Machines intelligentes, une théorie hérétique

A. M. Turing

"Vous ne pouvez pas faire qu'une machine pense pour vous". Ceci est un lieu commun qui est habituellement accepté sans questionnement. Ce sera l'objectif du présent article que de questionner cette phrase.

La plupart des machines fabriquées à des fins commerciales sont conçues pour effectuer une tâche très spécifique de façon sûre et extraordinairement vite. Très souvent, une telle machine fait la même série d'opérations de nombreuses fois sans aucune variété. Ce fait à propos des machines réelles est un argument puissant pour de nombreuses personnes de la phrase énoncée plus haut. Pour un mathématicien logicien, cet argument n'est pas valable, car il a été démontré qu'il est possible de fabriquer des machines qui feront quelque chose qui est très proche de la pensée. Elles pourront, par exemple, tester la validité d'une preuve formelle dans le système des *Principia Mathematica*, ou même dire d'une formule d'un tel système si elle est prouvable ou réfutable. Dans le cas où la formule n'est ni prouvable, ni réfutable,une telle machine ne se comportera vraisemblablement pas d'une manière très satisfaisante, car elle continuera à tourner indéfiniment, sans produire de résultat du tout, mais cela ne peut pas être considéré comme très différent comme attitude de la réaction des mathématiciens, qui ont par exemple travaillé des centaines d'années sur la question de savoir si le dernier théorème de Fermat est vrai ou faux. Dans le cas des machines de ce type, une sorte d'argument plus subtil est nécessaire. Par le fameux théorème de Gödel, ou par un argument similaire, on peut montrer que quelle que soit la manière dont une machine est construite, il y aura des cas où la machine échouera à donner une réponse, mais où un mathématicien pourra en donner une. D'un autre côté, la machine a certains avantages sur le mathématicien. On peut s'appuyer sur tout ce que la machine fait, si l'on suppose qu'il n'y aura pas de panne mécanique, tandis que le mathématicien commet parfois des erreurs, selon une certaine proportion. Je crois que ce danger du mathématicien faisant des erreurs est un corollaire inévitable de la possibilité qu'il utilise parfois de mettre en œuvre une méthode complètement nouvelle. Cela semble être confirmé par le fait bien connu que les personnes les plus fiables n'utilisent en général pas de méthodes vraiment nouvelles.

Ce que je prétends, c'est que des machines peuvent être construites qui simuleront la

Philosophia Mathematica (3) Vol. 4 (1996), pp. 256-260.
http://www.turingarchive.org/browse.php/B/4

Résumé. Dans cet essai posthume, Turing prétend qu'il pourrait être possible de construire une machine qui contiendrait un composant aléatoire et un analogue au principe de plaisir en psychologie, à qui on pourrait apprendre, et qui pourrait finalement devenir plus intelligente que les humains.

pensée humaine de façon très approchée. Parfois elles feront des erreurs, et parfois elles feront de nouvelles assertions très intéressantes, et globalement, les réponses qu'elles fourniront en sortie seront à première vue quasiment les mêmes réponses que celles fournies par un cerveau humain. Le contenu de mon assertion réside dans la grande fréquence attendue d'assertions vraies, et elle ne peut pas, je pense, être reçue comme une assertion vraie. Il ne pourrait pas, par exemple, être suffisant pour dire simplement qu'une machine exprimera une assertion vraie un jour ou l'autre, car un exemple d'une telle machine serait celui d'une machine qui exprime toutes les assertions un jour ou l'autre. Nous savons comment les construire, et comme elles devraient produire (probablement) des assertions vraies et des assertions fausses à peu près aussi fréquemment les unes que les autres, leurs verdicts seraient bien pire. Ce serait la réaction réelle de la machine aux circonstances qui prouverait ce que je prétends, si tant est qu'elle puisse être prouvée.

Voyons plus précisément la nature de cette argumentation. Il est clairement possible de fabriquer une machine qui fournirait un compte-rendu très précis à propos d'elle-même pour n'importe quel jeu de tests, si cette machine était suffisamment élaborée. Pourtant, ceci ne pourrait à nouveau que très difficilement être considéré comme une preuve adéquate. Une telle machine finirait par se perdre en faisant toujours la même sorte d'erreur encore et encore, et en étant quasiment incapable de se corriger elle-même, ou en étant corrigée par des arguments fournis de l'extérieur. Si la machine était capable d'une certaine manière d'"apprendre par expérience", ce serait beaucoup plus impressionnant. Si c'était le cas, il semblerait qu'il n'y ait aucune raison réelle pour que quelqu'un n'essaie pas de commencer avec une machine comparativement simple, et, en la soumettant à un certain nombre d'expériences adéquates, à la transformer en une autre machine qui serait plus élaborée, et qui serait capable de gérer un nombre plus élevé de contingences. Ce processus serait probablement accéléré par une sélection appropriée des expériences auxquelles la machine serait soumise. On pourrait appeler ce processus l'"éducation". Mais ici, nous devons être prudents. Ce serait assez facile d'arranger des expériences de telle manière qu'elles fassent que la structure de la machine se retrouve dans une forme particulière, et cela serait de façon évidente une grosse manière de tricher, presque autant que d'avoir un homme caché dans la machine. A nouveau ici, le critère exprimant ce qui devrait être considéré comme raisonnable en termes d'"éducation" ne peut pas être mis en termes mathématiques, mais je suggère que ce qui suit pourrait être considéré comme adéquat en pratique. Supposons que l'on souhaite que la machine comprenne l'anglais, et qu'elle n'ait ni mains ni pieds, et aucun besoin de se nourrir, aucun désir de cigarette, elle occupera son temps principalement à jouer à des jeux tels que les échecs ou le Go, et peut-être le Bridge. La machine est équipée d'un clavier de machine à écrire sur lequel on peut taper toute remarque qu'on souhaite lui faire, et elle écrit en retour toute chose qu'elle veut dire. Je suggère que l'éducation de la machine soit confiée à un maître d'école très compétent qui est intéressé par le projet mais à qui l'on n'a communiqué aucune connaissance détaillée du fonctionnement interne de la machine. Le mécanicien

qui a construit la machine, par contre, doit la maintenir en état de marche, et s'il suspecte que la machine n'a pas effectué ses tâches correctement, il a le doit de la remettre dans un état antérieur et de demander au professeur de répéter sa leçon à partir de cet endroit-là, mais il n'a pas le droit de prendre part en aucune manière au processus d'enseignement. Puisque cette procédure est uniquement destiné à tester la *bonne foi* de la mécanique, il faut que j'insiste sur le fait qu'elle ne pourrait pas être adoptée dans les étapes expérimentales. Comme je le vois, le processus d'enseignement serait en pratique essentiel à l'obtention d'une machine raisonnablement intelligente dans un intervalle de temps raisonnablement court. L'analogie humaine suggère cela.

Je peux maintenant donner quelques indications sur la manière de fonctionner que l'on peut attendre d'une telle machine. La machine devrait incorporer une mémoire. Cela ne nécessite pas beaucoup d'explication. Elle devrait consister en une liste de toutes les assertions qui existent pour elle ou bien qu'elle a faites, et de tous les mouvements qu'elle a faits, et de toutes les cartes qu'elle a jouées dans les jeux auxquels elle a participé. Tout ça sera listé dans l'ordre chronologique. Outre cette mémoire évidente, il y aura un certain nombre d'"index d'expériences". Pour expliquer cette idée, je suggèrerai la forme qu'un tel index pourrait prendre. Cela pourrait être un classement en ordre alphabétique des mots qui ont été utilisés en donnant l'"instant" auquel ils ont été utilisés, de telle manière qu'on puisse les retrouver dans la mémoire. Un autre tel index pourrait contenir les images des humains ou des parties des gobans qui auront été rencontrées. Aux périodes plus tardives de l'éducation, la machine pourrait être étendue pour inclure d'importantes parties de la configuration de la machine à tout moment, ou en d'autres termes, elle pourrait se rappeler quelles ont été les pensées qu'elle a eues. Cela donnerait naissance à de nouvelles formes d'indexation très fructueuses. Les nouvelles formes d'indexation pourraient être introduites en tenant compte de motifs spéciaux observés dans les index déjà utilisés. Les index seraient utilisés de la façon suivante : à chaque fois qu'un choix devrait être fait sur l'action à effectuer ensuite, les motifs de la situation présente seraient observés dans les index disponibles, et le choix précédent qui aurait été fait dans des situations similaires, et la sortie, bonne ou mauvaise, qui en aurait découlé seraient retrouvés.

Le nouveau choix sera fait selon toutes ces données. Cela entraîne un certain nombre de problèmes. Si certaines indications sont favorables et d'autres non favorables, que faire ? La réponse à cela diffèrera probablement d'une machine à l'autre et variera également en fonction du degré d'éducation de la machine. Initialement, probablement que quelques règles assez succinctes suffiront, e.g. faire l'action qui a le plus de votes en sa faveur. A un stade bien plus tardif de l'enseignement, la question entière de la procédure dans de tels cas aura probablement été étudiée par la machine elle-même, par les moyens d'une sorte d'index, et cela pourra résulter en quelque chose de beaucoup plus sophistiqué et, on espère, à des formes de règles bien plus satisfaisantes. Il semble probable pourtant

que les formes comparativement brutes des règles seront elles-mêmes raisonnablement satisfaisantes, de telle manière qu'un progrès global puisse être fait malgré le manque de précision du choix de règles. Cela semble être vérifié par le fait que les problèmes d'ingénierie sont parfois résolus de la manière la plus rustre par des procédures ad-hoc qui ne gèrent que les aspects les plus superficiels du problème, e.g. si une fonction croît ou décroît en l'une de ses variables. Un autre problème soulevé par cette image de la manière dont le comportement est déterminé est l'idée d'une "sortie favorable". Sans une telle idée, correspondant au "principe de plaisir" des psychologues, il est très difficile de voir comment procéder. Certainement qu'il serait plus naturel d'introduire quelques petites choses comme celles ci-après en machine. Je suggère qu'il pourrait y avoir deux clefs manipulables par le maître, et qui représente les idées de plaisir et déplaisir. A des stades ultérieurs de l'apprentissage, la machine reconnaîtrait certaines autres conditions comme étant désirables compte tenu du fait qu'elles auront par le passé constamment été associées au plaisir, et inversement, un certain nombre d'autres choses sont non désirables. Les expressions de colère de la part du maître pourraient, par exemple, être reconnues comme si déplaisantes qu'elles seraient négligées, de telle façon que le maître trouverait qu'il n'est plus nécessaire d'"employer la punition".

Faire de plus amples suggestions à propos de ces lignes serait probablement sans effets à ce niveau, dans la mesure où elles ne consisteront vraisemblablement en rien de plus qu'en une analyse des méthodes actuelles d'éducation appliquée aux enfants humains. Il y a, pourtant, une fonctionnalité dont j'aimerais suggérer qu'elle soit incorporée dans les machines, et c'est le "composant aléatoire". On devrait doter chaque machine d'une bande magnétique contenant une série aléatoire de figures, e.g., des 0 et des 1 à quantités égales, et cette série de figures devrait être utilisée dans les choix effectués par la machine. Cela aura pour conséquence que le comportement de la machine ne sera pas complètement déterminé par les expériences auxquelles elle a été sujette, et aura des utilisations précieuses quand on aura expérimenté ces nouvelles idées. En simulant les choix effectués, quelqu'un pourrait être capable de contrôler le développement de la machine dans une certaine mesure. On pourrait, par exemple, insister sur le fait que le choix effectué doit être un choix particulier, disons, en 10 endroits particuliers, et cela signifierait qu'une machine sur 1024 devrait se développer à un degré au moins aussi haut que celui qui aurait été simulé. Cela aura du mal à devenir une assertion précise à cause de la nature subjective de l'idée de "degré de développement" en ne disant même rien sur le fait que la machine qui avait été simulée aurait pu être aussi chanceuse dans ses choix non simulés.

Supposons maintenant, pour conforter l'argument, que ces machines soient une véritable possibilité, et regardons les conséquences de leur construction. Les construire effectivement devrait recontrer une grande opposition, à moins que nous n'ayions grandement avancé en terme de tolérance religieuse depuis le temps de Galilée. Il y aurait alors une

grande opposition des intellectuels, qui craindraient de perdre leur travail. Il est probable pourtant que les intellectuels se tromperaient à ce propos. Il y aurait beaucoup de choses à faire pour essayer, disons, de maintenir notre intelligence au niveau standard établi par les machines, car il semble probable qu'une fois que la méthode de pensée de la machine aura démarré, elle ne mettra pas longtemps à dépasser nos faibles possibilités. Il ne sera pas question de leur mort, et elles seront capables de converser les unes avec les autres pour aiguiser leurs esprits. Nous devons ainsi nous attendre à ce qu'à un moment, elles prennent le contrôle, comme cela est mentionné dans le livre de Samuel Butler *Erewhon*.