Alan Turing, le “ jeu de l’imitation ” et la première personne

Patrick Goutefangea

Alan Turing, le « jeu de l'imitation » et la première personne

Patrick Goutefangea

En 1950, dans l'article célèbre intitulé *Computing Machinery and Intelligence*[1], Turing imagine, en guise de réponse à la question « Les machines peuvent-elles penser ? », une expérience fictive : le « jeu de l'imitation ». La thèse développée ici consistera à montrer que l'expérience de Turing fait fond sur l'idée qu'entre une machine et des humains un *échange de paroles* est possible au même titre qu'entre un humain et ses semblables. L'échange de paroles exige l'expression à la première personne ; que la machine du jeu de l'imitation ait un tel échange avec des humains implique donc qu'elle puisse énoncer le signe de la première personne, par exemple le signe « je », et, surtout, que ce signe ait le même sens lorsqu'il est énoncé par elle que lorsqu'il est énoncé par un humain. Bref, Turing affirme, par le biais du jeu de l'imitation, qu'une machine peut accomplir un acte d'énonciation, et que cet acte peut avoir le même statut que celui effectué par des individus humains.

L'idée qu'il puisse y avoir une expression de la machine à la première personne apparaît notamment dans la dernière partie de *Computing Machinery and Intelligence*, consacrée, non plus au jeu de l'imitation en tant que tel, mais aux « machines qui apprennent ». Turing ne songe pas à n'importe quel type d'apprentissage : selon lui une machine peut être *éduquée* comme un petit d'homme. Or, on l'accordera, un apprentissage, lorsqu'il n'est pas un simple dressage, mais qu'il s'inscrit dans un processus d'éducation, présuppose la mise en œuvre, par l'entité éduquée, d'une première personne. Cette référence à la première personne est proprement ce qui lie l'un à l'autre, dans la réflexion de Turing, le développement sur les machines qui apprennent et celui sur le jeu de l'imitation. Elle est par là même ce qui assure la cohérence du texte.

On examinera tout d'abord comment la question de l'échange de paroles et du « je » de la machine surgit de la logique même du jeu de l'imitation ; on essaiera ensuite de déterminer quel est le statut du « je » de la machine dans le cadre particulier de la simulation mise en œuvre par le jeu ; on verra alors que c'est à travers « l'éducation » de la machine que ce statut rejoint celui du « je » humain.

Le jeu et la question du « je »

La machine que Turing entend faire participer au jeu de l'imitation est bien entendu une « machine de Turing », c'est-à-dire une machine logique, « à états discrets », définie par une table d'instructions. Turing avait montré, en 1936-37[2], qu'une telle machine pouvait calculer n'importe quel nombre calculable par un humain, et, bien plus, imiter le comportement de n'importe quelle autre machine discrète dont la

1 Alan Mathison Turing, « Computing Machinery and Intelligence », *Mind*, 59, octobrer 1950, p. 433-460. Publié in *Collected Works of A.M. Turing*, Londres, North-Holland, 1993, 3, *Mechanical Intelligence*. Publié en français sous le titre *Les ordinateurs et l'intelligence* in Jean-Yves Girard, *La machine de Turing*, trad. Patrice Blanchard, Paris, Seuil, 1995. Les citations feront référence à la traduction française.

2 Voir « On Computable Numbers with an Application to the Entscheidungsproblem », *Proceedings of the London Mathematical Society* , vol 42, 1937. Publié in *Collected Works of A. M. Turing, op. cit., 2, Mathematical Logic* . Publié en français sous le titre « Théorie des nombres calculables, suivie d'une application au problème de la décision », in Jean-Yves Girard, *La machine de Turing* , trad. par Julien Basch, *op. cit.*

configuration lui était fournie en entrée ; une « machine de Turing », parce qu'elle est programmable, peut être une « machine universelle ».

Quant au jeu de l'imitation, Turing en avait exposé le principe, à la fin des années quarante, dans un texte intitulé *Intelligent Machinery*[1] rédigé à l'intention du *National Physical Laboratory*, l'organisme de recherche auquel il était alors rattaché : supposons un assez médiocre joueur d'échec, affrontant deux adversaires de force à peu près égale à la sienne, et dont l'un serait une « machine de Turing » ; ce joueur aurait certainement, remarquait Turing, les plus grandes difficultés à décider lequel de ses adversaires est la machine[2]. Cet exemple servait de modèle au jeu décrit peu après dans *Computing Machinery and Intelligence* .

En tant que tel, le jeu de l'imitation oppose un homme ou une femme C à deux autres protagonistes : A - un homme – et B - une femme. C est isolé des deux autres. En posant des questions à A et B, il doit déterminer qui est l'homme, qui est la femme. L'homme A doit s'efforcer de tromper C, en se faisant passer pour la femme B, laquelle doit aider C. Les trois protagonistes communiquent par l'intermédiaire d'un téléscripteur et ne peuvent utiliser au cours du jeu de caractéristiques telles que l'apparence extérieure, la voix ou les performances physiques. Seul ce qui relève de l'échange linguistique est pris en compte. Il va de soi que les questions posées par C peuvent porter sur n'importe quel sujet.

Qu'arrivera-t-il, demande Turing, si A est remplacé par une machine, en l'occurrence par une « machine universelle » ? L'interrogateur, affirme-t-il, se trompera aussi souvent dans ce cas que lorsqu'il a affaire exclusivement à des humains. En d'autres termes, une machine, selon Turing, peut avoir autant de chances de faire bonne figure au jeu de l'imitation qu'un individu humain quelconque ; elle peut se montrer *l'égale* d'un être humain placé dans les conditions du jeu - réduit à la parole - c'est-à-dire l'égale d'un être dont les observateurs du jeu admettent qu'il pense.

Le jeu de l'imitation soulève naturellement de nombreuses questions. On en examinera deux qui déterminent directement l'intelligibilité même de la démarche de Turing. La première concerne la distinction qu'il convient de faire entre victoire de la machine au jeu - C s'est trompé et a déclaré que A était un homme - et réussite de la machine à ce qu'on a appelé le « test de Turing » - le comportement de la machine indique qu'elle « peut penser ». Une victoire de la machine n'est pas à elle seule, en effet, synonyme de réussite au test.

Quelles sont, dans le modèle strictement humain du jeu, les chances respectives de A d'un côté, de C et de B de l'autre ? Les trois protagonistes sont *grosso modo* de la même force puisqu'ils sont, par hypothèse, des individus humains quelconques. En conséquence, C, qui est aidé par B, a deux fois plus de chances de l'emporter que A. Inversement, celui-ci peut espérer s'imposer dans un tiers des cas. Tel est du moins le résultat qui doit prévaloir, à mesure que le nombre de parties augmente et si les joueurs sont choisis parmi un échantillon suffisamment large[3].

---

1 A. M. Turing, *Intelligent Machinery, Collected Works of A.M. Turing, Mechanical Intelligence, op. cit.*

2 op. cit., p.127. Turing s'était lui-même essayé à concevoir une machine « de papier « jouant, de manière élémentaire, aux échecs ; voir Andrew Hodges, *Alan Turing ou l'énigme de l'intelligence* , trad. par Nathalie Zimmermann, Paris, Editions Payot, 1988, p. 186 et sq.

3 « Je crois que dans une cinquantaine d'années il sera possible de programmer des ordinateurs [...] 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 ». *Les ordinateurs et l'intelligence* , op. cit. p. 148.

En fonction de ces considérations, on admettra qu'il y a réussite de la machine au test si le modèle où elle intervient reproduit les probabilités du modèle strictement humain. Il apparaîtra, alors, que le remplacement du protagoniste humain A par une machine ne change pas les résultats moyens du jeu, et donc que le comportement de la machine a les mêmes effets que celui de l'homme qu'elle a remplacé.

On peut certes discuter les présupposés du test : en quoi, par exemple, le fait que le comportement de A produise les mêmes effets lorsqu'il s'agit d'une machine que lorsqu'il s'agit d'un homme autorise-t-il à conclure que ces comportements sont identiques, ou encore que les entités agissantes sont identiques ? Cependant Turing se garde bien de conclure de manière aussi directe ; il se contente d'affirmer que la mesure des performances, sur le plan du discours, d'une machine et d'un être humain placés dans les mêmes conditions peut donner les mêmes résultats. Or, dans le cas de l'homme, le discours est rapporté à la pensée ; dès lors, dans le cas de la machine, s'il y a discours, ce discours ne peut-il être également rapporté à la pensée ?

La seconde question qu'il faut examiner est celle de savoir s'il s'agit encore, pour l'examinateur C, après substitution d'une machine à l'homme A, de dire qui, de A ou de B, est une femme[1]. Turing ne dit pas explicitement le contraire. Supposons donc que tel soit le but du jeu : la machine devra viser à se faire prendre pour une femme, et la meilleure stratégie pour elle consistera certainement à tenter d'imiter directement le comportement d'une femme plutôt que celui d'un homme cherchant à se faire passer pour une femme. Imaginons que la machine ne se montre pas assez « adroite » pour tromper C sur sa prétendue « féminité » : C conclura le jeu en énonçant quelque chose comme : « X (la machine) ne peut pas être une femme, donc X est un homme ». Suffit-il, toutefois, que la machine imite mal une femme pour que l'examinateur croit avoir affaire à un homme ? Il faut encore qu'imiter mal une femme ne l'empêche pas d'imiter de manière convaincante un *être humain* . En d'autres termes, ce n'est pas d'abord sur sa prétendue féminité, mais sur sa prétendue *humanité* que la machine doit tromper l'examinateur. Nous pouvons donc décrire le jeu sous la forme où il a été le plus souvent examiné : un joueur humain quelconque - homme ou femme - doit tenter de distinguer une machine - qui s'efforce de le tromper en se faisant prendre par lui pour un individu humain quelconque - d'un individu humain quelconque - qui l'aide en donnant des « réponses vraies ».

En outre, c'est l'humanité *telle que la manifeste la parole* que la machine doit simuler. Si, en effet, les observateurs du test (Turing, nous-mêmes...) concluent, à partir des erreurs de C, que la machine « peut penser », ce sera précisément parce que C lui-même aura commis ces erreurs en croyant reconnaître, dans la parole de A, « de la pensée ». A quoi reconnaîtrait-il en A un être humain, si ce n'était à ce qu'il interprète, chez celui-ci, comme la manifestation d'une pensée ? C, entendant une parole émise par A, considère que celui-ci pense, et, de là, conclut qu'il s'agit d'un être humain.

Dans le cadre de l'expérience fictive imaginée par Turing, l'important est donc que la machine fasse *jeu égal* avec un échantillon varié et le plus vaste possible d'individus humains. Il ne s'agit pas d'établir une

1 Ce point, souvent passé sous silence, a été souligné par plusieurs auteurs. Voir, par exemple, Peter Naur, « Thinking and Turing's Test », *Nordisk Tidskrift for Informations Behandling* , 26, 2, 1986 ; Jean Lassègue, « Le test de Turing et l'énigme de la différence des sexes », *Les contenants de pensée* , D. Anzieu, G. Haag éd., Paris, Dunod, 1993 ; W. Keith, « Artificial Intelligences, Feminist and Otherwise », *Social Epistemology* , vol. 8, 4, 1994 ; J. Genova, « Turing's Sexual Guessing Game », *Social Epistemology* , vol. 8, 4, 1994 ; Denis Vernant, « L'intelligence de la machine et sa capacité dialogique », *Penser l'esprit ; des sciences de la cognition à une philosophie cognitive* , V. Rialle et D. Fisette, éd., Grenoble, PUG, 1996.

quelconque supériorité du mécanique sur l'humain, ou l'inverse, mais de constater simplement qu'une entité mécanique et un individu humain peuvent se mesurer à un jeu comme celui de l'imitation, qu'ils peuvent y être, en droit, à égalité, les éventuels écarts entre eux ne relevant pas, dans le cadre du jeu, de leur différence de nature, mais d'une différence de « talent »...

Bref, la démarche de Turing dans *Computing Machinery and Intelligence* pourrait être formulée de la manière suivante : si une machine parle comme parle un homme, alors elle pense comme pense un homme ; or, une machine parle comme un homme si un homme lui parle comme il parle aux autres hommes, s'il a, autrement dit, avec elle, un *échange de paroles*. La fonction du jeu de l'imitation consiste donc à vérifier qu'un échange de paroles peut avoir lieu entre des hommes et une machine, d'une manière analogue aux échanges de paroles qui ont lieu entre les hommes.

Qu'une machine puisse « parler », Turing, d'une certaine façon, l'avait déjà démontré en élaborant dans les années trente la notion de « machine universelle » : une telle machine peut manipuler des symboles et énoncer des propositions, au sens logique du terme. Lui faire émettre ces propositions sous une certaine forme matérielle, « sonore » ou « écrite », est un problème de mécanique sans pertinence ici, et déjà résolu au moment où Turing publie *Computing Machinery and Intelligence*. Toutefois, les travaux de Turing des années trente démontraient seulement qu'une machine universelle, puisqu'elle est capable de mener un calcul et, par là, de manipuler des symboles, peut simuler la faculté humaine de discourir à la *troisième* personne, c'est-à-dire d'énoncer un discours purement descriptif, de l'ordre du « il y a », ou du « ça », où n'est pris en compte que ce que les linguistes nomment la « personne d'univers »[1] : l'expression de l'attachement de tout objet du discours à l'ordre des choses. Le prototype d'un tel discours est le discours scientifique, et son expression idéale la proposition mathématique, celle même qu'une « machine universelle », au moins pour ce qui concerne le « calculable », est en mesure d'énoncer. Dans un tel cadre, un échange de questions et de réponses entre locuteurs, comme celui qui doit avoir lieu pendant le jeu de l'imitation, est certainement possible[2] ; toutefois, on peut se demander entre *qui* et *qui* aurait lieu cet échange. C'est, en effet, le propre même du discours scientifique que « d'oublier », ou de mettre entre parenthèses, les actants du discours, en les identifiant à un unique locuteur abstrait, et on voit mal comment une machine « parlante » de ce type - qui ne connaîtrait que la troisième personne - pourrait faire bonne figure au jeu.

Pour que l'examinateur de celui-ci croie reconnaître l'humanité dans son adversaire mécanique, il doit de toute évidence avoir la conviction que le rapport établi entre lui et son interlocuteur est identique à celui qu'il entretiendrait avec un individu humain ; il faut, en d'autres termes, qu'il ne doute pas que ce qui a lieu entre lui et son adversaire est, non pas un échange abstrait de questions et de réponses d'où il serait lui-même absent en tant que sujet singulier, en tant que *personne*, mais un échange impliquant l'engagement existentiel des locuteurs. Il doit avoir la conviction d'être en présence d'un *semblable*. Bref, l'échange de paroles sur lequel repose le test exige que les protagonistes s'expriment à la *première* personne, que chacun d'eux fasse usage des déictiques *je* et *tu*. N'est-ce pas là, précisément, dans la démarche de Turing, tout le sens du passage du « test des échecs », ébauché dans *Intelligent Machinery*, à ce que nous pouvons appeler

1 L'expression est de G. Moignet, *Systématique de la langue française* , Paris, Klincksieck, 1981.
2 A la condition que les questions soient de « bonnes « questions, appelant des réponses purement descriptives.

un « test de la conversation », c'est-à-dire le jeu de l'imitation ? Pour que la machine l'emporte au jeu, l'examinateur C doit pouvoir dire « tu » à son interlocuteur A, ce qui implique qu'il croie à un « je » véritable de la part de celui-ci. Comment, selon Turing, une machine peut-elle « s'exprimer » à la première personne ?

Entre 1936 et 1939, Turing avait établi, non seulement qu'une « machine de Turing » est une « machine universelle », c'est-à-dire qu'elle peut simuler le fonctionnement propre de n'importe quelle autre machine « discrète », mais encore qu'il existe, pour toute machine de ce type, une formule critique qu'elle ne peut calculer. Il avait également montré que cette même formule est calculable par une autre machine, plus puissante parce que dotée, notamment, d'un axiome supplémentaire correspondant à la formule critique. Les résultats obtenus, alors, par Turing permettent, en somme, d'imaginer une série de machines universelles s'emboîtant les unes dans les autres : une machine A simulée par une machine B, capable de calculer la formule critique de A, cette machine B étant elle-même simulée par une machine C, capable de calculer la formule critique de B, et ainsi de suite. C'est ce principe que met en œuvre Turing pour concevoir une machine capable de faire bonne figure au jeu de l'imitation.

Son hypothèse implicite semble être, en effet, que rien n'interdit de diviser le jeu en une série infinie de « tests partiels », ou de « jeux partiels ». Quel que soit le type d'échange que l'on imagine entre la machine et ses interlocuteurs, quel que soit le moment de cet échange, il est, selon Turing, toujours possible de diviser le problème jusqu'à aboutir à des situations élémentaires dans lesquelles la machine peut faire jeu égal avec ses adversaires. Turing imagine, par exemple, un test portant sur l'expression artistique et ce qu'elle implique de conscience de soi, et fait appel pour cela à la situation type au cours de laquelle il est demandé à un élève, en guise d'exercice, de composer un sonnet, puis de le commenter[1]. Puisque le sonnet est une forme littéraire qui obéit à des règles précises, lesquelles commandent non seulement sa composition, mais son commentaire, rien, *a priori*, dans la définition de la « machine universelle », n'interdit de concevoir une « machine à sonnets ». De quels éléments disposerait l'examinateur du jeu de l'imitation, demande Turing, pour distinguer la machine de son propre partenaire humain, dans l'hypothèse où, à sa demande, cette machine composerait un sonnet - très probablement médiocre, mais cela n'importe pas... - et le commentait de manière simplement non absurde ? Ne serait-il pas aussi démuni que l'est le joueur d'échec d'*Intelligent Machinery* pour distinguer son adversaire mécanique de son adversaire humain ?

Or, quelle épreuve plus difficile que celle-ci pourrait avoir à affronter la machine ? Si elle peut reproduire les formes extérieures de la conscience et de la sensibilité artistique, ne sera-t-elle pas, *a fortiori* , capable de simuler tout autre comportement induit par le jeu ? Aussi bien pouvons-nous faire l'hypothèse qu'une session du jeu au cours de laquelle une machine l'emporterait serait constituée d'une série de jeux partiels tels que celui du sonnet - ou d'autres moins difficiles - et qu'il existe une machine universelle théorique capable de simuler chacune des machines l'emportant à un de ces jeux partiels, exécutant

1 « Le jeu, note-t-il, [NB le jeu de l'imitation] est fréquemment utilisé en pratique (en omettant le joueur B) 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...
Le *témoin* : Oui, mais personne n'a envie d'être comparé à un jour d'hiver.... « *Ibid.* p. 155.

autrement dit la série complète de jeux partiels dont se compose cette session du jeu.

C'est dans le cadre de cette hypothèse d'une machine théorique simulant un ensemble de machines simulant elles-mêmes des fragments de comportements humains que doit être située l'énonciation par la machine du jeu de l'imitation de signes tels que « je « et « tu », lesquels conduiront l'examinateur C, qui les interprétera comme des déictiques, à croire avoir affaire à un semblable, à un *sujet* comme lui-même, ou, plus exactement, comme il se voit lui-même.

Cependant, même si nous admettons que l'examinateur, face à une machine, et selon le mécanisme qui vient d'être décrit, peut être amené à croire qu'il a affaire à un être humain, sommes-nous assurés pour autant que la machine se sera « exprimée » à la première personne ? Certes, l'usage de la première personne aura été simulé, mais ne l'aura-t-il pas été par une *troisième* personne ? La situation de première personne n'aura-t-elle pas été simplement « calculée » par la personne (grammaticale) de la machine joueuse d'échecs, de la machine capable d'énoncer une proposition mathématique, bref, par un « il » ou un « ça » ? Suffit-il que l'examinateur dise, en toute bonne foi, « tu » à la machine pour que le « je » prononcé par celle-ci ne soit pas tout bonnement un « il » ?

La critique classique de la démarche de Turing s'est efforcée ainsi de montrer que le succès d'une machine au test n'établirait pas qu'elle « pense »[1] ; quand bien même une machine simulerait le discours humain à la première personne, quand bien même elle utiliserait le symbole *je*, ce *je* ne serait jamais pour elle un déictique. Quel est donc le véritable statut du « je » de la machine qui fait bonne figure au jeu de l'imitation ?

Le « je » de la simulation

Pour répondre à cette question il est nécessaire de rappeler le contexte dans lequel s'inscrit le « je » proféré par la machine et entendu par l'examinateur. Ce contexte est formé de ce que Turing regarde comme l'opinion commune à propos des rapports entre l'intelligence et la machine, à savoir l'idée qu'une machine ne peut pas penser car elle est parfaitement prévisible.

C'est, là encore, à *Intelligent Machinery* qu'il faut revenir. Turing y déclare : « Le point jusqu'où nous considérons que quelque chose se comporte de manière intelligente est déterminé autant par notre propre état d'esprit et formation que par les propriétés de l'objet considéré. Si nous sommes capables d'expliquer et de prévoir son comportement ou s'il semble y avoir le moindre plan sous-jacent, nous sommes peu tentés d'imaginer de l'intelligence »[2]. Pour l'opinion commune, telle que la comprend Turing, l'imprévisibilité serait une propriété de la pensée, comme la prévisibilité une propriété du mécanique.

Cette opinion commune a une expression philosophique, laquelle plonge ses racines notamment dans la métaphysique cartésienne et pose que la machine appartient à un autre ordre que l'homme. Que la machine ne puisse penser est, pour Descartes, une certitude métaphysique puisqu'il s'agit d'une conséquence de la distinction des substances : la machine relève de la substance étendue et non de la substance pensante. Descartes illustre cette idée, dans la cinquième partie du *Discours de la méthode*, à l'aide d'une

1 Voir en particulier l'argument dit « de la chambre chinoise « de John Searle ; « L'esprit est-il un programme d'ordinateur ? », *Pour la science* , 149, mars 1990 ; *Du cerveau au savoir* , Paris, Hermann, 1985 ; « Esprits, cerveaux et programmes », *Vues de l'esprit* , D. Hofstadter, D. Dennett éd., Paris, InterEditions, 1987.
2 *Intelligent Machinery* , op. cit., p. 127. C'est nous qui traduisons.

« expérience » fictive proche, dans sa structure, de celle proposée par Turing dans *Computing Machinery and Intelligence*. Imaginons un automate fabriqué par un artisan doué d'une habileté supérieure et imitant parfaitement l'apparence et le comportement d'un être humain ; nous aurions toujours, selon Descartes, deux moyens de ne pas confondre cet automate avec un homme véritable : la parole et l'action réfléchie [1]. Un tel automate pourrait, sans doute, émettre un discours, de la même façon qu'il serait en mesure d'effectuer certaines actions mieux qu'un être humain : on ne voit pas pourquoi la forme matérielle - donc, pour Descartes, mécanique - de la parole ne pourrait être reproduite par une machine ; pourtant, aussi parfait serait-il en son genre, cet automate ne pourrait « répondre au sens de tout ce qui se dit en sa présence comme le ferait l'homme le plus hébété »[2], de même qu'il ne saurait inventer une action pour lui encore inédite.

Chez Descartes, le discours humain, ainsi que l'action réfléchie, se distinguent irréductiblement de toute reproduction de leur seule forme matérielle par cela qu'ils expriment un *jugement*. Il y a jugement lorsqu'une volonté s'applique à ce qui est conçu[3] ; le jugement qu'énonce la parole ou que donne à voir l'action réfléchie manifeste la dualité de l'homme, union d'une âme et d'un corps, mode à la fois de la substance pensante et de la substance étendue. Par là, le discours ou l'action réfléchie, chez l'homme, ne sont pas seulement des actions mécaniques, comme celles que reproduirait un automate, mais *l'acte* d'une conscience inscrite dans un monde, expression de ce « je » qui, prononcé dans le « je pense », signifie toujours « je suis ».

Rien n'indique que Turing ait jamais lu de près le *Discours de la méthode* et encore moins qu'il l'ait eu à l'esprit en rédigeant *Computing Machinery and Intelligence*. Il n'en demeure pas moins que ce qu'il regarde comme l'opinion commune, à savoir l'idée que la machine et l'homme, c'est-à-dire la machine et la pensée, appartiennent à des ordres différents, est la cible qu'il vise. Qu'il ait lu ou non Descartes, Turing, dans *Computing Machinery and Intelligence*, refuse l'idée cartésienne d'une séparation radicale du mécanique et de l'humain, ce qui n'est pas sans déterminer fortement sa problématique : si l'examinateur C du jeu de l'imitation est amené à déclarer que son interlocuteur A - la machine - est un homme, ce sera pour avoir entendu, à travers la parole de cet interlocuteur, non pas simplement le son « je », non pas simplement, dans ce son, un signe indifférencié, mais bien le « je » du « je pense ». C peut-il se tromper à propos du « je » de son interlocuteur ? Un tel « je » aura-t-il été prononcé par A ou C aura-t-il simplement cru l'avoir entendu ?

La question peut être discutée en s'appuyant sur un raisonnement d'Hillary Putnam connu sous le nom d'argument des « cerveaux dans une cuve »[4]. Imaginons, propose Putnam, que le cerveau de chaque individu humain ait été placé dans une cuve, et que ses terminaisons nerveuses aient été reliées à un ordinateur qui reproduit les stimuli du monde extérieur. Tout individu humain sera privé, sans qu'il s'en rende compte, de contact avec le monde extérieur, et tout se passera de telle sorte que les cerveaux dans une

1 Descartes, *Discours de la méthode,oeuvres philosophiques* , I, Paris, Classiques Garnier, 1988, p. 628 et suivantes.
2 *Ibid.* p. 629.
3 Ce n'est que pour un jugement qu'il y a du vrai et du faux ; l'erreur découle d'un acte de volonté appliqué à ce qui est mal conçu, c'est-à-dire à ce qui n'est pas conçu clairement et distinctement. Il n'y a, en revanche, ni concevoir ni volonté pour l'automate, qui participe de la seule substance étendue. Un moulin à vent ne se trompe jamais, pas même lorsque son fonctionnement est défectueux.
4 Hillary Putnam, *Raison, vérité et histoire* , Paris, Les Editions de Minuit, 1984.

cuve se comportent, lorsqu'ils s'adresseront les uns aux autres, exactement comme s'ils étaient hébergés dans un corps. Pourtant, remarque Putnam, un cerveau ainsi traité ne pourra pas dire « je suis un cerveau dans une cuve » au sens où le dirait un cerveau « ordinaire » - qui ne serait pas dans une cuve mais dans un corps - car la référence de leur discours respectifs ne sera pas la même. Le cerveau dans une cuve fera référence « dans l'image » : lorsqu'il parlera d'un arbre, la référence de son discours ne sera pas l'arbre lui-même, mais l'image de l'arbre engendrée par les stimuli que produit l'ordinateur auquel il est relié. Ainsi, prononçant le mot « cuve », il fera référence, non pas à la cuve dans laquelle il se trouve, mais à l'image de celle-ci fournie par l'ordinateur. Bref, pour employer le mot « cuve » avec la même référence qu'un cerveau « ordinaire », il ne devrait pas être dans une cuve. Les cerveaux dans une cuve et ceux qui n'y sont pas ne partagent pas la référence. Rapportée au jeu de l'imitation, cette situation revient à montrer que la machine A ne peut pas partager la référence avec ses interlocuteurs humains B et C : en vérité, la machine et ses adversaires ne parlent jamais de la même chose ; ils ne communiquent pas.

Qu'en serait-il, cependant, du « je » prononcé par un cerveau dans une cuve disant « je suis ceci » ou « je suis cela », au cours d'une conversation avec des interlocuteurs humains, c'est-à-dire avec de « vrais cerveaux » ? Le statut de ce « je » ne serait-il pas le même que celui du « je » prononcé par ces interlocuteurs, à savoir le statut du « je » du « je suis » inhérent au *cogito* cartésien ? Jaako Hintikka a montré, à propos de ce dernier, que l'un des points d'appui de la démonstration de Descartes tenait à l'inconsistance existentielle de la proposition « je ne suis pas » : je ne puis la prononcer sans prouver par là même le contraire de ce qu'elle dit[1]. De quoi il découle que la proposition contraire « je suis » est consistante existentiellement. Le « je suis » que je ne puis prononcer sans être est un énoncé *performatif*. Supposons que le cerveau dans une cuve cherche à se faire passer auprès de « vrais » cerveaux - qui ne sont pas dans une cuve, mais dans un corps - pour un « vrai » cerveau. Il dira : « je suis un cerveau qui n'est pas dans une cuve », comme le dirait le cerveau qui n'est effectivement pas dans une cuve. Dans son cas, l'assertion sera fausse, mais le statut du « je suis » qu'il prononcera, revenant à dire : « je suis celui qui dit qu'il est ce qu'il n'est pas », sera un énoncé performatif comme dans le cas du « vrai » cerveau. Appliquée au jeu de l'imitation, cette constatation revient à considérer que tout se passe comme si la machine, énonçant « je suis », fût-ce en affirmant cette contre-vérité : « je suis un homme », partageait avec ses interlocuteurs humains la consistance existentielle. La machine victorieuse au jeu de l'imitation simule, en somme, la « performance » de A et de B disant « je suis », leur acte d'énonciation ; or, cette simulation de l'acte d'énonciation de A et B est sa « performance propre », son acte propre d'énonciation - qui implique la consistance existentielle.

La machine du jeu de l'imitation, toutefois, n'est pas un cerveau, même dans une cuve. Le raisonnement qui vient d'être suivi pourrait, en effet, s'appliquer à une machine programmée pour réussir un « test du sonnet » tel que celui imaginé par Turing, mais qui, à toute question posée par son examinateur, ne saurait fournir d'autres réponses que celles prévues pour ce test spécifique. Quand bien même elle l'emporterait au jeu, une telle machine ne pourrait être dite « pensante »[2]. Sa victoire s'expliquerait, non par

1 Jaakko Hintikka, « Cogito ergo sum : inférence ou performance ? », *Philosophical Review* , LXXI, 1962. Trad. de P. Le Quellec-Wolff.

2 Elle se montrerait incapable de « répondre au sens de tout ce qui se dit en sa présence comme le ferait l'homme le plus hébété ».

ses propres vertus, mais, sans doute, par le fait que le processus d'énonciation et de communication où est engagé l'examinateur détermine ce dernier à accorder au « je » de son interlocuteur le même statut qu'au sien propre. Dans le cadre du jeu, la situation exigerait de l'examinateur qu'il *prête* aux autres protagonistes la consistance existentielle, et c'est là ce qui rendrait possible la victoire de la machine. La consistance même du « je » de l'examinateur s'exprimerait à travers la postulation par celui-ci de la consistance du « je » de l'entité avec laquelle il communique. Sous peine de ne pouvoir s'exprimer lui-même, l'examinateur devrait, en somme, considérer la machine comme un *semblable.*

Dans une telle perspective, cependant, la machine ne l'emporterait-elle pas *par défaut* ? Son succès face à ses adversaires humains ne sanctionnerait-il pas la défaite de ceux-ci plutôt qu'il n'attesterait sa propre victoire ? Et, pour autant que le test fasse apparaître quelque chose, ne serait-ce pas à propos de l'énonciation telle que les humains la pratique, plutôt qu'à propos de la machine elle-même ? En somme, établir que l'examinateur humain de la machine est conduit à se comporter *comme si* cette dernière était un sujet ne suffit pas à démontrer qu'elle en est un, que les sons proférés par elle au cours du jeu ont le statut du discours à la première personne qu'émet un locuteur humain.

Tout a-t-il donc été dit ? Qu'en est-il, alors, de la dernière partie de *Computing Machinery and Intelligence*, consacrée à l'idée des « machines qui apprennent » ? En quoi cette seconde hypothèse - les machines peuvent apprendre - précise-t-elle la première - les machines peuvent faire jeu égal avec les hommes au jeu de l'imitation ? Comment s'y rattache-t-elle ? Quelle place tient-elle dans l'argumentation de Turing ? Ce dernier n'est guère explicite à ce sujet et, du reste, cet aspect de sa réflexion a été le plus souvent passé sous silence par les commentateurs. Or, le développement sur les machines qui apprennent ne répond-il pas précisément au problème que nous venons de rencontrer ?

Les machines qui apprennent

Il semble que, pour Turing, une machine capable de se tirer à son avantage du jeu de l'imitation soit en mesure d'apprendre. Mieux, la relation est symétrique. Tout se passe, en effet, comme si « éduquer » une machine - car c'est bien de cela qu'il s'agit - était la méthode préconisée par Turing pour réaliser la machine capable de faire bonne figure au jeu. A la question : comment construire une machine définie comme une série arbitrairement grande de machines réussissant des ‘ tests partiels’ ?, Turing répond : en concevant une « machine enfant » que l'on éduquera comme on éduque un petit d'homme. Si une machine peut apprendre, alors elle est en mesure de faire bonne figure au jeu. En vérité, l'intérêt du thème de l'apprentissage n'est-il pas, aux yeux de Turing, qu'apprendre implique un « je », une véritable première personne ?

Une fois de plus, c'est dans *Intelligent Machinery* que Turing expose les principes sur lesquels repose, selon lui, la possibilité d'éduquer une machine.

Plusieurs caractéristiques de la machine universelle sont ici sollicitées. En premier lieu, selon la définition donnée en 1937, une telle machine peut être non-déterministe : à certains moments de son mouvement plusieurs états sont possibles[1]. En second lieu, une machine de Turing universelle est modifiable puisqu'elle peut imiter diverses machines. Il est ainsi possible de concevoir une machine de telle sorte

1 Turing parle d'une « machine à choix », voir *Théorie des nombres calculables*, in *La machine de Turing* , op. cit., p.52.

qu'elle reçoive des informations de l'extérieur et qu'elle soit modifiée par ces « interférences ». Bien plus, une machine de Turing peut se modifier elle-même : elle stocke en mémoire des tables d'instructions et l'une de ces tables peut comporter des instructions spécifiques visant à modifier les autres tables[1].

Dans *Intelligent Machinery*, Turing décrivait une machine universelle non-déterministe et auto-modifiable en laquelle il voyait le « modèle le plus simple d'un système nerveux ayant un arrangement aléatoire de neurones »[2] : cette machine était constituée d'un nombre *n* d'unités semblables, comportant chacune deux entrées et une sortie, laquelle pouvait être connectée à une entrée d'une ou plusieurs autres unités, en fonction d'un nombre tiré au hasard dans l'ensemble des entiers compris entre 1 et *n*. Une fois mise en marche, une telle machine devait former aléatoirement des combinaisons de « neurones » ; son mouvement était périodique puisque ses états étaient en nombre fini, et, en l'absence « d'interférences », elle devait entrer rapidement dans un cycle répétitif. Or, affirmait Turing, en la soumettant à un système d'interférences conçu selon un modèle simulant des « punitions » et des « récompenses », il serait possible de modifier cette machine de manière à ce qu'elle s'auto-organise. L'interférence subie par elle pourrait constituer un analogue des stimuli de douleur ou de plaisir, le signal « douleur » forçant un changement aléatoire de configuration, et le signal « plaisir » entraînant au contraire le renforcement de la configuration associée. Les changements de la machine subiraient ainsi un processus de sélection. Si le système de punitions et de récompenses était établi en fonction de fins déterminées, les configurations non satisfaisantes du point de vue de ces fins devraient tendre statistiquement à disparaître et la machine se rapprocher du modèle visé.

Le cerveau d'un enfant, demande Turing dans *Intelligent Machinery*, ne peut-il être regardé comme une machine qui, par l'éducation, s'auto-organise ? Cette idée est celle qu'il développera, quelques mois plus tard, dans *Computing Machinery and Intelligence*. Plutôt que d'essayer de construire une machine imitant, à sa sortie de l'atelier, un cerveau humain adulte, Turing imagine une machine imitant un cerveau humain réduit à sa plus simple expression et *devenant* adulte par un processus d'éducation. L'hypothèse est ici que, si l'on dotait une « machine universelle », non-déterministe, auto-modifiable, et constituant un système logique élémentaire, d'un dispositif lui permettant de recevoir des signaux de l'extérieur, elle pourrait, à l'aide d'un système de punitions-récompenses tel que celui décrit dans *Intelligent Machinery* , « apprendre », et être « éduquée » comme l'est un petit d'homme. Il ne serait pas nécessaire pour cela que le système logique initial satisfasse les logiciens les plus exigeants[3] ; il suffirait que la machine soit capable d'inférence logique. Elle n'aurait pas non plus à être dotée d'un véritable « corps », au sens biologique du terme, ou d'un équivalent mécanique de celui-ci, pour autant qu'un moyen de communication réciproque entre elle et ses « maîtres », aussi rustique fût-il, puisse être mis en place[4]. Turing ajoute que le processus qu'il imagine est comparable à l'évolution biologique, la structure de la machine correspondant au « matériel héréditaire », les

1 « Nous pouvons, si nous voulons, diviser les opérations de la machine en deux classes, les opérations normales et celles d'auto-modification ». *Intelligent Machinery* , op. cit., p. 116. C'est nous qui traduisons.
2 *Ibid.* , p. 115.
3 Turing précise en particulier, qu'il n'est pas nécessaire que le système logique initial soit prémuni contre les paradoxes logiques par une hiérarchie des types.
4 Turing renvoie au cas d'Helen Keller : « l'exemple de Mlle 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 « *Les ordinateurs et l'intelligence* , op. cit. p. 170. Helen Keller était cette jeune américaine, devenue très jeune sourde, aveugle et muette et qui, confiée à Anne Mansfield Sullivan, apprit, à l'aide du seul sens du toucher, le langage des sourd-muets, puis l'écriture et enfin la parole, avant de faire des études supérieures et de consacrer un livre à son expérience.

changements subis par elle aux mutations et le jugement de l'expérimentateur, c'est-à-dire les choix faits par le « maître », à la sélection naturelle[1].

On remarquera que, si les tenants de l'intelligence artificielle (IA) classique ont pu voir dans l'argumentation de Turing accompagnant son hypothèse première - la possibilité pour une machine de faire bonne figure au jeu de l'imitation - un énoncé des principes sur lesquels ils fondèrent leur discipline au milieu des années cinquante, ce sont bien plutôt les recherches menées après l'abandon des espoirs mis dans l'IA classique, qu'annonce la dernière partie de *Computing Machinery and Intelligence*. Il est permis, en effet, de voir dans la « machine inorganisée », comme dans le recours au modèle épistémologique de l'évolution, une première formulation des principes sur lesquels reposent les « machines connexionnistes » ou les « algorithmes génétiques ». Turing insiste, en particulier, sur l'un des traits qui distingueront la démarche « connexioniste » ou « génétique » de celle de l'IA classique, à savoir le fait que la conception d'une machine « intelligente » ne passe pas par une formalisation *a priori* des réponses attendues de la machine, mais par la réalisation d'un système dynamique atteignant de lui-même un point d'équilibre, sans que la configuration correspondante ait à être spécifiée de manière formelle, sans qu'il soit besoin « d'avoir une représentation mentale claire de la machine à tout moment du calcul », comme c'est le cas lorsqu'il s'agit d'une machine à calculer classique[2].

Or, la prise en compte de cette dimension de la démarche de Turing n'est-elle pas susceptible de modifier l'interprétation examinée plus haut, selon laquelle la machine victorieuse au jeu resterait un dispositif du type « troisième personne » ? Turing ne s'efforce-t-il pas de montrer dans *Computing Machinery and Intelligence* que le processus qu'il décrivait un peu plus tôt comme le « modèle le plus simple d'un système nerveux » peut être la matrice de la situation de communication qui permet l'apprentissage ? De sorte que le ressort du jeu de l'imitation ne serait plus la situation de communication formalisée, par là même réduite et particulière, qui accompagne, par exemple, le jeu d'échecs, mais bien la relation humaine de communication en tant que telle. En vérité, Turing n'entend pas construire une machine capable de jouer aux échecs, *puis* la faire jouer contre un adversaire humain ; il n'entend pas davantage concevoir une machine « parlante », *puis* soumettre cette machine, par le biais du langage qu'elle vient d'acquérir, à un apprentissage ; il entend concevoir une machine qui *apprenne à parler*, c'est-à-dire une machine d'emblée plongée, comme l'enfant, dans la communication, et qui donne à cette communication, par son « évolution », une forme verbale.

Un processus d'éducation est essentiellement guidé par les fins et non par la connaissance complète de la structure de l'éduqué ; dès lors, tout ne se passe-t-il pas, dans l'hypothèse de Turing, comme si la machine qui apprend résolvait *par elle-même* des problèmes, puisque la manière dont elle réagit n'a pas à être prévue, et ne saurait l'être ? On peut certes relever combien Turing sous-estime les difficultés immenses que présente la réalisation d'une « machine parlante », ou combien trivial est le modèle d'apprentissage qu'il propose, fondé sur le seul système punitions-récompenses[3]. Cependant, entend-il faire plus qu'établir la

1 *Ibid.* , p. 169.

2 *Les ordinateurs et l'intelligence* , op. cit., p. 173. « 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... », ajoute Turing.

3 D. Andler évoque ainsi « la désinvolture de Turing » à propos de l'apprentissage du langage par la machine, désinvolture qui « semble aujourd'hui incroyable : rien n'indique que [Turing] entrevoie la difficulté pour la machine (ou son programmateur) de passer d'une ‘pensée’ ou d'une ‘intention’ communicative ou informative à une expression

compatibilité de son hypothèse avec la définition théorique de la machine universelle ? Peu lui importe au fond que l'on parvienne ou non à réaliser un jour les « expériences » qu'il imagine, s'il peut montrer que la définition admise de la machine universelle autorise la formulation de l'hypothèse selon laquelle une telle machine peut être « éduquée »[1].

Or, on voit bien que, si l'on faisait participer une telle « machine qui apprend » au jeu de l'imitation, celui-ci ne pourrait pas être considéré comme une simple extension de la partie d'échecs à trois joueurs décrite dans *Intelligent Machinery*. L'apprentissage de la parole, c'est-à-dire la transformation du processus de communication en échange verbal, n'implique-t-il pas, à travers l'émergence de la personne grammaticale, la maîtrise de la déixis, et ceci quelle que soit la nature de l'apprenant et la manière dont l'apprentissage s'effectue ? Une machine qui apprendrait à parler ne pourrait manquer de tenir bientôt un discours à la première personne, et ce serait précisément alors qu'elle pourrait, selon Turing, faire bonne figure au jeu de l'imitation. Dès lors, si une telle machine l'emportait au jeu environ une fois sur trois, comme l'homme qui, dans la situation initiale, cherche à se faire passer pour une femme, pourrait-on exclure qu'entre elle et ses interlocuteurs du jeu, ait eu lieu, non pas un simple échange de signaux, mais un échange de paroles ?

Résumons. Le test de Turing ne peut être valide qu'à la condition qu'il soit nécessaire de *penser*, au sens où les hommes pensent, pour le réussir. Que peut signifier, précisément, pour la machine, réussir le test de Turing ? Non pas tant gagner une partie de jeu de l'imitation contre des humains, que, plus simplement, *pouvoir disputer* une partie de jeu de l'imitation comme le fait l'homme de la version purement humaine du jeu. Le jeu de l'imitation implique, en effet, entre les joueurs, un certain type de communication, analogue à celui qui préside, chez les humains, au processus d'éducation : dans le contexte du jeu, la communication prend la forme d'un échange de paroles. La machine sera donc à même de participer à une partie de jeu de l'imitation si elle peut avoir avec des humains un échange de paroles. C'est alors qu'elle sera dite pensante.

Si l'hypothèse de Turing est valide, si, autrement dit, une machine peut s'exprimer à la première personne, il devient nécessaire, pour rendre compte du mécanique, de renvoyer à l'écart existant entre troisième et première personne grammaticale. Or, dans cet écart se glissent des notions à forte connotation philosophique, voire métaphysique, telles que celles de « sujet », « d'acte », ou de « personne morale ». D'où les conséquences philosophiques de l'hypothèse. Ces conséquences sont quelquefois ramenées à l'infirmation - qui serait en l'occurrence tardive - du dualisme cartésien sous ses diverses formes, mais elle vont sans doute plus loin : dans la perspective ouverte par Turing, si, entre la machine et l'homme, toute solution de continuité devait s'avérer fictive, ce ne serait pas parce que l'humain aurait été ramené au mécanique, mais bien plutôt parce que le mécanique devrait être élevé à la dignité de l'humain, ce qui conduirait, non à devoir renoncer à cette dignité, mais à devoir repenser ce qu'en termes kantiens on appellerait ses conditions de possibilité.

---

linguistique correcte et pragmatiquement adéquate » (« Turing : pensée du calcul, calcul de la pensée », *Le formalisme en question* , F. Nef, D. Vernant éd., Paris, Vrin, 1998).

1 Du reste, Turing ne prétend pas faire de la machine un modèle explicatif du comportement humain ; une « machine-enfant », précise-t-il, ne sera jamais l'exact équivalent d'un enfant humain, et si elle apprend effectivement à parler, on n'en saura pas davantage sur les mécanismes profonds qui auront permis cet apprentissage qu'on n'en sait dans le cas de l'enfant humain (Voir *Les ordinateurs et l'intelligence* , op. cit., p. 170).