sécurité" qui pourrait couvrir la perte de vitesse pouvant advenir de multiples façons. Notre problème alors est de trouver comment programmer ces machines pour qu'elles puissent jouer au jeu. A mon niveau de travail actuel, je produis environ un millier de digits de code par jour, ce qui fait qu'environ soixante personnes, travaillant pendant cinquante ans pourraient accomplir le travail, si aucune ligne n'est jamais mise à la poubelle. Des méthodes plus expéditives semblent souhaitables.

Dans la tentative d'imiter l'esprit d'un adulte humain, nous devons forcément répondre au défi de comprendre le processus qui l'a amené à l'état dans lequel il est. Nous pouvons considérer trois composants :

(a) L'état initial de l'esprit, disons à la naissance,
(b) L'éducation à laquelle il a été soumis,
(c) L'expérience, différente de son éducation, qu'il a vécue.

Plutôt que d'essayer de produire un programme pour simuler l'esprit adulte, pourquoi ne pas plutôt essayer d'en produire un qui simule celui de l'enfant ? Si celui-ci était soumis à une éducation adéquate, il permettrait d'obtenir un esprit adulte. On présume que le cerveau de l'enfant est une sorte de carnet de note qu'on achète chez le papetier. Plutôt peu de mécanismes, et beaucoup de papier blanc (mécanisme et écriture sont de notre point de vue presque synonymes). Notre espoir qu'il y ait peu de mécanisme dans le cerveau de l'enfant permettrait de le programmer facilement. La quantité de travail d'éducation que nous pouvons supposer pour notre système doit être environ la même que celle nécessaire pour l'éducation d'un cerveau d'enfant, en première approximation.

Nous avons ainsi divisé notre problème en deux parties : la programmation du cerveau d'enfant et le processus d'éducation. Ces deux aspects restent très intimement connectés. Nous ne pouvons espérer trouver une bonne machine-enfant à la première tentative. On doit alors expérimenter l'enseignement à une telle machine et voir si elle apprend bien. On peut alors en essayer une autre et voir si elle est meilleure ou pire. Il y a une connexion évidente entre ce processus et l'évolution, par les identifications suivantes :

$$\begin{aligned} \text{Structure de la machine enfant} &= \text{matériau héréditaire} \\ \text{Changements dans la machine enfant} &= \text{mutation} \\ \text{Sélection naturelle} &= \text{jugement de l'expérimentateur} \end{aligned}$$

On peut espérer, cependant, que ce processus sera plus rapide que l'évolution. Le fait

que le test d'adaptation perdure est dû à un lent processus de mesure des avantages. L'expérimentateur, en exerçant son intelligence, devrait être capable de l'accélérer. Le fait que le processus ne soit pas restreint à des mutations hasardeuses est aussi important. Si l'on peut retrouver la trace de certaines faiblesses, l'expérimentateur pourra probablement penser aux sortes de mutations qui l'amélioreront.

Il ne sera pas possible d'appliquer exactement le même processus d'apprentissage à une machine qu'à un enfant normal. Elle n'aura pas de jambes, et on ne pourra pas lui demander de sortir et de remplir le seau à charbon. Il est possible qu'elle n'ait pas d'yeux. Mais pourtant, ces déficiences peut être surmontées par une ingénierie intelligente, on ne peut pas envoyer la créature à l'école, les autres enfants ne pourront pas trop jouer avec elle. On doit payer des frais de scolarité. On n'a pas besoin de trop se préoccuper de ces problèmes de jambes, d'yeux, etc. L'exemple de Mademoiselle Helen Keller montre que l'éducation est possible du moment que la communication est possible dans les deux sens entre l'élève et l'enseignant et qu'elle peut être assurée d'une manière ou d'une autre.

Nous associons habituellement des récompenses et des punitions au processus d'enseignement. On peut construire des machines enfants simples ou les programmer selon cette sorte de principe. La machine doit être ainsi construite que les événements qui précèdent l'arrivée d'un signal de punition ne doivent pas être répétés alors que la probabilité de répétition de l'occurrence d'événements qui ont précédé un signal de récompense augmentera. Ces définitions ne présupposent aucun sentiment de la part de la machine, j'ai fait quelques expériences avec une telle machine enfant, et j'ai réussi à lui apprendre quelques petites choses, mais la méthode d'enseignement était trop peu orthodoxe pour qu'on puisse considérer cette expérimentation comme un réel succès.

L'utilisation de punitions et récompenses peut être au mieux une partie du processus d'enseignement. Pour parler grossièrement, si l'enseignant n'a pas d'autre moyen de communiquer avec l'élève, le montant d'information qui peut l'atteindre n'excède pas le nombre total de punitions et de récompenses appliquées. Quand un enfant apprendra à répéter "Casablanca", il sera sûrement très endolori, si ce mot peut seulement être découvert par la technique des "Vingt questions", chaque lettre "o" occasionnant un coup. Il est nécessaire par conséquent d'avoir des canaux de communication "non émotionnels". Si ceux-ci existent, on peut enseigner à une machine par des punitions et des récompenses à obéir à des ordres exprimés dans un certain langage, e.g. un langage symbolique. Ces ordres doivent être transmis à travers des canaux "non émotionnels". L'utilisation de ce langage ne diminuera pas beaucoup le nombre de punitions et récompenses requis.

Les opinions peuvent varier au sujet de la complexité qui est appropriée pour la machine enfant. On pourrait essayer de la rendre aussi simple que possible de façon à

satisfaire les principes généraux. De façon alternative, on pourrait avoir un système complet d'inférences logiques "pré-construites". Dans ce dernier cas, la mémoire devrait être largement occupée par les définitions et les propositions. Les propositions auraient plusieurs sortes de statuts, e.g. le statut de faits bien établis, de conjectures, de théorèmes mathématiquement prouvés, d'assertions données par une autorité, d'expressions ayant la forme logique de propositions mais qui ne sont pas des croyances. Certaines propositions peuvent être décrites comme "impératives" (les consignes). La machine pourrait être construite de telle manière que dès qu'un ordre est classé comme "bien établi", l'action appropriée serait automatiquement effectuée. Pour illustrer cela, supposons que l'enseignant dise à la machine "fais tes devoirs maintenant". Cela pourrait avoir comme conséquence l'inclusion de "L'enseignant dit "Fais tes devoirs" " dans les faits bien établis. Un autre tel fait pourrait être "Tout ce que dit l'enseignant est vrai". Combiner ces faits pourrait finalement amener la consigne "Fais tes devoirs maintenant", parmi les faits bien établis, et cela, par construction de la machine, signifierait que les devoirs seraient alors effectués, et cet effet est très satisfaisant. Les processus d'inférence utilisés par la machine n'ont pas besoin d'être tels qu'ils satisfassent les logiciens les plus sévères. Il pourrait par exemple n'y avoir aucune hiérarchie des types. Mais cela ne signifie pas pour autant que des erreurs de type auront lieu, ni que nous allons tomber dans des pièges. Les consignes adéquates (exprimées dans les systèmes, ne faisant pas partie des règles du système) telles que "N'utilisez pas une classe à moins que ce soit une sous-classe d'une classe qui a été mentionnée par l'enseignant" peuvent avoir un effet similaire à "Ne vous approchez pas du bord".

Les consignes auxquelles une machine qui n'a pas de membres peut obéir sont de nature plutôt intellectuelles, comme dans l'exemple donné ci-dessus (des devoirs). Les consignes importantes de l'ensemble des consignes seront les consignes qui déterminent l'ordre dans lequel les règles du système logique concerné doivent être appliquées. Car à chaque étape de la mise en œuvre d'un système logique, il y a un grand nombre d'étapes alternatives, chacune d'entre elles pouvant être appliquée, du moment qu'elle obéit aux règles du système logique. Ces choix font la différence entre un système qui raisonne vite et un système qui raisonne lentement, et non pas la différence entre un système qui raisonne juste, et un système qui raisonne faux. Les propositions amenant à des consignes de cette sorte peuvent être "Quand Socrate est mentionné, utilise le syllogisme de Barbara" ou bien "Si une méthode s'est avérée plus rapide qu'une autre, n'utilise pas la méthode lente.". Certaines de ces règles peuvent être "données par l'autorité" mais d'autres peuvent être produites par la machine elle-même, e.g. par induction scientifique.

L'idée d'une machine apprenante peut sembler paradoxale à certains lecteurs. Comment les règles opératoires de la machine peuvent-elles être modifiées ? Elles devraient décrire complètement comment la machine réagira quelle que soit les événements auxquels elle est soumise, quels que soient les changements qu'elle peut subir. Les règles sont ainsi

assez indépendantes du temps. C'est presque vrai. L'explication du paradoxe est que les règles qui seront modifiées par le processus d'apprentissage sont plutôt d'un type moins prétentieux, et n'ont qu'une validité éphémère. Le lecteur pourrait faire un parallèle avec la Constitution des Etats-Unis.

Une propriété importante d'une machine apprenante est que son enseignant sera souvent ignorant de ce qui se passe à l'intérieur de la machine, bien qu'il soit cependant capable dans une certaine mesure de prédire le comportement de son élève. Cela devrait s'appliquer d'autant plus à l'éducation d'une machine qui était précédemment une machine enfant avec un entraînement bien conçu (bien programmé). Ceci est en contraste clair avec la procédure normale qui est que quand on utilise une machine pour faire des calculs, on a une image mentale claire de l'état de la machine à tout moment durant l'exécution du calcul. Cet objectif est difficile à atteindre. Le point de vue qu'"une machine ne peut faire que ce qu'on lui a ordonné de faire" semble étrange face à cela. La plupart des programmes que nous pouvons entrer en machine auront comme résultat qu'elle fera quelque-chose à quoi nous ne pouvons donner du sens (ou du moins, auquel nous attribuerons un sens complètement hasardeux). Se comporter intelligemment consiste justement à s'éloigner du comportement complètement discipliné tel que celui utilisé lorsqu'on effectue un calcul, et d'effectuer plutôt de légers changements, ce qui ne signifie pas pour autant de se comporter de manière aléatoire, ou de faire des répétitions sans fin. Une autre conséquence importante de la préparation de notre machine pour qu'elle puisse exécuter sa partie dans le jeu de l'imitation en la lui enseignant est que la "faillibilité humaine" sera vraisemblablement omise de façon assez naturelle, i.e. sans "coaching" particulier (le lecteur devrait réconcilier cela avec le point de vue de certaines des pages précédentes). Les processus appris ne produisent pas un résultat sûr à cent pour cent ; si c'était le cas, ils ne pourraient pas être désappris.

Il est probablement sage d'inclure un élément aléatoire dans la machine apprenante. Un composant aléatoire est assez utile quand on cherche une solution à un problème. Supposons par exemple que nous voulions trouver un nombre entre 50 et 200 qui soit égal au carré de la somme de ses chiffres. On peut commencer par 51, puis essayer 52 et continuer jusqu'à trouver un nombre qui marche. On peut alternativement choisir des nombres au hasard jusqu'à en trouver un correct. Cette méthode a comme avantage qu'il n'est pas nécessaire de garder trace des valeurs qui ont été testées, mais le désavantage c'est qu'on peut tester deux fois le même nombre, mais ça n'est pas important s'il y a plusieurs solutions possible. La méthode systématique présente le désavantage qu'il peut y avoir un énorme bloc de nombres successifs qui ne sont pas solutions dans la région qu'on testera en premier. Maintenant le processus d'apprentissage peut être vu comme la recherche d'une forme de comportement qui satisfera l'enseignant (ou bien satisfera un autre critère). Puisqu'il y a probablement un très grand nombre de solutions satisfaisantes, la méthode aléatoire semble être meilleure que la méthode systématique.

On peut noter qu'elle est utilisée dans le processus analogue de l'évolution. Mais là, la méthode systématique n'est pas possible. Comment pourrait-il être gardé trace des différentes combinaisons génétiques qui ont été essayées, de manière à éviter de les tester à nouveau ?

Nous pouvons espérer que les machines finiront par entrer en compétition avec les hommes dans des domaines purement intellectuels. Mais quels sont les meilleurs domaines par lesquels commencer ? Même si c'est une question difficile, beaucoup de personnes pensent qu'une activité très abstraite, comme le fait de jouer aux échecs, serait la plus adaptée. On peut aussi arguer qu'il serait plus judicieux de pourvoir les machines des meilleurs organes des sens que l'on peut payer, et d'alors leur apprendre à parler anglais. Ce processus pourrait être identique à celui par lequel on enseigne habituellement à un enfant. On montrerait les choses du doigt et on dirait leur nom, etc. A nouveau je ne sais pas quelle est la bonne réponse, mais je pense que les deux approches devraient être tentées.

Nous ne pouvons que regarder à une petite distance temporelle dans le futur, mais nous voyons là qu'il y a beaucoup de choses à faire.