Alan Turing :la "pensée" de la machine et l'idée de pratique

Patrick Goutefangea

UNIVERSITÉ DE NANTES

DÉPARTEMENT DE PHILOSOPHIE

ALAN TURING : LA « PENSÉE » DE LA MACHINE ET L'IDÉE DE PRATIQUE

Thèse de doctorat

Discipline : philosophie

Présentée par

Patrick Goutefangea

Directeur : M. Jean-Michel Vienne, professeur à l'université de Nantes

Jury

M. Jean Mosconi, Président

M. Pierre Livet

M. Steve Torrance

13 mars 1999

Sommaire

Introduction

" Je propose de considérer la question : 'les machines peuvent-elles penser ?' ", annonce Alan Turing, au début de *Computing Machinery and Intelligence*, l'article quasi testamentaire qu'il publie dans *Mind* en 1950[1]. Il semble que tout se soit conjugué pour que ce texte célèbre hante les débats suscités par ce que l'on appelle commodément la " révolution technologique " de la deuxième moitié du 20e siècle : la question même qu'y aborde Turing – celle de la " pensée " des machines – au moment où apparaissent les premiers ordinateurs ; le fait qu'il y résume, peu de temps avant de mettre fin à ses jours, quinze années de réflexion sur la notion de machine ; la réponse provocante qu'il avance - les machines peuvent " penser ", et seules des raisons d'ordre psychologique empêchent de l'admettre ; enfin, la forme singulière de cette réponse. Dans *Computing Machinery*…, en effet, Turing propose de soumettre la machine à un test – le fameux jeu de l'imitation - et formule l'hypothèse que la " pensée " ne pourrait être refusée à la machine qui réussirait ce test, c'est-à-dire qui l'emporterait au jeu de l'imitation, face à des adversaires humains eux-mêmes considérés comme pensants.

Le lieu même où s'exprime Turing – la principale revue britannique de philosophie - montre assez que la question " Les machines peuvent-elles penser ? " n'est pas, à ses yeux, une question d'ingénieur, et que ce n'est pas à des ingénieurs qu'il l'adresse. Le problème de la " pensée " des machines porte sur ce qui, dans la machine, échappe à la technique proprement dite : sa relation à l'homme qui la construit, et elle concerne, à travers cela, le rapport que chaque homme entretient avec sa condition d'homme. Bref, il s'agit d'une question philosophique, et c'est à des philosophes que Turing destine son texte.

---

[1] " *I propose to consider the question, 'can machines think ?'* ", 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 in Jean-Yves Girard, *La machine de Turing*, trad. Patrice Blanchard, Paris, Seuil, 1995.

Qui sont ces philosophes ? Turing publie *Computing Machinery…* dans *Mind* au moment où se déroule, entre " cambridgiens " et " oxfordiens ", le débat à l'issue duquel sera renouvelée l'inspiration de la philosophie analytique. Dès lors, vouloir rendre compte de la portée philosophique de la démarche de Turing, n'est-ce pas, avant tout, tenter de fixer la place qu'elle occupe dans ce débat ? Sous cet angle même, et compte tenu du rôle joué par Wittgenstein, tout à la fois dans l'émergence du mouvement analytique et dans son évolution générale, ne s'agit-il pas d'élucider, tout particulièrement, le rapport de Turing à Wittgenstein ? De fait, nous verrons que l'échange intellectuel qu'eut Turing avec Wittgenstein, lors du *Cours sur les fondements des mathématiques*[2] donné par celui-ci en 1939, influença de manière déterminante sa propre réflexion sur la " pensée " des machines. Enfin, au regard des nombreuses références faites au jeu de l'imitation dans la littérature philosophique consacrée, à partir des années soixante, à la notion d'intelligence artificielle et suscitée par l'émergence des sciences cognitives, puis, un peu plus tard, par le développement de la " philosophie de l'esprit ", ne convient-il pas également de tenter de dégager le rôle joué par le texte de Turing à l'égard de ces réflexions plus récentes ? Quels sont les rapports de *Computing Machinery…* à la philosophie analytique formaliste de la première moitié du siècle, puis à la philosophie du " langage ordinaire " qui se développe à Oxford au moment même de sa publication ? Quels sont ses rapports au premier Wittgenstein, celui du *Tractatus*, puis au second, qui critique le *Tractatus* ? Quelle influence a eu la réflexion de Turing sur l'élaboration de la notion d'intelligence artificielle et le développement des sciences cognitives ? Comment caractériser l'usage fait de cette réflexion dans le cadre de la " philosophie de l'esprit " ? Telles pourraient être quelques-unes des questions auxquelles une analyse philosophique de la démarche de Turing aurait à répondre.

Cependant, est-ce bien de cette démarche que l'on traitera alors ? Le poids du contexte, l'importance reconnue, pour le philosophe, des questions qui y sont débattues, ne risquent-ils pas de confiner la réflexion de Turing " à la marge ", et de lui faire jouer, au mieux, un rôle utilitaire dans une problématique dont elle ne définit pas elle-même les principes ?

Le statut même de *Computing Machinery…*, du point de vue de la philosophie, reste, en effet, à éclaircir. A lui seul, l'intérêt philosophique général de la question de la " pensée "

[2] Ludwig Wittgenstein, *Cours sur les fondements des mathématiques*, trad. de Elisabeth Rigal, Paris, Editions TER, 1995.

des machines ne saurait rendre la réflexion de Turing justiciable d'une discussion philosophique, au sens " technique " du terme. Encore faudrait-il que cette réflexion propose une argumentation explicitement située sur le terrain de la philosophie entendue comme discipline, ou pouvant y être transposée sans artifice. Or, d'une part, Turing n'était pas philosophe - il s'était fait connaître par la publication en 1937 d'un article de logique mathématique consacré à la notion de " calculable "[3], dans lequel il élaborait la notion de " machine de Turing " ou de " machine universelle " - d'autre part, il n'entendait pas, semble-t-il, présenter, dans *Computing Machinery…*, une argumentation philosophique. Son objectif se limitait à la formulation d'une hypothèse, dans des conditions clairement définies : il s'agissait principalement, pour lui, de montrer que le test qu'il imaginait, et le résultat qu'il prévoyait – le succès de la machine au jeu de l'imitation – ne sont pas dénués de sens du point de vue de la définition logico-mathématique de la machine qu'il avait lui-même établie.

Du reste, la relative indétermination de la démarche de Turing vis à vis de la philosophie des philosophes semble attestée par son destin même. Ce n'est qu'avec l'émergence, à partir de 1956[4], de la notion d'intelligence artificielle, et l'ouverture des débats philosophiques suscités par celle-ci, que *Computing Machinery…* a rencontré ses premiers échos philosophiques. Sans doute ces échos n'ont-ils guère cessé, depuis, de se faire entendre, notamment dans la littérature philosophique anglo-saxonne ; c'est, cependant, le plus souvent, à titre d'illustration originale et stimulante, plutôt que comme véritable contribution théorique au débat philosophique, que la réflexion de Turing est évoquée par les philosophes. Allons plus loin : ne peut-on soutenir que, dans le cadre de l'interprétation la plus couramment admise, ce qui constitue la dimension proprement *philosophique* de la démarche de Turing est passé sous silence ?

Turing, en effet, est généralement regardé tout à la fois comme l'un des fondateurs de l'intelligence artificielle et comme l'un des théoriciens ayant permis l'émergence des sciences cognitives. Le rôle historique joué, sur ce plan, par *Computing Machinery…*, ne saurait être

[3] Alan Mathison Turing, " 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 in Jean-Yves Girard, *La machine de Turing*, trad. Julien Basch, *op. cit.*

[4] C'est lors d'une réunion tenue à Dartmouth en 1956, au cours de laquelle Newell, Shaw et Simon présentèrent un programme capable de démontrer des théorèmes simples de la logique des propositions, que le terme d'" intelligence artificielle " fut adopté pour nommer ce qui, aux yeux des participants, devait devenir une nouvelle discipline.

mis en doute. Dans la perspective ainsi tracée, l'affirmation de Turing selon laquelle les machines peuvent " penser " apparaît avant tout comme une thèse psychologique – la machine peut être considérée comme un modèle explicatif du comportement mental de l'être humain – et la discussion philosophique à laquelle elle donne lieu se résume à la question de savoir si cette thèse doit être appréhendée seulement comme un modèle épistémologique ou comme une explication à part entière des phénomènes de pensée chez l'homme.

Or, il nous semble qu'une telle lecture ne permet pas de prendre en compte, fût-ce pour la critiquer, ce que nous croyons pouvoir appeler la portée *spéculative* de la structure particulière du jeu de l'imitation, dans lequel la machine, pour l'emporter, doit *surprendre* un individu humain quelconque, *en se faisant passer elle-même, aux yeux de cet individu humain, pour un individu humain quelconque*. De là vient, peut-être, que la distance, quelquefois soulignée[5], entre l'interprétation classique énoncée ci-dessus et le texte même de Turing, soit en général attribuée à la situation historique de ce dernier, c'est-à-dire à son statut de pionnier des sciences cognitives : si l'on peut émettre des doutes sur le fait que Turing, dans *Computing Machinery*…, énoncerait comme telle l'idée que la machine constitue un modèle explicatif du comportement mental de l'être humain, ce serait principalement parce qu'il ne faisait qu'annoncer les sciences cognitives, et ne pouvait donc nourrir sa réflexion des résultats obtenus par celles-ci.

En vérité, dans *Computing Machinery*…, Turing choisit délibérément pour adversaire *l'opinion commune* : c'est contre celle-ci qu'il entend établir que les machines " peuvent penser ". On remarquera que cela même pourrait le faire reconnaître par les philosophes, sinon comme l'un d'entre eux, du moins comme un " aspirant philosophe ", déterminé à situer sa réflexion sur le terrain de la philosophie, laquelle se veut toujours, lors même qu'elle n'entend pas nécessairement la ruiner dans son contenu, distance critique à l'égard de l'opinion commune, lieu, par définition, non d'un savoir, mais d'une *doxa.*

Il est vrai que, compte tenu de la méthode suivie par Turing pour répondre à la question " Les machines peuvent-elles penser ? " - plutôt qu'une argumentation philosophique, une *expérience* purement imaginaire - tout se passe comme si, dans *Computing Machinery*…, le point de vue à partir duquel est discutée l'opinion commune à

[5] Voir, par exemple : Daniel Andler, " Turing, pensée du calcul, calcul de la pensée ", *Le formalisme en question*, F. Nef, D. Vernant, éd., Paris, Vrin, 1998.

propos de la " pensée " des machines, n'était pas davantage celui de la science que celui de la philosophie. Répondant à la question " Les machines peuvent-elles penser ? ", Turing formule une hypothèse, non pas tant scientifique que philosophique – les machines " peuvent penser " - qu'il s'efforce, pourtant, d'examiner, non à l'aide d'une argumentation philosophique qui aurait pris soin d'éclairer ses propres présupposés, mais par le biais d'une " expérience ", dont le statut apparaît, au mieux, comme indéterminé, au pire, comme relevant lui-même de ce qu'il s'agit de combattre à travers l'opinion commune : Turing, dans *Computing Machinery*…, parait vouloir opposer une *doxa* à une autre *doxa.*

C'est pourtant de ce curieux parti méthodologique que naît, selon nous, la portée philosophique de sa réflexion.

Telle qu'elle est évoquée par Turing, l'opinion commune visée dans *Computing Machinery*… repose sur l'idée que le fonctionnement d'une machine, déterministe dans son principe, peut être connu et théoriquement prévu, contrairement au comportement humain, présumé libre, et, par là, tenu pour essentiellement imprévisible. En conséquence, selon l'opinion commune, une machine ne saurait " penser ", au sens humain du terme " penser ", car elle ne peut *surprendre* un individu humain aussi radicalement qu'un individu humain pensant peut lui-même le faire. Le jeu de l'imitation a, quant à lui, précisément pour fonction de révéler la possibilité, pour une machine, de surprendre un individu humain quelconque. Or, c'est en se faisant passer elle-même, aux yeux de cet individu, *pour un autre individu humain*, que la machine victorieuse au jeu de l'imitation surprend ses adversaires humains. Par définition, toute opinion commune implique elle-même des individus humains quelconques, situés, autrement dit, sur le même plan que ceux qui participent au test de Turing. De sorte que, de " l'expérience " décrite dans *Computing Machinery*… – la victoire d'une machine au jeu de l'imitation - ressort que l'opinion commune peut être contredite à partir de ses propres prémisses. Le test de Turing, dès lors qu'il est réussi par la machine, *infirme* l'opinion commune, et, du fait que cette dernière revient à affirmer qu'un homme, dont il est admis qu'il pense, se distingue d'une machine en cela qu'il peut surprendre d'autres hommes, son infirmation - la réponse inattendue, au regard de l'opinion commune, à laquelle est conduit

l'adversaire humain de la machine - vaut, *par elle-même*, confirmation de la " pensée " de la machine[6].

Dans cette perspective, l'expérience de Turing met en évidence, à partir de l'opinion commune elle-même, le principe d'une prise de distance à l'égard de celle-ci, bref, d'une position dont le statut ne peut pas être celui de l'opinion commune. Cette position ne saurait être, on l'a vu, celle d'un savoir scientifique, puisque l'expérience envisagée est, en tant que telle, fictive. Son statut est d'abord celui d'une *critique interne* de l'opinion commune, et c'est en ce sens qu'on le qualifiera de philosophique ; par là, en effet, l'hypothèse de Turing ne renvoie pas seulement à la pensée comme phénomène mental, mais à la question, par excellence spéculative, posée par l'idée d'une transparence de la pensée à elle-même, c'est-à-dire à l'idée de pensée comme principe intelligible.

C'est à cette dimension de la démarche de Turing que ce travail est consacré. A cet égard, l'hypothèse proposée sera que les idées de Turing prennent leur véritable sens de ce qu'elles impliquent une réflexion sur la notion philosophique de *pratique* ; qu'elles s'inscrivent, par là, profondément, en la prolongeant, dans l'histoire de la philosophie occidentale ; que les écrits, enfin, dans lesquels Turing les exprime[7] doivent être traités, non comme des exemples sollicités dans le cadre de réflexions philosophiques élaborées en dehors d'eux, mais comme des textes philosophiques à part entière. C'est enfin pourquoi il nous semble qu'examiner les idées énoncées dans *Computing Machinery…* directement à la lumière de leur contexte philosophique immédiat, non seulement ne permettrait pas de rendre pleinement justice à leur intérêt philosophique propre, mais en outre, que, faute d'avoir

---

[6] On voit que la position défendue ici repose sur l'interprétation du jeu de l'imitation selon laquelle l'examinateur humain, dès lors qu'une machine participe au jeu à la place d'un homme, doit nécessairement se prononcer sur la distinction entre *une machine et un individu humain*, et non pas seulement, comme le veut la règle explicite du jeu, sur la distinction entre un homme et une femme. Les raisons précises pour lesquelles cette interprétation nous semble devoir être retenue seront exposées au chapitre consacré au " jeu de l'imitation (voir plus loin, partie II, chapitre I, section I).

[7] Turing rédige, entre 1945 et 1948, plusieurs rapports destinés au *National Physical Laboratory*, le principal organisme de recherche britannique, qui lui avait demandé de proposer des plans et un programme de travail pour la construction d'un calculateur électronique. Ces rapports sont autant de textes préparatoires au grand article de *Mind*, *Computing Machinery and Intelligence*. Voir *Collected Works of A.M. Turing*, *op. cit.* 3, *Mechanical Intelligence*. On peut ajouter à cet ensemble de textes le script des émissions de radio auxquelles Turing a participé en 1952 sur ce même sujet de la " pensée " des machines. Voir le BBC Written Archives Centre, émissions du 15 Mai 1951 – *Can digital computer's think ?* – et du 10 Janvier 1952 - *Can calculating machines be said to think ?*

préalablement dégagé leur portée philosophique, leur relation à ce contexte ne pourrait être déterminée de manière pertinente. Aussi bien les questions formulées plus haut ne seront-elles pas, ici, traitées en tant que telles, quand bien même certaines d'entre elles seront évoquées - ainsi, par exemple, de l'influence de la discussion de Turing avec Wittgenstein sur la formulation de la question " Les machines peuvent-elles penser ? ", ou encore, du rapport de Turing à l'intelligence artificielle.

Du point de vue où nous nous plaçons, la démarche de Turing dans *Computing Machinery…* met en jeu deux thèmes essentiels. D'une part, l'idée de pensée à laquelle fait référence, ici, l'hypothèse qu'une machine " peut penser ", est bien celle-là même qui sous-tend l'opinion commune, et selon laquelle la pensée, au sens humain du terme, parce qu'elle implique une imprévisibilité fondamentale, ne peut être réduite au seul déterminisme logique, dont la machine est, au contraire, au regard de la même opinion commune, une figure privilégiée ; dans la question " Les machines peuvent-elles penser ? ", telle que Turing l'envisage, le terme " penser " renvoie à l'idée d'une dimension extra-logique, ou ultra-logique de la pensée, et *l'hypothèse affirmant que les machines " peuvent penser " signifie que les machines peuvent partager cette dimension*.

D'autre part, l'affirmation inverse selon laquelle les machines ne peuvent penser, revêt, au-delà de son expression commune, une forme directement et pleinement philosophique : l'idée que les machines ne sauraient penser, au sens humain du terme, et que la pensée, en l'homme, se distingue irréductiblement du " mécanique ", appartient de plein droit à l'histoire de la philosophie occidentale. De sorte que l'infirmation de l'opinion commune qui donne son sens au test de Turing se révèle également porteuse d'une critique, sur le terrain spécifique de la philosophie, de l'idée selon laquelle la pensée relève d'un autre ordre que le mécanique. L'analyse de ces deux thèmes détermine la lecture de *Computing Machinery…* que nous voulons proposer.

Le sens du terme " penser " dans la question " Les machines peuvent-elles penser ? " et dans l'hypothèse selon laquelle une machine qui l'emporte au jeu de l'imitation peut penser.

L'un des aspects du premier thème envisagé ici concerne les rapports de Turing à l'intelligence artificielle, et le malentendu entretenu, selon nous, à cet égard.

Il est, en effet, généralement admis comme allant de soi que *Computing Machinery…* doit être compris, sinon comme l'acte de naissance de l'intelligence artificielle (IA) - qui voit officiellement le jour en 1956, quelques années après la disparition de Turing – du moins comme son annonce. Aussi est-ce souvent à la lumière des réalisations de l'intelligence artificielle, ainsi qu'à travers la discussion des principes de celle-ci, que l'hypothèse de Turing a été jugée, comme si elle trouvait en ce lieu sa véritable expression. Or, dans un tel contexte, il est clair que la démarche de Turing sera justiciable des critiques qui ont en grande partie ruiné l'intelligence artificielle dans sa version classique[8], sinon en tant que spécialité de la science informatique, du moins au regard de ses ambitions originelles.

L'intelligence artificielle classique s'est constituée en discipline à partir de l'idée que les actions humaines pouvaient être représentées par des systèmes de symboles, régis par des règles établies en fonction de l'analyse, supposée pouvoir être exhaustive, du comportement effectif d'un individu humain devant un problème déterminé. Dans cette perspective, les promoteurs de l'intelligence artificielle se sont intéressés principalement à quelques thèmes privilégiés, tels les jeux logiques, le maniement du langage naturel, la reconnaissance des formes, l'apprentissage... Cependant, au-delà des résultats spectaculaires atteints dans ces différents domaines[9], des limites sont assez vite apparues, et, pour l'essentiel, les prédictions aventureuses faites à la fin des années cinquante n'ont pas été vérifiées[10]. Tout se passe, en

[8] Nous entendons par là l'intelligence artificielle " computationnelle " ou " symbolique " - à laquelle s'adressent principalement, par exemple, les critiques minutieuses d'Hubert Dreyfus (Hubert Dreyfus, *Intelligence artificielle : mythes et limites*, Paris, Flammarion, 1984) – et non l'intelligence artificielle " connexioniste " développée depuis les années 1980.

[9] Le premier programme de traduction automatique (un programme traduisant - mal - du russe en anglais) fut élaboré dès 1956. Les premiers programmes de jeu d'échecs satisfaisants, c'est-à-dire capables de battre un amateur occasionnel, apparurent à la fin des années soixante. A peu près au même moment le *Stanford Research Institute* mettait au point le premier robot capable d'orienter son déplacement en fonction de formes qu'il reconnaissait.

[10] En 1958, Newell et Simon prédisaient que dix ans plus tard un programme d'échecs serait capable de devenir champion du monde. Ce n'est qu'en 1996 qu'un ordinateur s'est montré capable de vaincre le champion du monde en titre, résultat qui a, du reste,

effet, comme si la compétence des programmes d'intelligence artificielle était inversement proportionnelle à l'étendue de leur champ d'application : la machine excelle dans des tâches précises, dont les contours sont parfaitement circonscrits, mais toute une part du comportement humain, qui ne semble pas pouvoir être interprétée sous la forme d'un système de règles explicites, lui demeure inaccessible. De l'entreprise originelle reste aujourd'hui une thèse, que l'on énoncera, ici, en reprenant la formulation qu'en a proposée Jean Mosconi : " Tout comportement humain qui peut être décrit avec précision peut être simulé par un ordinateur convenablement programmé "[11].

Cette thèse donne lieu elle-même, comme l'a montré John Searle, à deux versions, l'une " faible ", l'autre " forte " :

> D'après l'IA faible, la principale valeur de l'ordinateur dans l'étude de l'esprit, c'est qu'il est pour nous un outil très puissant. Ainsi il nous permet de formuler et de tester des hypothèses de façon plus rigoureuse et plus précise. D'après l'IA forte en revanche, l'ordinateur n'est pas simplement un outil d'étude de l'esprit ; l'ordinateur convenablement programmé est véritablement un esprit, en ce sens que des ordinateurs munis des bons programmes *comprennent* et ont d'autres états cognitifs. En IA forte, l'ordinateur programmé ayant des états cognitifs, les programmes ne sont pas simplement des outils nous permettant de tester des explications ; ils sont eux-mêmes les explications[12].

La distinction proposée par John Searle correspond assez bien à la réalité actuelle de l'intelligence artificielle classique, laquelle a, certes, conquis un véritable statut, celui d'une science appliquée - qui fait d'elle une technique informatique parmi d'autres - mais au prix du renoncement progressif à son objectif initial de simulation du comportement général de l'homme. En vérité, l'histoire de l'intelligence artificielle a suivi celle de la machine spécifique à laquelle elle est liée - l'ordinateur - et sa reconnaissance comme " IA faible " ou encore comme science appliquée, accompagne la banalisation de ce type de machine. On remarquera, à cet égard, que la description du fonctionnement de la machine dont se sert

été obtenu davantage en faisant appel aux progrès réalisés en matière de puissance de calcul qu'aux techniques d'intelligence artificielle proprement dites.

[11] Jean Mosconi. " Sur quelques capacités et incapacités des machines ". *Bulletin de la Société Française de Philosophie*, 3, juillet-septembre 1991, p. 86.

[12] John Searle. " Esprits, cerveaux et programmes ", *Vues de l'esprit*, D. Hofstadter, D. Dennett, éd., Paris, InterEditions, 1987, p. 354.

l'intelligence artificielle fait appel à un usage bien défini du terme " action " : " l'action " de l'ordinateur se caractérise par le fait qu'elle peut être entièrement définie par le programme qui lui correspond. De celui-ci, entendu comme suite d'instructions énoncée au moyen d'un langage formel, à son " exécution ", la machine ne change pas de statut : elle coïncide toujours, y compris dans son " action ", avec sa définition formelle. Ses erreurs mêmes, ou les " ratés " de son action, en relèvent. En tout état de cause, il doit toujours être possible, au moins *a posteriori*, de rendre compte de l'action de la machine à partir de sa description formelle. Une fois la validité du programme établie au regard des résultats visés, il apparaît, au fond, indifférent, sur le plan théorique, que " l'action " qui lui correspond ait effectivement lieu ; cela même est, paradoxalement, au fondement du rôle technique de l'ordinateur, et de la portée de l'intelligence artificielle en tant que science appliquée.

On voit par là que la conception de la machine qui sous-tend l'idée " d'action " telle qu'elle est appliquée à la machine informatique, est, peu ou prou, celle de l'opinion commune visée par Turing dans *Computing Machinery…* C'est bien de la machine dont " l'action " est entièrement décrite par un programme que l'opinion commune soutient qu'elle ne peut " penser ", au sens humain du terme, car elle ne saurait jamais surprendre un homme comme peut le faire un autre homme, dont le comportement, présumé libre, est, pour cette même opinion commune, imprévisible dans son principe. Aux yeux de l'opinion commune, la notion " d'action ", telle qu'elle est appliquée à la machine, ne saurait rendre compte du comportement humain. On opposera, ici, au *fonctionnement* de la machine, l'*agir* de l'homme, et, à la notion " d'action ", celle d'*acte*, entendue comme ce qui pose, dans son avènement même, sa propre nécessité.

Or, Turing le précise dans *Computing Machinery…*, la machine du jeu de l'imitation est un " ordinateur digital ", c'est-à-dire l'ordinateur au sens moderne du terme ; en somme, la machine du test de Turing est celle de l'intelligence artificielle. Aussi bien l'hypothèse de *Computing Machinery…* - une machine victorieuse au jeu de l'imitation " pense ", c'est-à-dire " comprend " – est-elle généralement regardée comme une première formulation de la thèse de " l'IA forte ". Celle-ci faisant fonds sur l'idée que le fonctionnement d'une machine implique une méthode définie - la machine peut simuler, comme le dit Jean Mosconi, ce qui est décrit " avec précision " - l'assimilation de l'hypothèse de Turing à la thèse de " l'IA forte " conduit à poser que, pour Turing, puisqu'une machine peut " penser ", la pensée en

tant que telle tient tout entière dans l'idée de méthode définie. La machine serait en mesure de l'emporter au jeu de l'imitation car le comportement du vainqueur de celui-ci pourrait être décrit " avec précision ". Par là même, *Computing Machinery...* semble appeler les critiques, fondées sur les échecs de l'intelligence artificielle au regard de ses intentions originelles, qui ruinent la thèse de " l'IA forte ", à savoir que, précisément, le comportement général des êtres humains n'est pas définissable en termes de méthodes définies, et ne saurait être ramené à " l'action " d'un ordinateur. De sorte, enfin, que l'histoire même de l'intelligence artificielle devrait être regardée, à travers ses échecs, comme l'infirmation de l'hypothèse de Turing : ou bien la " victoire " d'une machine au jeu de l'imitation relève d'un effet de surprise dû à la mise en œuvre, au cours de la construction de la machine, d'une méthode définie, c'est-à-dire d'un effet de surprise qui n'est pas de même nature que celui qu'un homme peut lui-même produire, ou bien la machine, contrairement à ce que veut montrer Turing, est simplement incapable de réussir le test. L'histoire de l'intelligence artificielle, en infirmant l'hypothèse de Turing, confirmerait, en somme, l'opinion commune.

Il nous semble, cependant, que la notion " d'action " mise en jeu dans le cadre de cette interprétation, ne permet pas de rendre compte de ce que fait la machine qui réussit le test, c'est-à-dire de ce qui se passe, de ce qui advient, selon Turing, au cours du jeu de l'imitation.

On notera ainsi que, dans le contexte de la notion classique d'intelligence artificielle, le sens même de la question " les machines peuvent-elles penser ? ", telle que la pose Turing, est singulièrement obscurci. Au moment, en effet, où paraît *Computing Machinery…*, Turing est avant tout connu comme logicien et mathématicien, et plus particulièrement pour avoir proposé, dans *On Computable Numbers with an Application to the Entscheidungsproblem*, à partir des concepts de " machine de Turing " et de " machine universelle ", une définition scientifique de la notion de machine. Dès lors, le premier problème d'interprétation posé par *Computing Machinery…* est, précisément, celui de son rapport à la notion même de machine universelle : pourquoi Turing pose-t-il la question de la " pensée " des machines à propos d'une notion qui appartient de plein droit à la logique mathématique, et dont la validité ne peut être établie que dans le cadre de celle-ci ? Or, la réponse à cette question, écartée, semble-t-il, par l'interprétation classique de la démarche de Turing, interdit d'admettre pour cette dernière les présupposés sur lesquels s'appuie l'idée d'intelligence artificielle.

La notion de machine universelle définit, dans *On Computable Numbers…*, une machine à " états discrets ", capable de simuler l'action de n'importe quelle autre machine, dès lors que celle-ci peut être décrite comme discrète. Sous un certain angle, une " machine de Turing universelle " se présente donc, vis à vis des " machines de Turing simples " qu'elle est en mesure de reproduire, comme une sorte de métamachine, une machine qui " connaîtrait " les machines qu'elle imite, à travers le symbolisme, décrivant ces machines, qu'elle utilise pour les imiter. De sorte que tout se passe comme si la notion de machine universelle enveloppait un rapport à la pensée qui ne peut être réduit à la forme logique du calcul. Turing insiste, en particulier, dans *On Computable Numbers…*, sur le fait que la possibilité, pour une " machine universelle ", de simuler n'importe quelle autre machine décrite sous la forme d'une " machine de Turing simple ", repose sur sa capacité à simuler, en quelque sorte à l'autre bout de la chaîne des simulations, les conditions intuitives du calcul pour un homme, c'est-à-dire la part qui, dans le calcul humain, échappe au formalisme. Conformément à la description qu'en donnait, au début du siècle, David Hilbert, ces conditions intuitives sont celles qui président à la perception du signe mathématique comme figure matérielle. Or, du fait qu'il s'agit ici, d'une part, non de n'importe quelle créature percevante, mais de l'homme, et, d'autre part, de la perception, non de n'importe quelle figure matérielle, mais du signe mathématique, la simulation par la machine de la perception du signe comme condition du calcul renvoie aux conditions mêmes de sa construction ; elle renvoie, en d'autres termes, à l'homme, non seulement en tant qu'il perçoit, mais, au-delà, en tant qu'il est le *constructeur* de la machine. Ce que simulerait, en définitive, la " machine universelle ", c'est l'étrange propriété sur laquelle repose la possibilité, pour la langue de son constructeur, de se constituer en métalangue. La notion de " machine universelle " renvoie à *l'acte* de pensée par lequel est posé, dans sa dynamique spécifique, le signe mathématique. C'est, selon nous, cette dimension de la notion de " machine universelle " que Turing se propose d'examiner dans *Computing Machinery…*, en formulant l'hypothèse que les machines " peuvent penser ". De sorte qu'en posant la question " les machines peuvent-elles penser ? ", et en y répondant par l'affirmative, Turing n'entend pas réduire le " penser ", au sens humain du terme, à ce qui, dans le comportement de l'homme qui calcule, est logiquement formalisable, mais plutôt élever " l'action " de la machine au niveau de ce qui, dans le comportement de l'homme qui calcule, est condition du calcul. L'hypothèse de Turing, selon laquelle les machines " peuvent

penser ", n'affirme pas que l'homme est l'égal de la machine, mais bien que la machine est l'égale de l'homme.

C'est enfin pourquoi toute la démonstration de Turing dans *Computing Machinery*... vise à établir que le succès d'une " machine universelle " au test proposé – sa victoire face à un adversaire humain au jeu de l'imitation - tient à la possibilité pour elle d'effectuer des actions qui, précisément, ne sont pas susceptibles d'être décrites sous la forme de méthodes définies. Le comportement même de la machine victorieuse, c'est-à-dire de celle qui se montre apte à reproduire l'action d'un individu humain capable lui-même de l'emporter au jeu de l'imitation, ne peut être décrit " avec précision ". La seconde hypothèse envisagée par Turing, dans *Computing Machinery*…, après celle de la victoire d'une machine au jeu de l'imitation, à savoir l'hypothèse des " machines qui apprennent ", souligne tout particulièrement cet aspect. La machine qui réussit le test de Turing doit être entendue comme une machine qui a la possibilité *d'apprendre*, et soumettre une telle machine à un processus d'éducation analogue à celui auquel le petit d'homme est lui-même soumis, apparaît, aux yeux de Turing, comme la seule manière de la construire effectivement. En ce sens, la machine victorieuse au jeu de l'imitation n'est pas l'exact équivalent d'un programme qui peut toujours être considéré indépendamment de sa mise en œuvre ; tout se passe, en vérité, comme si certains moments de l'action de cette machine n'étaient pas strictement déterminés par les moments précédents, comme si, dans certains cas, la description formelle de l'état actuel de la machine ne rendait pas compte de tout ce qu'elle peut faire au moment suivant. La machine qui réussit le test de Turing est, certes, déterminée par des programmes – sa configuration, à un moment quelconque de son action, peut toujours être décrite – mais il n'est jamais possible de ramener ces programmes à une unité formellement descriptible, car la machine est construite par son action même, c'est-à-dire par un devenir, qui la définit comme *agent*.

Turing, en somme, établit, dans *Computing Machinery*…, la possibilité pour la " machine universelle ", de manifester l'imprévisibilité même qui caractérise, aux yeux de l'opinion commune, le comportement humain, c'est-à-dire cette imprévisibilité qui relève d'un *agir* plutôt que d'un fonctionnement, qui appelle, plutôt que l'idée " d'action ", celle d'*acte*.

La démarche de Turing comme critique de l'idée philosophique selon laquelle la pensée appartient à un autre ordre que le mécanique.

Puisque l'expérience proposée par Turing est conçue de telle sorte que l'infirmation de l'opinion commune constitue, en elle-même, la confirmation de l'idée selon laquelle la machine se situe sur le même plan que l'individu humain, en tant que celui-ci est considéré comme pensant, on pourrait arguer, sans doute, de ce que l'infirmation de l'opinion commune, ici, démontre seulement que la machine " pense " au sens du terme " penser " pour l'opinion commune. On constate, cependant, que cette dernière rejoint une théorie philosophique : l'idée qu'une machine ne peut pas penser, c'est-à-dire que le " penser " appartient à un autre ordre que le " mécanique ", s'inscrit dans une conception du penser développée au sein de la philosophie occidentale moderne.

Du point de vue de la démarche adoptée par Turing – un homme peut-il distinguer un homme d'une machine ? - il paraît légitime de s'attarder plus particulièrement sur la manière dont la distinction radicale de l'homme et de la machine est affirmée dans le cadre de deux moments particulièrement significatifs de l'histoire de la philosophie occidentale : le cartésianisme et le kantisme. Si la machine l'emporte au jeu de l'imitation – si elle est prise, par un homme, pour un autre homme – c'est, en effet, montre Turing, que son adversaire humain doit, dans le cadre du jeu, se comporter " comme si " son interlocuteur mécanique était, pour lui, un *semblable*, c'est-à-dire un *autrui*. Or, la problématique dans laquelle s'inscrit le " comme si " de l'adversaire humain de la machine, sollicite, d'une part, la structure conceptuelle qui, appuyée sur le *cogito,* définit, chez Descartes, le penser comme *jugement*, d'autre part, l'analyse dans le cadre de laquelle sont dégagés, chez Kant, les éléments d'une théorie des conditions de possibilité de la reconnaissance, par l'homme, d'un autrui.

Exposant les principes de la célèbre théorie de " l'animal-machine ", Descartes, dans la " Cinquième Partie " du *Discours de la Méthode*, évoque déjà, bien avant Turing, une situation analogue à celle imaginée dans *Computing Machinery…* Un homme, affirme-t-il, ne peut jamais être confondu avec un automate, en raison notamment du fait qu'il use de la parole. La pensée, en lui, se manifeste à travers un acte : le jugement, lequel est l'expression d'une volonté libre, qui porte en elle-même la certitude première du *cogito* – puisque le

“ doute méthodique ” implique la liberté. Or, *cogito* et jugement n’ont pas de sens au regard de la notion de machine, qui relève de la seule “ substance étendue ” ; il n’y a pas d’acte de pensée pour une machine. C’est pourquoi, déclare Descartes, dans l’hypothèse purement imaginaire où nous devrions distinguer un automate imitant l’homme d’un homme véritable, la parole, en tant qu’elle énonce le jugement, serait l’un des moyens infaillibles de reconnaître l’homme : une parole de même nature que la parole humaine est inconcevable pour l’automate, aussi parfait soit-il en son genre. Nous verrons, cependant, que, dans *Computing Machinery*…, le “ comme si ” de l’adversaire humain de la machine, qui rend possible la victoire de celle-ci au jeu de l’imitation, est déterminé par la nécessité dans laquelle le jeu même place cet adversaire de prêter au discours à la première personne énoncé, pour le tromper, par son interlocuteur mécanique, un statut, fondé sur l’idée de “ consistance existentielle ”, équivalent au sien propre. On sait que la notion de “ consistance existentielle ” rend compte, dans la lecture proposée par Jaakko Hintikka, de la dynamique du *cogito* entendu comme “ performance ”.

Cette question du statut du discours à la première personne énoncé par la machine victorieuse au jeu de l’imitation prend toute sa signification à la lumière de la confrontation de la démarche de Turing avec la réflexion développée, à propos de l’automate, par Kant. Celui-ci dénonce, dans l’“ Analytique ” de la *Critique de la raison pratique*, l’usage abusif qui est fait du concept de liberté lorsqu’il est appliqué à l’automate. Dans le contexte spécifique de la problématique kantienne de la liberté, la question des rapports de la machine et de l’homme telle que la pose Turing – la question de ce qui distingue l’homme de la machine – concerne les conditions de la reconnaissance dans l’autre, qui me parle et à qui je parle, d’un autrui. L’acte de pensée, défini, là encore, par le jugement et mettant en jeu le “ je pense ” sous la forme de “ l’aperception pure ”, renvoie à l’idée de sujet, comme sujet transcendantal, et, à travers celle-ci, à la *personne*. La personne, parce qu’elle relève d’un ordre intelligible, ne saurait, chez Kant, être connue ; je peux cependant la penser. Ainsi, dans le cadre de *l’échange de parole*, et parce qu’à travers la parole se manifeste la faculté de juger, je peux me comporter, et je me comporte effectivement, *comme si* j’étais assuré, par une certitude de connaissance, que l’autre qui me parle et à qui je parle est une personne.

Sous cet angle, la démarche kantienne – le “ comme si ” mis en oeuvre par la reconnaissance d’un autrui en tant que personne – constitue une sorte de modèle

d'interprétation de la victoire possible d'une machine au jeu de l'imitation : la machine l'emporte, en effet, car son adversaire humain est conduit, au cours d'un échange de parole, à se comporter à son égard comme il se comporterait à l'égard d'un autre individu humain, c'est-à-dire, en termes kantiens, *comme si* elle était une personne ; c'est parce qu'il croit voir en la machine ce qui est explicité dans le kantisme sous l'idée de personne que l'adversaire humain de la machine confond celle-ci avec un être humain. En d'autres termes, dès lors qu'il repose sur le " comme si " de l'adversaire humain de la machine, le succès de celle-ci au jeu de l'imitation ne saurait se concevoir sans la mise en œuvre de ce qui, dans la problématique kantienne, renvoie à la détermination des conditions de la reconnaissance d'un autrui comme personne, par l'homme considéré comme " être raisonnable ".

Or, il ne saurait y avoir, chez Kant, de jugement de la machine, laquelle relève, non d'une " logique transcendantale ", impliquant la postulation d'un ordre intelligible inconnaissable, mais de la causalité naturelle ; quand bien même une machine serait suffisamment bien conçue pour imiter la parole humaine dans sa forme phénoménale, elle ne " parlerait " pas au sens où un homme parle. Il n'y a pas d'échange de parole possible avec une machine ; on ne saurait donc reconnaître, en elle, un sujet, ni penser une personne, pour la raison que l'on n'a jamais à le faire.

On voit par là que, ni chez Kant ni chez Descartes, la question de la distinction de l'homme et de l'automate n'est posée sous la forme d'un problème : la définition même de la machine repose sur l'idée qu'elle appartient à un ordre différent de celui du penser. Turing, dans *Computing Machinery…* rompt avec cette approche par le fait même qu'il inscrit la question de la distinction de l'homme et de la machine dans la logique d'une expérience. La situation imaginée avec le jeu de l'imitation - l'acte consistant, pour un homme, à distinguer un homme d'un automate – prend sens, en tant qu'expérience, de ce qu'elle n'est pas univoque : il est possible, entend montrer Turing, que, dans une telle situation, un homme se trompe. En d'autres termes, il y a, pour Turing un *problème* de la reconnaissance, par un homme, des autres hommes à travers l'échange de parole. Aussi bien l'acte de pensée est-il, ici, inséparable d'un procès de communication ; dans le cadre du jeu de l'imitation le penser est public, alors qu'il est, chez Descartes et Kant, le fait d'une conscience - un " je pense " - qui s'adresse d'abord à soi, et qui, par là, se trouve définie, en dernière instance - chez Kant - par l'idée de sujet, entendu comme sujet transcendantal. L'acte de pensée, pour le kantisme

comme pour le cartésianisme, est dirigé vers lui-même. Dès lors, la problématique dans laquelle s'inscrit la démarche de Turing n'est-elle pas d'emblée étrangère à celle où s'exprime, chez Descartes et Kant, au-delà de ce qui les séparent, l'idée de pensée ? En vérité, là est précisément, à nos yeux, l'enjeu philosophique de la réflexion de Turing : de celle-ci ressort, si elle a un sens, que l'absence de prise en compte, pour elle-même, de la dimension publique de l'acte de pensée constituerait la limite même de la problématique générale de l'idée de pensée chez Descartes et Kant ; c'est parce qu'elle est d'abord dialogue, procès d'énonciation-communication, que la pensée est acte et ne peut être réduite au formalisme logique.

Nous l'avons vu, l'expérience proposée par Turing est imaginaire : il ne saurait être question, en 1950, d'envisager sa réalisation. Aussi bien Turing s'efforce-t-il, à défaut de pouvoir vérifier son hypothèse, d'établir la validité de celle-ci en montrant que rien, dans la notion théorique de machine qu'il avait lui-même élaborée dans *On Computable Numbers*…, n'interdit de poser la possibilité pour une machine, d'une part, de se bien comporter au jeu de l'imitation, d'autre part " d'apprendre " comme le fait un individu humain. Tout se passe, en somme, comme si, pour Turing, rien, dans la définition de la machine comme machine universelle, ne s'opposait à ce qu'il y ait, au cours du jeu, un *échange de parole* entre des individus humains et une machine, ni à ce que cette dernière, dans ce cadre, soit tenue, par un individu humain, pour ce que Kant explicite à travers l'idée de personne.

Devrions-nous en conclure qu'admettre la validité de l'hypothèse de Turing, entraîne à considérer que la machine participe d'une " logique transcendantale ", et non pas seulement de la causalité naturelle ? Etablir la possibilité, pour une machine conforme à la définition qu'en donne Turing, d'être prise pour une personne, ne nous conduit-il pas à la nécessité d'élever la machine à la dignité du sujet moral, c'est-à-dire à l'affirmation d'une *liberté* de la machine ? Tout ne se passe-t-il pas comme si la définition de la machine élaborée par Turing élargissait la notion de machine de telle sorte qu'il conviendrait simplement d'adapter cette notion – par exemple en définissant la machine, au même titre que l'être humain, comme un " être raisonnable " - tout en tentant de conserver l'appareil kantien rendant compte des conditions de possibilité de la reconnaissance d'un autrui ?

Une telle conclusion, cependant, dépasserait l'hypothèse de Turing. Admettre la validité de celle-ci n'impose pas la mobilisation de l'ensemble de la problématique kantienne.

S'il est vrai, en effet, que la machine réussit le test de Turing parce que son adversaire humain postule à son propos, à travers le " comme si " par lequel il s'exprime, l'ordre intelligible où s'inscrit la notion de personne, nous ne sommes pas autorisés pour autant à rendre compte de l'attitude de l'adversaire humain de la machine à partir de cette postulation même. Ce serait oublier ce qu'implique le fait même que Turing entende répondre à la question " Les machines peuvent-elles penser ? " par une " expérience " plutôt que par une discussion de concepts, à savoir qu'à ses yeux le cadre de cette expérience est délimité, non par une définition - introuvable - de l'idée de pensée, mais par celle – fixée dans *On Computable Numbers*… - de la notion de machine. Dès lors, nous ne pouvons nous donner le droit d'interpréter " l'expérience " du jeu de l'imitation en nous appuyant sur une approche conceptuelle préétablie de l'idée de pensée ; la seule donnée conceptuelle à laquelle fasse appel l'hypothèse de Turing *stricto sensu* est la notion de " machine universelle ". En d'autres termes, dans le cadre strict où se place Turing, nous pouvons seulement conclure, en premier lieu, qu'un échange de parole est possible entre des individus humains et une machine, qui conduit à l'erreur commise par l'adversaire humain de celle-ci, et, en second lieu, non pas que le procès d'énonciation-communication dans lequel l'individu humain est ainsi engagé, serait fondé, même " problématiquement ", pour reprendre l'idée kantienne, par l'ordre dont il exprime la postulation, mais, plutôt, et de manière plus limitée, que l'erreur de l'adversaire de la machine, exprimée sous la forme de la postulation d'un intelligible " problématique ", est *constitutive* du procès d'énonciation-communication même. Le fait qu'admettre la validité de l'hypothèse de Turing implique la mise en jeu de la structure catégorielle du " comme si ", ne saurait conduire, en toute rigueur, à poser la distinction kantienne d'un sensible, objet de connaissance, et d'un intelligible " problématique ", représentant l'ordre dont relève, en dernière instance, l'acte de pensée accompli par l'homme. Or, cela même bouleverse la problématique kantienne : considérée dans son extension stricte, l'hypothèse de Turing conduit à *renverser* la démarche kantienne, en faisant de l'intelligible hypothétique inscrit dans celle-ci, non plus un *fondement* posé par sa postulation même, mais seulement un *moment* du procès – en l'occurrence l'échange de parole – au cours duquel un semblable, c'est-à-dire un autrui, est affirmé par l'individu humain.

La singulière spécificité de l'argumentation de Turing dans *Computing Machinery*…, où l'infirmation de l'opinion commune vaut par elle-même confirmation de la " pensée " de la

machine, est porteuse, en ce sens, d'une critique du dualisme qui, chez Descartes comme chez Kant, marque l'approche de l'idée de pensée. Nous ne saurions tenter d'élucider, dans toute leur portée, les implications de cette dimension de la réflexion de Turing sans dépasser nous-mêmes les limites que celui-ci lui a donné. Aussi bien nous efforcerons-nous seulement de dégager ce qui pourrait en constituer le principe. Notre thèse sera, de ce point de vue, que la réflexion critique inscrite dans *Computing Machinery…*, à l'encontre, non plus seulement de l'opinion commune, mais de l'idée philosophique selon laquelle une machine ne peut penser, telle que l'expriment, dans leurs contextes respectifs, cartésianisme et kantisme, se noue autour de la notion de *pratique*, dont la réflexion kantienne montre, précisément, comment elle détermine l'idée d'acte de pensée, à travers une théorie de la subjectivité qui définit cette dernière comme subjectivité transcendantale. Nous remarquerons, à cet égard, que la démarche de Turing enveloppe un *acte* critique à l'égard de la notion de pratique élaborée au cours de l'histoire de la philosophie occidentale jusqu'à Kant, et que cet acte se révèle, ici, comme étant lui-même constitutif, en tant que tel, de la notion de pratique. A la lumière de la réflexion de Turing dans *Computing Machinery…*, tout se passe comme si l'idée philosophique de pratique apparaissait comme " fondée " - mais peut-il encore s'agir d'un " fondement " ? - sur la mise en jeu de sa propre critique.

Telle est la lecture de la démarche de Turing que nous voudrions défendre. Cela nous conduira, tout d'abord, à tenter de dégager le sens donné par Turing à la question " les machines peuvent-elles penser ? ", laquelle surgit, nous semble-t-il, dans l'environnement du problème théorique posé par la notion de " machine universelle " : celle-ci doit simuler l'acte même dont le signe qu'elle manipule est inséparable. Nous aurons ensuite à étudier la double hypothèse formulée par Turing dans *Computing Machinery…* : une machine universelle peut l'emporter, face à des adversaires humains, au jeu de l'imitation, et une telle machine " peut apprendre ". Nous devrons, en particulier, nous efforcer de comprendre sur quoi repose, aux yeux de Turing, la validité de ces hypothèses, à savoir que rien dans la notion théorique de machine n'interdit la réussite d'une machine au jeu de l'imitation, ni qu'une machine puisse " apprendre ". Il s'agira, alors, de déterminer le rapport critique, à notre sens organique, de la réflexion de Turing à l'idée philosophique selon laquelle le penser appartient à un autre ordre que le mécanique. Admettre la validité des hypothèses de Turing implique que l'adversaire

humain de la machine accorde au discours de celle-ci un statut analogue à celui de son propre discours, et, qu'il mette en jeu, à travers cela, une problématique pour laquelle la réflexion kantienne, par le biais de la question de la reconnaissance d'autrui comme personne, impose un modèle. Les hypothèses de Turing sont ainsi porteuses d'un bouleversement de cette problématique même, qui peut s'exposer dans la notion philosophique de pratique, définie comme la dynamique de sa propre critique.

Première partie : la question « Les machines peuvent-elles penser ? »

Quel est le sens donné par Turing à la question " Les machines peuvent-elles penser ? " ? Turing précise, au début de *Computing Machinery...*[13], que les machines dont il va être question dans ce texte sont exclusivement les calculateurs électroniques - les ordinateurs - considérés indépendamment de ce que la technique rend possible en 1950, et en tant qu'ils peuvent être définis comme des " machines universelles de Turing ". Or, la notion de " machine de Turing " appartient à la logique mathématique, ce qui n'est certainement pas le cas de la question de la " pensée " des machines. Bien plus, il est permis de considérer que, dans son champ spécifique, la définition de la " machine de Turing " est précisément construite sur l'exclusion de toute une part de l'idée de " pensée " au sens usuel - c'est-à-dire humain - du terme. Certes la machine " calcule ", et Turing affirme qu'elle calcule comme le fait un individu humain ; cependant, la portée de l'idée de " procédé mécanique ", du point de vue des problèmes de logique mathématique envisagés par Turing, repose fondamentalement sur l'idée que le calcul opéré par la machine exclut toute *invention* mathématique, toute " ingéniosité ". Quel est donc le lien théorique qui unit la question de la " pensée " des machines à la notion de " machine universelle " ?

Ce problème détermine l'intelligibilité de la démarche de Turing dans *Computing Machinery...*, et son examen exige que soit tout d'abord rappelées les grandes lignes du travail de logique mathématique, aboutissant à la définition de la " machine universelle ", par lequel Turing se fit connaître en 1937[14]. La notion de " machine universelle ", au-delà de sa portée

[13] Aux § 3, 4 et 5 de la première partie du texte.

[14] L'exposé qui suit ne vise qu'à rappeler ce sans quoi la réflexion de Turing à propos de la " pensée " des machines ne pourrait être comprise. Pour un exposé historique et technique de la notion de " machine de Turing ", nous renvoyons le lecteur intéressé à la thèse de Jean Mosconi (Jean Mosconi, *La constitution de la théorie des automates*, thèse soutenue à l'université de Paris I, 1989) et à l'importante bibliographie qui y est établie.

proprement mathématique, renforce la plausibilité de l'équivalence intuitive entre procédure de calcul et procédé mécanique. Sa définition met en jeu la théorie hilbertienne du signe mathématique et implique la possibilité, pour la machine, de simuler ce que nous appellerons ici les conditions intuitives du calcul pour un individu humain. Or, nous verrons que la discussion que Turing eut avec Wittgenstein, lors des *Cours sur les fondements des mathématiques* données par ce dernier en 1939 à Cambridge, relativise l'équivalence intuitive entre procédure de calcul et procédé mécanique. En vérité, la question " Les machines peuvent-elles penser ? ", si elle ne relève pas, en tant que telle, du domaine de définition de la notion de machine universelle, à savoir la logique mathématique, semble être, non pas requise, mais appelée par la définition scientifique donnée de cette notion par Turing. C'est à partir de la simulation des conditions intuitives du calcul pour un individu humain, et de la perspective ouverte par la définition de ces conditions intuitives, que se pose, à propos de la machine, la question de la " pensée ", au sens humain du terme, c'est-à-dire au sens où tout individu humain est d'abord considéré comme pensant. Le signe mathématique, en effet, comme le montrait Jean Cavaillès, à propos de la théorie hilbertienne du signe, au moment même où Turing élaborait la notion de machine universelle, est inséparable de l'*acte* de pensée qui le pose, et la notion de machine universelle doit elle-même renvoyer à cet acte.

Chapitre I : La “ machine universelle ”

Turing, au printemps de 1935[15], suit un cours de M.H.A. Newman, sur les fondements des mathématiques et étudie, dans ce cadre, les théorèmes d'incomplétude de Gödel. L'article qu'il publie en 1937, *On Computable Numbers, with an Application to the Entscheidungsproblem*[16], poursuit la réflexion qu'il avait engagée à cette occasion. Le travail de Turing porte alors sur le “ problème de la décision ”, auquel il entend étendre la démonstration de Gödel, et s'inscrit donc dans le contexte des recherches suscitées par la fameuse “ crise des fondements ” : la théorie cantorienne des ensembles, qui apparaissait, à la fin du 19e siècle, comme le principe d'une synthèse possible des diverses branches des mathématiques, conduisait à des antinomies, lesquelles, du fait même du statut fondateur de la théorie, ébranlaient l'édifice mathématique tout entier. Le caractère non-contradictoire des mathématiques – leur consistance – jusque là non problématique, demandait à être démontré. Le programme des travaux à accomplir en ce sens, parmi lesquels l'examen du “ problème de la décision ” - *das Entscheidungsproblem* - avait été établi par Hilbert, notamment au cours du Congrès de Paris en 1928.

Le “ problème de la décision ” met en jeu la notion essentielle de “ procédure effective ” de calcul. L'outil théorique forgé par Turing dans *On Computable Numbers*…, la “ machine de Turing ”, a d'abord pour fonction de clarifier cette notion.

Les antinomies sur lesquelles débouchait la théorie cantorienne des ensembles étaient liées à la nécessité de raisonner sur des ensembles infinis, c'est-à-dire de poser un infini actuel. Or, il semblait qu'abolir toute référence à l'idée d'un infini actuel imposât de rejeter

[15] Andrew Hodges, *Alan Turing ou l'énigme de l'intelligence*, Paris, Payot, 1988, p. 85.
[16] A. M. Turing, *On Computable Numbers…, op. cit.*

une bonne partie de l'Analyse classique. Hilbert, qui s'y refusait[17], propose, vers 1920, une théorie du signe mathématique - principe d'une axiomatisation générale – dont il espère qu'elle permettra une démonstration rigoureuse de la consistance des mathématiques. Il remarque ainsi que le contenu intuitif des raisonnements peut être réduit à la seule perception des formes matérielles des signes mathématiques : il est possible de considérer une démonstration mathématique comme une suite de figures matérielles – des symboles écrits sur du papier - comparables aux pièces d'un jeu d'échecs ; peu importe, pour le déroulement d'une partie d'échecs, que le " cavalier " représente un cavalier dans le monde réel : son rôle dans le jeu est défini uniquement par une règle de déplacement spécifique. On considérera, de la même façon, que la manipulation des signes utilisés dans une démonstration mathématique est entièrement déterminée par des règles, et non par une signification qui leur serait attribuée vis à vis du monde réel. Pour reprendre un exemple proposé par Jean Dieudonné, l'énoncé " x + y = y + x " sera regardé uniquement comme un certain assemblage des signes " x ", " y ", " + " et " = "[18]. Tout raisonnement mathématique devient par là un ensemble fini d'éléments parfaitement déterminés : des signes matériels et des règles. Il en est ainsi y compris lorsque sont utilisées des expressions telles que " quel que soit ", ou " il existe ", qui, si leur contenu était pris en compte, feraient nécessairement référence à un infini actuel : dans le cadre de la démarche hilbertienne, de telles expressions ne sont plus elles-mêmes que des formes matérielles. La théorie du signe mathématique défendue par Hilbert constitue, nous le verrons, l'une des conditions de la définition de la notion de machine par Turing.

Le principe mis en avant par Hilbert permettait d'examiner en tant que tels les procédés de démonstration utilisés dans les raisonnements, sous la forme des règles définissant les assemblages de signes. Il devenait possible d'élaborer une " théorie de la démonstration ", qui dégage des mathématiques les procédés déductifs utilisés, et en examine les propriétés. Les mathématiques devenaient elles-mêmes objet pour une théorie de second niveau, la métamathématique. La théorie de la démonstration donnait corps, en particulier, à la notion de système formel, laquelle permet de formaliser le processus de dérivation de théorèmes à partir d'axiomes ou à partir d'autres théorèmes eux-mêmes dérivés d'axiomes.

[17] Hilbert n'acceptait pas, disait-il, d'être " chassé du paradis cantorien ".
[18] Jean Dieudonné, " Les méthodes axiomatiques modernes et les fondements des mathématiques ", in *Les grands courants de la pensée mathématique*, Paris, Librairie scientifique et technique Albert Blanchard, 1962, p. 550.

Or, par cela même, la notion de système formel implique que l'on dispose d'une définition suffisamment précise de ce qu'est une " procédure effective " : comment caractériser de manière rigoureuse, sous la forme d'une méthode définie, le processus même de dérivation, c'est-à-dire la mise en oeuvre d'une règle ? En 1934, Gödel, reprenant une suggestion de Herbrand, proposait l'idée d'une relation entre procédure effective et fonctions récursives ; Kleene, en 1936, démontra l'équivalence entre le concept de " lambda-définissabilité " élaborée par Church quelques années plus tôt et la récursivité ; Turing, pour sa part, s'attache à l'idée, jusque là admise implicitement, qu'une " procédure effective " peut être assimilée à un procédé mécanique : il imagine une machine théorique - la " machine de Turing " - à partir de laquelle une fonction calculable peut être définie comme une fonction susceptible d'être calculée par une telle machine. La notion de " machine de Turing " constitue une formalisation de l'équivalence admise entre procédure de calcul et procédé mécanique. Turing montrera ensuite, en 1937, dans un appendice à *On Computable Numbers*..., que sa définition et celle de Church font référence aux mêmes classes de fonctions. Est ainsi formulée la " thèse de Church-Turing ", selon laquelle les fonctions calculables sont celles que calcule une machine de Turing, c'est-à-dire des fonctions récursives générales.

La conceptualisation proposée par Turing dans *On Computable Numbers*… sera clarifiée à partir de la fin des années quarante, notamment par Post[19], puis développée sous la forme d'une théorie des automates infinis complétant la théorie des automates finis mise au point par Kleene. Turing lui-même, en 1950, modifiera la présentation de sa machine en tenant compte des remarques de Post[20]. Cependant, la validité générale de la définition du " calculable " donnée par Turing au moyen de la notion de " machine de Turing " ne sera pas mise en question. De sorte que, lorsqu'il pose, dans *Computing Machinery*..., la question " Les machines peuvent-elles penser ? ", c'est bien, sur le fond, à la machine décrite dans *On Computable Numbers*... que Turing fait référence.

---

19 Post avait développé, en 1936, une théorie des ensembles récursifs dans laquelle il présentait une formalisation différente de celle de Turing, et qu'il réutilisa en 1947 pour proposer une révision de la machine de Turing. (E. Post. " Finite Combinatory Processes, Formulation I ", *The Undecidable : Basic Papers on Undecidable. Propositions, Unsolvable Problems and Computable Functions*, M. Davis, éd., New-York, Raven Press,1965).

20 A. M Turing, " The word Problem in Semi-Groups with Cancellation ", *Ann. of Math.*, 52, 2, 1950. Réédité in *Collected Works of A.M. Turing*, *op. cit.*, 1, *Pure Mathematics*.

La " machine de Turing "

L'idée d'une équivalence entre calcul et machine, exploitée par Turing dans *On Computable Numbers…* ne renvoie pas au calcul simplement en tant que celui-ci constitue une procédure, mais en tant qu'il est effectué par un être humain. La machine envisagée par Turing doit être capable de calculer tout nombre calculable par un individu humain. Ce point, nous le verrons, est décisif, et, en tout état de cause, conduit à définir, d'une manière générale, la " machine de Turing " comme un procédé mécanique destiné à " lire " et " écrire " sur un support des suites de figures matérielles ayant valeur de signes, de telle sorte que certaines de ces suites soient les nombres " calculés " par elle.

La " machine de Turing " peut être envisagée à deux niveaux ; en tant que " machine simple de Turing ", et en tant que " machine universelle ". C'est par le biais de ce second niveau, qu'elle permet à Turing de proposer une solution au " problème de la décision ".

On la décrira tout d'abord comme un dispositif constitué, d'une part, d'une tête de lecture-écriture et d'un ruban de longueur infinie divisé en cases sur chacune desquelles peut figurer un symbole et un seul, et d'autre part, d'" un nombre fini de conditions q1, q2, ..., qk [...] appelées des ' m-configurations ' ... "[21]. Les actions de ce dispositif sont déterminées à tout moment par la combinaison de sa " m-configuration " et du symbole figurant sur la case " observée " (" scanned ") à ce moment par sa " tête ". Ces actions peuvent consister à écrire un symbole sur la case " observée ", effacer un symbole, déplacer la " tête " d'une case vers la droite ou vers la gauche, enfin changer de " m-configuration ".

Turing distingue deux catégories parmi les symboles utilisés par la machine. La première comporte uniquement les symboles " 0 " et " 1 " - elle correspond à ce que nous appellerions aujourd'hui les " données " sur lesquelles travaille la machine ; la seconde peut comporter des symboles quelconques, et correspond à peu près à ce que nous appellerions des " codes de contrôle "[22]. Une séquence de symboles de la première catégorie constitue la

[21] " … a finite number of conditions q1, q2, …qk which will be called 'm-configurations' ". A. M. Turing. *On Computable Numbers..., op. cit.,* p. 231.

[22] " Certains des symboles écrits formeront la séquence de chiffres qui est le décimal du nombre en train d'être calculé. Les autres sont simplement des notes sommaires pour 'assister la mémoire'. Ce seront seulement ces notes sommaires qui seront susceptibles d'effacement. " (*Some of the symbols written down will form the sequence of figures which is the decimal of the real number which is being computed. The others are just rough notes to 'assist the memory'. It will only be these rough notes which will be liable to erasure*). *Ibid.,* p. 232.

représentation " en notation décimale binaire " de la suite infinie de chiffres située après la virgule d'un nombre réel. Il suffit, autrement dit, de " préfixer " la séquence de symboles de la première catégorie à l'aide d'une virgule pour obtenir le " nombre calculé " par la machine[23].

Turing montre que les " configurations ", c'est-à-dire l'ensemble des éléments qui déterminent le comportement de la machine à chaque étape, peuvent être décrites sous formes de tables. Si l'on admet, par exemple, comme le propose Turing, qu'" il est pratique d'inscrire les chiffres en sautant une case sur deux "[24], une machine construite pour calculer une séquence infinie de 0 et de 1 alternés, comportera quatre états (quatre " m-configurations ") - b, c, e, f - et un alphabet composé des deux symboles " 0 " et " 1 " ; elle effectuera les opérations " déplacement d'une case à droite " (D), " déplacement d'une case à gauche " (G), " effacement " (E), " impression " (I) et elle pourra être décrite sous la forme :

Configuration		Comportement	
« m-config » (ou état)	symbole	opération	état suivant
b	Aucun	I0, D	c
c	Aucun	D	e
e	Aucun	I1, D	f
f	Aucun	D	b

La partie gauche de la table - intitulée " Configuration " - indiquera la manière dont la machine est définie aux différentes étapes de son fonctionnement ; la partie droite - intitulée " comportement " - indiquera ce que la machine fait lorsqu'elle est déterminée par la conjonction d'une configuration et d'un symbole donné.

[23] " if an a-machine prints two kinds of symbols, of which the first kind (called figures) consists entirely of 0 and 1 (the other being called symbols of the second kind), then the machine will be called a computing machine. If the machine is supplied with a blank tape and set in motion, starting from the correct initial m-configuration, the subsequence of the symbols printed by it which are called of the first kind will be called the sequence computed by the machine. The real number whose expression as a binary decimal is obtained by prefacing this sequence by a decimal point is called the number computed by the machine ", *ibid.*, p. 232. Le choix de travailler sur des nombres décimaux sera plus tard critiqué, en particulier par Post, et il est d'usage aujourd'hui de considérer des machines calculant des fonctions d'entiers.

[24] Convention qui sera plus tard abandonnée, là encore à l'instigation de Post.

Une machine de ce type, ajoute Turing, “ qui n’écrit jamais qu’un nombre fini de symboles de la première famille est dite cyclique ”[25], et, par conséquent, une machine capable, au contraire, d’écrire une suite infinie de symboles sans séquence répétée sera dite “ acyclique ”.

Écrire la table de description d’une machine de Turing consiste donc à indiquer comment les configurations qui définissent la machine sont dérivées les unes des autres. Turing affirmait que les opérations ainsi décrites “ incluent toutes celles qui sont utilisées dans le calcul d’un nombre ”[26].

Or, une telle machine présente une propriété singulière : elle peut être conçue comme une “ machine universelle ”.

Certaines opérations telles que “ copier des séquences de symboles, comparer des séquences, effacer tous les symboles d’une certaine forme, etc ”[27] sont effectuées de manière répétitive par toute “ machine de Turing ” ; les tables impliquant ces opérations peuvent donc être abrégées sous forme de “ tables types ” (“ skeleton table ”), constituées de variables - notées par exemple par des lettres majuscules pour celles qui correspondent aux “ m-configurations ” et par des lettres grecques minuscules pour celles qui correspondent aux symboles - auxquelles seront substituées les valeurs réclamées par l’opération que la machine visée doit effectuer. Turing nommait ces “ skeleton table ”, des “ fonctions d’état ” ou “ m-fonctions ” (“ ‘m-configuration function’ or ‘m-function’ ”).

La table définissant une machine peut être écrite sous la forme de la suite des lignes décrivant la configuration et le comportement de la machine à chaque étape. Si l’on attribue des numéros aux signes figurant dans la colonne “ état ” : q1, q2, ..., qr (l’état initial étant toujours q1), ainsi qu’à ceux figurant dans la colonne “ symboles ” : S1, ..., Sm (avec pour toutes les tables, S0 = blanc, S1 = 0, S2 = 1), une ligne telle que la première de l’exemple considéré ci-dessus :

---

[25] Nous reprenons ici la traduction de “ *circular* ” et “ *circle-free* ” proposée par J. Basch in Jean-Yves Girard, *La machine de Turing*, *op. cit*. Cette traduction rend bien compte du comportement de la machine qui entre dans une “ boucle ” et imprime de manière cyclique une même séquence de symboles.

[26] “ it is my contention that these operations include all those which are used in the computation of a number ”. A. M. Turing, *On Computable Numbers..., ibid.*, p. 232.

[27] “ there are certain types of process used by nearly all machines, and these, in some machines, are used in many connections. These processes include copying down sequences of symbols, comparing sequences, erasing all symbols of a given form, etc. ”. A. M. Turing, *On Computable Numbers..., op. cit.,* p. 235.

b aucun I0, D c

pourra être écrite (avec i, j, k, m quelconques) sous la forme générale :

qi Sj ISk, D qm

et une description complète de la machine pourra être donnée en écrivant les unes à la suite des autres des lignes de cette forme, séparées par un signe tel que " ; ".

De même, les variables " q " et " S " peuvent être remplacées par des capitales ou des combinaisons de capitales telles que " C " suivi de i fois " A ", pour qi ou " C " suivi de j fois " B " pour Sj, etc., de telle sorte que les lignes composant les tables soient constituées uniquement de capitales telles que " A ", " B ", " C ", " D ", " G ", " E ". Nous obtenons ainsi ce que Turing appelait une " description standard " (" standard description ") de table.

Enfin, les signes constituant les lignes d'une " description standard " - les capitales, plus le signe " ; " - peuvent à leur tour être remplacés par des entiers : 1 pour A, 2 pour B, 3 pour C, 4 pour D, 5 pour E, 6 pour G, 7 pour " ; ", de telle sorte que chaque " description standard " soit représentée par un entier, que Turing appelle un " nombre descriptif " (" description number "). Une table telle que celle donnée plus haut pourra ainsi être ramenée à la " description standard " :

q1S0S1Dq2;q2S0Dq3;q3S0S0Dq4;q4S0Dq1;

ou :

CABCBDCAA;CAABDCAAA;CAAABCBBDCAAAA;CAAAABCBDCA;

et décrite par le " nombre descriptif " :

3123243117311243111731112322431111731111232431 7

Ainsi, déclare Turing, " à une séquence calculable correspond au moins un nombre descriptif, tandis qu'à un nombre descriptif correspond au plus une séquence calculable " [28], et " un nombre qui est le nombre descriptif d'une machine acyclique sera dit un nombre satisfaisant "[29].

Ces transformations complexes permettent à Turing de montrer qu'une " machine de Turing " peut simuler le comportement d'une autre " machine de Turing " dont le nombre

[28] " To each computable sequence there corresponds at least one description number, while to no description number does there correspond more than one computable sequence ", *ibid.*, p. 241.

[29] " A number which is a description number of a circle-free machine will be called a satisfactory number ", *ibid*. Un nombre " satisfaisant " calcule une suite infinie de décimales après la virgule d'un nombre réel.

descriptif lui est fourni en entrée sur son ruban. Nous sommes assurés en effet que si une machine M peut être construite, une machine M' sachant traduire son nombre descriptif en description standard - c'est-à-dire en notation par capitales - pourra l'être également. M' " traduira " en description standard la première configuration notée sous forme d'entiers (la première suite de chiffres jusqu'à 7), puis calculera la séquence correspondante ; elle traduira la seconde configuration puis calculera la séquence correspondante, et ainsi de suite.

Il est donc possible, déclare Turing, " d'inventer une unique machine U utilisable pour calculer n'importe quelle séquence calculable "[30].

Turing parvenait ainsi à la définition d'une " machine universelle " (" the universal computing machine "), et fournissait le principe qui sera plus tard utilisé par Von Neumann pour élaborer l'architecture des futurs ordinateurs : les instructions sont représentées dans la machine de la même manière que les données, et peuvent y être stockées, sous la forme de " tables types " aux variables desquelles la machine attribuera les valeurs exigées par les calculs particuliers qu'elle doit faire. Il ressort de là, enfin, que les instructions peuvent être soumises aux mêmes traitements arithmétiques que les données, c'est-à-dire qu'elles sont susceptibles d'être modifiées par la machine elle-même.

La caractérisation de la " machine de Turing " comme " machine universelle " était indispensable à Turing pour qu'il puisse appliquer sa définition mécanique de la notion de procédure effective au problème de la décision - *das Entscheidungsproblem*. La représentation de la configuration d'une machine donnée par un nombre descriptif, lui-même manipulable par une " machine de Turing " équivalait à traduire en termes de procédure mécanique l'arithmétisation des éléments d'un système formel tel que le calcul des prédicats, arithmétisation qui constituait le principe de la démonstration des théorèmes d'incomplétude de Gödel. Dans ce cadre, les techniques utilisées par Gödel, en particulier la " diagonalisation "[31] devenaient accessibles à une " machine de Turing ".

---

[30] " It is possible to invent a single machine which can be used to compute any computable sequence ", *ibid.,* p. 241.

[31] On sait que Cantor utilise le procédé de la diagonale pour démontrer l'impossibilité de dénombrer les nombres réels compris entre 0 et 1. Le problème consiste à examiner s'il est possible d'établir une correspondance biunivoque entre l'ensemble des entiers et l'ensemble des nombres réels compris entre 0 et 1. On peut constituer pour cela une liste de nombres réels mis sous la forme de fractions décimales illimitées dont la partie entière est nulle :

1/2 1/3 1/4 2/3 1/5 1/6 2/5...

ou :

Le “ problème de la décision ”

Le “ problème de la décision ”, d’un point de vue général, porte sur l’existence ou non, pour tout objet appartenant à un certain ensemble, d’un procédé effectif permettant de déterminer si cet objet possède une certaine propriété. Si l’on introduit la notion de système formel, il portera sur l’existence ou non d’un procédé général - un algorithme - permettant de déterminer si une formule quelconque d’un système formel est un théorème de ce système, c’est-à-dire est démontrable dans ce système. Le problème de la décision équivaut par conséquent à déterminer s’il existe une procédure générale permettant d’aboutir à une réponse pour toute question d’une classe de questions posées en termes de “ vrai ” ou de “ faux ”. En l’absence d’une connaissance de la totalité des algorithmes possibles, il est permis de traduire le problème en termes de calcul d’un nombre : soit une relation P(n) admettant deux valeurs possibles : “ vrai ” ou “ faux ”. Cette relation P(n) peut être considérée comme une fonction admettant deux valeurs possibles : “ 0 ” ou “ 1 ” - cette fonction est généralement appelée “ fonction caractéristique ” du prédicat P. La relation P(n) sera décidable si la fonction

---

0,5000000000...
0,3333333333...
0,2500000000...
0,6666666666...
...

Cette liste peut être ordonnée selon l’ordre des entiers :

1 0,5000000000...
2 0,3333333333...
3 0,2500000000...
4 0,6666666666...
...

Or, il est possible de définir, à partir du tableau ainsi obtenu, une fraction décimale illimitée du même type, c’est-à-dire un nombre réel compris entre 0 et 1, qui pourtant ne fait pas partie du tableau. On ajoute pour cela un entier à chaque décimale prise dans la diagonale du tableau :

1 0,**5**000000000...
2 0,3**3**33333333...
3 0,25**0**0000000...
4 0,666**6**666666...

ou :

0,(5+1)(3+1)(0+1)(6+1)...

Cette fraction est nécessairement différente de chacune de celles du tableau puisqu’elle diffère de la première au moins par la première décimale, de la seconde au moins par le seconde décimale. Voir : Jean Ladrière, *Les limitations internes des formalismes*, Paris, Gauthier-Villars, 1957, p. 77.

caractéristique du prédicat P est récursive, ou encore si elle peut être calculée par une " machine de Turing ".

A partir de là, le problème de la décision devient celui de savoir si la fonction caractéristique du prédicat " être démontrable " est calculable par une " machine de Turing ". Turing va établir qu'il n'en est rien.

Il démontre tout d'abord, en traduisant dans les termes de sa machine le procédé de la " diagonale ", qu'il n'existe pas de machine M capable de dénombrer la liste des machines de Turing acycliques (c'est le fameux " problème de l'arrêt "). Il montre alors que s'il existait une machine E capable de déterminer qu'une machine M écrit le symbole " 0 " " infiniment souvent " (" infinitely often "), cette machine E serait l'équivalent de la machine M.

Turing définit ensuite un " calcul fonctionnel K ", équivalent au calcul des prédicats du premier ordre[32], pour lequel existe une machine de Turing susceptible d'en calculer toutes les formules démontrables.

Dés lors, il s'agit d'établir que le calcul des prédicats du premier ordre n'est pas décidable en démontrant " qu'il n'existe pas de procédure générale permettant de déterminer si une formule U du calcul fonctionnel K est démontrable, c'est-à-dire qu'il n'existe pas de machine qui pourra, pour toute formule U, dire si U est démontrable "[33].

Turing montre qu'à partir de la définition d'une machine M, une formule Un(M) peut être construite, signifiant : " dans une configuration complète de M, S1 (c'est-à-dire 0) apparaît sur la bande "[34].

Puis il montre, à l'aide de deux lemmes, l'équivalence entre cette formule et la formule " Un(M) est démontrable " ; 1) si le symbole S1 apparaît sur la bande dans une configuration complète de M, alors Un(M) est démontrable ; 2) si Un(M) est démontrable, alors le symbole S1 apparaît sur la bande dans l'une des configurations complètes de M. En d'autres termes, si Un(M) est démontrable, M imprime au moins une fois le symbole 0.

---

[32] Turing modifie le " calcul fonctionnel restreint " de Hilbert " de façon à le rendre systématique et à ce qu'il n'utilise qu'un nombre fini de symboles " (" ... the Hilbert functional calculus is modified so as to be systematic, and so as to involve only a finite number of symbols... "). A. M. Turing, *On Computable Numbers..., op. cit.* p. 252.

[33] " ... there can be no general process for determining whether a given formula U of the functional calculus K is provable, i.e. that there can be no machine which, supplied with any one of these formulae, will eventually say whether it is provable ", *ibid.,* p. 259.

[34] " In some complete configuration of M, S1 (i.e. 0) appears on the tape ". *Ibid.,* p. 260.

Or, en fonction des démonstrations précédentes sur le problème de l'arrêt, nous savons qu'il n'existe pas de machine capable de déterminer si une machine donnée imprime ou non un symbole donné ; il n'existe donc pas de machine qui puisse, pour toute formule U, dire si elle est démontrable.

Ainsi, la notion de machine universelle permettait à Turing d'établir l'un des " théorèmes de limitation " - pour reprendre l'expression de Jean Ladrière[35] - qui définissent les limites de l'effort de formalisation par lequel Hilbert pensait pouvoir sortir les mathématiques de la " crise des fondements " entraînée par les antinomies de la théorie cantorienne des ensembles.

Cependant, dès lors que la notion de machine universelle donne une définition théorique de l'idée de machine et que sa mise au point se trouve d'abord justifiée par son rôle dans l'établissement du théorème d'indécidabilité, il est permis de se demander quel peut être, au regard de cette notion, le sens de la question posée quelques années plus tard par Turing dans *Computing Machinery…* - " Les machines peuvent-elles penser ? " - et de l'hypothèse formulée à ce propos.

Le problème posé par la question " Les machines peuvent-elles penser ? "

En quoi la validité des " théorèmes de limitation ", et parmi eux, de celui de Turing, qui circonscrivent avec précision les conditions dans lesquelles les mathématiques peuvent satisfaire à leur exigence fondamentale de consistance, est-elle concernée par l'hypothèse que les machines " pensent " ? Certes, le travail effectué par Turing en 1936-37 ne constituait le dernier mot ni de la question des limitations des systèmes formels, ni d'une théorie des automates qui restait alors à élaborer ; toutefois, on ne voit pas immédiatement de quelle façon l'idée de la " pensée " des machines pouvait contribuer au développement des recherches sur ces questions. Il est intéressant, à cet égard, de comparer la démarche de Turing avec celle de Von Neumann, dont on connaît le rôle décisif, à la fin des années quarante, aux Etats-Unis, dans la construction des premiers grands calculateurs électroniques[36]. Von Neumann énonce en 1945 les principes gouvernant l'organisation logique[37] des différents

[35] Jean Ladrière, *Les limittions internes des formalismes, op. cit.*
[36] En particulier l'EDVAC (Electronic Discrete variable Calculator).
[37] La fameuse " architecture de Von Neumann ", toujours en vigueur aujourd'hui.

organes composant une machine universelle réelle, et s'inspire pour cela, comme il l'a lui-même déclaré, du travail de Turing dans *On Computable Numbers...*, avant de proposer, dans une série de textes et de conférences[38], les éléments d'une théorie systématique des automates. Comme Turing dans *On Computable Numbers...*, il établit un rapport entre la machine, en l'occurrence le calculateur électronique, et l'être humain, au niveau du système nerveux central de celui-ci. Il justifie ce parallèle en en soulignant la fécondité sur le plan épistémologique : l'étude comparative du cerveau humain et du calculateur artificiel devait permettre, à ses yeux, tout à la fois de résoudre certains des problèmes posés par la construction des grands calculateurs et de comprendre certains aspects du fonctionnement du système nerveux central chez l'homme. Ce sont, cependant, les avancées plus spécifiquement mathématiques qu'il escomptait de cette réflexion qui l'intéressait : Von Neumann espérait ainsi préciser les fondements d'une logique intégrée aux mathématiques du continu et atteignant le même niveau de puissance que l'Analyse. Or jamais ce prolongement de ses recherches mathématiques à partir d'un parallèle entre machine et cerveau ne le conduisît à examiner la notion d'automate dans les termes de la question " Les machines peuvent-elles penser ? ". La réflexion de Von Neumann comporte certes une dimension qui peut être qualifiée de spéculative, mais elle ne sort pas pour autant du champ scientifique proprement dit : il s'agit de trouver une structure mathématique commune aux phénomènes du vivant et aux réalisations matérielles de systèmes logiques. Turing au contraire, en posant la question de la " pensée " des machines, paraît s'éloigner du domaine scientifiquement défini dans lequel il a élaboré la notion même de machine. Il est, du reste, permis de penser que cette orientation a joué un rôle dans le malentendu qui s'est rapidement installé entre lui et les ingénieurs avec lesquels, à partir de 1945, il travaille au projet de calculateur électronique britannique, malentendu qui devait finalement le conduire à quitter le *National Physical Laboratory* ; c'est en effet, rappelons-le, dans le cadre même des travaux officiels qui lui sont commandités par celui-ci que Turing aborde pour la première fois la question de la " pensée " des machines[39].

---

[38] Reproduit in John Von Neumann, *Theory of Self-Reproducing Automata*, Londres, University of Illinois Press, 1966.

[39] Tout d'abord dans une conférence qu'il prononce devant la *London Mathematical Society* en tant que membre du *National Physics Laboratory – Lecture to the London Mathematical Society on 20 February 1947* - puis dans le second rapport officiel qu'il rédige à l'intention de celui-ci – *Intelligent Machinery* (1948) (*Collected Works of A. M.*

Bien plus, il peut paraître paradoxal que Turing formule son hypothèse de la " pensée " des machines à propos d'une notion dont les limites logiques viennent d'être démontrées, d'autant que la notion de machine de Turing définit le " calculable " en excluant clairement de celui-ci toute idée *d'invention* mathématique : la machine fonctionne en mettant en jeu des états - les " m-configurations " - assignés à l'avance, et les modifications mêmes qu'elle pourra faire subir à ces états sont dérivées de ceux-ci ; c'est là une condition de sa validité formelle. Turing avait lui-même souligné fortement ce point dans une thèse intitulée *Systems of Logic Based on Ordinals*[40], rédigée sous la direction de Church et soutenue aux Etats-Unis deux ans après la publication de *On Computable Numbers*....

Il ne saurait être question, ici, de commenter la portée mathématique de la thèse américaine de Turing, mais uniquement d'essayer d'en dégager les éléments d'ordre périmathématique qui influeront, plus tard, sur la rédaction de *Computing Machinery*… La réflexion de Turing s'appuyait, dans ce second travail, sur le fait que, si le théorème de Gödel établit que tout système logique est incomplet, " il indique également les moyens par lesquels, à partir d'un système de logique L, un système L' plus complet peut être obtenu [et que] en répétant le processus, nous obtenons une séquence [de systèmes] L, L1 = L', L2 = L1', ... chacun plus complet que le précédent "[41]. De cette façon, continuait-il, " une logique Lω peut être construite dans laquelle les théorèmes prouvables sont la totalité des théorèmes prouvables avec l'aide des logiques L, L1, L2, ... "[42], et " nous pouvons alors former L2ω en relation avec Lω de la même façon que Lω est en relation avec L "[43].

Turing se proposait d'examiner le problème de la complétude dans ce cadre, en utilisant la suite des nombres ordinaux pour principe d'" arithmétisation " des logiques Lω. Le passage d'un système logique à un système plus fort, c'est-à-dire de L à L', se faisant par l'ajout d'un axiome, le problème revenait à déterminer si l'ensemble constitué par les axiomes

---

Turing, 2, *Mechanical Intelligence*, *op. cit.*).

[40] A. M. Turing, " Systems of Logic Based on Ordinals ", *Proceedings of the London Mathematical Society*, 2, 45, 1939.

[41] " ... but, at the same time, it indicates means whereby from a system L of logic a more complete system L' may be obtained. By repeating the process we get a sequence L, L1=L', L2=L1', ... each more complete than the preceding ", *ibid.,* p. 161.

[42] " A logic Lω may then be constructed in wich the provable theorems are the totality of theorems provable with the help of the logics L, L1, L2, ... ", *ibid.,* p. 161.

[43] " We may then form L2ω related to Lω in the same way as Lω was related to L. ", *ibid.*, p. 161.

ajoutés était fini. Turing démontrait qu'il n'en était rien, et remarquait, dans la conclusion de son travail, que tout raisonnement mathématique fait appel à deux " facultés " (" faculty "), " l'intuition " et " l'ingéniosité " (" ingenuity "). L'intuition consiste, expliquait-il, à " construire des jugements spontanés qui ne sont pas le résultat de suites conscientes de raisonnements "[44], l'ingéniosité ayant, quant à elle, pour fonction d'" aider l'intuition par des arrangements adéquats de propositions... "[45] ; un système logique " sera en général construit de façon à admettre une variété considérable de pas possibles à chaque étape d'une preuve. L'ingéniosité, alors, déterminera quels pas sont les mieux adaptés à la preuve d'une proposition particulière "[46].

L'ingéniosité permet en somme de traduire l'intuition en un processus effectif d'inférences. En elle-même, considérée indépendamment de l'intuition qui la dirige, l'ingéniosité est assimilable à un ensemble de règles ; dans le cadre d'un système fini de telles règles, on pouvait se demander si l'ingéniosité ne peut pas être remplacée par l'inventaire systématique des propositions démontrables dans ce système, si, en d'autres termes, et pour reprendre la formule de Turing, " l'ingéniosité peut être remplacée par la patience "[47] ? Traduit dans les termes de l'article sur les " nombres calculables ", le système Lω consistait en une suite de " machines de Turing " représentées chacune par un nombre ordinal. Pour représenter le passage d'une machine à une autre, Turing imaginait le recours à un " oracle " (" oracle ") : chaque machine devait disposer d'un état dans lequel elle s'en remettait à cet " oracle " pour connaître la configuration de la machine suivante. Si le système Lω avait été fini, l'oracle aurait pu être remplacé par l'examen systématique de toutes les propositions démontrables dans Lω. Turing établissait qu'au contraire, en vertu des " théorèmes d'incomplétude ", la " patience " ne permettrait jamais à la machine qui doit avoir

---

44 " The activity of intuition consists in making spontaneous judgments which are not the result of conscious trains of reasoning ", *ibid.,* p. 215.

45 The exercise of ingenuity in mathematics consists in aiding the intuition through suitable arrangements of propositions... ", *ibid.,* p. 215.

46 " In general a formal logic will be framed so as to admit a considerable variety of possible steps in any stage in a proof. Ingenuity will then determine which steps are the more profitable for the purpose of proving a particular proposition ", *ibid.,* p. 215.

47 " ... ingenuity may be replaced by patience ", *ibid.,* p. 215.

normalement recours à un oracle de trouver quelle doit être la configuration de la machine suivante ; il n'y a pas d'ingéniosité de la machine universelle[48].

Bref, si l'on admet que l'insistance de Turing sur le thème de la " pensée " des machines, dans les travaux officiels qu'il effectue pour le compte du *National Physical Laboratory*, ne laisse guère de doute quant au fait qu'il y a bien là, pour lui, un véritable enjeu théorique, le problème se pose de savoir quel est le sens, dans *Computing Machinery*..., de la question " Les machines peuvent-elles penser ? ". Quel est le lien théorique de la question " Les machines peuvent-elles penser ? " aux travaux antérieurs de Turing, au cours desquels est élaborée la notion de " machine de Turing " ? Telle est la difficulté d'interprétation immédiatement présentée par le texte publié dans *Mind* en 1950.

[48] Diverses solutions seront plus tard proposées (par Wang et Kleene notamment) pour mécaniser l'oracle. Cela ne signifie cependant pas que la machine soit " ingénieuse ", ou qu'elle " invente ", mais plutôt que les " inventions " peuvent en quelque sorte lui être initialement fournies. La définition la plus adéquate de la " machine à oracle " pourrait être celle de Davis, selon laquelle il s'agit d'une machine pouvant communiquer avec l'extérieur. Voir : Jean Mosconi, *La constitution de la théorie des automates, op. cit.,* p. 441.

Chapitre II : Les conditions intuitives du calcul pour un individu humain ; la plausibilité de l'équivalence entre procédure effective de calcul et procédé mécanique.

L'élément décisif du problème d'interprétation qui vient d'être posé est sans aucun doute l'idée d'une équivalence entre procédure effective de calcul et procédé mécanique, formalisée par la " thèse de Church-Turing ", ainsi nommée une fois démontré par Turing que sa définition du calculable par la notion de machine, et celle de Church par la théorie des fonctions récursives, mettent en jeu les mêmes classes de fonctions. L'idée d'une équivalence entre procédure de calcul et procédé mécanique clarifie celle de " procédure effective " ; la " thèse de Church-Turing ", selon laquelle les fonctions calculables par une procédure effective sont celles que calcule une machine de Turing, et donc sont des fonctions récursives générales, formalise cette équivalence et en propose une définition proprement mathématique. Cependant, elle ne la démontre pas, puisque sa relation à l'idée d'une équivalence entre procédure effective de calcul et procédé mécanique est celle d'une notion formelle à une notion intuitive[49]. L'équivalence entre procédure de calcul et procédé mécanique est formalisée, sans doute, par la description d'une machine de Turing, dont l'action consiste à calculer tout nombre calculable par un individu humain, mais elle ne saurait être conclue de cette description. L'équivalence entre procédure effective de calcul et procédé mécanique peut, au mieux, être considérée, pour reprendre un terme de Gödel, comme un principe

[49] Church écrit en 1936 : " This definition is thought to be justified... so far as positive justification can ever be obtained for the selection of a formal definition to correspond to an intuitive notion " (cité in Robin Gandy. " The Confluence of Ideas in 1936 ". *The Universal Turing Machine, a Half-Century Survey*, Rolf Herken, éd., Oxford, Oxford University Press, 1988, p. 71)

heuristique[50]. A ce titre, il ne peut être exigé à son égard davantage qu'une *plausibilité* intuitive.

Il en va ici de même que dans la théorie hilbertienne du signe mathématique : aux yeux de Hilbert, la perception du signe mathématique comme figure matérielle constitue, pour toute démarche mathématique, une condition intuitive irréductible ; or, cela ne peut, par définition, être démontré formellement. Toutefois, sous l'angle de la démarche mathématique, il importe seulement que l'on ait acquis l'assurance de ne pouvoir aller plus loin qu'Hilbert lui-même dans l'isolement du fondement intuitif du raisonnement mathématique. Ce fondement intuitif échappe nécessairement aux mathématiques proprement dites et renvoie d'abord aux conditions psychologiques dont l'effort du mathématicien est inséparable, à savoir la perception externe, la vision organisée du signe comme objet matériel.

Aussi bien Turing, dans *On Computable Numbers*..., a-t-il lui-même recours, pour justifier l'identification qu'il propose entre procédure effective de calcul et machine de Turing, à " un appel direct à l'intuition "[51], qui consiste à mettre au jour une équivalence intuitive entre la machine de Turing et ce que nous proposons d'appeler les conditions intuitives du calcul pour un individu humain, telles que ces conditions sont dégagées, précisément, par Hilbert. La plausibilité de l'équivalence entre procédure effective de calcul et procédé mécaniquee s'exprime, pour Turing, à travers le fait que la machine de Turing reproduit les conditions, chez l'homme, de la perception du signe mathématique comme figure matérielle.

Cependant, l'idée de procédure effective de calcul déborde le seul cadre des conditions de la perception externe du signe mathématique : il ne suffit pas que celui-ci puisse être perçu comme figure matérielle pour qu'un calcul ait lieu ; il faut encore qu'un *acte* soit effectué. C'est là ce que montrait, par exemple, Jean Cavaillès, à propos de la théorie hilbertienne du signe, au moment même où Turing publiait *On Computable Numbers…* : le signe mathématique perçu comme figure matérielle appartient à un espace déjà abstrait, déjà lui-même mathématique ; le signe mathématique est, en vérité, inséparable de l'acte de pensée qui le pose en tant que signe mathématique. Or, c'est au problème même qui découle de là, quant à la plausibilité de l'équivalence entre procédure de calcul et procédé mécanique, que

[50] Cité par Gandy, *ibid.,* p. 67 et 73.
[51] " A direct appeal to intuition ". A. M. Turing, *On Computable Numbers…, op. cit.,* p. 249.

Turing sera confronté, au cours de la discussion qu'il engage avec Wittgenstein, en 1939, sur ce que c'est qu' " appliquer une règle " : la plausibilité, ici, conduit à l'idée que la machine de Turing reproduit, non seulement les conditions de la perception externe du signe, mais *l'acte* humain de calcul.

La simulation par la machine des conditions intuitives du calcul

Une fois démontré qu'une machine de Turing peut simuler le calcul de certains nombres calculables par un homme, il reste à s'assurer qu'elle peut simuler le calcul de *tout* nombre calculable par un homme. Il est certes possible de présenter, comme, du reste, s'y emploie Turing, des " exemples de grandes classes de nombres qui sont calculables "[52] par une machine de Turing, mais on ne peut procéder ainsi pour l'ensemble infini des nombres calculables par un homme. La solution consiste à montrer que la machine calcule en faisant appel exactement aux mêmes éléments qu'un individu humain quelconque. La machine dispose, pour construire les séquences correspondant au calcul d'un nombre, des mêmes moyens que le calculateur humain ; le " calcul " effectué par une " machine de Turing " met en jeu, en particulier, un équivalent mécanique des conditions intuitives du calcul pour un individu humain, telles que ces conditions avaient été définies, quelques années auparavant, par Hilbert.

Celui-ci déclarait en effet que

> « la condition préalable à l'application des raisonnements logiques et à la mise en oeuvre des opérations logiques, c'est que quelque chose soit déjà donné à la représentation : à savoir certains objets concrets, extra-logiques, qui sont présents dans l'intuition en tant que données vécues immédiatement, préalablement à toute activité de pensée. Pour que le raisonnement logique soit doué de solidité, il faut que l'on puisse embrasser ces objets du regard de façon complète dans toutes leurs parties et que l'on puisse reconnaître par intuition immédiate, en même temps que ces objets eux-mêmes, comme des données qui ne se laissent plus réduire à quelque chose d'autre ou qui en tout cas n'ont pas besoin d'une telle réduction, comment ils se présentent, comment ils se distinguent les uns des autres, comment ils se suivent ou comment ils sont rangés les uns à côté des autres [...] En mathématiques en particulier, l'objet de notre examen ce sont les signes concrets eux-mêmes dont la forme nous apparaît immédiatement avec évidence,

[52] " ... examples of large classes of numbers which are computable. ". *Ibid.* p. 249.

> conformément à notre position fondamentale, et demeure parfaitement reconnaissable. »[53]

Construire une machine de Turing consiste, en somme, à préciser ce que sont les " signes concrets " dont parle Hilbert - ces " objets " qu'on doit pouvoir " embrasser du regard de façon complète dans toutes leurs parties " - et de quelle façon il est possible de reconnaître " comment ils se distinguent les uns des autres, comment ils se suivent ou comment ils sont rangés les uns à côté des autres ".

Un être humain qui calcule, remarque Turing, utilise du papier divisé en cases, par exemple un cahier quadrillé[54]. Il dessine sur les cases de la feuille de papier des symboles, lesquels doivent être en nombre fini : la quantité de papier dont dispose celui qui calcule étant toujours finie, s'il y avait un nombre infini de symboles, la différence matérielle entre deux d'entre eux ne pourrait être évaluée[55]. D'autre part, il est toujours possible d'obtenir un symbole qui s'avérerait manquant par une combinaison des symboles admis[56]. De la même façon, les limites de la perception humaine font que le nombre de cases de la feuille de papier que celui qui calcule peut observer simultanément est fini. Ces éléments définissent ce que Turing appelle, reprenant les notions utilisées par Hilbert, la " reconnaissabilité immédiate " (" immediate recognisability)[57].

Par ailleurs, explique Turing, le comportement d'un être humain qui calcule est déterminé à tout moment, non seulement par le ou les symboles qu'il observe, mais aussi par son " état d'esprit " (" state of mind ") à ce moment. Turing fournit une description concrète de ce qu'il entend, ici, par " état d'esprit ". D'une part, nous pouvons considérer que chaque

---

[53] David Hilbert. " Uber das Unendliche ", *Mathematischen Annalen*, 95, p. 170-171, trad. de Jean Ladrière. Cité par Jean Ladrière, *op. cit,.* p.3.

[54] Turing parle d'un " livre d'arithmétique pour enfant " (*a child's arithmetic book*). A. M. Turing, *On Computable Numbers…*, *op. cit.,* p. 249.

[55] " Si nous devions permettre une infinité de symboles, alors il y aurait des symboles qui différeraient selon une mesure arbitrairement petite.... " (*If we were to allow an infinity of symbols, then there would be symbols differing to an arbitrary small extent*), *ibid.,* p. 249.

[56] " Il est toujours possible d'utiliser des séquences de symboles à la place de symboles uniques. Ainsi un nombre arabe tel que 17 ou 999999999999999 est traité normalement comme un symbole unique. " (*It is always possible to use sequences of symbols in the place of single symbole. Thus an arabic numeral such as 17 or 999999999999999 is normally treated as a single symbol*), *ibid.,* p. 250.

[57] Hilbert évoque " l'intuition immédiate " et le fait pour les signes de devoir être " reconnaissables ". " Uber das Unendliche ", *op. cit.,* p. 171.

état d'esprit correspond à une configuration neurophysiologique de celui qui calcule, comme le montre le fait que les " états d'esprits " doivent être, comme les symboles, et pour les mêmes raisons qu'eux, en nombre fini : si nous admettions un nombre infini d'" états d'esprit ", " certains d'entre eux seraient ' arbitrairement proches ' et confondus "[58] ; en somme, l'élément matériel dont se sert le calculateur humain - son cerveau - étant fini, on ne pourra éviter d'introduire des états " arbitrairement proches " qu'en admettant un nombre fini d'états[59].

D'autre part, Turing explique que la notion d'" état d'esprit " peut être comprise par référence à la situation d'un homme en train d'effectuer un calcul, et contraint de s'arrêter alors que celui-ci n'est pas terminé. Avant d'abandonner son travail, cet homme rédige une note qui décrit le stade auquel il est parvenu et qui lui permettra de reprendre son travail là où il l'avait laissé ; cette " note d'instructions " (" note of instructions "), dit Turing, est " le pendant de l'état d'esprit " (" The counterpart of the ' state of mind ' "), et on comprend qu'elle détermine, comme l'" état d'esprit ", l'étape suivante du calcul.

Un moment quelconque de la démarche humaine de calcul peut donc toujours être défini comme une " configuration ", constituée des symboles observés à ce moment sur la partie du papier appréhendée par celui qui calcule - c'est-à-dire dans l'ensemble des cases observées simultanément - et de son " état d'esprit " à ce moment. A partir de cette analyse de ce qu'est un " signe concret " et de ce qu'il implique, il devient possible de caractériser la démarche de celui qui calcule comme une procédure effective, en définissant ce que Turing appelle les " opérations simples " (" simple operations "), c'est-à-dire les opérations auxquelles toutes les autres peuvent être ramenées, et en montrant que la démarche d'un être humain qui calcule peut être décrite comme une suite ordonnée de ces " opérations simples ". Pour qu'un homme puisse calculer, il doit pouvoir changer d'état d'esprit, de symbole lu, et de cases observées ; en ce sens, toute opération implique la possibilité, lorsqu'on change

[58] " We will also suppose that the number of states of mind which need to be taken into account is finite. The reasons for this are of the same character as those which restrict the number of symbols. If we admitted an infinity of states of mind, some of them will be ' arbitrarily close ' and will be confused ". A. M. Turing, *On Computable Numbers..., op. cit.,* p. 250.

[59] Là encore, du reste, il est possible de définir un état d'esprit plus complexe en combinant entre eux certains des symboles figurant sur le papier.

d'état d'esprit, de changer de symbole lu et la possibilité, lorsqu'on passe d'un ensemble de cases observées à un autre, de changer d'état d'esprit[60].

Imaginons maintenant que celui qui calcule le fasse de la manière la plus discontinue qui soit ; à chaque étape de son calcul, c'est-à-dire à chaque symbole lu, le calculateur humain s'arrête et rédige une " note d'instructions " qui lui indique où il en est et ce qu'il devra faire pour avancer d'une étape lorsqu'il reprendra son travail. Nous avons alors la description de la procédure effective décrite par la définition de la machine universelle.

Sous cet angle, il est possible de considérer qu'une machine de Turing simule les conditions intuitives et le comportement absolument nécessaires à tout calcul, et tels que ces conditions et ce comportement peuvent être définis pour un calculateur humain, à partir de la perception par celui-ci du signe mathématique comme figure matérielle. Aux yeux de Turing, tout se passe comme si une correspondance explicite pouvait être établie entre les éléments constituant la machine - un ruban, une tête de lecture-écriture, des symboles et des " m-configurations " - et ceux d'un second dispositif matériel composé d'un " cahier d'écolier quadrillé ", d'un crayon, d'une gomme, et d'un organisme comportant lui-même un appareil sensitif, un appareil moteur et des " états d'esprits ". En ce sens, la thèse mathématique de Turing avait pour corollaire une hypothèse, identifiant le calculable au mécanique, qui pouvait être regardée comme plausible car fondée sur une analyse du comportement observable d'un individu humain en train de calculer. Le caractère de plausibilité de l'identification d'une

---

[60] " Les opérations simples doivent par conséquent inclure :

a) des changements de symboles sur l'une des cases observées ;

b) des changements de l'une des cases observées pour une autre cases dans l'étendue de L cases pour l'une des cases observées précédemment.

Il est possible que certains de ces changements impliquent nécessairement un changement d'état d'esprit. L'opération déterminée la plus générale doit donc être considérée comme l'une des suivantes :

un changement possible (a) de symbole en même temps qu'un changement possible d'état d'esprit ;

un changment possible (b) de cases observées en même temps qu'un changement possible d'état d'esprit. "

(*The simple operations must therefore include :*

a) Changes of the symbol on one of the observed squares.

b) Changes of one of the squares observed to another square within L squares of one of the previously observed squares.

It may be that some of these changes necessarily involve a change of state of mind. The most general single operation must therefore be taken to be one of the following :

A possible change (a) of symbol together with a possible change of state of mind.

A possible change (b) of observed squares, together with a possible change of state of mind), *ibid.,* p. 251.

procédure effective de calcul à un procédé mécanique défini par une machine de Turing renvoyait à certaines des conditions psychologiques du calcul[61], celles qui, en particulier, ont trait à la perception du signe mathématique comme figure matérielle, et c'est sans doute la raison pour laquelle elle fut généralement bien accueillie.

Or, c'est pourtant au niveau de cette plausibilité de l'équivalence entre procédure effective et procédé mécanque que va se poser le problème qui conduira Turing à envisager la question de la " pensée " des machines. La description de la simulation, par la machine universelle, des conditions intuitives du calcul pour un individu humain, telle qu'elle est proposée par Turing dans *On Computable Numbers*…, c'est-à-dire sous la forme d'une simulation de la perception, au sens psychologique du terme, du signe comme figure matérielle, reste en quelque sorte trop extérieure au calcul pour permettre une formulation intuitivement pleinement satisfaisante de l'idée d'une équivalence entre procédure de calcul et procédé mécanique. De la seule perception d'une figure matérielle quelconque ne saurait découler, en effet, un calcul effectif ; il faut encore que cette figure matérielle soit bien interprétée comme un signe *mathématique*. " L'appel direct à l'intuition " fait, ici, par Turing, dès lors qu'il prend la forme d'une mise en correspondance des éléments constitutifs de la machine et des éléments non formels qui entrent dans le calcul d'un nombre par un individu humain, ne suffit pas à garantir l'idée d'équivalence entre procédure effective de calcul et procédé mécanique, non parce que l'on ne peut être certain que l'énumération des éléments considérés soit bien complète, mais parce que n'y est pas pris en compte ce qui est, en définitive, au principe de l'effectivité du calcul effectué par un homme, à savoir le *geste* humain par lequel les éléments non formels et la procédure formelle de calcul sont unifiés dans la production d'un nombre.

Comme le montrait, en effet, Jean Cavaillès dès sa thèse de 1938 – *Méthode axiomatique et formalisme*[62] - au moment même, autrement dit, où Turing publie *On Computable Numbers…,* l'interprétation de la théorie hilbertienne du signe comme figure matérielle sous l'angle de la seule perception externe du signe est réductrice : pour Hilbert, le

[61] " Turing et Post avec lui ont vu [...] qu'il y avait place, entre la psychologie du calculateur et la simple considération des traces du calcul, pour une analyse de l'acte de calcul qui avait une portée mathématique " (Jean Mosconi, *La constitution de la théorie des automates*, *op. cit.,* p. 36).

[62] Jean Cavaillès, *Oeuvres complètes de philosophie des sciences*, Paris, Hermann, 1994.

signe, comme figure matérielle, est *inséparable de l'acte qui le pose en tant que signe mathématique*.

Cavaillès et la théorie hilbertienne du signe

Quelques années après la publication de *Méthode axiomatique et formalisme*, Cavaillès devait reprendre et approfondir son analyse de la théorie hilbertienne du signe dans le texte, rédigé en prison et publié après sa mort, qu'il consacra à l'idée de " théorie de la science "[63]. En vérité, plutôt que d'élaborer une théorie de la science en tant que telle, il s'agissait, pour Cavaillès, d'établir que la connaissance scientifique implique par elle-même une telle théorie et de dégager, à partir de là, les principes qui lui donnent son contenu.

Cavaillès rappelait ce que, s'agissant des mathématiques, Kant avait montré, à savoir que le problème philosophique posé par les mathématiques ne consiste pas seulement à rendre compte de leur nécessité, mais aussi de leur développement, de leur devenir, de leur *progrès* : la connaissance mathématique peut être étendue ; elle n'est pas analytique, mais synthétique. Aux yeux de Cavaillès, toutefois, le criticisme kantien souffre d'un défaut dont hériteront plus tard l'intuitionnisme de Brouwer ou l'épistémologie brunschvicgienne[64] : tous trois rendent compte de la construction mathématique, du *progrès* mathématique, à partir d'un principe externe, conscience originaire, ou auto-intelligibilité absolue[65], au sein duquel le mouvement mathématique comme tel disparaît. Comment rendre compte de la nécessité propre aux mathématiques du sein de la construction mathématique elle-même, sans perdre, en les rapportant à un absolu " caractérisable par ailleurs "[66], la spécificité de cette construction, c'est-à-dire le *progrès* qui marque la connaissance mathématique ? Cavaillès, en examinant cette question, entendait poursuivre l'effort de Bolzano, pour qui, disait-il, " il s'agit à la fois de déterminer ce qui constitue une science comme telle et le moteur de son développement "[67].

Chez Bolzano, remarquait Cavaillès, la nécessité scientifique se confond avec le mouvement même de la science. Celle-ci est ainsi saisie dans son autonomie, sans pour autant

[63] Jean Cavaillès, *Sur la logique et la théorie de la science*, in *Oeuvres complètes de philosophie des sciences*, *op. cit.* Ce texte, écrit en prison en 1942, peut être considéré comme le testament philosophique de Cavaillès.
[64] C'est Brunschvicg qui avait poussé Cavaillès à s'intéresser à l'histoire des mathématiques et qui avait dirigé sa grande thèse de 1938.
[65] l'idée de l'idée spinoziste chez Brunschvicg, par exemple.
[66] Jean Cavaillès, *Sur la logique…*, *op. cit.*, p. 498.
[67] *Ibid.*, p. 503.

être considérée comme un absolu, car elle n'est pas débarrassée de la finitude : la science doit être appréhendée dans son mouvement concret, comme science réalisée et non comme idéal. L'incomplétude fait partie de sa définition même[68]. Dès lors, notait Cavaillès, " il faut... que ses énoncés [NB : de la science] ne soient pas constitutifs d'un développement particulier mais apparaissent immédiatement dans une auto-illumination du mouvement scientifique, se distinguant de lui pourtant par leur permanente émergence "[69]. L'activité scientifique elle-même contribue, en d'autres termes, à mettre au jour une " structure de la science qui n'est que manifestation à elle-même de ce qu'elle est "[70]. La " doctrine de la science " est un moment de la science elle-même.

Or, cette structure ne peut être que celle de la *démonstration*[71], et tout le problème philosophique posé par la connaissance scientifique consiste à saisir la doctrine de la science au sein de la démonstration scientifique elle-même, c'est-à-dire à appréhender le progrès scientifique à travers le processus démonstratif. Les sciences en général, et les mathématiques en particulier, effacent leur propre histoire dans l'apodictique ; il s'agit de retrouver cette histoire dans l'apodictique. Comment cela est-il possible ?

Une démonstration, remarquait Cavaillès, est toujours unique et singulière : elle vaut par ce qu'elle démontre ; elle n'est pas " renouvelable dans son intégrité "[72], c'est-à-dire qu'elle ne saurait être séparée de son résultat singulier. Cependant, une démonstration est " enchaînement ", et comme telle, toujours liée à d'autres démonstrations, de sorte que les démonstrations " présentent par groupes une parenté de type, marque de l'unité de mouvement et manifestable dans l'abstrait "[73]. Cette double dimension de la démonstration donne lieu, chez Cavaillès, au développement de deux notions : celle de " paradigme ", et celle de " thématisation ", empruntée à Husserl. Le " paradigme " exprime la parenté de type repérable entre démonstrations singulières ; il renvoie à la dimension *d'enchaînement* de la

[68] " Une théorie de la science ne peut être que théorie de l'unité de la science. Cette unité est mouvement : comme il ne s'agit pas ici d'un idéal scientifique mais de la science réalisée, l'incomplétude et l'exigence de progrès font partie de la définition ". *Ibid.*, p. 504.
[69] *Ibid.*, p. 506.
[70] *Ibid.*
[71] " il n'est qu'une façon de s'imposer par une autorité qui n'emprunte rien au dehors, il n'est qu'un mode d'affirmation inconditionnelle, la démonstration. La structure de la science non seulement est démonstration, mais se confond avec la démonstration ". *Ibid.*
[72] *Ibid.*, p. 509.
[73] *Ibid.* " Ainsi apparaît, ajoute Cavaillès, la notion de forme logique indispensable à une théorie de la démonstration. "

démonstration. Le paradigme rend compte ainsi de la *variable* comme structure essentielle de la démonstration : la notion de variable met en évidence le fait que l'enchaînement, le lien avec d'autres démonstrations, est une propriété interne de toute démonstration singulière[74].

A travers le " paradigme " se dégage donc une structure. Un niveau supplémentaire d'abstraction consiste à raisonner sur cette structure, à en énoncer les propriétés et les règles. C'est là qu'intervient la notion husserlienne de " thématisation " : " La pensée ne va plus vers le terme créé, mais part de la façon de créer pour en donner le principe par une abstraction de même nature que l'autre, mais dirigée transversalement "[75]. La thématisation repose sur la dualité qui apparaît en particulier dans le mouvement de la démonstration mathématique, " prototype ", chez Cavaillès, comme le signale Benis-Sinaceur, " de l'activité rationnelle essentiellement progressive "[76] ; elle consiste à transformer le " sens posant " en " sens posé " :

> Dans le moment qui dégage le sens apparaît une dualité entre sens posant et sens posé, entre la signification d'une opération en tant qu'elle est opérée - abstraction faite du problème dont l'exigence a créé sa singularité active, et dépouillée par la substitution possible de l'accidentel de son départ - et sa signification en tant qu'opérante...[77].

Or, les formes " transversales " dégagées par la thématisation appartiennent elles-mêmes au mouvement de la démonstration, elles sont elles-mêmes des démonstrations singulières, enchaînements d'où se dégagent de nouveaux paradigmes, objets à leur tour d'une thématisation. " Paradigmes " et " thématisation ", corollaires l'un de l'autre, décrivent ainsi l'effort de formalisation qui caractérise la science, et, au premier chef, les mathématiques. Ils conduisent aux théories du formalisme qui voient le jour au cours du 19e siècle, à travers les travaux de Boole et de Frege, puis à travers la définition de la notion de système formel, et qui aboutissent à la distinction des deux niveaux de formel que recouvre l'idée de métalangue : " Tandis que précédemment la syntaxe se trouvait définie en langage courant

---

[74] " Ce qui importe ici est le décrochage opéré à chaque suppression de singularité : c'est ce qui dans le calcul logique est représenté par la règle de substitution, savoir la possibilité de remplacer dans le nouvel élément celui dont il procède effectivement par un quelque autre, équivalent à lui du nouveau point de vue atteint ". *Ibid.*, p. 511.

[75] *Ibid.*, p. 512.

[76] H. Benis-Sinaceur, " Thématisation ", *Encyclopédie philosophique universelle*, Paris, PUF, 1990-1991

[77] Jean Cavaillès, *Sur la logique...*, *op. cit.*, p. 513.

sans appel à un type déterminé de démonstration pour ses lois déduites, l'approfondissement même de son étude exige sa formalisation. Mais il faudra, pour définir le système formel S1 de la syntaxe S, définir la syntaxe de S1, ce qui ne s'accomplira qu'au sein d'un système S2, etc. "[78]

Dès lors, le problème posé par l'idée de théorie de la science est celui du terme - ou du support, de la " base ", dit Cavaillès - de ce mouvement d'approfondissement que rien ne vient arrêter. Or, il apparaît que le mouvement mathématique possède, à travers l'arithmétique, la propriété de pouvoir donner une forme à *tous* les moments de l'approfondissement : les règles qui gouvernent une démonstration arithmétique peuvent être formalisées dans le langage de l'arithmétique ; les paradigmes révélés par un ensemble de démonstrations arithmétiques donnent lieu à une thématisation dont le moyen d'expression est encore l'arithmétique[79]. Est-ce à dire que l'on pourrait voir dans l'arithmétique une sorte d'absolu mathématique, que l'arithmétique devrait être considérée comme le " système initial ", la " base " à partir de laquelle est engendré, à travers paradigmes et thématisations, le mouvement de la démonstration, le progrès mathématique ?

En vérité, le système " base ", conçu comme mathématique absolue, consiste en une " position simultanée de tous les possibles "[80] qui n'est pas compatible avec le mouvement même de la démonstration tel qu'il a été analysé par Cavaillès. Entendu comme absolu du mathématique, le " système initial " correspond au dépassement des systèmes formels singuliers et successifs ; or, saisir ce dépassement comme un absolu est en contradiction avec son sens même : il faut poser un système formel pour pouvoir en extraire la syntaxe ; le paradigme précède toujours la thématisation. Il y a, dans la propriété qu'a l'arithmétique de formaliser sa propre syntaxe, " une sorte de retour sur soi de la pensée formelle qu'il était impossible de prévoir avant son accomplissement et qui ne prend qu'en lui sa véritable portée "[81]. L'idée d'un primat de la " base ", à titre de " système initial " ou de fin du mathématique, doit être rejetée, puisque le sens véritable de cette " base " s'avère être le

[78] *Ibid.,* p. 516.
[79] " c'est un résultat obtenu dans l'acte effectif du formalisme que tout système contenant l'arithmétique peut formaliser sa propre syntaxe. L'acte secondaire pour lequel le sens posant de l'acte primaire est sens posé coïncide alors avec lui ". *Ibid.,* p. 517. Le théorème de Gödel exploite cette dimension de l'arithmétique.
[80] *Ibid.*
[81] *Ibid.*

mouvement même de dépassement. “ Que tout ne soit pas d’un seul coup... est la caractéristique de l’intelligible... ”[82].

En outre, c’est au point de vue syntaxique que le mouvement mathématique présente la propriété de pouvoir se retourner sur lui-même. Cependant, la syntaxe ne suffit pas à le définir : comme Tarski l’a montré, tout formalisme comporte aussi une dimension sémantique[83], de sorte que la question à laquelle doit répondre une doctrine de la science, et qui la fonde en tant qu’elle doit décrire le mouvement de la démonstration de l’intérieur même de ce mouvement, est celle de l’incomplétude de la description du “ sens posant ” en “ sens posé ”. Le problème philosophique soulevé par la connaissance scientifique tient à ce que la description de la forme dégagée par le paradigme est à la fois nécessaire - elle constitue le moteur même du procès de démonstration - et nécessairement incomplète. Telle est la caractéristique singulière dont il s’agit de rendre compte.

Or, note Cavaillès, l’interprétation logiciste du formalisme mathématique y échoue ; le logicisme interprète l’idée hilbertienne selon laquelle une démonstration mathématique doit être considérée comme une suite de figures dessinées sur un support matériel, “ en confondant les actes primaires avec leur représentation sensible, objets figurables, dont sinon la totalité, au moins les modes de construction paraissent délimitables exhaustivement ”[84]. C’est oublier que les “ objets figurables ”, ici, sont des *signes*, et que c’est en ce sens que l’entendait Hilbert. Cavaillès ne manquait pas de rappeler, dans *Méthode Axiomatique et Formalisme*, la phrase célèbre de celui-ci : “ Au commencement est le signe ”[85]. Certes, le signe mathématique, figure matérielle définie comme les pièces du jeu d’échec par un ensemble de règles, ne représente rien d’autre que lui-même - il ne “ renvoie pas à autre chose dont il serait le représentant ”[86] - mais il “ renvoie pourtant aux actes qui l’utilisent ”[87]. En vérité, aux yeux

[82] *Ibid.*, p. 518.
[83] “ D’autre part, la définition formelle d’un système n’est pas complète avec l’énoncé de la syntaxe. C’est encore le mérite de Tarski d’avoir constitué dans son originalité la sémantique à côté de la syntaxe proprement dite. Il s’agit en effet non seulement de donner la forme des énoncés pourvus de sens (règles de structure) et les modes de passage d’un groupe de propositions à un autre (règles de consécution), mais de définir les objets mêmes, éléments et composés intervenant avec leurs propriétés dans les enchaînements : variables, fonctions, individus, démonstrations, définitions, bref d’introduire les concepts du système ”. *Ibid.*, p. 520.
[84]
[85] “ am Anfang, so heisst es hier, ist das Zeichen ” (Hilbert IX, p. 163) ”. Jean Cavaillès, *Méthode axiomatique et formalisme*, in *Œuvres complètes..., op. cit.* p. Mm102.
[86] *Ibid.*
[87] *Ibid.*

de Cavaillès, dans la théorie hilbertienne, la condition nécessaire de tout raisonnement mathématique n'est pas tant le signe comme " objet du monde ", que sa perception, à travers laquelle il est appréhendé directement comme rapport, comme système de relation, c'est-à-dire comme abstraction. Ce qui est perçu sous la forme du signe mathématique, au sens hilbertien du terme, est un espace *déjà abstrait*[88]. En somme, le point décisif, ici, n'est pas tant qu'il y ait un donné à représenter que le fait même que l'acte de représentation soit possible. Le symbole mathématique - " chiffre, figure, même bâton " - présente " ce caractère fondamental [...] de n'être là qu'en tant que partie intégrante ou base d'application d'une activité déjà mathématique : *le symbole est intérieur à l'acte*, il n'en peut être ni le départ, ni le véritable aboutissement (qui est engendrement d'autres actes) "[89]. Ainsi, pour Cavaillès, le signe, en tant qu'" objet figurable ", n'est pas séparable de l'acte par lequel il est posé ; bien plus, c'est d'abord cet acte même qu'il *représente*.

L'idée " d'acte " mise en jeu, ici, renvoyait à une problématique de l'intentionalité que Cavaillès entendait détacher de la réflexion husserlienne, dont il soulignait certaines des difficultés, notamment à propos de la notion de " logique transcendantale ". Au regard, cependant, de la démarche de Turing, ce qui importe, à notre sens, dans l'idée d'acte mise en jeu par l'analyse du signe mathématique, en deçà même de l'orientation spécifique de la réflexion menée par Cavaillès, est le fait que l'acte y apparaisse comme porteur d'une nécessité propre, *irréductible tout en même temps à l'effectivité formelle et aux conditions psychologiques de la perception du signe*. De là ressort, en effet, qu'il ne suffit plus, pour que l'on puisse admettre l'identification d'une procédure effective de calcul à un procédé mécanique qu'existe, pour une machine, un équivalent des conditions de la perception externe d'une figure matérielle chez l'homme, ni même que cet équivalent soit défini de telle sorte que la figure matérielle " perçue " par la machine soit un signe mathématique, appartenant comme tel à un espace mathématique existant indépendamment de l'action de la machine, il faut encore que cet espace soit posé et défini comme espace mathématique par cette action même, c'est-à-dire par cette " perception " que l'on attribue à la machine.

---

[88] " l'essence même de la mathématique est jeu réglé de symboles, ceux-ci n'étant pas un adjuvant pour la mémoire, mais définissant une sorte d'espace abstrait avec autant de dimensions qu'il y a de degrés de liberté dans l'opération concrète et imprévisible de la combinaison ". *Ibid.*, p. 101.

[89] Jean Cavaillès, *Sur la logique…*, *op. cit.*, p. 521. C'est nous qui soulignons.

Or, s'agissant de la prise de conscience, par Turing, d'un tel problème, la discussion qui l'oppose à Wittgenstein, au cours des leçons sur la question des fondements des mathématiques données par celui-ci en 1939, joue, nous semble-t-il, un rôle déterminant. Wittgenstein, en effet, d'une part, rappelle à Turing, au cours de cette discussion, qu'une procédure effective de calcul implique, outre la perception du signe, la mise en jeu d'un système de règles, et lui montre, d'autre part, que le formalisme récursif ne suffit pas à rendre compte de ce qu'est *l'application d'une règle*, laquelle renvoie à un geste irréductible, sorte de décision collective déterminant un usage.

La rencontre de Turing et Wittgenstein. Qu'est-ce qu'appliquer une règle ?

En 1939, de retour à Cambridge après deux années passées aux Etats-Unis, où il a rédigé *Systems of Logic based on ordinals*, Turing participe au séminaire, comme nous dirions aujourd'hui, tenu par Wittgenstein sur la question des fondements des mathématiques. Wittgenstein y traite le sujet sous un angle philosophique, et, à sa manière habituelle, sous la forme d'une discussion, parfois rude, avec ses auditeurs, et, tout particulièrement, parmi ceux-ci, avec Turing. Son argumentation ne laisse rien subsister de l'idée de nécessité mathématique telle que celle-ci pouvait être entendue par un mathématicien, et quelle qu'en soit l'interprétation admise : Wittgenstein rejette la conception " platonicienne " de la nécessité mathématique, selon laquelle celle-ci relèverait d'un monde autonome des objets mathématiques, mais il refuse tout autant la conception formaliste selon laquelle la nécessité mathématique pourrait être représentée sous la forme d'un mouvement mécanique. C'est pourquoi, si Wittgenstein, au cours de ses " Leçons ", fait de Turing son interlocuteur privilégié, ce n'est pas seulement que ce dernier se trouve être le seul mathématicien en exercice de l'assistance, et qu'il lui revient, par là, d'exprimer et de défendre ce que l'on pourrait appeler le " sens commun mathématique ", c'est-à-dire l'idée même d'une nécessité mathématique, c'est aussi que Turing est le mathématicien qui vient de donner à la métaphore mécaniste son expression théorique avec la notion de machine universelle.

On sait que Wittgenstein, dans son *Cours sur les fondements des mathématiques*, s'interrogeait sur ce que veut dire appliquer une règle. Nous pouvons admettre, dit-il, que, pour appliquer une règle, nous devions l'avoir présente à l'esprit, dans son intégralité. Toutefois, cela ne suffit pas : " Supposez, demande-t-il, que nous ayons, vous et moi, la même

page de règles à l'esprit, cela garantirait-il que nous les appliquions, vous et moi, de la même manière ? "[90]. En outre, rien ne garantit non plus que tel individu ayant à l'esprit, à deux moments différents, la même " page de règles ", l'applique chaque fois identiquement. L'un des exemples utilisés par Wittgenstein vise directement l'idée de " procédure effective ". Supposons que l'on écrive au tableau les nombres :

1, 4, 9, 16

et que l'on montre à quelqu'un en situation d'élève qu'il s'agit d'une suite correspondant à la formule " $y = x^2$ ", où x représente la suite des nombres entiers. A l'aide de cette règle, l'élève pourra développer la suite. Toutefois, la règle garantit-elle qu'il continuera toujours correctement ? Rien ne permet d'affirmer qu'elle sera appliquée, à chaque étape, de la même façon ; n'importe quelle raison pourra conduire l'élève à ne pas répéter à telle étape ce qu'il a fait précédemment. Comment pouvons-nous, par exemple, préjuger de ce qu'il fera après 100, dans le cas où nous lui aurons montré des exemples allant seulement jusqu'à 100 ? L'acte de " répéter " n'est pas, en l'occurrence, parfaitement déterminé, puisque chaque étape est, sous un certain angle, différente de la précédente : " on pourrait mettre en évidence ce que je veux montrer, explique Wittgenstein, en disant : '99 est de toutes façons différent de 100 ; alors comment pouvons-nous dire que ce que nous faisons avec 99 est la même chose que ce que nous faisons avec 100 ?' "[91]. En d'autres termes, rien dans la règle elle-même, telle qu'elle est énoncée, n'est susceptible de nous contraindre à la suivre : l'application de la règle n'est pas mécanique. C'est pourquoi Wittgenstein pourra déclarer : " Je m'élève contre l'idée de 'machinerie logique'. Je tiens à dire qu'il n'existe rien de tel "[92].

L'idée de machinerie logique présuppose, en effet, qu'il y aurait quelque chose *derrière* nos symboles. Wittgenstein prend l'exemple du réveil : supposons que nous fassions tourner, devant un groupe d'élèves, l'aiguille des minutes de trois tours. Les élèves peuvent constater que l'aiguille des heures a tourné d'un quart de tour. Il leur est possible, à partir de là, d'anticiper la situation suivante et de prévoir ce que fera l'aiguille des heures lorsque nous ferons tourner l'aiguille des minutes. Toutefois, il se peut aussi qu'ils ne comprennent pas la situation et ne sachent pas anticiper ; ou encore, qu'ils comprennent ce que nous voulons leur

---

90 Ludwig Wittgenstein, *Cours sur les fondements des mathématiques, op. cit.,* p. 11.

91 *Ibid.,* p. 15.

92 *Ibid.,* p. 199.

montrer, à savoir qu'il existe un certain rapport mathématique entre le mouvement des deux aiguilles, mais qu'ils refusent de voir là une explication du mouvement des aiguilles. Bien entendu, dans ce cas, nous pouvons démonter le réveil et leur montrer le mécanisme ; sans doute se convaincront-ils, alors, d'une part, qu'il y a une certaine nécessité dans le mouvement des aiguilles, d'autre part, que cette nécessité à son principe dans le mécanisme. Pourtant, nous n'aurons fait ainsi que déplacer le problème : en montrant le mécanisme, qu'aurons-nous montré d'autre que lorsque nous montrions le mouvement des aiguilles ? Or, si les élèves se sont montrés sceptiques à l'égard de la nécessité qui gouverne celui-ci, ne devraient-ils pas l'être autant devant le mécanisme ? Ce qui leur a été montré, dans les deux cas, c'est une certaine régularité, mais la seconde n'est pas d'une autre nature que la première[93].

Nous le savons, si l'identification de l'effectivité d'une procédure logique à une machine, par le biais de la machine de Turing, a été facilement admise, c'est qu'on a vu dans la machine de Turing le mécanisme gouvernant l'enchaînement des formules au cours d'un calcul. Pourtant, si l'on suit Wittgenstein, il n'y a rien de plus, du point de vue de l'effectivité, dans le mouvement de la machine que dans la suite des symboles arithmétiques eux-mêmes. L'identification proposée par Turing entre la machine et le calcul est, certes, parfaitement légitime, en ce sens que la machine de Turing et le calcul tel qu'un individu humain le mène sur un cahier quadrillé représentent bien la même chose, mais ce qui, en définitive, rend cette identification légitime, c'est le fait, précisément, que la machine constitue elle-même un symbolisme pour le calcul. Comme la suite des symboles arithmétiques, la machine donne une *image* de la contrainte spécifique attachée à la notion d'effectivité. Bref, si le principe de cette contrainte spécifique doit être cherché ailleurs que dans l'énoncé mathématique en tant que tel, il doit également être cherché ailleurs que dans la machine en tant qu'elle est une machine ; ce qui, dans la machine, rend compte de l'effectivité ne peut pas être le mécanique.

Au cours des mêmes discussions, Wittgenstein s'efforçait de mettre ses interlocuteurs, et, parmi eux, Turing, devant l'idée que la contrainte propre à une règle mathématique relève, en vérité, non d'une nécessité que la règle porterait en elle-même, mais d'une sorte de " décision " que nous prenons collectivement de la considérer comme devant être suivie. Une

[93] Voir à ce sujet le problème de la " preuve " in François Schmitz, *Wittgenstein, la philosophie et les mathématiques,* Paris, PUF, 1988.

règle mathématique est d'abord ce que Wittgenstein appelle une " relation interne ", c'est-à-dire la perception d'une certaine configuration de termes au sein de laquelle chacun de ceux-ci est saisi non en lui-même, mais dans son rapport aux autres. Cette relation interne devient règle parce que nous adoptons à son égard un usage spécifique, par lequel nous la traitons en paradigme. C'est parce que l'emploi d'une règle consiste, en quelque sorte par convention, à reproduire une certaine relation interne à chaque étape d'une procédure, que celle-ci est effective. A travers la règle, l'effectivité de la procédure mathématique est celle d'un usage, acquis par un apprentissage, usage qui est à lui-même son propre principe, puisque l'on ne saurait, pour saisir la nécessité qui devrait le fonder, quitter le territoire qu'il délimite.

Wittgenstein mettait ainsi à jour la fragilité théorique, sur un certain plan, de la notion d'effectivité telle qu'elle est définie à partir de l'idée de machine. Bien plus, la critique wittgensteinienne de la catégorie de nécessité faisait ressortir une dimension particulière de la notion de machine universelle : la simulation par celle-ci des conditions intuitives du calcul pour un individu humain implique que la machine, dès lors qu'elle calcule effectivement tout nombre calculable par un homme, simule, au-delà des conditions en quelque sorte externes de l'énoncé mathématique – les conditions constitutives de la perception externe du signe comme figure matérielle, et l'ensemble des règles régissant la construction de l'énoncé – *l'application* des règles, c'est-à-dire *l'acte* humain de calcul, l'espèce de " décision " impliquée, selon Wittgenstein, par l'idée même de règle. C'est en effet à travers la dynamique de cet acte que les éléments avec lesquels les organes de la machine universelle sont mis en correspondance – " papier quadrillé ", crayon, gomme, appareil sensitif, appareil moteur, " états d'esprits " - deviennent les conditions intuitives du calcul pour un individu humain.

La notion de machine universelle et le problème de la " pensée " des machines

Dès lors que l'élaboration par Turing de la notion de machine fait appel à la description hilbertienne des conditions intuitives ultimes de toute démarche mathématique, le statut particulier, dans cette description, du signe mathématique, défini comme acte par Cavaillès, et mis en évidence par Wittgenstein à travers la discussion de l'idée de règle, ne peut manquer de retentir sur la plausibilité intuitive admise par Turing de l'équivalence entre procédure effective de calcul et procédé mécanique. On soulignera tout particulièrement, à cet égard, que la simulation, décrite dans *On Computable Numbers*…, des conditions intuitives du calcul

pour un individu humain par une machine universelle, n'est pas strictement équivalente à la simulation d'une machine de Turing par une autre machine de Turing, c'est-à-dire par une machine universelle. Lorsqu'une machine de Turing N simule une machine de Turing M, la relation d'identité entre M et la partie de N qui simule celle-ci, repose sur le fait que N reproduit la configuration de M. En revanche, dans le cas de la simulation des conditions intuitives du calcul pour un individu humain par une machine universelle U, la configuration de U est celle de la machine qui représente les conditions intuitives du calcul, c'est-à-dire qu'il n'y a pas dans ce cas de configuration manipulée par U. Le rapport de U à ce qu'elle simule n'est pas d'ordre logique. Les conditions intuitives du calcul pour un individu humain constituent un point d'arrêt dans l'imbrication des machines simulant d'autres machines ; elles renvoient à l'homme non seulement en tant qu'il calcule, mais, bien plus, en tant qu'il est le *constructeur* des machines. De sorte que ce sont *les conditions de sa construction même* que la machine universelle, à travers les conditions intuitives du calcul pour un individu humain, doit simuler.

Aussi bien ne suffit-il pas, pour pouvoir considérer comme intuitivement satisfaisante l'équivalence entre une procédure humaine de calcul et un procédé mécanique tel que la machine universelle, de démontrer, d'une part, que la machine peut formellement " calculer " certains nombres calculables par un individu humain, et d'établir, d'autre part, une correspondance purement descriptive entre les éléments constitutifs de la dite machine et les éléments décrivant la perception externe du signe par un individu humain qui calcule ; il faut encore s'assurer que la machine effectue bien, lorsqu'elle " calcule ", le *geste* par lequel, chez l'homme, la procédure formelle de calcul et les éléments non formels de la perception du signe sur lesquels cette procédure s'appuie sont unifiés dans la production d'un nombre. En d'autres termes, il faut encore montrer que le signe mathématique est bien, lorsqu'il est considéré sous l'angle de la machine de Turing, le même que dans le cadre du calcul humain, c'est-à-dire que la machine simule en effet ce qui, au-delà de la procédure de calcul proprement dite, rend celle-ci possible pour l'homme : non seulement ce qui a trait à la perception externe du signe matériel, au sens psychologique du terme, mais l'acte complet dans lequel, chez l'homme, la perception en tant que telle s'inscrit.

C'est ce problème de l'extension du champ de la plausibilité intuitive de l'équivalence entre procédure de calcul et procédé mécanique que Turing, dans *Computing Machinery…*,

pose sous la forme de la question " Les machines peuvent-elles penser ? ". De la discussion avec Wittgenstein ressort que la mise en correspondance descriptive des conditions intuitives du calcul pour un individu humain avec les éléments constitutifs d'une machine universelle ne suffit pas à rendre compte de la possibilité même d'une machine capable de calculer tout nombre calculable par un homme. " Quelque chose " en l'homme, au-delà de l'énumération des éléments constituant la perception externe du signe, rend le calcul possible, et c'est ce " quelque chose " que doit simuler la machine universelle. Il est naturel, enfin, de nommer ce " quelque chose " la " pensée ". En effet, du strict point de vue de la question de la plausibilité de l'équivalence entre procédure de calcul et procédé mécanique, telle qu'elle se pose à Turing, ce qui est exigé est une donnée d'intuition, attestant de ce que la machine universelle simule ce que l'on pourrait appeler la dimension ultra-logique de l'acte de calcul ; or, intuitivement, la dimension ultra -logique du calcul renvoie en l'homme au " penser " : aucun homme qui calcule ne doute de la pensée en lui -même, ni de la pensée chez les autres hommes qui calculent. Intuitivement, pour l'homme, l'homme calcule parce qu'il " pense ".

On sait, par ailleurs, que, dans le cas de l'usage mathématique de la notion de machine, l'idée d'intuition qui prévaut est celle qui organise la perception externe du signe mathématique comme figure matérielle. Il découle de là que la question posée par le renvoi du calcul à la pensée en l'homme est celle de l'attestation de la pensée, de la trace de la pensée, de la manifestation de la pensée à travers ce qui constitue la perception externe, bref, la question posée est celle du *signe* de la pensée. A cet égard, l'affirmation de l'équivalence de la machine universelle avec l'acte humain de calcul doit, aux yeux de Turing, pouvoir s'appuyer sur une observation, sur la reconnaissance d'un fait, bref, elle doit être corroborée, non pas tant par une théorie - en l'occurrence philosophique - que par une *expérience*. Une expérience doit pouvoir être proposée, qui établira comme une donnée d'observation que la machine universelle simule bien les conditions intuitives du calcul pour un individu humain, c'est-à-dire l'acte à travers lequel les éléments constitutifs du calcul, tels qu'ils sont décrits par Hilbert, deviennent effectivement des conditions intuitives. Cette expérience sera le test de Turing, le fameux jeu de l'imitation.

Or, on voit immédiatement que le statut de l'expérience, ici, est problématique. L'intuition recherchée est constituée de la perception et de l'interprétation de certains signes comme renvoyant à la pensée - ces signes seront rapportés à la certitude qu'a tout homme qui

calcule de sa propre pensée et de la pensée chez tout homme qui calcule. Pour qui, cependant, ou pour quoi y aura -t -il intuition, sinon pour le " quelque chose " qui en l'homme perçoit et interprète, c'est-à-dire pour cela même qui est appelé ici " pensée " ? Le référent du signe perçu et interprété sera la " pensée ", autrement dit ce pour quoi il peut y avoir signe. En d'autres termes, il s'agit, dans le test proposé par Turing, de reconnaître la pensée, laquelle, en toute rigueur, ne peut être reconnue que par elle-même. Bref, la question du " quelque chose " qui agit renvoie, ici, à l'idée d'*intelligibilité* : ce qui, au-delà de l'appréhension par les sens – la perception externe – fonde le comprendre par cela qu'il serait transparent à soi-même. De sorte, enfin, que " l'expérience " envisagée par Turing apparaît déterminée elle-même par une question philosophique – celle de la reconnaissance de la pensée – dont on peut dire qu'elle traverse toute l'histoire de la philosophie occidentale, au moins depuis Descartes.

On peut douter, dès lors, qu'une " expérience " suffise, par elle -même, à fournir un appui pleinement satisfaisant à l'affirmation de la plausibilité de l'équivalence entre procédure de calcul et procédé mécanique, tout au moins dans les termes où la question est abordée par Turing, à savoir sous la forme de l'idée d'une " pensée " des machines. Plus que l'expérience en tant que telle, considérée dans sa dimension factuelle, ce qui importe ici semble être, en effet, l'interprétation qui lui donne sens, comme si la réflexion de Turing devait être confrontée aux positions prises historiquement par la question de la reconnaissance de la pensée, et, éventuellement, ramenée à l'une ou l'autre de ces positions. A moins que cette interprétation soit celle prévalant dans le contexte de la philosophie analytique, et qui consisterait à considérer la question de la reconnaissance de la pensée comme une pseudo-question, renvoyant au mieux à une tautologie, et face à laquelle nous en serions réduits en quelque sorte au " grognement ", pour reprendre une allusion de Wittgenstein. Cela même pourrait sans doute être regardé comme conforme à la position de Turing, dans la mesure où celui-ci entend limiter sa réflexion aux conditions de plausibilité de la démarche mathématique : Turing, en somme, s'efforcerait, par l'expérience qu'il propose, de formuler le " grognement " appelé par l'usage mathématique de la notion de machine…

Dans ce cas, cependant, " l'expérience " envisagée est-elle nécessaire ? Plutôt que par une expérience, le problème examiné par Turing ne devrait-il pas être résolu par une analyse visant à clarifier l'usage des termes en présence - ceux de " machine " et de " pensée " ? Or, nous le verrons, c'est précisément parce qu'il refuse une telle démarche que Turing propose

une expérience : celle-ci, sous la forme du jeu de l'imitation, doit se substituer, dira-t-il, à l'analyse des termes de la question " les machines peuvent-elles penser ? ". Bien plus, nous le verrons également, le contenu même de l'expérience du jeu de l'imitation, met directement en œuvre, dans la description qu'en donne Turing, des catégories relevant de la réflexion sur la question de la reconnaissance de la pensée telle que celle-ci a été historiquement élaborée au travers de l'histoire de la philosophie occidentale. En un mot, l'expérience imaginée par Turing – le test de Turing - a bien pour objet une question qui traverse toute l'histoire de la philosophie occidentale, et la portée philosophique de *Computing Machinery…* se joue sur la capacité de cette expérience à traiter philosophiquement, *en tant qu'elle est une expérience* et non une discussion philosophique, une telle question.

Deuxième partie : « Computing Machinery and Intelligence »

Une fois dégagé le sens de la question " Les machines peuvent-elles penser ? ", nous pouvons envisager l'examen de la démarche adoptée par Turing pour y répondre. Turing traduit la question " les machines peuvent-elles penser ? " dans les termes du jeu de l'imitation. Ce dernier consiste, dans son principe, à vérifier qu'une machine universelle peut tromper, sur son statut de machine, un individu humain quelconque, en se faisant passer, aux yeux de cet individu, *pour un autre individu humain*. A cet effet, la machine doit, dans les conditions du jeu tel qu'il est imaginé par Turing, simuler le comportement linguistique, ou langagier, d'un individu humain : la machine et ses adversaires humains du jeu de l'imitation doivent " converser ". Turing s'efforce de démontrer que rien, dans la définition de la notion de machine qu'il a lui-même donnée dans *On Computable Numbers...* n'invalide l'hypothèse qu'une machine puisse satisfaire à un tel test, c'est-à-dire qu'une machine puisse l'emporter au jeu de l'imitation face à des adversaires humains.

En quoi, cependant, le jeu de l'imitation, dans sa structure spécifique, peut-il devenir une " confirmation expérimentale " de l'hypothèse de la " pensée " des machines ? Certes, on comprend aisément que Turing imagine une " expérience " mettant en jeu la compétence linguistique, puisque la question à laquelle l'hypothèse de la " pensée " des machines est censée apporter une réponse est celle de l'acte humain de calcul, c'est-à-dire celle des propriétés fondatrices de la " langue " manipulée par la machine universelle. On comprend également, à ce même titre, pourquoi Turing propose un test au cours duquel la machine doive simuler ce qui confère à l'être humain, aux yeux mêmes de celui-ci, son statut d'être humain : l'hypothèse est que la machine, simulant les conditions intuitives du calcul pour un individu

humain, simule, en vérité, le principe, en l'homme, de sa propre construction ; la machine doit se montrer l'égale de son constructeur. Toutefois, en quoi est-il décisif que le test consiste en un *affrontement* entre une machine et des hommes ? C'est qu'en vérité, l'hypothèse de la " pensée " des machines, n'est pas neutre : elle se heurte à l'opinion commune des hommes, pour laquelle une machine ne peut pas, en quelque sorte par définition, être " pensante ". Autrement dit, l'hypothèse ne peut être établie et " vérifiée " que *contre* l'opinion commune. Sa confirmation ne peut être apportée que par l'infirmation de l'opinion commune, laquelle sera acquise si une machine universelle l'emporte face à des hommes, porteurs de l'opinion commune, sur le terrain même que l'opinion commune des hommes refuse à la machine. Cette structure profonde de l'argument de Turing détermine, nous le verrons, sa portée spécifiquement philosophique.

La démonstration par laquelle Turing entend établir qu'une machine universelle peut l'emporter au jeu de l'imitation et qu'une telle performance atteste la " pensée " de la machine met en œuvre deux idées essentielles. En premier lieu, la victoire de la machine au jeu est rendue possible, selon Turing, par le fait que son principal adversaire humain est contraint, au cours du jeu, de faire *comme si* son interlocuteur mécanique était un autre individu humain. L'action de la machine théoriquement autorisée par la définition même de la machine universelle peut être telle que, dans le cadre du test, l'examinateur humain ne dispose d'aucun moyen de distinguer son comportement de celui d'un autre individu humain.

En second lieu, l'action de la machine ainsi envisagée ne saurait être décrite comme une procédure purement formelle. Le jeu de l'imitation est conçu de telle manière qu'un succès de la machine à celui-ci implique qu'elle simule, non pas tel ou tel moment du comportement humain susceptible d'être formalisé par ses programmeurs, mais le comportement humain en tant, précisément, qu'il ne peut être réduit à l'un de ces moments, c'est-à-dire en tant qu'il surpasse, par hypothèse, tout effort de formalisation de la part des constructeurs humains de la machine. C'est pourquoi l'hypothèse de la victoire possible d'une machine au jeu de l'imitation s'appuie elle-même, dans la réflexion de Turing, sur une seconde hypothèse, celle des " machines qui apprennent ". Turing s'efforce d'établir, là encore, que rien, dans la définition de la notion de machine comme machine universelle, n'invalide l'hypothèse qu'une machine puisse être soumise à un processus d'éducation analogue, quant à son sens, à celui qui préside au devenir adulte d'un petit d'homme.

L'hypothèse des " machines qui apprennent " situe la machine théorique envisagée par Turing dans un autre registre que celui de l'idée de machine qui a servi de modèle aux fondateurs de l'intelligence artificielle, et qui a présidé au développement de celle-ci dans sa version classique : la machine susceptible de l'emporter au jeu de l'imitation, machine capable d'" apprendre ", n'est pas définie par un programme, qui, dans son principe, conserve son sens indépendamment de son exécution, comme c'est le cas de l'ordinateur classique, mais par son devenir, par les " expériences " qu'elle traverse, bref, par son histoire propre, que l'on pourrait dire *individuelle*.

Nous étudierons tout d'abord la méthode élaborée par Turing pour répondre à la question " Les machines peuvent-elles penser ? ", avant de suivre la mise en place successive des deux hypothèses qui structurent son argumentation : celle de la victoire possible d'une machine au jeu de l'imitation, et celle des " machines qui apprennent ". Nous discuterons ensuite certaines des critiques auxquelles a donné lieu la démarche de Turing, en nous efforçant de dégager, à partir de là, la solidarité qui lie l'une à l'autre ces deux hypothèses.

Chapitre I : Les deux hypothèses de Turing

Section I : La méthode : le jeu de l'imitation et l'infirmation de l'opinion commune

Turing se demande immédiatement, dans *Computing Machinery*…, si l'examen de la question " Les machines peuvent-elles penser ? " ne devrait (should) pas commencer par la définition des termes " machine " et " penser ", définitions qui pourraient (might), en l'occurrence, être formées de telle sorte qu'elles " reflètent autant que possible l'utilisation normale des mots... "[94]. On notera qu'en abordant le problème sous cet angle, en 1950, dans *Mind*, principale revue britannique de philosophie, Turing s'inscrit délibérément dans le débat philosophique propre, alors, à la philosophie anglo-saxonne : il s'adresse manifestement, en premier lieu, à Wittgenstein, mais aussi, sans doute, aux philosophes qui, à la suite d'Austin et de Ryle[95], et en s'inscrivant dans le champ ouvert par Wittgenstein, élaborent ce que l'on appelle déjà la philosophie du " langage ordinaire ". Tout se passe comme si Turing entendait poursuivre la discussion avec Wittgenstein commencée en 1939, et choisissait d'entamer sa réflexion en se plaçant dans le cadre de la méthode imposée alors par celui-ci à ses interlocuteurs : la question " les machines peuvent-elles penser ? " ne met pas d'abord en jeu des notions constituées dans le cadre d'un appareil technique spécifique, mais des termes de langage qu'il s'agit d'analyser dans leur usage " normal ", dans leur usage " ordinaire ", dans la mesure où cet usage est une condition de tous les autres.

Or, en l'occurrence, cette méthode, affirme Turing, est " dangereuse ", et ne peut être suivie naïvement :

> Si on doit trouver la signification des mots " machine " et " penser " en examinant comment ils sont communément utilisés, il est difficile d'échapper à la conclusion que la

94 " The definitions might be framed so as to reflect so far as possible the normal use of the words ". A. M. Turing, *Computing Machinery..., op. cit.,* p. 133.

95 *The Concept of Mind*, de Ryle, paraît en 1949 ; le premier texte important d'Austin, *The Other Minds*, est publié en 1946.

> signification de la question ‘ Les machines pensent-elles ? ’ et la réponse à cette question doivent être recherchées dans une étude statistique telle que le sondage d'opinion. Mais cela est absurde[96].

Turing ne prend pas la peine de s’expliquer davantage sur ce qu’il qualifie ici d’“ absurde ”, renvoyant à ce qu’il estime, sans doute, relever de l’évidence : une enquête statistique fournirait, certes, ce que l’on pourrait tenir pour la réponse “ commune ”, fondée sur l’usage “ ordinaire ” des termes “ machine ” et “ penser ”, à la question “ les machines peuvent-elles penser ? ”, mais, outre qu’en lui-même, le fait qu’une réponse soit “ commune ” ne constitue rien de décisif, la fin visée dans *Computing Machinery...* n’est pas de découvrir quelle est la réponse du “ sens commun ” à la question de la “ pensée ” des machines. Il ne s’agit pas, pour Turing, de mener une étude lexicographique ou sociolinguistique qui, plutôt qu’à la question “ Les machines peuvent-elles penser ? ”, fournirait une réponse à la question “ Quelle réponse apportent la majorité des gens interrogés à la question ‘Les machines peuvent-elles penser ?’ ”.

Est-il possible, cependant, s’agissant du terme “ penser ”, de ne pas recourir à “ l’utilisation normale des mots ”, dès lors que l’on décide de faire abstraction de toute définition du “ penser ” renvoyant à l’armature théorique d’un système philosophique préétabli ? Certes, une fois écarté l’examen théorique, en préalable à l’analyse de la question de la “ pensée ” des machines, des rapports de la notion générale de “ pensée ” aux notions connexes de “ conscience ”, “ raison ”, “ entendement ”, “ imagination ”, “ perception ”, “ sensation ”, “ volonté ”, “ liberté ”… reste ce que nous avons déjà mentionné comme le “ fait ” du “ penser ”, à savoir la certitude intuitive, indépendante de toute connaissance et de toute justification - par cela même qu’elle est la condition de toute connaissance et de toute justification - que tout homme a de sa propre pensée et de la pensée en tout homme. Cependant, cette certitude, précisément parce qu’elle précède toute définition, quel que soit le statut de celle-ci, précisément parce qu’elle constitue le fait premier de tout usage du terme “ penser ”, ne s’exprime-t-elle pas d’abord à travers “ l’usage ordinaire ”, à travers “ l’utilisation normale ” du mot “ penser ” ?

[96] “ If the meaning of the words ‘machine’ and ‘think ‘ are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question ‘ Can machines think ? ’ is to be sought in a statistical survey such as a Gallup poll ”. A. M. Turing, *Computing Machinery..., op. cit.,* p. 133.

En outre, si " l'usage ordinaire " du terme " penser " ne peut être évité, est-il théoriquement " absurde " que la question " Les machines peuvent-elles penser ? " puisse être ramenée à la question " Quelle réponse apporte la majorité des gens interrogés à la question de la 'pensée' des machines ? ". La réponse " commune " à la question " Les machines peuvent-elles penser ? " ne pourrait-elle être considérée comme la seule réponse légitime à cette question ?

En vérité, le danger que souligne Turing, d'une méthode s'appuyant sur " l'utilisation normale des mots ", ne tient pas tant à " l'utilisation normale " du mot " penser " qu'à celle du mot " machine ". Celui-ci renvoie, pour Turing, à la notion de machine universelle, et, par là, interdit que la deuxième question - " Quelle réponse apporte la majorité des gens interrogés à la question de la 'pensée' des machines ? " - puisse être substituée à la première - " Les machines peuvent-elles penser ? ". Pour Turing, la signification de la question " Les machines peuvent-elles penser ? ", ne peut pas être, du fait, ici, du terme " machine ", celle que déterminerait l'usage " ordinaire " des termes " machine " et " penser ". De sorte qu'en ce qui concerne la question de la " pensée " des machines, ce n'est pas le *statut* du " sens commun ", ou de l'opinion commune, mais bien le *contenu* de la réponse donnée par ceux-ci, que refuse Turing.

L'opinion commune et la question de la " pensée " des machines

Dans un texte de 1948, lui-même intitulé *Intelligent Machinery*, Turing faisait, en effet, remarquer que " l'opinion commune ", à propos d'une question telle que celle de savoir si une machine peut montrer un comportement intelligent, " suppose habituellement sans argument que cela n'est pas possible "[97]. L'attitude commune, souligne Turing, met en jeu des éléments d'ordre " émotionnel " (" emotional ") qu'il n'y a pas lieu de discuter comme tels. Il en est ainsi, par exemple, de la " ... réticence à admettre la possibilité que l'humanité puisse avoir quelque rival que ce soit quant à son pouvoir intellectuel " [98], ou encore de " la croyance religieuse que toute tentative de construire de telles machines [NB : des machines pensantes]

[97] " Common catch phrases such as ' acting like a machine ', ' purely mechanical behaviour ' reveal this common attitude ". A. M. Turing, *Intelligent Machinery, Collected Works of A.M. Turing, Mechanical Intelligence, op. cit.*, p. 107.
[98] " An unwillingness to admit the possibility that mankind can have any rivals in intellectual power ", *ibid.*, p. 107.

est une sorte d'irrévérence prométhéenne "[99]. Reste que, pour le sens commun, " penser " est quelque chose qui appartient à un autre ordre que ce qu'une machine est susceptible de faire[100] ; au regard de l'usage " ordinaire " du terme " machine ", le comportement d'une machine est entièrement défini par des règles explicites et c'est précisément à cause de cela qu'il ne saurait être identifié au comportement intelligent d'un être pensant :

> 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 (la moindre visée sous-jacente), nous sommes peu tentés d'imaginer de l'intelligence. S'agissant du même objet, par conséquent, il est possible qu'un homme le considère comme intelligent et un autre non ; le second aura découvert les règles de son fonctionnement.[101].

Pour le sens commun, la pensée, conformément au fait premier de " l'usage ordinaire " du terme " penser ", est d'abord de l'ordre du mystère, elle outrepasse la règle et implique, par là, une forme d'imprévisible – celle de *l'invention*. La pensée impose sa présence au-delà de toute explication, elle est d'abord ce qui ne se laisse enfermer dans aucune définition positive, ce qui transcende toute détermination, et se situe, par suite, selon " l'usage ordinaire " du

[99] " A religious belief that any attempt to construct such machines is a sort of Promothean irreverence ", *ibid.*

[100] *Ibid.*

[101] " The extent to which we regard something as behaving in an intelligent manner is determined as much by our own state of mind and training as by the properties of the object under consideration. If we are able to explain and predict its behaviour or if there seems to be little underlying plan, we have little temptation to imagine intelligence. With the same object, therefore, it is possible that one man would have consider it as intelligent and another would not ; the second man would have found out the rules of its behaviour ", *ibid.,* p. 127. On rapprochera ce texte du récit célèbre d'Edgar Poe à propos du " joueur d'échecs de Maelzel ", cet automate, construit par le baron hongrois Kempelen en 1769 et repris par Maelzel à la fin du 18e siècle, qui se montrait capable de jouer aux échecs et de gagner contre la plupart de ceux qui acceptaient de se mesurer à lui. Poe remarque qu'" aucun coup, dans le jeu d'échecs ne résulte nécessairement d'un autre coup quelconque ", et que, par conséquent, les opérations du joueur d'échecs de Maelzel ne peuvent être assimilées à celles d'une machine à calculer, c'est-à-dire à celles d'un automate : " il est tout à fait certain que les opérations de l'automate [de Maelzel] sont réglées par l'esprit et non par autre chose ". D'où l'hypothèse toute simple, qui s'avérera exacte, que la prétendue " machine " cache un (excellent) joueur humain. Une machine ne peut effectuer d'actions impliquant des choix indéterminés, dès lors, une structure qui effectue de tels choix ne peut être une machine. La seule hypothèse acceptable est qu'il s'agit d'un être disposant d'un " esprit ", d'une pensée, c'est-à-dire, très probablement, d'un homme... (Edgar Poe, " Le joueur d'échecs de Maelzel ", *Histoires grotesques et sérieuses, Œuvres imaginatives et poétiques complètes*, Paris, éd. Vialetey, 1966).

terme “ machine ”, à l’opposé de la “ machine ” ; il y a une étrangeté radicale de la machine vis à vis de l’homme, étrangeté définie à partir du fait que l’homme est considéré *a priori* comme “ pensant ”. En un mot, sur le plan de l’usage “ ordinaire ” des termes “ machine ” et “ penser ”, la question “ les machines peuvent-elles penser ? ” ne se pose pas. De sorte que la réponse apportée par la majorité des gens interrogés à la question de la “ pensée ” des machines consiste, non seulement à dire que la machine ne peut pas penser, mais que la question même “ Les machines peuvent-elles penser ? ” n’a pas de sens. Or, nous l’avons vu, c’est précisément à partir de la notion de machine, dans sa définition scientifique comme “ machine universelle ”, que Turing est, quant à lui, amené à la poser. La question de la “ pensée ” des machines a, pour Turing, un sens objectif, inscrit, en définitive, non dans l’usage “ ordinaire ” des termes “ machine ” et “ penser ”, mais dans l’usage mathématique du terme “ machine ”. L’usage scientifique du terme “ machine ” implique une relation au terme “ penser ”, y compris dans ce que recèle “ l’usage ordinaire ” de celui-ci. La notion logique de machine ne s’oppose pas à la “ pensée ” comme à quelque chose qui appartiendrait à un ordre radicalement différent ; l’usage logique du terme “ machine ” conduit Turing, non pas à affirmer l’opposition de la machine à la “ pensée ”, mais à postuler une relation de la machine à ce que nous avons appelé une dimension ultra-logique de la pensée. De sorte que répondre à la question “ Les machines peuvent-elles penser ? ” suppose, non que l’on fasse abstraction du “ sens commun ”, mais que l’on mette celui-ci en question. Dans *Computing Machinery...*, il s’agit avant tout, pour Turing, de montrer que “ l’opinion commune ”, à propos de la question de la “ pensée ” des machines, se trompe, lorsqu’elle refuse que la notion de machine, considérée comme ce qui relève de la précision logique et de la prévision déterministe, puisse être mise sur le même plan que celle de pensée. L’opinion commune ne sera donc pas mise entre parenthèses, Turing ne la laissera pas à la porte de son laboratoire : *en tant qu’elle peut être contredite*, elle constituera un terme de la démonstration.

Puisque, pour cette “ opinion commune ”, un comportement à l’égard duquel des règles peuvent être découvertes, qui permettent d’en prévoir l’évolution, n’est pas intelligent, l’une des conditions - celle relevant de l’usage ordinaire du terme “ penser ” - pour qu’un comportement soit considéré comme manifestant une “ pensée ”, sera que ce comportement *surprenne* des individus humains. C’est pourquoi Turing propose, en guise de réponse à la question “ Les machines peuvent-elles penser ? ”, un *test* dont le principe consistera à vérifier

que la machine peut *surprendre* un individu humain, dont il est admis qu'il pense, en dehors de toute définition du " penser ". Il s'agit, en somme de mener une expérience, au sens épistémologique du terme : le test de Turing entre dans la constitution d'un syllogisme prévisionnel visant à vérifier une hypothèse. A partir de la définition scientifique de la machine comme machine universelle, Turing formule l'hypothèse que la machine, en tant que machine universelle, c'est-à-dire en tant qu'elle est capable de simuler toute machine construite par un homme, et en tant qu'elle est elle-même construite par un homme, peut *penser* au sens humain du terme : nous l'avons vu, la machine universelle simule, en dernière instance, le principe même, en l'homme, de sa propre construction. Puis il imagine un cas permettant de vérifier l'hypothèse : une machine universelle qui simulerait le comportement d'un homme de telle manière qu'elle surprenne une série d'individus humains - pour qui la machine, selon l'usage ordinaire des termes, est nécessairement autre chose qu'un homme en tant que celui-ci est " pensant " - en se faisant passer aux yeux de ces individus humains, *pour un autre individu humain*. De sorte que la démarche de Turing peut être exprimée selon le schéma suivant : toute entité qui, en se faisant passer pour un individu humain considéré comme pensant, *trompe* une série d'examinateurs humains, est " pensante " au même titre que tout individu humain ; une machine peut tromper une série d'individus humains en se faisant passer pour un individu humain ; donc, la machine est " pensante " au sens humain du terme. En ce sens, l'expérience envisagée par Turing confirmera l'hypothèse de la " pensée " des machines en *infirmant* l'assertion commune selon laquelle les machines ne peuvent pas " penser ". La " vérification " de la justesse de l'hypothèse de la " pensée " des machines ne fera qu'un avec celle de la fausseté de l'opinion commune. L'infirmation de l'opinion commune constituera le ressort même de l'argumentation ; elle ne sera ni un préalable à l'affirmation de la " pensée " des machines, ni une conséquence de cette affirmation, mais sa forme première et immédiate. Si l'on montre qu'une machine peut surprendre sur son statut des individus humains en se faisant passer pour un individu humain, non seulement la conclusion de l'opinion commune devra être considérée comme fausse, mais elle devra l'être à partir de ses prémisses mêmes, et c'est à partir de celles-ci que la " pensée " de la machine sera conclue. Bref, il aura été établi que la machine " pense " à partir de l'usage ordinaire du terme " penser ", dans ce que cet usage a, nous l'avons vu, d'inévitable parce que considéré,

dans la démarche de Turing, comme premier. Nous verrons quel rôle fondamental joue, quant à la portée de *Computing Machinery*…, cette structure spécifique du raisonnement de Turing.

Le jeu de l'imitation

C'est dans *Intelligent Machinery*, qui précède de quelques mois *Computing Machinery*…, que Turing expose le principe de ce qui deviendra le jeu de l'imitation. Il imagine, alors, un test impliquant trois joueurs d'échec, A, B, C, de force sensiblement égale, celle, en l'occurrence très limitée, que l'on était susceptible d'atteindre avec un programme d'ordinateur en 1947. A et C étaient des joueurs humains, B une machine. Turing estimait qu'il serait très difficile à C, s'il jouait à la fois contre A et B, et s'il n'était pas plus expérimenté qu'eux, de dire contre lequel, de A - le joueur humain - ou de B - la machine - il est en train de jouer.

D'*Intelligent Machinery* à *Computing Machinery*…, le principe demeure – l'hypothèse d'une confusion possible, de la part d'un examinateur humain, entre un homme et une machine – mais la situation dans laquelle l'homme et la machine sont opposés change de nature : ce n'est plus la seule compétence particulière des participants à un jeu tel que les échecs qui est sollicitée - compétence dont il est intuitivement admis qu'elle peut être dans une mesure significative définie à l'aide de méthodes formelles - mais l'ensemble, au contraire mal défini, des compétences que recouvre la notion de " comportement humain ". Le saut qualitatif qui est ainsi accompli par Turing est au cœur du problème d'interprétation soulevé par le jeu de l'imitation.

Celui-ci oppose un joueur A à deux partenaires B et C. A est un homme, B, une femme, C, enfin, un examinateur, homme ou femme. Chacun d'eux se trouve dans une pièce isolée des deux autres. C, en posant des questions à A et B, doit déterminer qui est l'homme, qui est la femme. L'homme A doit s'efforcer de tromper C, en se faisant prendre pour la femme B ; B doit aider C. Le jeu est conçu de telle sorte qu'il permette de " tracer une ligne assez nette entre les capacités physiques et intellectuelles d'un homme "[102] : rien de ce qui concerne les premières ne doit jouer un rôle dans l'expérience[103] ; les trois protagonistes ne

[102] " The new problem has the advantage of drawing a fairly sharp line between the physical and intellectual capacities of a man ". A. M. Turing, *Computing Machinery..., op. cit.,* p. 134.

[103] " Nous ne souhaitons pas pénaliser la machine pour son incapacité à briller dans des concours de beauté, ni pénaliser l'homme parce qu'il perd dans une course contre un

communiquent pas directement : ils ne se voient pas et ne s'adressent les uns aux autres que par l'intermédiaire d'un téléscripteur ; ils ne peuvent donc pas 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 lors du test ; A, B et C sont en quelque sorte réduits à l'expression désincarnée de leur pensée.

Il est, sans doute, permis de considérer que sur une série significative de sessions du jeu, les trois protagonistes - les deux hommes et la femme – seront de " force " sensiblement égale, puisqu'ils sont tous trois des êtres humains, et, dès lors, que A aura environ trente pour cent de chances de l'emporter. Par ailleurs, la meilleure stratégie pour la joueuse B sera " probablement de donner des réponses vraies ", puisqu'aussi bien son véritable atout réside dans le fait qu'elle est femme, c'est-à-dire qu'elle n'a pas, contrairement à l'homme A, à imiter la féminité pour exprimer celle-ci.

Supposons que A soit remplacé par une machine ; " qu'arrive-t-il, demande Turing, si une machine prend la place de A dans le jeu ? L'interrogateur se trompera-t-il aussi souvent que lorsque le jeu se déroule entre un homme et une femme ? Ces questions remplacent la question originale : " Les machines peuvent-elles penser ? "[104] Si l'interrogateur se trompe en effet aussi souvent, si, autrement dit, la machine l'emporte dans environ trente pour cent des cas[105], nous devrons constater qu'elle se sera comportée comme un homme placé dans les conditions du jeu - réduit à l'expression de sa pensée – c'est-à-dire comme un homme dont il est admis qu'il pense.

---

avion. " (*We do not wish to penalise the machine for its inability to shine in beauty competitions, nor to penalise a man for losing in a race against an aeroplane*), *ibid.,* p. 135.

[104] " What will happen when a machine takes the part of A in the game ? Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman ? These questions replace our original, 'Can machines think ?' ", *ibid.*, p. 134.

[105] Turing estimait, au moment où il écrivait, que, dans un délai de cinquante ans, il serait possible de construire une machine face à laquelle l'examinateur A, aidé de B, n'aurait que soixante-dix pour cent de chances de l'emporter au cours d'un test de cinq minutes (" Je crois que dans environ cinquante ans, il sera possible de programmer des ordinateurs ayant une capacité de stockage d'environ 10^9, pour les faire si bien jouer au jeu de l'imitation qu'un interrogateur moyen n'aura pas plus de soixante dix pour cents de chances de faire la bonne identification après cinq minutes d'interrogation " - *I believe that in about fifty years' time it will be possible to programme computers, with a storage capacity of about 10^9, to make them play the imitation game so well that an average interrogator will not have more than 70 per cent chance of making the right identification after five minutes of questioning*, *ibid.,* p. 136.

La machine considérée ici sera, précise Turing, un " ordinateur digital ", c'est-à-dire une machine universelle[106], laquelle, en tant que machine manipulant des symboles, peut être programmée pour répondre à des " questions ". Turing savait, toutefois, mieux que quiconque, que les toutes premières " machines universelles " alors disponibles étaient incapables, non seulement de réussir le test qu'il imaginait, mais, bien plus, d'y participer. Leur " mémoire ", en particulier, était très insuffisante, et le problème technique consistant à mettre au point une mémoire leur permettant d'accomplir, de manière satisfaisante, les tâches, encore strictement de calcul, qu'on leur demandait, était précisément celui auquel les ingénieurs en charge de leur construction consacraient l'essentiel de leurs efforts. C'est pourquoi Turing précise : " nous ne nous demandons pas si tous les ordinateurs digitaux feraient bonne figure dans le jeu, ni si les ordinateurs actuellement disponibles feraient bonne figure, mais s'il existe des ordinateurs imaginables qui feraient bonne figure "[107]. Ce qui importe, en définitive, c'est que la machine qui participe au jeu soit bien une machine universelle, et, en particulier, eu égard au problème initial, que, comme telle, elle puisse " effectuer n'importe quelle opération qui pourrait être faite par un calculateur humain "[108].

Une description complète de " l'expérience " imaginée par Turing exige, cependant, que soit mise au jour la structure profonde sur laquelle s'appuie la procédure proposée. Il semble bien, en effet, que, pour l'examinateur C, une fois l'homme A remplacé par une machine, le but du jeu soit toujours de dire qui, de A ou de B, est la femme[109]. La machine devra donc viser, quant à elle, à se faire prendre, par l'examinateur C, non pour un homme –

[106] Le fait que A ne doive pas être un individu humain exclut non seulement que sa place soit tenue par un être humain " né de la manière habituelle " (... *born in the usual manner*), mais également par un être humain conçu artificiellement, par un procédé génétique : " ... il est certainement possible de créer un individu complet à partir d'une seule cellule (disons) de la peau d'un homme." (... *it is probably possible to rear a complete individual from a single cell of the skin (say) of a man*), *ibid.,* p. 136.

[107] " ... we are not asking wether all digital computers would do well in the game nor wether the computers at present available would do well, but wether there are imaginable computers which would do well ", *ibid.,* p. 136.

[108] " ... these machines are intended to carry out any operations which could be done by a human computer ", *ibid.,* p. 136.

[109] 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, N° 4, 1994 ; J. Genova, " Turing's Sexual Guessing Game ", *Social Epistemology*, vol. 8, N° 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.

un représentant mâle de l'espèce humaine - mais pour une femme… Dès lors, il est permis de supposer que la meilleure stratégie pour la machine consistera, non pas à imiter le comportement d'un homme cherchant à se faire passer pour une femme, mais à tenter d'imiter directement le comportement d'une femme. Nous pouvons imaginer que si la machine se montre insuffisamment " adroite " pour tromper l'examinateur C sur sa prétendue " féminité ", celui-ci sera amené à conclure le jeu en énonçant quelque chose comme : " X (la machine) ne peut pas être une femme, donc X est un homme ". Toutefois, si la machine n'imite pas assez bien la femme pour l'emporter, l'examinateur croira-t-il pour autant qu'elle est un homme ? Pour qu'il en soit ainsi, la machine devra, tout en étant insuffisamment adroite pour imiter une femme, imiter un homme de manière convaincante. Aussi bien est-ce le " comportement humain " que la machine doit avant tout pouvoir imiter : " on supposera, dit Turing, que la meilleure stratégie est d'essayer de fournir des réponses qui seraient naturellement données par un homme ", le terme " homme " signifiant, ici, d'après le contexte, " l'être humain "[110]. En d'autres termes, la condition initiale de la victoire de la machine est que l'examinateur C ne puisse pas prendre celle-ci pour une machine… La machine n'a de chance de l'emporter que si elle est d'abord prise pour un être humain, homme ou femme. De sorte que la première signification du fait qu'un examinateur se trompe aussi souvent, lorsque la place de A dans le jeu est occupée par une machine, que lorsqu'elle est occupée par un homme, sera que la machine a été confondue par lui avec un *être humain*. Qu'elle l'emporte dans environs trente pour cent des cas attestera que l'examinateur a implicitement accordé à la machine " l'humanité ". " L'humanité " est, en somme, le point sur lequel la machine doit en premier lieu tromper l'examinateur. Qu'elle y parvienne ne signifiera sans doute pas qu'elle l'aura emporté au test en tant que tel, c'est-à-dire qu'elle aura réussi à tromper l'examinateur sur sa prétendue " féminité ". Elle aura néanmoins passé avec succès un *premier* test, consistant pour elle à se faire prendre pour un être humain. Ainsi, à partir du moment où la machine est introduite dans le jeu, tout se passe comme si l'examinateur était amené à se prononcer sur la question de savoir qui est la machine, et qui est l'homme (l'être humain).

[110] " answers that would naturally be given by a man ", *Ibid.*, p. 135. Ainsi, P. Blanchard traduit-il *a man* par " l'homme ", plutôt que par " un homme ", Jean-Yves Girard, *La machine de Turing*, *op. cit.*, p. 138.

C’est, nous semble-t-il, ce que tend à prouver la nouvelle formulation de la question proposée par Turing :

> Fixons notre attention sur un ordinateur particulier C. Est-il vrai que, en modifiant cet ordinateur pour avoir une capacité de mémoire adéquate, en accroissant de manière satisfaisante sa vitesse de travail, et en lui fournissant un programme approprié, on peut faire jouer à C le rôle de A dans le jeu de l’imitation, le rôle de B étant tenu par un homme ?[111]

Comment, en effet, pouvons-nous interpréter, ici, l’expression : “ le rôle de B étant tenu par un homme (a man) ” ? Supposons que le terme “ man ” signifie l’homme entendu comme le mâle de l’espèce humaine. Dans ce cas, que devra faire cet homme au cours du jeu ? Son rôle n’est pas de tromper l’examinateur C, mais de l’aider. Il ne s’agit donc pas pour lui de se faire passer pour une femme, mais de “ donner des réponses vraies ”, c’est-à-dire d’exprimer, de toutes les manières possibles, le fait qu’il est bien un homme. Que devra faire, quant à elle, la machine A ? Devra-t-elle se faire prendre pour un homme ou pour une femme ? Si l’on respectait strictement la lettre du jeu, c’est pour un homme qu’elle aurait à se faire prendre, puisque le rôle de A, dont elle occupe la place, consiste, dans le jeu, à tromper C en se faisant passer pour B. Nous sommes, de toutes façons, ramenés, ici, au problème examiné plus haut, à savoir qu’il s’agit d’abord pour la machine de se faire passer pour un être humain. En d’autres termes, Turing semble considérer qu’il est indifférent que B soit un homme ou une femme, et qu’il est, de même, indifférent, que la machine se fasse passer pour un homme ou pour une femme ; elle doit d’abord être confondue avec un être humain. Bref, tout se passe comme si le terme “ man ” devait être interprété, dans la nouvelle formulation de la question, comme signifiant “ être humain ”.

C’est, du reste, nous le verrons, sous cet angle que la question est discutée par Turing : les arguments qu’il examine sont ceux au regard desquels une machine ne peut pas imiter le comportement humain en général, et non ceux au regard desquels un homme, ou une machine cherchant à imiter son comportement dans le jeu, pourrait éprouver des difficultés à se faire passer pour une femme.

---

[111] “ Let us fix our attention on one particular digital computer C. Is it true that by modifying this computer to have an adequate storage, suitably increasing its speed of action, and providing it with an appropriate programme, C can be made to play satisfactorily the part of A in the imitation game, the part of B being taken by a man ? ”, *ibid.*, p. 142.

Au terme de l'examen, en revanche, et si l'on admet les raisons avancées par Turing à l'encontre des arguments selon lesquels une machine ne peut pas imiter le comportement général de l'homme, il sera permis de considérer qu'une machine, puisqu'elle peut imiter le comportement *humain*, est en mesure de participer au jeu de l'imitation en tant que tel, c'est-à-dire d'essayer de se faire passer, aux yeux d'un examinateur humain, pour une femme… De sorte que la véritable question à laquelle il s'agit, pour Turing, dans *Computing Machinery*…, de répondre, est celle de savoir si une machine peut *participer*, comme le ferait un être humain - au jeu de l'imitation. Présentée sous une autre forme, la question revient à se demander si une machine est capable de l'emporter à un test dont nous sommes assurés qu'un être humain, considéré *a priori* comme pensant, peut, dans environ trente pour cent des cas, le réussir. Est-il possible, autrement dit, de concevoir une machine qui, comme telle, aura, au moment où elle doit participer au jeu de l'imitation, environ trente pour cent de chances de l'emporter ? C'est enfin pourquoi nous pouvons, au niveau de la discussion menée ici par Turing, traduire le jeu sous la forme où il a été le plus souvent examiné : un joueur humain quelconque – homme ou femme – désormais noté A par Turing, doit tenter de distinguer un individu humain quelconque – homme ou femme – noté B, d'une machine notée C, C devant s'efforcer de le tromper en se faisant prendre par lui pour un individu humain quelconque, et B de l'aider, en donnant des " réponses vraies ".

Par là, le jeu de l'imitation constitue le protocole par lequel est représentée la majeure du syllogisme prévisionnel. Si la machine triomphe dans environ trente pour cent des cas lorsqu'elle prend la place de A dans le jeu, il sera acquis qu'elle aura effectivement imité le comportement intellectuel d'un être humain. Le test, s'agissant des " capacités intellectuelles ", fait, en effet, appel à l'ensemble des compétences qui sont celles d'un individu humain quelconque : aucun champ d'investigation n'est *a priori* interdit aux protagonistes ; toute question, quel que soit le domaine sur lequel elle porte, peut être posée. En d'autres termes, et ce point, nous le verrons, est d'une importance capitale, il s'agit, pour la machine, de simuler un comportement dont il n'est pas présupposé, du fait de sa complexité même, qu'il soit formalisable. C'est à cette condition que la réussite du test par la machine établira l'erreur de l'opinion commune, pour qui la notion de machine, parce qu'elle implique un ensemble de règles explicites, n'a pas de sens là où de telles règles ne peuvent être réunies.

Telle est la méthode originale proposée par Turing pour répondre à la question “ Les machines peuvent-elles penser ? ” Elle consiste à traduire l’hypothèse initiale - une machine peut “ penser ” au sens humain du terme - dans les termes d’une seconde hypothèse, plus “ manipulable ” aux yeux de Turing, et qui constitue, en quelque sorte, la mineure du syllogisme prévisionnel : une machine peut l’emporter au jeu de l’imitation face à des adversaires humains. Certes, pour des raisons technologiques, cette hypothèse, au moment où Turing l’envisage, n’est pas vérifiable. Pour reprendre les termes de Carnap, elle n’est pas “ testable ”. Cependant, il est permis de la discuter sur le plan de sa “ confirmabilité ”. A défaut d’une vérification directe et immédiate de l’hypothèse de la victoire d’une machine au jeu de l’imitation, il s’agira, autrement dit, de montrer qu’elle détermine avec toute la rigueur concevable le champ du *possible* conformément, d’une part, à la définition scientifique de la machine comme “ machine universelle ”, d’autre part à l’usage *a priori* - qui fonde toute discussion - du terme “ penser ”.

Section II : L'hypothèse de la victoire d'une machine universelle au jeu de l'imitation : le " comme si " de l'examinateur

Le problème posé par l'hypothèse de la victoire d'une machine au jeu de l'imitation peut être abordé sous deux angles : on l'a vu, le test met en jeu l'ensemble des compétences d'un individu humain quelconque, et, par là, exige de la machine qu'elle simule un comportement dont la complexité est telle que sa formalisation peut être, par hypothèse, considérée comme hors de portée de ses concepteurs ; il s'agit donc, en premier lieu, d'établir que ce caractère n'entre pas directement en contradiction avec la définition de la machine universelle, laquelle est formelle. Il s'agira ensuite d'indiquer comment une machine capable de simuler un comportement non formalisé peut être effectivement conçue. C'est pourquoi Turing mène sa réflexion en deux temps. Dans une première partie, intitulée " Vues contradictoires sur la question ", il raisonne de manière négative, en examinant, à la lumière de l'hypothèse de la victoire d'une machine au jeu de l'imitation, les raisons communément avancées pour refuser l'idée d'une " pensée " des machines, raisons qui, toutes, renvoient au caractère non formalisable du comportement d'un individu humain considéré comme pensant ; Turing montre que, dans le cadre du " jeu ", ces raisons perdent leur force. Deux catégories de raisons peuvent, ici, être distinguées ; la première rassemble des " objections " (" objections "), la seconde, des " arguments " (" arguments "). Ces deux catégories diffèrent par leur portée. Sous leurs formes respectives, les " objections " contestent l'hypothèse comme telle de la victoire d'une machine au jeu de l'imitation, qu'elles jugent directement contradictoire avec l'idée de machine. Dans leur optique, le test n'a pas même à être mené, car il est d'emblée exclu que la machine puisse le réussir. Les objections renvoient ainsi à

“ l’opinion commune ”, pour qui, on l’a vu, la question de la “ pensée ” des machines ne se pose pas.

Les “ arguments ”, quant à eux, soulèvent les difficultés concrètes que rencontrerait une machine universelle pour réussir le test, dans le cas où celui-ci serait effectivement mené. La méthode de discussion suivie par Turing repose sur cette différence de portée entre “ objections ” et “ arguments ” : il s’agit dans tous les cas de montrer, soit que “ l’objection ” n’est pas recevable en tant que telle, soit qu’elle peut être ramenée à un “ argument ”, lequel, dès lors qu’il est discuté dans le cadre strict du test, c’est-à-dire sous l’angle de la question “ Une machine universelle peut-elle faire bonne figure au jeu de l’imitation ? ”, s’avère non pertinent, faute de déterminer les moyens dont peut disposer l’examinateur A pour distinguer C de B. Turing s’efforce, ainsi, dans cette première partie, d’établir que rien ne s’oppose à ce qu’une machine universelle puisse, en vertu de sa définition même, l’emporter au jeu de l’imitation contre un examinateur humain, car celui-ci est amené, au cours de l’épreuve, à faire *comme si* son interlocuteur C était un individu humain.

Dans la seconde partie, intitulée “ Les machines qui apprennent ”, Turing complète son argumentation en raisonnant directement sur les capacités de la machine, et non plus en les considérant à travers la critique des différentes formules par lesquelles elles sont en général niées. Il s’agit ici de montrer comment la réalisation d’une machine susceptible de l’emporter au jeu peut être envisagée : une telle machine doit être conçue comme une “ machine qui apprend ” ; Turing prolonge l’hypothèse de la victoire de la machine au jeu de l’imitation par celle de la possibilité de soumettre la machine universelle à un processus “ d’apprentissage ” analogue à celui par lequel un individu humain est lui-même éduqué. La “ confirmabilité ” de cette dernière hypothèse est discutée de la même manière que, précédemment, celle de l’hypothèse de la victoire de la machine au jeu : l’argumentation de Turing vise à établir qu’il n’est pas contradictoire avec la notion de machine universelle qu’une telle machine “ apprenne ”. Nous verrons que l’articulation de ces deux moments - la discussion des “ objections ” et “ arguments ”, et la démonstration que la machine universelle peut “ apprendre ” - qui se situent apparemment sur deux plans différents, constitue le cœur de la réflexion de Turing.

I - Les " objections "

Turing examine en premier lieu trois " objections " : " l'objection théologique ", celle dite " de l'autruche ", ou encore de " la tête dans le sable ", et enfin " l'objection mathématique "[112].

L'objection théologique et l'objection " de l'autruche "

" L'objection théologique " consiste à soutenir que " penser est une fonction de l'âme immortelle de l'homme. Dieu a donné une âme immortelle à tout homme ou femme, mais à aucun animal ni à aucune machine. En conséquence, ni l'animal ni la machine ne peuvent penser "[113]. " L'objection de la tête dans le sable " se ramène à l'idée selon laquelle " le fait que les machines pensent aurait des conséquences trop terribles. Il vaut mieux croire et espérer qu'elles ne peuvent pas le faire "[114]. Ces deux objections sont manifestement liées : la seconde, qui consiste à refuser de considérer l'idée que l'homme pourrait ne pas être " nécessairement supérieur " au " reste de la création ", s'appuie sur la première[115]. Aussi bien Turing se contente-t-il de critiquer " l'objection théologique "[116]. Il le fait en arguant tout d'abord de ce qu'elle peut être discutée sur son propre terrain : peut-on, à partir de l'idée que Dieu n'a accordé une âme ni aux animaux ni aux machines, refuser, sans porter atteinte à sa toute puissance, l'idée qu'il a la liberté et la puissance de le faire ?[117] Enfin, sur un plan plus général, Turing rappelle que l'argument théologique peut faire obstacle à une réflexion de type scientifique :

> De tels arguments se sont souvent montrés peu satisfaisants dans le passé. Au temps de Galilée, on disait que les textes " Et le soleil s'arrêta [...] et ne se hâta pas de se cacher pendant

[112] " The theological objection ", " The 'Heads in the sand' objection', " The mathematical objection ". A.M. Turing, *Computing Machinery..., op. cit.*

[113] " thinking is a function of man's immortal soul. God has given an immortal soul to every man and woman, but not to any other animal or to machines. Hence no animal or machine can think. ", *ibid.*, p. 143.

[114] " The consequences of machines thinking would be too dreadful. Let us hope and believe that they cannot do so ", *ibid.*, p. 144.

[115] " la popularité de l'argument théologique est clairement liée à ce sentiment " (*The popularity of the theological objection is clearly connected with this feeling*), *ibid.*

[116] " Je ne pense pas que cet argument soit suffisamment substantiel, dit Turing à propos de l'objection de la tête dans le sable, pour rendre nécessaire une réfutation " (*I do not think that this argument is sufficiently substantial to require refutation*), *ibid.*

[117] " Il me semble que l'argument cité ci-dessus implique une grave restriction de l'omnipotence du tout-puissant " (*It appears to me that the argument quoted above implies a serious restriction of the omnipotence of the Almighty*), *ibid.*, p. 143.

> toute une journée ” (Jos X, 13) et “ Il posa les fondations de la Terre pour qu’elle ne bouge à aucun moment ” (Ps CV, 5) étaient une réfutation appropriée de la théorie copernicienne[118].

Bref, posée en termes théologiques, la question de la “ pensée ” des machines n’a certainement plus le sens de celle que Turing entend poser à partir de la notion même de machine universelle, telle qu’il l’a définie, comme notion logico-mathématique, dans *On Computable Numbers*...

L’objection mathématique

“ L’objection mathématique ”, quant à elle, doit retenir plus particulièrement l’attention, dans la mesure où elle met en jeu précisément la définition scientifique de la machine universelle ; elle s’appuie, en effet, sur la soumission, que Turing avait lui-même mise en évidence, de la machine universelle, en tant que notion logique, aux “ théorèmes de limitation ”. Sa discussion permet à Turing de préciser tout à la fois les termes de la question à laquelle il entend répondre et le principe de l’argumentation qu’il va suivre. Le problème abordé ici est celui de savoir si les théorèmes de limitation n’établissent pas, par eux-mêmes, un fossé infranchissable entre l’idée de “ procédé mécanique ”, considérée dans sa précision logique, et celle de “ pensée ”, dans sa dimension ultra-logique. Turing rappelle ainsi

> [qu’]un certain nombre de résultats de la logique mathématique peuvent être utilisés pour montrer qu’il y a des limites aux pouvoirs des machines à états discrets. Le plus connu de ces résultats est connu sous le nom de théorème de Gödel et montre que dans tout système logique suffisamment puissant, on peut formuler des affirmations qui ne peuvent ni être prouvées, ni être réfutées à l’intérieur du système, à moins que le système lui-même ne soit inconsistant[119].

Le sens de cette objection a été clairement exposé par J. R. Lucas une dizaine d’années après la publication de *Computing Machinery*..., dans un article visant les premiers

[118] “ Such arguments have often been found unsatisfactory in the past. In the time of Galileo it was argued that the texts, ‘And the sun stood still... and hasted not to go down about a whole day’ (Joshua x. 13) and ‘he laid the foudations of the earth, that it should not move at any time’ (Psalm cv. 5) were an adequate refutation of the Copernican theory ”, *ibid.*

[119] “ There are a number of results of mathematical logic which can be used to show that there are limitations to the power of discrete-state machines. The best known of these results is known as Gödel’s theorem (1931) and shows that in any sufficiently powerful logical system statements can be formulated which can neither be proved nor disproved within the system, unless possibly the system itself is inconsistent ”, *ibid.,* p. 144.

développements de l'intelligence artificielle :

> Le théorème de Gödel me semble prouver que la théorie mécaniste est fausse, c'est à dire que l'esprit humain ne peut être compris comme une machine. Beaucoup de gens ont pensé de même ; presque tous les logiciens en mathématiques auxquels j'ai proposé le sujet ont avoué des positions similaires...[120].

Le théorème d'incomplétude de Gödel établit en effet qu'il existe pour tout système formel une formule vraie dans ce système qui ne peut y être démontrée. Or, le sens même du théorème réside dans le fait que ce que le système formel ne permet pas de faire, un observateur extérieur, situé dans une position métathéorique par rapport à lui, le peut : s'il y a un théorème de Gödel, c'est que Gödel pouvait, quant à lui, déclarer vraie la formule indérivable dans son système. Puisqu'une " machine de Turing " est la réalisation d'un système formel, pour chaque machine, poursuit Lucas,

> il y a une vérité qu'elle ne peut pas produire comme étant vraie, alors qu'un esprit le peut. Voilà qui montre qu'une machine ne peut être un modèle complet et adéquat de l'esprit. Elle ne peut pas faire *tout* ce qu'un esprit peut faire puisque, quelles que soient ses capacités, il y a toujours quelque chose qu'elle ne peut pas faire et qu'un esprit peut faire. Ceci ne veut pas dire que nous ne puissions pas construire une machine à même de simuler *n'importe quelle partie* d'un comportement semblable à celui de l'esprit. Mais nous ne pouvons pas construire une machine à même de simuler *toutes* les parties d'un comportement similaire à l'activité de l'esprit "[121].

Ainsi formulée, l'objection renvoie, non seulement à l'article de Turing sur les " nombres calculables ", mais également à sa thèse américaine, *Systems of logic based on ordinals*, dans laquelle il montrait que le passage d'une machine donnée à une machine " plus puissante " ne peut être lui-même mis sous la forme d'une procédure effective, c'est-à-dire d'un procédé mécanique. Ainsi, poursuit Lucas,

> même si nous ajoutons à un système formel l'ensemble infini d'axiomes constitué par les formules de Gödel successives, le système résultant est encore incomplet et contient une formule

[120] J.R. Lucas, " L'esprit humain, la machine et Gödel ", *Pensée et machine*, sous la dir. de Ross Anderson, Paris, Champ Vallon, 1983. p. 81. Ed. américaine de 1964.
[121] *Ibid.*, p. 85. C'est l'auteur qui souligne.

> qui ne peut pas être prouvée-dans-le-système bien qu'un être rationnel puisse, se trouvant en dehors du système, établir qu'elle est vraie[122].

En vérité, l'argumentation de Lucas fait écho à la position antimécaniste soutenue par Gödel lui-même. Celui-ci admettait l'un des principes de la démonstration de Turing dans *On Computable Numbers...*, à savoir que l'esprit humain comporte un nombre fini d'états internes, mais il professait un " optimisme rationaliste ", soutenant que " l'intelligence dans son usage n'est pas statique, mais se développe sans cesse "[123] : rien n'interdit de considérer que l'esprit humain soit fini à chaque instant, mais que le nombre de ses états internes puisse converger à l'infini dans le cadre du développement historique. En d'autres termes, aux yeux de Gödel, la validité des théorèmes d'incomplétude qu'il avait établis pour l'arithmétique ne concernait pas le caractère essentiellement dynamique de l'esprit humain, c'est-à-dire la capacité de celui-ci à toujours surpasser sa propre limite à un instant donné.

Il y a " objection ", ici, en un double sens. D'une part, le " test " proposé par Turing pour répondre à la question " Les machines peuvent-elles penser ? " ne s'imposerait pas, dans la mesure où les théorèmes de Gödel sont interprétés comme fournissant une réponse directe à la question ; le fait que l'esprit humain puisse toujours répondre à la " question critique " qui arrête une machine, indiquerait une supériorité irréductible de l'homme sur la machine, et, de ce point de vue, cette supériorité serait ce qui atteste la pensée en l'homme. On peut affirmer ainsi que la machine ne " pense " pas, sans qu'il soit besoin, pour en décider, de lui faire subir un test.

C'est, d'autre part, l'idée même selon laquelle la simulation, par la machine universelle, des conditions intuitives du calcul pour un individu humain renvoie à l'" esprit humain " comme tel, qui est niée ; la machine universelle ne simulerait pas les conditions mêmes de sa construction, c'est-à-dire la part qui, dans l'acte humain de calcul, échappe au formalisme.

Nous l'avons vu, cette argumentation a été développée par Lucas en 1961, bien après la mort de Turing ; elle participe cependant d'un débat ouvert dans le cadre de la réflexion sur

[122] *Ibid.*, p. 87.
[123] Voir à ce sujet l'article de Jacques Dubrucs, " Réalisme et antimécanisme chez Gödel ", *Dialectica*, 40, 4, 1986, ainsi que les ouvrages de Hao Wang, *From Mathematics to Philosophy*, Cambridge, Massachusetts, The MIT Press, 1974 ; *Reflections on Kurt Gödel*, Cambridge, Massachussetts, The MIT Press, 1988.

les fondements des mathématiques et qui accompagne, de manière plus ou moins explicite, l'ensemble de la recherche menée sur les théorèmes de limitation[124]. Turing l'ignorait d'autant moins que *Computing Machinery...* constitue, en tant que tel, la réponse qu'il entendait proposer aux questions soulevées dans le cadre de ce débat. C'est précisément pour clarifier la relation de simulation de la machine à l'esprit humain qu'il propose la méthode originale du jeu de l'imitation, censée permettre une véritable confrontation des termes en présence. Autrement dit, aux yeux de Turing, la validité de " l'objection mathématique " ne pourrait être établie qu'*au terme* de la démarche qu'il entreprend. Le statut de cette objection est, pour lui, le même que celui des deux précédentes, " l'objection théologique " et " l'objection de l'autruche " : elle est, en définitive, une expression de cette opinion commune qu'il s'agit précisément de critiquer parce qu'elle occulte le problème même posé, aux yeux de Turing, par la notion de machine.

Dès lors, si l'on fait abstraction de la réponse qu'elle apporte *a priori* à la question " Les machines peuvent-elles penser ? ", l'objection mathématique doit être considérée, non plus comme une " objection ", mais plutôt comme un " argument " : il s'agit, non pas de conclure *directement* des théorèmes de limitation à l'absence de sens de la question de la " pensée " des machines, mais de discuter effectivement la question de savoir si la limite logique de la machine, en tant que machine universelle, rend théoriquement impossible sa " victoire " au jeu de l'imitation.

Nous sommes d'ores et déjà certains qu'en vertu des théorèmes d'incomplétude, une machine universelle programmée pour participer au jeu de l'imitation ne pourra pas répondre à certaines des questions qui lui seront posées par son adversaire humain. Il en sera ainsi, note Turing, pour toute question du type : " Considérez la machine spécifiée comme suit... Cette

[124] L'idée que l'on ne saurait rendre compte de la réflexion mathématique comme telle à l'aide de la notion de machine traverse par exemple la contribution propre de Post aux recherches sur la " calculabilité effective " : " Cela fait du mathématicien bien plus qu'une sorte d'être intelligent qui peut effectuer vite ce qu'une machine finit par effectuer. Nous voyons qu'une machine ne fournirait jamais une logique complète : car, une fois la machine construite, nous pourrions prouver un théorème qu'elle ne prouve pas ", (*It makes of the mathematician much more than a kind of clever being who can do quickly what a machine could do ultimately. We see that a machine would never give a complete logic : for once the machine is made we could prove a theorem that it does not prove*). E. L. Post, " Absolutely Unsolvable Problems and Relatively Undecidable Propositions : Account of an Anticipation ", *The Undecidable...,* M. Davis, éd., *op. cit.,* p. 417. Cité in R. Gandy, " The Confluence of Ideas in 1936 ", *The Universal Turing Machine, a Half-Century Survey,* Rolf Herken, éd., *op. cit.,* p. 88.

machine répondra-t-elle ‘ oui ’ à n’importe quelle question? ”[125]. Nous savons, d’après *On Computable Numbers*..., qu’“ il ne peut pas y avoir de machine ε qui, lorsqu’elle est équipée de la description standard d’une machine quelconque M, déterminera si M imprime un symbole donné (disons 0) ”[126]. En d’autres termes, la question évoquée ici par Turing est indécidable.

Cependant, note Turing, “ ... bien qu’il soit établi qu’il y a des limites à la puissance de n’importe quelle machine, il a seulement été affirmé [NB : dans le cadre de l’objection mathématique], sans aucune sorte de preuve, que de telles limites ne s’appliqueraient pas à l’esprit humain ”[127]. Pour que l’impuissance de machines données devant certaines “ questions critiques ” prouve la supériorité de l’esprit humain sur la machine universelle, il faudrait que les hommes eux-mêmes ne donnent jamais de réponses erronées, ou bien ne se trouvent eux-mêmes jamais dans une situation où ils sont incapables de répondre, ce qui n’est évidemment pas le cas[128]. Surtout, poursuit Turing, ce qui est vrai d’une machine par rapport à une autre machine, à savoir qu’une machine plus forte peut toujours “ prendre le relais ” de la machine arrêtée par sa question critique, l’est aussi d’une machine, non par rapport à l’homme “ en général ”, ou, pour parler comme Lucas, par rapport à “ l’esprit humain ” comme tel, mais par rapport à un individu humain donné : si l’on admet l’argument même de Lucas, selon lequel un esprit humain, parce qu’il se trouve en dehors du système représenté par la machine, peut établir que la proposition devant laquelle s’arrête celle-ci est vraie, alors il sera toujours possible, du point de vue logique, de construire une machine s’acquittant de la tâche que tel individu humain n’aura pas lui-même su mener, mais qu’un autre individu humain aura accomplie. Pour affirmer qu’il existe un état à partir duquel la capacité pour une machine d’en suppléer une autre ou de suppléer tel individu humain ne s’applique plus et que l’homme l’emporte définitivement, nous devons quitter le point de vue logique, ou mathématique. Dans

---

125 “ Consider the machine specified as follows... Will this machine ever answer ‘ yes ’ to any question ? ”. A. M. Turing, *Computing Machinery..., op. cit.,* p. 145.

126 Voir plus haut, 1e partie, chapitre I.

127 “ ... although it is established that there are limitations to the powers of any particular machine, it has only been stated, without any sort og proof, that no such limitations apply to the human intellect ”, *Ibid.,* p. 145.

128 “ Nous donnons nous-mêmes trop souvent des réponses fausses à des questions pour que nous ayons le droit de nous réjouir d’une telle preuve de la faillibilité des machines ” (*We too often give wrong*
answers to questions ourselves to be justified in being very pleased at such evidence of fallibility on the part of the machines ”), *ibid.*, p.145.

le cadre de l'objection mathématique, une telle affirmation reviendrait à nier cela même sur quoi elle s'appuie, à savoir, d'une part, les théorèmes d'incomplétude, d'autre part, la capacité humaine à répondre à la question qui arrête telle machine. La limite qui a été établie pour la machine l'a, en effet, d'abord été pour l'homme. Si l'on parle, comme le fait J.R. Lucas, de l'homme en général, ou de " l'esprit humain " comme tel, et de la machine en général, affirmer la supériorité de l'un sur l'autre au nom du théorème de Gödel est contradictoire : il ne peut y avoir de supériorité que d'un individu humain déterminé sur une machine déterminée[129]. De sorte que, si rien n'autorise à dire qu'il n'y aura pas toujours un individu humain capable de l'emporter, lors d'un test de Turing, sur une machine déterminée, rien non plus n'autorise à soutenir qu'il n'y aura pas toujours une machine déterminée capable de l'emporter sur un individu humain. La machine et son adversaire humain ont pour Turing le même statut logique, c'est-à-dire la même limite. Or, un homme peut l'emporter à un jeu tel que le jeu de l'imitation - en tout état de cause, " l'objection mathématique " le suppose. La limite logique de la machine ne saurait donc interdire à celle-ci de l'emporter. Pour le dire autrement, si une machine ne pouvait réussir le test, ce ne pourrait être, aux yeux de Turing, à cause de son statut logique.

Le principe de l'argumentation de Turing, ainsi que la méthode de discussion qu'il entend suivre sont ainsi bien établis : il s'agit, d'une part, non plus d'examiner directement la question " Les machines peuvent-elles penser ? ", mais la question " Une machine universelle, ou un ordinateur digital, peuvent-ils faire bonne figure au jeu de l'imitation ? " ; d'autre part, de transformer les " objections " en " arguments ", qui seront considérés dans le cadre délimité par la nouvelle question. Cette méthode est celle que va suivre Turing à l'égard des raisons communément avancées à l'encontre de l'idée d'une " pensée " des machines.

II - Les " arguments "

Ces raisons renvoient, d'une manière générale, au problème de " l'informalité du comportement " humain. Pour l'opinion commune, la " pensée ", puisqu'elle est manifestée

[129] " notre supériorité, en de telles occasions, ne peut être ressentie que par rapport à la machine particulière sur laquelle nous avons remporté un triomphe insignifiant " (... *our superiority can only be felt on such an occasion in relation to the one machine over which we have scored our petty triumph*), *ibid*. p. 145.

par la capacité de surprendre, suppose l'imprévisible. Cet imprévisible, attendu de toute entité pensante par l'opinion commune, est, pour celle-ci, d'ordre, non pas infra-logique, mais ultra-logique : il ne relève pas du hasard, mais de *l'invention*. C'est en ce sens qu'il est, au regard de l'opinion commune, contradictoire avec l'idée de machine. Selon l'opinion commune, une entité ne peut être une machine que si l'on sait précisément ce qu'elle va faire. Elle ne peut donc " surprendre " que par rapport à l'action spécifique que l'on attend d'elle, de sorte qu'il n'y a, en vérité, de surprise à attendre, dans l'ordre du mécanique, que d'une machine mal conçue, en quelque façon " ratée ". Le statut de la machine est nécessairement dérivé de celui de l'outil ; certes, la machine est plus que le simple outil, dans la mesure où elle se substitue pour une part à l'action de l'agent qui manie l'outil, cependant, cela n'est possible que parce que l'action effectuée par la machine en lieu et place de l'agent est parfaitement connue. Bref, il ne saurait y avoir d'*invention* de la part de la machine.

Turing discute les raisons ainsi considérées en s'efforçant de montrer que la machine peut, non seulement présenter une imprévisibilité purement accidentelle, mais également simuler l'imprévisible tenu pour spécifiquement humain par l'opinion commune, et que, dans le cadre du jeu de l'imitation, l'examinateur A ne disposera d'aucun moyen de distinguer l'imprévisible ainsi simulé par la machine de celui qu'il attend spontanément d'un interlocuteur humain. Turing s'efforce, en d'autres termes, de montrer que, s'agissant de la notion de machine universelle, les raisons communément avancées pour refuser l'idée d'une " pensée " des machines perdent leur force, lorsqu'elles sont examinées dans le cadre de la participation d'une telle machine au jeu de l'imitation

L'objection-argument " de la conscience "[130]

Le premier argument discuté par Turing, celui dit " de la conscience ", s'appuie, en ce sens, sur un imprévisible qui renvoie, non à l'aléatoire, mais au singulier. L'expression artistique, toujours singulière, et qui met en jeu l'émotion et la conscience, est à ce titre ce qui paraît le plus irréductiblement éloigné de la machine. Turing adopte la formulation donnée de

[130] " The Argument from Consciousness ", *ibid*. p. 145.

cet argument, dans une conférence prononcée en 1949, par le professeur Jefferson[131], lequel déclarait :

> Nous ne pourrons pas accepter l'idée que la machine égale le cerveau jusqu'à ce qu'une machine puisse écrire un sonnet ou composer un concerto à partir de pensées ou d'émotions ressenties, et non pas en choisissant des symboles au hasard, et non seulement l'écrire, mais savoir qu'elle l'a écrit.[132]

Supposons que l'on demande à une machine de nous proposer un sonnet de son invention ; si elle s'exécute, ce sera, au mieux, à l'aide d'un " artifice " (" contrivance ") - on peut imaginer, par exemple, qu'elle " retourne ", selon le terme utilisé en informatique, au moment voulu, l'enregistrement sur support magnétique d'un sonnet écrit par un individu humain. La machine n'aura fait, au mieux, que " produire artificiellement un signal "[133]. Certes, si l'examinateur humain de la machine reconnaît un sonnet dans cette production de signal, un message aura été transmis, mais la machine n'aura fait, précisément, que le transmettre ; ce message lui restera à jamais extérieur et inintelligible. Le fait même de *composer* un sonnet - et non pas simplement de le présenter lorsqu'il est demandé - ne suffira pas davantage à nous convaincre que la machine " pense " : la forme poétique en général et la forme " sonnet " en particulier renvoient à un système de règles, et, sous cet angle, il n'est certainement pas impossible de faire simuler par une machine l'écriture ou la " composition " d'un sonnet ; cependant, celui-ci ne sera rien d'autre, du " point de vue " de la machine, si l'on peut s'exprimer ainsi, qu'une combinaison de symboles, qui, par définition, toujours du même " point de vue ", seront dépourvus de sens. Bref, si une machine " écrit " un sonnet, elle ne saura pas qu'elle l'a écrit, car elle ne saura ni de quoi elle " parle ", ni même qu'elle " parle ". A cet égard, continue le professeur Jefferson,

> aucun mécanisme ne pourrait ressentir... du plaisir quand il réussit, du chagrin quand ses lampes grillent ; il ne serait pas ému par la flatterie, rendu malheureux par ses erreurs, charmé

---

[131] Il s'agit de sir Geoffrey Jefferson, professeur de neuro-chirurgie à l'Université de Manchester, avec lequel Turing participe, en janvier 1952, à une émission de la BBC sur la question de la " pensée " des machines. Voir le BBC Written Archives Centre, *op. cit.*

[132] " Not until a machine can write a sonnet or compose a concerto because of thoughts and emotions felt, and not by the chance fall of symbols, could we agree that machine equals brain, that is, not only write it, but know that it had written it... ". A. M. Turing, *Computing Machinery..., op.cit.,* p. 145.

[133] " ... not merely artificial signal... ", *ibid.*, p. 146.

> par le sexe, et ne se mettrait pas en colère ou ne se sentirait pas déprimé quand il ne peut pas obtenir ce qu'il veut.[134]

Pour que nous puissions admettre " l'idée que la machine égale le cerveau ", il faudrait encore que le sonnet proposé par la machine soit bien la forme réfléchie d'une émotion, qu'il ne renvoie pas à celle-ci simplement comme à la référence théorique d'un discours, mais qu'il exprime, d'une part, cette émotion en tant qu'elle est vécue, et d'autre part, l'acte conscient de réflexion de cette émotion dans une forme.

Par là, l'argument dit " de la conscience " apparaît tout d'abord comme une " objection " : il invalide le principe même du jeu de l'imitation. C'est ce que relève Turing en poussant la logique de la déclaration du professeur Jefferson jusqu'à repérer en elle un point de vue solipsiste ; l'affirmation du professeur Jefferson repose, dit-il, sur l'idée que " la seule manière dont on pourrait s'assurer qu'une machine pense serait d'être la machine et de ressentir qu'on pense […] On pourrait alors décrire ces sentiments au monde, mais bien sûr personne n'aurait de raison d'en tenir compte "[135]. Comme tel, l'argument revient à déclarer que, quand bien même la machine satisferait en apparence à ce qui lui est demandé, rien ne nous autoriserait à affirmer qu'elle pense. Cependant, fait remarquer Turing, le prix à payer sera particulièrement lourd, car, dans la perspective adoptée ici, " la seule manière de savoir qu'un *homme* pense est d'être cet homme lui-même "[136], de telle sorte que, si le professeur Jefferson refuse, selon ce principe, la " pensée " à la machine qui aura " écrit " un sonnet, il devra également la refuser à tout interlocuteur, et notamment à tous les individus humains auxquels il s'est adressé jusque là et auxquels il va désormais s'adresser. Le solipsisme serait-il juste, poursuit Turing, reprenant une critique traditionnelle, qu'il s'opposerait aux exigences de la pratique de communication entre les hommes : " Il se peut que ce soit la position la plus logique à tenir, mais cela rend difficile la communication des idées. A est enclin à croire que ' A pense, mais B ne pense pas ', pendant que B croit ' B pense mais A ne pense pas ' "[137].

[134] " No mechanism could feel (and not merely artificially signal, an easy contrivance) pleasure at its successes, grief when its valves fuse, be warmed by flattery, be made miserable by its mistakes, be charmed by sex, be angry or depressed when it cannot get what it wants ", *ibid*.
[135] " One could then describe these feelings to the world, but of course, no one would be justified in taking any notice ", *ibid*.
[136] " ... the only way to know that a *man* thinks is to be that particular man ", *ibid*. (c'est Turing qui souligne).
[137] " It may be the most logical view to hold but it makes communication of ideas difficult. A is liable to believe ' A thinks but B does not ' whilst B believes ' B thinks but A does

Aussi Turing remarque-t-il qu'" au lieu de discuter continuellement ce point, on adopte habituellement la convention polie stipulant que tout le monde pense "[138]. Or, si l'on isole le raisonnement de son assise solipsiste, d'une " objection ", on passe à un " argument " : découle-t-il des remarques faites par le professeur Jefferson une impossibilité théorique pour la machine à " faire bonne figure " au jeu de l'imitation ?

Dans le cadre du jeu, la question n'est plus de savoir si la machine C éprouve des émotions ou si elle sait qu'elle compose un sonnet, mais de savoir si elle peut se comporter de telle sorte que l'examinateur A se trompe à son sujet, *y compris en lui accordant de véritables émotions et une conscience*. Turing évoque à ce propos la situation de l'examinateur qui, lors d'une épreuve orale, pour déterminer si un candidat a réellement compris ce qu'il énonce ou bien se contente de répéter son cours, pose à ce candidat des questions complémentaires. Dans le cadre d'un jeu tel que celui du sonnet, l'examinateur, afin de vérifier que le texte proposé par son interlocuteur a bien un *sens* pour celui-ci, lui demandera, par exemple, d'expliquer ses choix. Un dialogue s'instaurera alors entre eux, qui pourrait être, nous dit Turing :

> *L'examinateur* : Dans le premier vers de votre sonnet qui dit : ' Te comparerai-je à un jour d'été ? ', est-ce qu' ' un jour de printemps ' serait aussi bien ou mieux ?
> *Le candidat* (" *witness* ") : Cela ne rimerait pas[139].
> *L'examinateur* : Et ' un jour d'hiver ' ? Cela rimerait très bien.
> *Le candidat* : Oui, mais personne n'a envie d'être comparé à un jour d'hiver.
> *L'examinateur* : Diriez-vous que M. Pickwick vous fait penser à Noël ?
> *Le candidat* : D'une certaine manière, oui.
> *L'examinateur* : Pourtant, Noël est un jour d'hiver, et je ne pense pas que la comparaison ennuierait M. Pickwick.
> *Le candidat* : Je ne pense pas que vous soyez sérieux. Par un jour d'hiver, on veut dire un jour d'hiver typique, plutôt qu'une journée spéciale comme Noël.[140]

---

not ' ", *ibid*.
[138] " Instead of arguing continually over this point it is usual to have the polite convention that everyone thinks ", *ibid*.
[139] " Winter " rime avec " summer ", mais pas avec " spring ".
[140] " Interrogator : In the first part of your sonnet which reads ' Shall I compare thee to a summer's day ', would not ' a spring day ' do as well or better ?
Witness : It wouldn't scan.
Interrogator : How about ' a winter's day '. That would scan all right.
Witness : Yes, but nobody wants to be compared to a winter's day.
Interrogator : Would you say Mr. Pickwick reminded you of Christmas ?
Witness : In a way.

Si la machine fournissait les réponses du " candidat " dans un tel dialogue, serions-nous en mesure de la distinguer, selon le critère de " l'émotion ressentie et consciente ", d'un individu humain ? Quel moyen aurions-nous de faire coïncider le comportement de la machine au cours du jeu avec ce que recouvre l'usage commun du terme " machine " ? Aussi bien, demande Turing

> que dirait le professeur Jefferson si la machine à écrire des sonnets était capable de répondre ainsi *in viva voce* ? Je ne sais pas s'il considérerait que la machine ' produit simplement et artificiellement un signal ' avec ces réponses, mais si les réponses étaient aussi satisfaisantes et fermes que dans le passage ci-dessus, je ne pense pas qu'il la décrirait comme ' un artifice facile '[141].

Du reste, pourrions-nous ajouter, tout homme sera-t-il en mesure d'exprimer ses émotions de manière suffisamment convaincante pour qu'aucun doute ne soit permis quant au fait qu'il les " éprouve " effectivement ? On ne saurait, en l'occurrence, demander à la machine de faire mieux qu'un élève simplement moyen…

Turing, en vérité, transforme l'argument dit " de la conscience ", en argument " du sonnet " : ce qui importe, ce n'est pas la " conscience " de la machine, mais le point de vue de l'examinateur de celle-ci, tel qu'il est induit par la problématique du jeu de l'imitation. La machine l'emportera si, au cours de celui-ci, l'examinateur humain doit adopter " la convention polie " stipulant qu'elle pense, s'il doit faire *comme si* la machine était un homme, c'est-à-dire un être postulé comme pensant. En d'autres termes, si l'on n'est pas prêt à payer le prix exigé par la position solipsiste, non seulement il n'y a plus de raison de refuser le principe du jeu de l'imitation, mais il faut encore admettre que les résultats de celui-ci seront fonction, non plus d'un critère absolu - dont l'usage commun, qui n'exprime pas un savoir, mais une simple *doxa*, ne fait que refléter le caractère introuvable - mais de la seule *habileté* des protagonistes, la machine universelle comprise. L'unique question qui importe consiste,

---

Interrogator : Yet Christmas is a winter's day, and I do not think Mr. Pickwick would mind the comparison.
Witness : I don't think you're serious. By a winter's day one means a typical winter's day, rather than a special one like Christmas. *", ibid.*

[141] " What would Professor Jefferson say if the sonnet-writing machine was able to answer like this in the viva voce ? I do not know wether he would regard the machine as 'merely artificially signalling ' these answers, but if the answers were as satisfactory and sustained as in the above passage I do not think he would describe it as ' an easy contrivance ' ", *ibid.*, p. 147.

sous cet angle, à montrer que la machine peut être aussi " habile " que ses deux adversaires humains, c'est-à-dire qu'elle peut simuler le comportement attendu d'un individu humain. Auquel cas, à habileté à peu près égale, les résultats du jeu relèveront de la statistique, et la machine aura environ trente pour cent de chances de l'emporter.

Il s'agit dès lors de montrer que rien, dans la notion théorique de machine universelle ne s'oppose à ce que son comportement présente les traits attendus par un examinateur humain pour qu'elle puisse l'emporter à un test tel que celui du sonnet. Compte tenu, en effet, de la nature même de celui-ci, au regard de l'opinion commune, il est permis de penser que, si la machine le réussit, elle sera en mesure de réussir tout autre test du même ordre – ce que l'on pourrait appeler des " tests partiels de Turing ".

***Les arguments " des diverses incapacités "*[142]**

C'est pourquoi Turing examine ensuite l'argument qu'il nomme " des diverses incapacités ", dont il propose la formulation suivante : " Je vous concède que vous pouvez fabriquer des machines qui fassent tout ce que vous avez mentionné, mais vous ne serez jamais capable d'en fabriquer une qui fasse X "[143]. Et Turing énumère " différents traits de X " (" Numerous features X ") : on ne saurait construire une machine

> gentille, débrouillarde, belle, amicale, [qui] ait de l'initiative, un sens de l'humour, [qui] fasse la différence entre le bien et le mal, fasse des erreurs, tombe amoureuse, aime les fraises à la crème, rende quelqu'un amoureux d'elle, apprenne à partir de son expérience, utilise les mots correctement, soit l'objet de ses propres pensées...[144].

Il s'agit là, remarque Turing, de " formes déguisées de l'argument tiré de la conscience "[145] : il est, en quelque sorte, communément présumé qu'une machine qui réussirait le test du sonnet, devrait être capable de " faire X ", tel que cet " X " vient d'être décrit. Turing va donc

---

[142] " Arguments from various disabilities ", *ibid*. p. 147.
[143] " I grant you that you can make machines do all the things you have mentioned but you will never be able to make one to do X ", *ibid.,* p. 147.
[144] " Be kind, resourceful, beautiful, friendly, have intiative, have a sens of humour, tell right from wrong, make mistakes, fall in love, enjoy strawberries and cream, make someone fall in love with it, learn from experience, use words properly, be the subject of its own thought... ", *ibid*.
[145] " ... disguised forms of the argument from consciousness ", *ibid.,* p. 149.

s'efforcer de montrer que l'examinateur A ne peut guère davantage que dans le cas du test du sonnet s'appuyer sur ces " différents traits " pour vaincre la machine au jeu de l'imitation.

Il en est ainsi, notamment, à propos de l'erreur[146]. Ne suffirait-il pas que, dans le cadre du jeu, l'examinateur humain pose à la machine des questions d'arithmétique pour qu'elle soit démasquée, " à cause de son exactitude implacable "[147] ? Aux yeux de Turing, l'argument confond deux types d'erreurs : les " erreurs de fonctionnement "[148] et les " erreurs de conclusion "[149]. Les " erreurs de fonctionnement " résultent de " quelque faute mécanique ou électrique qui fait que la machine ne se comporte pas comme elle est conçue pour le faire "[150]. Les " erreurs de conclusion ", quant à elles, " apparaissent seulement quand une signification est attribuée aux signaux de sortie de la machine "[151]. Les erreurs du premier type sont en général écartées " dans les discussions philosophiques ", où l'on préfère discuter " de ' machines abstraites ' [qui] sont des fictions mathématiques plutôt que des objets physiques "[152]. Il ne s'agit là, cependant, que d'hypothèses d'école ; l'expérience de Turing en matière de construction de machines réelles était suffisante pour qu'il ait pu se convaincre que les machines ne sont pas à l'abri " d'erreurs de fonctionnement ". En outre, à supposer que l'on ne prenne en compte la machine que sous sa forme de machine idéale, de " fiction mathématique ", son infaillibilité théorique même permettrait de faire en sorte qu'elle introduise " délibérément des erreurs d'une manière calculée pour dérouter l'interrogateur "[153]. La machine pourra être expressément programmée pour commettre, avec son " exactitude implacable ", des erreurs de conclusion ; elle pourra, sans " erreur de fonctionnement ", simuler des " erreurs de conclusion ", c'est-à-dire émettre des propositions qui seront considérées comme telles du point de vue d'un observateur humain extérieur. Il y a même,

---

[146] " Le fait de revendiquer que ' les machines ne peuvent pas faire d'erreurs ' semble curieux, commence par souligner Turing. On est tenté de répondre : ' En sont-elles pires pour cela ? ' ". (*The claim that ' machines cannot make mistakes ' seems a curious one. One is tempted to retort, ' are they any the worse for that? '*), *ibid.*, p. 148.

[147] " ... because of its deadly accuracy ", *ibid.*

[148] " Errors of functioning ", *ibid.*, p. 149.

[149] " Errors of conclusion ", *ibid.*

[150] " Errors of functioning are due to some mechanical or electrical fault which causes the machine to behave otherwise than it was designed to do ", *ibid.*.

[151] " Errors of conclusion can only arise when some meaning is attached to the output signals of the machine ", *ibid.*

[152] " In philosophical discussions one likes to ignore the possibility of such errors ; one is therefore discussing ' abstract machines '. These abstract machines are mathematical fictions rather than physical objects ", *ibid.*

[153] " It would deliberately introduce mistakes in a manner calculated to confuse the interrogator ", *ibid.*, p. 148.

ajoute Turing un peu obscurément, une forte probabilité pour que, dans le cas de la machine concrète et non plus idéale, cela entraîne des " erreurs de fonctionnement " : " une erreur mécanique se révélerait probablement à cause d'une décision inopportune à propos du type d'erreur à commettre... "[154].

On peut, du reste, aller plus loin : la probabilité est également très forte que la simulation de l'erreur entraîne la machine à commettre, de manière non délibérée, des " erreurs de conclusion ", ne serait-ce que parce qu'ajouter à un système une règle qui contredit l'une ou l'autre de celles qui le définissent originellement peut en affaiblir la consistance. Sur un autre plan, c'est à peu près ce qui se passera, affirme Turing, si on dote la machine " d'une méthode pour tirer des conclusions par induction scientifique "[155]. Turing appelle ainsi, de manière très classique, la formulation de conclusions générales à partir d'un certain nombre d'observations. Il est des domaines pour lesquelles, dit-il, la conclusion générale ne peut être juste que si " une grande partie d'espace-temps "[156] est étudiée, " sinon, nous pouvons décider (comme le font la plupart des enfants anglais) que tout le monde parle anglais et qu'il est idiot d'apprendre le français "[157]. Une machine qui disposerait d'une règle d'induction l'appliquerait implacablement sans pour autant disposer d'un système de règles consistant pour tout domaine, et " nous devons nous attendre à ce qu'une telle méthode conduise occasionnellement à des résultats erronés "[158]. En un mot, l'adversaire humain de la machine ne peut certainement pas compter sur la prétendue infaillibilité de celle-ci pour la distinguer, dans le cadre du test, de son partenaire humain.

Il en est ainsi également de l'affirmation qu'une machine ne peut être " l'objet de ses propres pensées " et ne peut " apprendre à partir de son expérience ". Turing remarque tout d'abord que l'objection suppose que la machine ait des pensées et que celles-ci aient un objet. Pour éviter lui-même une telle pétition de principe, il ramène le problème à la question : " la machine peut-elle être son propre objet ? ". On ne niera pas, fait-il remarquer, que la machine fasse des opérations et que celles-ci, " du moins pour les gens qui travaillent dessus [sur la

---

154 " A mechanical fault would probably show itself through an unsuitable decision as to what sort of a mistake to make... ", *ibid.*
155 " ... some method for drawing conclusions by scientific induction ", *ibid.,* p. 149.
156 " A very large part of space-time... ", *ibid.,* p. 148.
157 " Otherwise we may (as most english children do) decide that everybody speaks English, and that it is silly to learn French ", *ibid.*
158 " We must expect such a method to lead occasionally to erroneous results ", *ibid.,* p. 149.

machine] "[159] aient un objet. C'est pourquoi, " si, par exemple, la machine essayait de trouver une solution à l'équation $x^2 - 40x - 11 = 0$, on serait tenté de décrire l'équation comme une partie de l'objet de la machine à ce moment-là "[160]. En ce sens même, si l'on admet que les opérations de la machine portent sur quelque chose, alors, il ne fait pas de doute que la machine peut être son propre objet : là est, précisément, le propre de la machine universelle, qui peut simuler le comportement d'une autre machine, dont elle est capable de calculer la configuration ; de sorte que cette autre machine sera, selon la définition donnée plus haut, " une partie de l'objet de la machine à ce moment-là ". Or, la machine dont la configuration est calculée par la machine universelle est constitutive de cette dernière - les deux machines partagent le même " ruban ". C'est ainsi, explique Turing, que la machine universelle peut être " utilisée pour aider à la confection de ses propres programmes ou pour prévoir les effets de modifications de sa propre structure "[161]. Dans les limites du calcul de configurations dont elle est capable, c'est-à-dire jusqu'au calcul de sa propre configuration non compris, il est possible de programmer la machine universelle de manière à ce qu'elle teste les résultats des machines dont elle simule le comportement, et modifie la démarche de celles-ci[162] ; par là enfin, il est possible de faire simuler à une machine universelle l'apprentissage " à partir de son expérience ". Ce point constituera, nous le verrons, le cœur de l'argumentation de Turing dans la dernière partie de *Computing Machinery*...

L'objection-argument " de Lady Lovelace "[163]

Turing examine ensuite une autre forme de l'argument fondé sur " l'exactitude implacable " de la machine, forme à laquelle il donne le nom de la collaboratrice de Charles

---

[159] " ... at least to the people who deal with it ", *ibid.*

[160] " If, for instance, the machine was trying to find a solution of the equation $x^2 - 40x - 11 = 0$ one would be tempted to describe this equation as part of the machine's subject matter at that moment ", *ibid.*

[161] " It may be used to help in making up its own programmes, or to predict the effect of alteration in its own structure ", *ibid.*

[162] " En observant les résultats de son propre comportement, elle peut modifier ses propres programmes pour atteindre un but de manière plus efficace " (*By observing the results of its own behaviour it can modify its own programmes so as to achieve some purpose more effectively*), *ibid.* Turing ajoute qu'au moment où il écrit, " il s'agit là de possibilités du futur proche... " (*these are possibilities of the near future, rather than utopian dreams*), *ibid.* Les machines cybernétiques que Turing, qui avait rencontré Wiener, pouvait déjà connaître, réalisaient sous une forme élémentaire la capacité pour une machine de modifier son comportement en fonction de " stimuli " extérieurs.

[163] " Lady Lovelace's objection ", *ibid.* p. 150.

Babbage, Lady Lovelace. On sait que Babbage avait conçu, au milieu du 19e siècle, les plans d'une " Machine Analytique " qui ne devait jamais voir le jour, du fait de sa complexité et de la difficulté que présentait sa réalisation avec les moyens techniques de l'époque. Cette machine est souvent considérée comme l'ancêtre des ordinateurs modernes, dans la mesure où elle intégrait des programmes, sous la forme de cartes perforées analogues à celles des métiers à tisser Jacquard, et permettait de stocker les données numériques qu'elle manipulait. Lady Lovelace[164] déclarait, à propos de la machine de Babbage : " La Machine Analytique n'a pas la prétention de donner naissance à quoi que ce soit. Elle peut effectuer *tout ce que nous savons* lui ordonner de faire "[165]. Cette formule peut conduire à une objection que Turing exprime de la manière suivante : "...une machine ' ne peut jamais rien faire de vraiment nouveau ' "[166]. En d'autres termes, la machine est incapable *d'invention* et ne saurait donc jamais nous surprendre[167].

Mis sous cette forme, l'argument devient une véritable " objection ", c'est-à-dire qu'il invalide le principe même du test : une machine ne pourra jamais réussir celui-ci puisque cela suppose qu'elle surprenne son examinateur humain ; or, nous avons vu que si la machine n'a aucune chance de l'emporter, le test n'a plus à être mené. L'objection repose manifestement sur l'idée que le nouveau renvoie à l'absolument indéterminé, et, comme telle, peut être considérée, à l'instar des objections précédemment discutées par Turing, comme une expression de l'opinion commune à propos des machines. S'il renvoie à l'indéterminé, le nouveau, comme le miracle, échappe à tout discours ; il n'est l'objet que d'une expérience incommunicable - hors d'une foi - et l'on ne voit pas, en effet, comment la machine envisagée pour le test pourrait prétendre l'atteindre.

Comme dans le cas de l'argument " de la conscience ", cependant, les conséquences dépassent le cadre mis en place par la problématique du test. Par l'idée que la machine ne fera jamais autre chose que ce pourquoi elle a été programmée, nous entendons en général, soit qu'elle n'est en mesure d'effectuer que ce que nous avons explicitement prévu de lui faire

[164] Lady Lovelace était la fille de Lord Byron.
[165] " The Analytical Engine has no pretentions to *originate* anything. It can do *whatever we know how to order it to perform* " Cité par Turing, *ibid.*, p. 150. C'est Lady Lovelace qui souligne.
[166] " ... a machine can ' never do anything really new ' ", *ibid.*
[167] Turing examine à nouveau l'argument en mai 1951 et en janvier 1952, au cours des deux émissions consacrées à la question des " machines pensantes " auxquelles la BBC l'avait invité.

faire, soit qu'il est toujours possible de rendre raison, au moins *a posteriori*, de ce qu'elle fait à partir du programme qui la définit. Toutefois, pouvons-nous réellement prévoir toutes les conséquences de ce que nous avons ordonné à une machine de faire ? La discussion du problème de l'erreur n'a-t-elle pas, précisément, montré que non[168] ? " Les machines me prennent très fréquemment par surprise, note Turing. Ceci est dû pour une large part à ce que je ne fais pas de calculs suffisants pour décider de ce à quoi je peux m'attendre de leur part... "[169]. Il se peut que ce que nous entendions faire faire à la machine, ce pour quoi nous l'avons programmée, conduise celle-ci à des résultats qui nous surprennent. Or, si l'on parvient à rendre raison *a posteriori* des résultats imprévus de l'action d'une machine, en montrant, par une démarche analytique, qu'ils étaient inscrits dans le programme qui la décrit, aura-t-on, par là, rendu compte de *l'événement* que constituent ces résultats imprévus ? Ceux-ci n'en auront pas moins été appréhendés comme surprenants. En d'autres termes, il existe bien une forme de " nouveau " qui renvoie au déterminisme caractéristique de la machine, et rien ne permet, dans le cadre du test, de distinguer la surprise ainsi provoquée par la machine de celle que peut produire un comportement humain ; affirmer que le nouveau, dans le cas de l'homme, est d'une autre nature, car il ne relève pas de ce déterminisme, n'indique pas les moyens dont peut disposer l'examinateur A du jeu de l'imitation pour distinguer ce nouveau spécifiquement humain de celui que produit la machine : " qui peut être certain, demande Turing, que le ' travail original ' qu'il a effectué n'était pas simplement la croissance de la semence plantée en lui par l'enseignement, ou la conséquence du fait d'avoir des principes généraux bien connus ? "[170]. C'est pourquoi, Turing s'estime en droit de renvoyer l'objection au simple dicton selon lequel " il n'y a rien de nouveau sous le soleil "[171].

---

[168] " Il arrive qu'une machine à calculer fasse quelque chose de plutôt étrange que nous n'attendions pas. En principe, nous aurions pu le prévoir, mais en pratique, c'est habituellement trop compliqué. Il est manifeste que si l'on pouvait prévoir tout ce qu'un ordinateur allait faire, on pourrait tout aussi bien le faire sans lui " (*Sometimes a computing machine does do something rather weird that we hadn't expected. In principle one could have predicted it, but in practice it's usually too much trouble. Obviously if one were to predict everything a computer was going to do one might just as well do without it.*) Emission du 14 janvier 1952. BBC Written Archives Centre, *op. cit.*

[169] " Machines take me by surprise with a great frequency. This is largely besause I do not do sufficient calculation to decide what to expect them to do... ". . A. M. Turing, *Computing Machinery..., op. cit.,* p.150.

[170] " Who can be certain that ' original work that he has done was not simply the growth of the seed plantes in him by teaching, or the effect of following well-known general principles ? ", *ibid.*

[171] " There is nothing new under the sun ", *ibid.*

***L'argument " de la continuité dans le système nerveux "*[172]**

Est-ce à dire que, dans le contexte spécifique du " jeu ", toute différence *a priori* doit être refusée entre l'homme et la machine ? L'une des vertus du test, aux yeux de Turing, est, sans aucun doute, qu'il permette d'écarter, face à la question " les machines peuvent-elles penser ? ", l'usage de tout principe d'ordre métaphysique ; cependant, indépendamment de cette dimension particulière, et bien que le test soit conçu de manière à ce que le " corporel ", hormis une certaine forme matérielle du langage, n'y intervienne pas en tant que tel, son résultat ne sera-t-il pas déterminé par la simple différence d'ordre physique qu'introduit entre l'homme et la machine le fait que, contrairement à celle-ci, le système nerveux ne soit probablement pas un organe " à états discrets " ?[173] Turing précise, par exemple, que dans le cas du cerveau, une

> petite erreur dans l'information, à peu près de la taille d'une impulsion nerveuse heurtant un neurone, peut faire une grande différence quant à la taille de l'impulsion de sortie. On peut soutenir que, puisqu'il en est ainsi, il ne faut pas s'attendre à pouvoir imiter le comportement du système nerveux avec un système à états discrets.[174]

Comme les précédents, cet argument doit, toutefois, être examiné du point de vue du jeu de l'imitation : le fait que la machine universelle soit à états discrets peut-il garantir que l'interrogateur humain A sache la distinguer, au cours du jeu, du protagoniste B ? Turing fait remarquer à ce sujet que l'écart introduit, entre les réponses de la machine et celle d'un individu humain, par cette différence objective, ne sera pas perceptible par un homme : " si nous acceptons les conditions du jeu de l'imitation, l'interrogateur ne pourra pas tirer avantage de cette différence "[175].

Supposons, imagine Turing, que l'on demande à une machine de donner la valeur de π. Nous sommes capables de construire " un analyseur différentiel ", c'est-à-dire ce que nous

---

[172] " Argument from continuity in the nervous system ", *ibid.* p. 151.
[173] " le système nerveux n'est certainement pas une machine à états discrets " (*The nervous system is certainly not a discrete-state machine*), *ibid.*
[174] " A small error in the information about the size of a nervous impulse impining on a neuron, may make a large difference to the size of the outgoing impulse. It may be argued that, being so, one cannot expect to be able to mimic the behaviour of the nervous system with a discrete-state system ", *ibid.*
[175] " ... if we adhere to the conditions of the imitation game, the interrogator will not be able to take any advantage of this difference ", *ibid.*

appelons aujourd'hui une machine analogique[176], qui donnera une valeur précise, par exemple 3,1416, qu'une machine à états discrets ne pourra jamais reproduire. Or la machine de Turing ne peut être universelle que parce qu'elle est à états discrets ; cela seul lui permet de " connaître " - c'est-à-dire de " calculer " - les autres machines par le biais des configurations qui les définissent. Cependant, il est possible de faire en sorte que la machine à états discrets sache choisir entre différentes valeurs approchées, par exemple 3,12 ; 3,13 ; 3,14 ; 3,15 ; 3,16, en tenant compte de l'aspect statistique - " avec, précise Turing, des probabilités, disons, de 0,05 ; 0,15 ; 0,55 ; 0,19 ; 0,06 "[177]. " Dans ces circonstances, souligne-t-il, il serait très difficile pour l'interrogateur de distinguer l'analyseur différentiel de l'ordinateur digital "[178]. Aux yeux de l'examinateur humain, la différence entre le continu et le discontinu ne sera pas suffisamment " discrète " pour qu'une signification puisse lui être accordée… Le principe est ici le même que celui mis en avant dans *On Computable Numbers*... pour admettre un nombre fini d'" états d'esprit " chez l'individu humain qui calcule[179]. Le fait que cela soit établi à propos du calcul ne réduit pas la portée de l'argument : si la conclusion est juste dans le cas du calcul, où les réponses données sont précises, elle le sera *a fortiori* dans des domaines où l'on n'attend pas de réponses aussi univoques.

L'argument " de l'informalité du comportement "[180]

Plus fondamentalement, le problème examiné ici - une machine est-elle capable de simuler le comportement humain dans ce qui en est communément attendu, à savoir la manifestation d'une appartenance à un ordre autre que celui du déterminisme mécanique ? - peut être envisagé sous l'angle de la question générale de " l'informalité du comportement ".

---

[176] " un analyseur différentiel est un type de machine qui n'est pas à états discrets, et qu'on utilise pour certains types de calculs " (*A differential analyser is a certain kind of machine not of the discrete-state type used for some kinds of calculation*), *ibid.* Le signal est traité par un " analyseur différetiel " - une machine analogique - non à partir de la détermination de deux états, mais à partir des variations continues d'une grandeur physique.

[177] " ... with the probabilities of 0,05, 0,15, 0,55, 0,19, 0,06 (say)... ", *ibid.*, p. 152.

[178] " Under these circumstances it would be very difficult for the interrogator to distinguish the diferential analyser from the digital computer ", *ibid.*

[179] Voir ci-dessus, 1e partie, chapitre II.

[180] " The argument from informality of behaviour ". A. M. Turing, *Computing Machinery..., op. cit.*, p. 152. La traduction de Patrice Blanchard, " L'argument du comportement informalisable " (Alan Turing, " Les ordinateurs et l'intelligence ", in Jean-Yves Girard, *La machine de Turing*, *op. cit.*), exprime très directement le sens qu'a l'argument pour Turing, mais elle s'éloigne du texte.

“ Il n’est pas possible de créer un ensemble de règles qui ait la prétention de décrire ce qu’un homme devrait faire dans tout ensemble concevable de circonstances ”[181], admet Turing. Il en est généralement conclu, poursuit-il, que “ si chaque homme disposait d’un ensemble défini de règles de conduite d’après lesquelles il organiserait sa vie, il ne serait pas supérieur à la machine. Mais de telles règles n’existent pas, ainsi les hommes ne peuvent pas être des machines ”[182].

L’argument repose, comme le souligne Turing, sur le fait que nous sommes incapables d’énumérer complètement les “ règles de conduites ” qui déterminent le comportement humain. Cependant, remarque-t-il, plutôt que de “ règles de conduites ”, ne pourrait-on parler, ici, de “ lois du comportement ” ? “ Par ‘ lois du comportement ’, précise Turing, j’entends des lois de la nature comme celles qui s’appliquent au corps humain... ”[183]. Si nous substituons, dans la formulation de l’argument, l’expression “ lois du comportement qui règlent sa vie ” à l’expression “ règles de conduite d’après lesquelles il règle sa vie ”, l’argument devient : “ si chaque homme disposait d’un ensemble défini de *lois du comportement qui règlent sa vie*, il ne serait pas supérieur à la machine. Mais de telles *lois* n’existent pas, ainsi les hommes ne peuvent pas être des machines ”. Qu’en est-il, dans ce cas, de la première proposition : “ si chaque homme disposait d’un ensemble défini de lois du comportement qui règlent sa vie, il ne serait pas supérieur à la machine ” ? Cette proposition est fondée sur l’idée d’“ ensemble défini ” : la conviction de l’opinion commune à propos de la machine est que celle-ci a pour caractéristique d’être déterminée par un “ ensemble défini ” de règles. Cependant, la valeur de vérité de la proposition, au regard de l’opinion commune, sera la même, que cet ensemble défini soit constitué de “ lois ” ou de “ règles ”.

Examinons maintenant la proposition : “ Mais de telles lois n’existent pas ”. S’il s’agit de lois et non plus de règles, elle est illégitime ; “ nous ne pouvons pas nous convaincre aussi facilement de l’absence de lois complètes du comportement que de règles complètes de conduite ”[184]. Rien n’interdit, en effet, de supposer qu’il existe un ensemble complet de

[181] “ It is not possible to produce a set of rules purporting to describe what a man should do in every conceivable set of circumstances ”, *ibid.*
[182] “ If each man had a definite set of rules of conduct by which he regulated his life he would be no better than a machine. But there are no such rules, so men cannot be machines ”, *ibid.*
[183] “ By ‘ laws of behaviour ’ I mean laws of nature as applied to a man’s body... ”, *ibid.*
[184] “ ... we cannot so easily convince ourselves of the absence of complete laws of behaviour as of complete rules of conduct ”, *ibid.*.

“ lois ” : “ la seule manière que nous connaissions de découvrir de telles lois est l’observation scientifique, et nous ne pouvons certainement pas imaginer de circonstances nous permettant de dire : ‘nous avons assez cherché, de telles lois n’existent pas’ ”[185]. Cette idée sera sans doute facilement acceptée lorsqu’il s’agit du comportement humain. Or, elle doit l’être aussi dans le cas du comportement de la machine : “ supposons, déclare Turing, que nous puissions être sûrs de découvrir de telles lois si elles existent. Alors, à partir d’une machine à états discrets donnée, il devrait être possible de découvrir, par l’observation, assez d’éléments à son sujet pour prédire son comportement futur, et ceci dans une période de temps raisonnable, disons mille ans [!] ”[186]. Cependant, ajoute-t-il, “ ... il ne semble pas que ce soit le cas ”[187]. Se fondant sur sa propre expérience de programmeur, Turing met en effet quiconque au défi d’en apprendre suffisamment, par l’observation, au sujet d’un programme même simple, pour pouvoir prédire la réponse que fera ce programme “ pour des valeurs non encore utilisées ”[188]. En d’autres termes, nous ne pouvons pas, par l’observation, connaître les *lois du comportement* de la machine mieux que celles du comportement de l’homme. La conclusion de l’argument - “ les hommes ne peuvent pas être des machines ” - ne peut donc pas être tirée.

Bref, dans la position où il est, l’examinateur du jeu de l’imitation n’a aucun moyen de considérer que le comportement du protagoniste C - la machine - relève de “ règles ” susceptibles d’être dénombrées complètement, et non pas de “ lois ”, pour lesquelles il acceptera un dénombrement incomplet ; rien ne lui permet d’utiliser un critère différent pour C et pour B, et donc de les distinguer. Si C simule le comportement attendu d’un individu humain - et la discussion des arguments précédents tend à montrer que cela est possible - A sera donc dans la situation de devoir faire *comme si* C était un individu humain.

---

185 “ The only way we know for finding such laws is scientific observation and we certainly know no circumstances under which we could say, ‘we have searched enough. There are no such laws’ ”, *ibid.*

186 “ ... suppose we could be sure of finding such laws if they exist. Then given a discrete-state machine it shoud certainly be possible to discover by observation sufficient about it to predict its future behaviour, and this within a reasonable time, say a thousand years ”, *ibid.,* p. 153.

187 “ ... this does not seem to be the case ”, *ibid.*

188 “ I have set up on the Manchester computer a small programm using only 1000 units of storage, whereby the machine supplied with one sixteen-figure number replies with another within two seconds. I would defy anyone to learn from these replies sufficient about the programme to be able to predict any replies to untried values ”, *ibid.*

La question examinée consistait, nous l'avons vu, à se demander s'il était théoriquement possible qu'une machine universelle, fasse preuve, dans le cadre du jeu de l'imitation, d'une habileté suffisante pour l'emporter face à son adversaire humain. La discussion menée par Turing des " objections " et " arguments " examinés ci-dessus établit qu'il n'est nullement contradictoire avec la notion scientifique de machine universelle qu'une machine de ce type simule les effets par lesquels une telle habileté peut se manifester dans un contexte comme celui du jeu de l'imitation. Il n'est rien, dans la notion de machine universelle qui interdise, sur le plan théorique, qu'une machine puisse être " l'objet de ses propres pensées " et qu'elle " apprenne à partir de son expérience ", ou qu'elle commette des erreurs, et agisse d'une manière qui, aux yeux de l'examinateur A, présente un aspect déconcertant analogue à celui qui peut être attendu du comportement d'un individu humain. De là ressort qu'il n'est pas, aux yeux de Turing, théoriquement contradictoire avec la notion de machine universelle qu'une machine de ce type " fasse bonne figure " à un test tel que celui du sonnet, et, *a fortiori*, compte tenu de la nature même de ce test, à tout autre " test partiel " que l'on pourrait imaginer. Or, dès lors qu'une machine peut l'emporter à un " test partiel de Turing ", son comportement peut être simulé par une autre machine qui l'emportera elle-même à un autre " test partiel ". Une machine universelle théorique peut donc être imaginée qui soit capable de simuler le comportement de toutes les machines l'emportant à des " tests partiels ", ou, plus concrètement, sous l'angle du résultat à atteindre, qui soit capable de simuler le comportement d'un nombre suffisant, pour vaincre A, de machines l'emportant à des " tests partiels ".

Plusieurs problèmes, cependant, demeurent. On remarquera tout d'abord que la discussion menée jusque là par Turing a seulement mis en évidence la possibilité théorique, pour l'examinateur A du jeu de l'imitation, de se tromper, au cours de celui-ci, sur le statut de C, C étant une machine universelle ; il n'a pas été prouvé pour autant que le statut de la machine, parce qu'elle " réussit " le test, appartienne à un autre registre que celui d'un *outil*. Rien n'interdit de considérer que ce ne soit pas véritablement la machine qui gagne au jeu de l'imitation, mais plutôt l'homme qui perde, parce qu'il commet une erreur de manipulation. Certes, la machine l'emporte parce que l'examinateur A se trompe, mais celui-ci ne se trompe-t-il pas à peu près comme l'automobiliste qui appuie sur la pédale d'accélérateur

plutôt que sur celle de frein ? Mal se servir d'un outil ne signifie pas que celui-ci ait une action propre à laquelle l'erreur que l'on commet pourrait être attribuée.

On notera ensuite que, si la discussion des " objections " et des " arguments " établit qu'il n'y a pas de contradiction théorique entre la notion de machine universelle et les conditions d'une victoire de C au jeu de l'imitation, elle ne dit rien quant à la manière dont C, en tant que machine universelle, peut être effectivement conçue.

Ces deux difficultés sont traitées ensemble par Turing à l'aide d'une hypothèse auxiliaire, celle des " machines qui apprennent " : la machine susceptible de l'emporter au jeu de l'imitation doit être conçue comme une " machine qui apprend " ; la réalisation effective d'une telle machine devra consister, à partir de la construction de la machine simple constituant son " état initial ", à la soumettre à un processus d'apprentissage analogue, dans son ampleur et sa complexité, à celui que suit un individu humain quelconque, lui-même susceptible de l'emporter au jeu. Comme un individu humain, la machine devra être " éduquée ". Dès lors, rien n'autorisera à affirmer une différence de nature entre la victoire d'une machine ainsi conçue au jeu de l'imitation, et celle, dont la possibilité a été établie, d'un individu humain.

Section III : L'hypothèse des " machines qui apprennent "

Après avoir discuté objections et arguments opposés à l'hypothèse de la victoire d'une machine au jeu de l'imitation, Turing déclare : " Le lecteur aura compris, que je n'ai pas d'argument positif très convaincant pour soutenir mon point de vue. Si j'en avais, je n'aurais pas pris tant de peine à montrer les erreurs des points de vue opposés au mien. Les preuves que j'ai, je vais maintenant les donner "[189]. Pour administrer ces " preuves ", Turing commence par préciser ce que l'expérience du jeu de l'imitation est censée faire apparaître quant à la machine victorieuse, et, à ce propos, évoque de nouveau l'argument dit " de Lady Lovelace ". Il s'agit en vérité, pour Turing, d'introduire , ici, l'idée que l'action de la machine peut, dans certaines conditions, franchir un " seuil critique ", au-delà duquel cette action ne peut plus être contrôlée par les constructeurs de la machine à l'aide des moyens logico-mathématiques – les " m-configurations " de la " machine de Turing " - qui définissent la notion de " programme ".

La machine " surcritique " et l'hypothèse des " machines qui apprennent "

Une machine, déclarait Lady Lovelace - en l'occurrence la " machine analytique " de Babbage – " peut effectuer *tout ce que nous savons lui ordonner* de faire "[190]. Examinant, dans la partie précédente[191], l'argument construit sur cette remarque, Turing citait un commentaire d'Hartree pour qui " cela n'implique pas qu'il ne soit pas possible de construire des machines

[189] " The reader will have anticipated that I have no very convincing argument of a positive nature to support my views. If I had I should not have taken such pains to point out the fallacies in contraries views. Such evidence as I have I shall now give ". A. M. Turing, *Computing Machinery..., op. cit.*, p. 154.
[190] " It can do *whatever we know how to order it to perform* ", *ibid.*, p. 150. Les italiques, précise Turing, sont de Lady Lovelace.
[191] Voir ci-dessus, 2e partie, chapitre 1, section 2.

électroniques qui ‘penseront par elles-mêmes’... ”[192]. Il fait remarquer, par ailleurs, que la déclaration de lady Lovelace ne contient pas le mot “ seulement ”, contrairement à l’argument généralement utilisé - celui qu’il discutait - selon lequel “ la machine peut faire *seulement* ce que nous savons lui ordonner de faire ”[193]. Pour illustrer la différence qu’il veut voir entre “ *tout* ce que nous savons lui ordonner de faire ” et “ *seulement* ce que nous savons lui ordonner de faire ”, Turing suggère que l’“ on pourrait dire qu’un homme peut ‘injecter’ une idée dans la machine, laquelle réagira jusqu'à un certain point, puis retournera à l’immobilité, comme une corde de piano frappée par un marteau ”[194]. Dans certaines circonstances, l’idée “ injectée ” pourra provoquer une réaction sans commune mesure avec sa portée propre, comme l’illustre l’exemple de la “ pile atomique ” :

> “ un autre point de comparaison serait une pile atomique d’une masse inférieure à la masse critique : une idée injectée correspondra à un neutron entrant dans la pile, en provenance de l’extérieur. Tout neutron de ce type produira une certaine perturbation qui finira par cesser. Si toutefois la masse de la pile est suffisamment accrue, la perturbation créée par l’entrée d’un tel neutron continuera probablement à s’accroître jusqu’à ce que toute la pile soit détruite ”[195].

Turing imagine qu’un tel “ point critique ” existe pour l’esprit humain. “ La majorité d’entre eux [NB : les humains] paraissent ‘ souscritiques ’ ”[196] et semblent correspondre, dans cette analogie, aux piles à masse souscritique. Une idée proposée à de tels esprits, déclare Turing, “ donnera lieu en moyenne à l’apparition de moins d’une idée en réponse ”[197]. Une faible proportion d’êtres humains est “ surcritique ” : une idée proposée à ceux-ci “ pourra donner

---

[192] “ This does not imply that it may not be possible to construct electronic equipment which will ‘think for itself’... ”, *ibid.,* p. 150. Turing ne précise pas la source de cette citation.

[193] “ The machine can only do what we know to order it to do. ”. En note, Turing précise : “ Compare Lady lovelace’s statement which does not contain the word ‘only’ ”, *ibid.,* p. 159.

[194] “ One could say that a man can ‘inject’ an idea into the machine, and that it will respond to a certain extent and then drop into quiescence, like a piano string struck by a hammer ”, *ibid.,* p. 154.

[195] “ Another simile would be an atomic pile of less than critical size : an injected idea is to correspond to a neutron entering the pile from without. Each such neutron will cause a certain disturbance which eventually dies away. If, however, the size of the pile is sufficiently increased, the disturbance caused by such an incoming neutron will very likely go on and on increasing until the whole pile is destroyed ”, *ibid*.

[196] “ The majority of them seem to be ‘ subcritical ’ ”, *ibid.*

[197] “ An idea presented to such a mind will on average give rise to less than one idea in reply ”, *ibid*.

lieu à l'apparition de toute une ' théorie ' constituée d'idées secondaires, tertiaires ou encore plus éloignées "[198]. Ce n'est pas ici la distinction entre une catégorie d'humains et une autre qui importe ; nous devons plutôt comprendre qu'un esprit humain, tant qu'il est " souscritique ", n'a pas beaucoup d'idées, alors que, sitôt qu'il devient " surcritique ", nous ne sommes plus en mesure de fixer une limite à l'enchaînement de ses idées.

Le principe énoncé ici est éclairé à partir d'une autre analogie : celle dite de " la peau de l'oignon ". Il a été montré lors des discussions précédentes qu'en considérant " les fonctions de l'esprit ou du cerveau, nous trouvons certaines opérations qui peuvent s'expliquer en termes purement mécaniques "[199] ; à ce constat, fait par Turing au cours de la première partie de son argumentation, il est généralement objecté que nous n'appréhendons pas par là " l'esprit réel "[200], mais seulement " une espèce de peau que nous devons enlever si nous voulons trouver l'esprit réel "[201]. Cependant, ajoute Turing, la question est de savoir ce que nous rencontrons après avoir enlevé cette première peau. Sous la première peau, nous pouvons trouver d'autres fonctions susceptibles, en tant que telles, d'être expliquées également en termes mécaniques ; une seconde peau devra être enlevée. Aussi, demande Turing, " en continuant de cette manière, arrivons-nous jamais à l'esprit 'réel', ou arrivons-nous finalement à la peau qui ne contient rien ? "[202]. Si, en particulier, nous sommes en mesure de montrer qu'une machine universelle peut devenir " surcritique ", ne serons-nous pas, précisément, " en continuant de cette manière ", devant une " peau qui ne contient rien " ? Le caractère " surcritique " de l'esprit humain ne pourra plus, en effet, être attribué à un principe situé au-delà de l'explication mécanique donnée pour chacun des niveaux de fonctions qui auront été examinés. Le problème du test est ainsi précisé : l'entité susceptible

---

[198] " An idea presented to such a mind that may give rise to a whole ' theory ' consisting of secondary, tertiary and more remote ideas ", *ibid.* La distinction faite, ici, entre humains " souscritiques " et humains " surcritiques " ne signifie naturellement pas qu'il y ait aux yeux de Turing deux espèces différentes d'intelligence humaine, comme si la " faible proportion " d'individus humains " surcritiques " appartenaient à une autre catégorie que " la majorité " de ceux qui " paraissent souscritiques ", mais plutôt que le fait d'être humain implique que l'on puisse être " surcritique ".

[199] " In considering the functions of the mind or the brain we find certain operations which we can explain in purely mechanical terms ", *ibid.*

[200] " the real mind ", *ibid.*

[201] " ... a sort of skin we must strip off if we are to find the real mind ", *ibid.*

[202] " Proceeding in this way do we ever come to the 'real' mind or do we eventually come to the skin which has nothing in it ? ", *ibid.* p. 155.

de l'emporter au jeu de l'imitation doit pouvoir être " surcritique ", et la question désormais posée est celle de savoir si l'on peut " rendre une machine surcritique "[203].

Turing convient que " le seul élément vraiment satisfaisant qui puisse soutenir le point de vue exprimé au début de la section 6 [NB : l'hypothèse de la victoire de la machine au jeu de l'imitation] nous sera fourni par la réalisation […] de l'expérience décrite ".[204] Cependant, à partir du moment où l'on est assuré qu'il n'y a pas de contradiction entre l'hypothèse et la définition scientifique de la machine, la non-vérification de l'hypothèse n'infirme pas celle-ci en tant que telle. Certes, tant qu'une machine n'aura pas effectivement réussi le test, rien n'aura été prouvé, ni dans un sens ni dans un autre, mais l'hypothèse n'en restera pas moins valide. Il ne s'agit pas, en effet, d'établir que la machine ne saurait penser, mais au contraire d'infirmer cette idée ; aussi, tant qu'il n'aura pas été prouvé, par exemple à l'aide d'une autre expérience à déterminer, qu'une machine ne peut penser, l'hypothèse de la victoire de la machine au jeu de l'imitation, et la signification que lui accorde Turing, pourront être admises. Il sera donc permis de raisonner à partir de cette hypothèse[205]. Il est légitime, autrement dit, de se demander, sans attendre la réalisation effective de l'expérience, quelles implications comporte l'hypothèse d'une victoire de la machine au jeu de l'imitation[206].

A travers la notion de machine " surcritique " apparaît l'idée que la machine n'est pas nécessairement une entité statique, définie *a priori* par son programme : la machine " surcritique " peut évoluer, et la limite de son évolution ne peut, quant à elle, être fixée *a priori*. Cette évolution, toutefois, n'en est pas moins susceptible d'être ordonnée : si, on l'a vu, la machine, comme l'individu humain, est soumise à l'erreur, elle peut, comme celui-ci,

[203] " can a machine be made to be supercritical ? ", *ibid.*

[204] " The only really satisfactory support that can be given for the view expressed at the beginning of § 6, will be that provided by waiting for the end of the century and then doing the experiment described… ", *ibid.* p. 156.

[205] " L'idée populaire selon laquelle les savants avancent inexorablement d'un fait bien établi à un autre, sans être influencés par des hypothèses non vérifiées, est absolument fausse, déclare Turing. Pourvu que nous sachions clairement quels sont les faits prouvés et quelles sont les hypothèses, aucun mal ne peut en résulter. Les hypothèses sont de grande importance puisqu'elles suggèrent d'utiles voies de recherches " (*The popular view that scientists proceed inexorably from well-established fact to well-established fact, never being influenced by any unproved conjecture, is quite mistaken. Provided it is made clear which are proved facts and wich are conjectures, no harm can result. Conjectures are of great importance since they suggest useful lines of research*). *Ibid.*, p. 149.

[206] " Mais que pouvons-nous dire en attendant ? [NB : que l'expérience du " jeu " soit réalisée] Quelle démarche devrions-nous entreprendre maintenant si l'expérience devait être couronnée de succès ? ", (*But what can we say in the meantime ? What steps should be taken now if the experiment is to be successful ?*). *Ibid.*, p. 155.

corriger les erreurs qu'elle commet ; ce qu'une machine donnée ne peut faire, une machine plus puissante, capable en outre de la simuler, sera en mesure de le faire. Une machine qui commettra des erreurs, c'est-à-dire une machine dont le comportement est en quelque manière imprévisible, sera ainsi capable d'" apprendre ". C'est là, précisément, on s'en souvient, l'un des " différents traits de X " dont la discussion précédente a montré qu'il n'entrait pas en contradiction avec la notion de machine universelle : cette machine peut " se prendre pour objet " et " apprendre à partir de son expérience ". C'est pourquoi Turing va désormais reprendre et développer ce point : si une machine, en effet, peut " apprendre ", alors, elle doit pouvoir être " surcritique ".

Les machines " inorganisées " et les " machines qui apprennent "

Il s'agit tout d'abord de dégager le terrain en montrant que les contraintes d'ordre physique ne sont pas déterminantes pour la question théorique soulevée par la notion de " machine qui apprend ". Turing souligne, par exemple, que " les pièces des machines modernes qui peuvent être considérées comme analogues aux cellules nerveuses fonctionnent à peu près mille fois plus vite que ces dernières "[207]. Il montre, d'autre part, que, s'agissant des capacités de stockage nécessaires - la " mémoire " de la machine – les machines existant en 1950 sont déjà très certainement satisfaisantes[208]. En d'autres termes, du simple point de vue des performances physiques, en l'état même de la technique à l'époque où Turing rédige

[207] " Parts of modern machines which can be regarded as analogs of nerve cells work about a thousand times faster than the latter ", *ibid.* Turing n'ignorait pas les recherches menées dans les années quarante à Chicago sur la simulation des neurones par McCulloch et Pitts. C'est à partir du travail de McCulloch et Pitts que Kleene mit au point le concept d'automate fini (voir à ce sujet : Jean Mosconi, *La constitution de la théorie des automates*, *op. cit.*).

[208] " Les estimations de la capacité de stockage du cerveau varient de 10^{10} à 10^{15} chiffres binaires. Je penche pour les valeurs les plus basses, et je crois que seule une très petite partie est utilisée pour les types les plus élevés de pensée. La plus grande partie sert probablement à la conservation des impressions visuelles. Je serais surpris que plus de 10^9 soit nécessaire pour jouer de manière satisfaisante au jeu de l'imitation, du moins contre un aveugle (note : la capacité de l'Encyclopaedia Britannica, 11e édition, est de 2 x 10^9). Une capacité de stockage de 10^7 serait une possibilité très réalisable, même avec les techniques actuelles ", (*Estimates of the storage capacity of the brain vary from 10^{10} to 10^{15} binary digits. I incline to the lower values and believe that only a very small fraction is used for the higher types of thinking. Most of it is probably used for the retention of visual impressions. I should be surprised if more than 10^9 was required for satisfactory playing of the imitation game, at any rate against a blind man. (Note : The capacity of the Encyclopaedia Britannica, 11th edition, is 2 X 10^9.) A storage capacity of 10^7 would be a very practicable possibility even by present techniques)*. A. M. Turing, *Computing Machinery..., op. cit.,* p. 155.

Computing Machinery..., on peut estimer que la machine n'est guère inférieure à l'homme. Le problème soulevé par l'idée d'une " machine qui apprend " est donc " surtout un problème de programmation "[209].

Sous cet angle, les réponses données précédemment à la question de l'" invention " et du " nouveau ", doivent être précisées. En effet, admet Turing,

> l'idée d'une machine qui apprend peut paraître paradoxale à certains lecteurs. Comment les règles d'opération de la machine peuvent-elles changer ? Elles devraient décrire complètement la manière dont la machine réagira, quelle que soit son histoire, quels que soient les changements qu'elle puisse subir. Les règles ne varient donc pas du tout dans le temps[210].

Par définition, tout processus d'apprentissage suppose l'évolution du comportement du sujet apprenant ; si la machine " apprend ", son comportement doit être soumis à des changements. Or, l'action de la machine est déterminée par un système de règles et tout changement dans son comportement renverra à ce système de règles. En tant que telles, ces dernières ne sauraient donc " évoluer ". Bref, comment la machine pourrait-elle " apprendre ", dès lors qu'elle est définie par un programme ? Turing avait déjà répondu à cette question dans le rapport intitulé *Intelligent Machinery*[211], qu'il avait rédigé en 1947 à l'intention du *National Physical Laboratory*.

Il y reprenait la possibilité, mentionnée dans *On Computable Numbers*..., de concevoir une " machine de Turing " non déterministe, notion qu'il utilisait dans sa thèse américaine - *Systems of Logic Based on Ordinals* - pour raisonner sur la " machine à oracle " :

> " Pour certains objectifs, nous pouvons utiliser des machines (machines à choix ou *c*-machine), dont le mouvement n'est que partiellement déterminé par sa configuration... Lorsqu'une telle machine atteint l'une de ces configurations ambiguës, elle ne peut continuer tant qu'un choix arbitraire n'a pas été fait par un opérateur extérieur. "[212].

---

[209] " ... the problem is mainly one of programming ", *ibid.*

[210] " The idea of a learning machine may appear paradoxical to some readers. How can the rules of operation of the machine change ? They should describe completely how the machine will react whatever its history might be, whatever changes it might undergo. The rules are thus quite time-invariant. This is quite true ", *ibid.*, p. 158.

[211] A. M. Turing, *Intelligent Machinery*, *op. cit.*

[212] " For some purposes we might use machines (choice machines or c-machines) whose motion is only partially determined by the configuration. When such a machine reaches one of these ambiguous configurations, it cannot go on until some arbitrary choice has been made by an external operator ". A. M. Turing, *On Computable Numbers*..., *op. cit.*,

Dans *Intelligent Machinery*, Turing utilise cette même notion pour décrire une machine " aléatoire " (" random machine "). A certaines étapes du processus suivi par une telle machine, explique-t-il, deux choix, au moins, sont possibles, et la machine est conçue de manière à ce que l'un ou l'autre choix soit déterminé par une procédure faisant intervenir le hasard - un équivalent électronique du coup de dé. Turing distingue deux types de machines aléatoires : la machine réellement aléatoire et la machine " partiellement aléatoire " (" partially random machine "), où le " coup de dé électronique " est remplacé par une procédure bien définie, telle que, par exemple, un calcul fondé sur le nombre π. Lorsqu'une machine aléatoire n'est pas conçue en vue d'un but déterminé, Turing lui donne le nom de " machine inorganisée " (" unorganized machine "). Il propose également un exemple de ce que l'on peut entendre par là. Imaginons une machine constituée d'un nombre *n* d'unités semblables. Chaque unité dispose de deux entrées et une sortie. Celle-ci peut être ou non connectée à une entrée d'une ou plusieurs autres unités. Pour chaque entier r ($1 \leq r \leq n$), deux nombres, i(r) et j(r), sont choisis au hasard dans l'ensemble des entiers compris entre 1 et *n*. L'unité r est connectée aux unités i(r) et j(r). Les unités sont coordonnées par un dispositif qui émet des impulsions à des intervalles égaux, lesquels définissent des " moments ". A chaque moment, chaque unité peut avoir deux états. Chaque état est déterminé par le produit des états respectifs des unités à laquelle l'unité courante est connectée en entrée[213]. Puisque les états d'une telle machine sont en nombre fini, son mouvement sera périodique, et sa période ne pourra pas être supérieure à 2^n. Ce type de machine peut enfin être compliqué en imaginant que les liens entre les unités soient eux mêmes constitués de telles unités[214].

Une machine inorganisée peut être soumise à des " interférences ", c'est-à-dire à certaines conditions imposées de l'extérieur, qui entraînent une modification de son comportement. Turing indique, en particulier, qu'il serait possible d'obtenir ces modifications

---

p. 232.

[213] La formule exacte est : " The state is determined by the rule that the states of the units from which the input leads come are to be taken at the previous moment, multiplied together and the result substracted from 1 ". A. M. Turing, *Intelligent Machinery, op. cit.,* p. 114.

[214] On notera que la construction de Turing anticipe, ici, sur la conception " connexionniste " qui inspire les recherches actuelles en intelligence artificielle (voir : Margaret A. Boden (éd.), *The Philosophy of Artificial Intelligence,* Oxford, Oxford University Press, 1990).

en simulant le système " punitions-récompenses " qui, peu ou prou, est utilisé dans tout processus d'éducation d'un enfant. Sous cet angle, l'action d'une machine sera déterminée par trois éléments : sa " m-configuration ", c'est-à-dire la table la décrivant, le signal d'entrée qui lui est fourni, autrement dit le contact avec l'extérieur qui peut donner lieu à interférence, et, enfin, un " stimulus de douleur " ou de " plaisir ", correspondant à une " punition " ou à une " récompense ". Dans le cas où l'action sera considérée comme inadaptée au signal d'entrée, un " stimulus de douleur ", c'est-à-dire un signal " punition ", sera envoyé à la machine et entraînera un changement aléatoire de la m-configuration de celle-ci ; dans le cas où au contraire l'action sera considérée comme adaptée, un " stimulus de plaisir, c'est-à-dire un signal " récompense " sera envoyé à la machine, entraînant la fixation de la m-configuration de celle-ci. Le comportement que l'on vise à faire adopter par la machine peut être défini sous la forme d'un ensemble de signaux d'entrée (correspondant aux différents cas qui peuvent être rencontrés dans la situation correspondant à ce comportement). Pour chacun de ces signaux d'entrée, la machine tendra statistiquement, par essais et erreurs sanctionnés par un signal de peine ou de plaisir, à atteindre une configuration fixe. Sans doute n'est-il pas possible de définir *a priori* l'ensemble des signaux d'entrée constituant un comportement relativement complexe, même si cet ensemble est considéré comme fini ; la machine peut à tout moment se trouver face à un signal d'entrée jusque là non essayé et déterminant une réponse qui n'a pas encore été évaluée. Cependant, sur une durée suffisamment longue, la machine tendra à donner une réponse considérée comme " bonne " du point de vue de l'ensemble des signaux d'entrée définissant le comportement visé.

Par là, en particulier, une machine inorganisée peut être, en fonction des lois de la statistique, transformée en machine organisée, conçue dans un but déterminé. Plus précisément, Turing estime qu'une machine inorganisée peut être transformée en machine universelle, susceptible de simuler le comportement d'autres machines.

Une machine inorganisée peut donc évoluer. Sa transformation en machine universelle n'est pas directement inscrite dans sa structure ; elle peut être transformée en autre chose qu'une machine universelle. Son devenir dépend, en réalité, du système de punitions-récompenses qui est appliqué, ou, plus exactement, de " l'histoire " des punitions et récompenses effectivement " vécues " par elle. Sa structure fondamentale - c'est-à-dire, dans l'exemple donné par Turing, l'ensemble de *n* unités à deux entrées et une sortie, et le principe

de leur liaison - ne sera jamais modifiée par son évolution, mais les liaisons, quant à elles, seront, au départ, indéterminées. Le même principe peut être adopté à l'égard de la machine lorsqu'elle est devenue une machine universelle puisque celle-ci présente, on le sait, la particularité de pouvoir se modifier elle-même : elle peut être conçue de telle sorte que le résultat auquel elle parvient entraîne une modification de la configuration de la machine simulée. On peut donc imaginer de pourvoir une machine universelle d'un système initial qui pourra être modifié à l'aide du processus de " punitions-récompenses " ; les modifications ne porteront pas sur la structure de la machine universelle en tant que telle, mais seulement sur ce système initial.

Or, si l'on admet la possibilité théorique d'une " machine qui apprend ", une solution apparaît au problème posé par la réalisation effective d'une machine susceptible de l'emporter au jeu de l'imitation : cette solution consiste à soumettre la machine à un processus d'éducation analogue à celui qui est suivi par un individu humain, lui-même capable, on le sait, de " bien se comporter " au jeu.

La " machine qui apprend "

" Au lieu de produire un programme qui simule l'esprit de l'adulte, pourquoi ne pas essayer plutôt d'en produire un qui simule celui de l'enfant ? "[215], demande Turing. Il est en effet raisonnable de penser que le cerveau d'un enfant sera plus facile à simuler que celui d'un adulte, dans la mesure où une grande partie des processus que celui-ci met en œuvre sont acquis au cours de l'éducation et non pas innés[216]. Turing distingue, en effet, trois composantes dans l'" esprit humain adulte "[217] : " (a) L'état initial de l'esprit, disons à la naissance. (b) L'éducation à laquelle il a été soumis. (c) D'autres expériences, que l'on ne

---

215 " Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child's ? ". A. M. Turing, *Computing Machinery..., op. cit.,* p. 156.

216 " Il est probable que le cerveau d'un enfant est quelque chose comme un carnet tel qu'on en achète chez le papetier : assez peu de mécanisme et beaucoup de feuilles blanches... Notre espoir est qu'il y ait si peu de mécanisme dans le cerveau d'un enfant que quelque chose comme cela soit très facile à programmer. " (*Presumably the child brain is something like a notebook as one buys it from the stationer's. Rather little mechanism, and lots of blank sheets... Our hope is that there is so little mechanism in the child brain that something like it can be easily programmed*), *ibid.*

217 " ... adult human mind ", *ibid.,* p. 155.

peut pas décrire comme éducatives, auxquelles il a été soumis "[218]. C'est pourquoi le problème peut être divisé " en deux parties : le programme-enfant et le processus d'éducation "[219], et la tâche de programmation proprement dite concernera uniquement la première partie.

Avant d'examiner celle-ci, Turing précise ce qu'il entend ici par " processus d'éducation ". L'" éducation " de la machine, on le sait, sera conduite à l'aide d'une simulation du système " punitions-récompenses ", de telle sorte que la probabilité associée aux " bonnes " réponses deviennent dominante[220]. Cependant, un enseignement strictement fondé sur le système punitions-récompenses, poserait un problème : " en gros, si le maître n'a pas d'autres moyens de communiquer avec l'élève [NB : que le système punitions-récompenses] la quantité d'information qui peut lui parvenir ne dépassera pas le nombre total de récompenses et punitions utilisées "[221]. Si la machine-enfant ne pouvait " apprendre " que par le biais d'un tel système, elle serait, sans doute, le plus souvent dans l'incapacité de décider si elle doit ou non faire telle chose, les réponses possibles n'ayant pas encore été significativement pondérées par le système punitions-récompenses. C'est pourquoi, déclare Turing, " il est [...] nécessaire d'avoir d'autres canaux de communication ' non-émotionnels ' " [222] : la machine-enfant devra être aidée. Lorsque cela sera possible, son constructeur devra lui fournir, sous forme d'instructions formelles, les connaissances dont elle peut avoir besoin dans telle ou telle circonstance ; il s'agira là d'un procédé " autoritaire " assez proche de l'apprentissage " par coeur " pour les individus humains. Par là, le processus d'éducation peut être rapproché, déclare Turing, de celui de l'évolution biologique :

> « Il y a un lien évident entre ce processus et l'évolution, à travers les identités suivantes :

[218] " (a) The initial state of the mind, say at birth, (b) The education to which it has been subjected, (c) Other eperience, not to be described as education, to which it has been subjected. ", *ibid*.
[219] " We have thus divided our problem into two parts. The child programme and the education process ", *ibid*.
[220] " La machine doit être construite de telle manière que les événements qui précèdent immédiatement l'apparition d'un signal-punition soient peu susceptibles de se reproduire, tandis qu'un signal-récompense accroisse la probabilité de répétition des événements qui l'ont provoqué " (*The machine has to be so constructed that events wich shortly preceded the occurrence of a punishment signal are unlikely to be repreated, whereas a reward signal increased the probability of repetition of the events which led up to it*), *ibid.*, p. 157.
[221] " Roughly speaking, if the teacher has no other means of communicating to the pupil, the amount of information which can reach him does not exceed the total number of rewards and punishments applied ", *ibid*.
[222] " It is necessary therefore to have some other ' unemotional ' channels of communication ", *ibid*.

> Structure de la machine-enfant = matériel héréditaire.
> Changement dans la machine-enfant = mutations.
> Sélection naturelle = jugement de l'expérimentateur. »[223]

L'expérimentateur sélectionnera ou non les changements intervenus dans la machine-enfant. Le processus ne sera donc pas entièrement soumis au hasard : l'expérimentateur sélectionnera les " mutations " en fonction d'un but déterminé. Le modèle mis en oeuvre est, en quelque sorte, celui d'une évolution idéale, orientée par une finalité explicite[224]. L'expérimentateur choisira, entre plusieurs " mutations ", celle qu'il jugera la plus " utile " ; il pourra, du reste, introduire lui-même dans le programme une modification destinée à provoquer la " mutation " souhaitée - laquelle pourra entraîner d'autres mutations elles-mêmes imprévues.

Cependant, la question décisive demeure celle de savoir quel doit être " l'état initial " de la " machine-enfant ". Quel " système initial " doit être fourni " en entrée " à la machine universelle ? " On pourrait, propose Turing, [...] intégrer [à la machine] un système complet d'inférences logiques "[225]. On remarquera immédiatement que, si le système initial fourni en entrée à la machine universelle est un système d'inférences logiques, le problème théorique soulevé par la notion de " machine qui apprend " se pose désormais au niveau de ce système : comment celui-ci pourra-t-il être soumis à évolution sans que ses règles soient altérées ? Dans le cas où le système initial serait un " système complet d'inférences logiques ", précise Turing,

> « la mémoire serait largement occupée par des définitions et des propositions. Les propositions auraient différents types de statuts, par exemple : des faits bien établis, des hypothèses, des théorèmes mathématiquement prouvés, des affirmations provenant d'une autorité, des expressions ayant la forme logique de propositions mais sans valeur de croyance. »[226]

---

[223] " There is an obvious connection between this process and evolution, by the identifications
Structure of the child machine = hereditary materiel
Changes of the child machine = mutations
Natural selection = judgement of the experimenter ", *ibid.,* p. 156.

[224] " s'il sait repérer la cause d'une faiblesse, il [NB : l'expérimentateur] est probablement en mesure d'imaginer le type de mutation qui l'améliorera [NB : la machine]. ", (*If he can trace a cause for some weakness he probably think of the kind of mutation which will improve it*). *Ibid*.

[225] " ... one might have a complete system of logical inference ' built in '. *ibid.,* p. 157. Dans *Intelligent Machinery,* Turing précise que ce système pourrait être celui des *Principia* de Russell et Whitehead.

Ce que Turing appelle un “ fait bien établi ” constitue ici l'élément important. C'est cela, en effet, qui réglera l'action de la machine : “ La machine devrait être construite de manière à ce que, dès qu'un impératif est classé comme ‘ bien établi ’, l'action appropriée ait automatiquement lieu ”[227]. Une hypothèse pourra être formulée sous la forme d'“ une expression logique sans valeur de croyance ” ; comme tel, ce statut ne déterminera aucune action de la part de la machine. Celle-ci n'agira que si quelque chose, par exemple le fait qu'elle soit émise par une “ autorité ”, fait passer l'expression dans la classe des “ faits bien établis ”. Il en ira de même si l'expression fait l'objet d'une “ preuve mathématique ”. De ce point de vue, note Turing, “ les processus d'inférence utilisés par la machine n'ont pas besoin d'être de nature à satisfaire les logiciens les plus exigeants. Il se pourrait par exemple qu'il n'y ait pas de hiérarchie des types ”[228]. Il est établi, pourtant, qu'en l'absence d'une théorie des types, les paradoxes logiques ne peuvent pas être évités[229]. Cependant, explique Turing, “ ceci ne veut pas obligatoirement dire que les erreurs de type auront lieu, pas plus que nous ne sommes obligés de tomber du haut de falaises non protégées ”[230] ; il ne s'agit pas de mettre la machine à l'abri des paradoxes logiques, mais de faire en sorte qu'elle parvienne, comme l'homme, soit à les éviter, soit à circonscrire leurs effets, dans l'exercice du raisonnement tel qu'il est mis en œuvre, non pas d'abord chez le logicien, mais chez un individu quelconque. Lorsque la machine, dans des circonstances déterminées, entrera en conflit avec elle-même, la hiérarchie nécessaire pour éviter ce conflit devra être construite dans le cadre de l'“ éducation ”, non pas, autrement dit, par l'ajout au système d'inférence d'une règle supplémentaire, mais par un “ impératif ” :

> « Des impératifs adéquats (exprimés *à l'intérieur* du système, ne faisant pas partie des règles *du* système), tels que : ‘ n'utilisez

[226] “ ... the store would be largely occupied with definitions and propositions. The propositions would have various kinds of status, e.g., well-established facts, conjectures, mathematically proved theorems, statements given by an authority, expressions having the logical form of propositions but no belief-value ”, *ibid.*

[227] “ The machine should be so constructed that as soon as an imperative is classed as ‘ well established ’ the appropriate action automatically takes place ”, *ibid.*

[228] “ the processes of inference used by the machine need not be such as would satisfy the most exacting logicians. There might for instance be no hierarchy of types ”, *ibid.,* p. 158.

[229] Russell propose, pour éviter le paradoxe de l'ensemble de tous les ensembles qui ne s'appartiennent pas à eux-mêmes, qu'une classe ne puisse contenir que des éléments situés un niveau au-dessous d'elle.

[230] “ But this need not mean that type of fallacies will occur, any more that we are bound to fall over unfenced cliffs ”, *ibid.*

> pas une classe, à moins qu'elle ne soit une sous-classe de l'une de celles que le maître a mentionnées ', peuvent avoir un effet similaire à : ' ne t'approche pas trop près du bord ' ».[231]

L'objectif poursuivi ne consiste pas, en effet, à construire une machine à calculer, mais une machine effectuant, dans des circonstances données, des " choix " qui, par le biais des " impératifs ", soient de moins en moins arbitraires :

> « Parmi ces impératifs, seront importants ceux qui régleront l'ordre dans lequel les règles du système logique concerné devront être appliquées. Car, à chaque étape, lorsque l'on utilise un système logique, il y a un très grand nombre de pas alternatifs, chacun d'eux pouvant être utilisé, en ce qui concerne l'obéissance aux règles du système logique. Ces choix font la différence entre un brillant ou un piètre raisonneur, mais non pas entre quelqu'un qui raisonne juste et quelqu'un qui raisonne faux ".[232]

Turing ne suppose pas que les êtres humains soient eux-mêmes dotés, à leur naissance, c'est-à-dire dans leur " état initial ", d'un système d'inférence logique susceptible de satisfaire les logiciens les plus exigeants. Il fait fonds au contraire sur l'idée que, pour l'être humain, un système logique satisfaisant, tel celui qui doit, par exemple, équiper une machine destinée au calcul, machine dont on doit avoir " une représentation mentale claire... à tout moment du calcul "[233], ne peut être " atteint qu'à l'issue d'une lutte "[234]. A ses yeux, un système d'inférence satisfaisant sur le plan logique ne peut précisément relever, chez l'homme, que de l'acquis, non de l'inné.

Telle est la méthode suggérée par Turing pour construire la machine du jeu de l'imitation. Turing ne se dissimule pas le caractère éminemment spéculatif de l'hypothèse

[231] " Suitable imperatives (expressed within the systems, not forming part of the rules of the system) such as 'Do not use a class unless it is a subclass of one which has been mentioned by teacher ' can have a similar effect to ' Do not go too near the edge ' ", *ibid.* C'est Turing qui souligne.
[232] " Important amongst such imperatives will be ones which regulate the order in which the rules of the logical system concerned are to be applied. For at each stage when one is using a logical system, there is a very large number of alternative steps, any of which one is permitted to apply, so far as obedience to the rules of the logical system is concerned. These choices make the difference between a brilliant and a fooling reasoner, not the difference between a sound and a fallacious one ", *ibid.*
[233] " ... one's object is then to have a clear mental picture of the state of the machine at each moment in the computation ", *ibid.,* p. 159.
[234] " This object can only be achieved with a struggle ", *ibid.*

ainsi énoncée : en attendant la réalisation effective d'une machine telle que celle qu'il décrit et de l'expérience du jeu de l'imitation, les considérations exposées sont, dit-il, des sortes de " déclamations tendant à produire une croyance "[235]. De fait, on a souligné combien Turing semblait sous estimer la complexité, du point de vue de l'étude psychologique, des processus mis en jeu au cours de l'éducation de l'enfant. On remarquera, cependant, qu'il ne s'agit pas, ici, de fournir des procédures précises d'apprentissage pour la machine, calquées sur ce qui se passe effectivement lors de l'apprentissage humain, et présupposant un corpus scientifiquement établi de connaissances psychologiques, mais seulement de déterminer s'il y a ou non contradiction entre la notion de machine universelle et l'idée d'apprentissage ramenée à un processus statistique. L'hypothèse des " machines qui apprennent " ne fait strictement intervenir que la définition logico-mathématique de la machine universelle et les lois établies scientifiquement – puisque ce sont des lois mathématiques - de la statistique[236].

Un point important de la démarche suivie est ainsi mis en évidence : la machine considérée, bien qu'elle soit une machine universelle, se distingue de la " machine à calculer " qui constitue le modèle admis de " l'ordinateur digital " - " Cela s'oppose clairement à la procédure normale d'utilisation d'une machine opérant des calculs : dans ce cas, l'objet est d'avoir une représentation mentale claire de la machine à tout moment du calcul "[237]. La

---

[235] " recitations tending to produce belief ", *ibid.*

[236] 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 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). On peut soutenir, cependant que le problème soulevé à cet égard se situe sur un autre plan que celui envisagé par Turing. Sans doute ce dernier fait-il preuve de " désinvolture ", mais en ce sens que la question spécifique de l'apprentissage du langage par la machine est, pour lui, réglée dans son principe avec celle de la programmation générale de la " machine enfant ". Certes, il s'agit bien, déclare Turing, de faire subir à la machine un apprentissage analogue, dans son ampleur et dans ses finalités, à celui que connaît l'enfant, mais cela ne signifie pas que les chose se passeront, ou devront se passer, de la même façon pour la machine que pour l'enfant. Etant donné une définition de la machine – la " machine universelle " - et un certain processus d'apprentissage – celui suivi par l'enfant ; étant donné, enfin, un outil théorique sûr, car mathématique – la statistique - est-il théoriquement possible de simuler, sur une machine universelle, le même processus à l'aide de cet outil théorique ? L'apprentissage de la parole se situe dans ce cadre. Si l'hypothèse est valide, alors, on ne peut, selon Turing, exclure la possibilité, pour la machine, d' " apprendre " à parler, comme apprend à le faire un enfant.

[237] " This is in clear contrast with normal procedure when using a machine to do computations : one's object is then to have a clear mental picture of the state of the machine at each moment in the computation ", A. M. Turing, *Computing Machinery...*,

“ machine à calculer ”, ou l’ordinateur tel que nous le connaissons encore aujourd’hui, sous l’espèce de la machine conforme au modèle de Von Neuman, peut toujours être ramenée à son “ programme ”, c’est-à-dire à une suite bien définie d’instructions régissant analytiquement son action ; le modèle admis de “ l’ordinateur digital ” peut toujours être, en théorie, une “ machine de papier ”, une machine dont il est théoriquement possible, sans qu’elle soit effectivement construite, de simuler l’action “ à la main ”. Il en est ainsi, en particulier, du modèle de machine qui a historiquement servi de référence à la constitution de l’intelligence artificielle en discipline spécifique. Or, il n’en va pas de même de la “ machine qui apprend ”. On ne peut, ici, parler de programme que pour l’état initial de la machine, et l’action de celle-ci est bien davantage fonction de son devenir que de cet état initial. Pour l’essentiel, les procédures mises en œuvre par la machine qui participe au jeu de l’imitation n’auront pas à être formulées telles quelles, dans un langage “ humain ” - fût-ce un langage formel. Les procédures régissant le succès de la machine, par exemple, à un “ test partiel ” de Turing, ne seront connues *a priori* ni par elle, ni, bien plus, par son programmeur - “ une caractéristique importante de la machine qui apprend est que son maître ne saura souvent que très peu de choses sur ce qui se passe à l’intérieur, bien qu’il puisse dans une certaine mesure prévoir la conduite de son élève ”[238]. Les “ maîtres ” de la machine, comme la plupart des éducateurs, ne connaîtront que deux choses : le but qu’il s’agit d’atteindre, et une méthode, fondée en l’occurrence sur le système “ punitions-récompenses ”. Quant à leur capacité à prévoir le comportement de leur “ élève ”, elle restera, comme dans le cas d’élèves humains, largement empirique. Ainsi, pour réaliser une machine capable de “ bien se comporter ” au “ test du sonnet ”, le travail des constructeurs de la machine ne consistera pas à formaliser, par un

op. cit., p. 159.

[238] “ An important feature of a learning machine is that its teacher will often be very ignorant of quite what is going on inside, although he may still be able to some extent to predict his pupil’s behavior ”, *ibid.,* p. 158. Cette “ importante caractéristique ” a été de nouveau mise en évidence par les modèles connexionnistes en intelligence artificielle, lesquels se montrent en certains domaines – par exemple la reconnaissance des formes – capables de performances inaccessibles aux modèles classiques. Ces capacités émergent naturellement des systèmes connexionnistes sans avoir à être spécifiées de manière formelle. On a fait remarquer que, par cela même, ces modèles ne pouvaient constituer une approche épistémologique valide en psychologie de l’intelligence, puisqu’ils “ fonctionnent ” sans qu’il soit besoin de fournir une description théorique de ce qu’ils font. Or, la thèse de “ l’IA forte ” postule que les propriétés de la machine “ intelligente ” expliquent la fonction effectuée par cette machine (voir Andy Clark, “ Connectionism, Competence and Explantion ”, *The Philosophy of Artificial Intelligence,* Margaret A. Boden éd., *op. cit.*

ensemble de règles énoncées explicitement, le problème " écrire un sonnet, puis le commenter ", mais à amener la machine-enfant, par une suite d'expériences, d'essais et erreurs, à être en mesure de composer un sonnet - probablement mauvais - puis de le commenter à la demande de l'examinateur - sans doute maladroitement - comme le ferait un élève quelconque. La machine victorieuse au jeu de l'imitation construira, en somme, *elle-même*, à partir des modifications subies par son " système initial " au cours du " processus d'éducation ", la configuration des machines " singulières " - participant à des " tests partiels " de Turing - qu'elle simulera. Enfin, le " processus d'éducation " reposant sur un principe aléatoire, le franchissement d'un seuil " surcritique " ne peut, statistiquement, être exclu[239].

Une difficulté apparaît cependant : la possibilité d'une victoire de la machine au jeu de l'imitation, telle qu'elle a été dégagée dans le premier moment de l'argumentation de Turing, par la discussion des " objections " et des " arguments ", ne prouve nullement que la machine qui peut l'emporter doive, pour cela, être " surcritique ". On l'a vu, le principe du test met en jeu tout autant le comportement de l'examinateur que celui de la machine, et, sous cet angle, le point décisif est le fait que l'examinateur, à un certain moment, se comporte *comme si*, à ses yeux, la machine était un individu humain. Peu importe, autrement dit, la manière dont ce résultat sera obtenu par la machine. Il faut, pour que celle-ci réussisse le test, qu'elle soit conçue pour émettre un " discours " signifiant pour l'examinateur : " ton interlocuteur est un homme ", et qu'elle l'émette de telle manière qu'elle soit crue par cet examinateur ; or, puisque sa victoire repose sur l'attitude de celui-ci, ne suffira-t-il pas, pour qu'elle l'obtienne, que, dans les conditions du jeu, elle énonce la proposition : " ton interlocuteur est un homme ", ou ce qui en tient lieu, comme une proposition purement formelle ? Ce que la machine " dira ", ou sera censée " dire ", n'aura, en définitive, de *sens* que pour l'examinateur. De ce point de vue, la victoire de la machine au jeu de l'imitation, c'est-à-dire l'erreur commise à l'égard de la machine par l'examinateur, démontrera peut-être quelque chose à propos de ce dernier, mais certainement pas que la machine elle-même a des " idées ", et, par là, est " surcritique ". On voit que ce qui est ainsi contesté, c'est la majeure du syllogisme prévisionnel, qui affirme que toute entité réussissant le test est " pensante ". Certes, le test est

[239] Turing, en somme, étend, ici, à la machine, la thèse de " l'optimisme rationaliste " professée par Gödel à l'égard de l'esprit humain : si la machine peut être " éduquée ", c'est qu'elle est à même de surpasser sa propre limite à un instant donné, et rien n'interdit, dès lors, de considérer que son développement, comme celui de l'esprit humain, puisse converger vers l'infini.

conçu par Turing de telle sorte que nous soyons assurés qu'il peut être réussi par un individu humain : c'est à quoi tend sa présentation initiale, où il n'est joué que par des humains. Cependant, le fait qu'il sera, alors, nécessairement réussi par un individu humain, considéré *a priori* comme pensant, ne prouve nullement qu'il ne pourra être réussi *que* par des entités pensantes. Comme l'ont précisément montré un certain nombre d'auteurs, il semble tout à fait concevable que le test soit passé avec succès par une machine dont on sait avec certitude qu'elle n'est pas intelligente, ou dont il serait acquis, pour reprendre les termes de Turing, qu'elle ne peut pas être " surcritique ". Tel est le point que nous devons désormais examiner.

Chapitre II : La solidarité des deux hypothèses de Turing

Section I : La critique du jeu de l'imitation

Les critiques que nous allons examiner ici ne portent pas sur la démarche d'ensemble de Turing dans *Computing Machinery*…, mais sur le jeu de l'imitation en tant que tel. Leur visée consiste principalement à montrer que le test du jeu de l'imitation peut être passé avec succès par une machine universelle alors même que l'action de celle-ci reste strictement formelle et, par là, ne peut en aucun cas être assimilée à un comportement de *compréhension*. Selon ces critiques, la participation au jeu de l'imitation d'une machine victorieuse n'est rien d'autre que la construction d'une suite d'éléments qui, s'ils sont bien, pour l'examinateur humain du jeu, des symboles, ne sont, quant à la machine, que des configurations matérielles quelconques.

A travers le jeu de l'imitation, c'est principalement l'intelligence artificielle, dans sa version classique, que ces critiques visent ; en ce sens, la machine à laquelle elles font référence est l'ordinateur que nous connaissons aujourd'hui, dont nous sommes censés savoir à tout moment quel ensemble explicite d'instructions définit son comportement – ou, plutôt, son action - et dont les instructions peuvent, précisément, être considérées comme de simples configurations spatiales, analogues à des figures " sur le papier ". C'est pourquoi ces analyses ne disent rien, en particulier, de l'hypothèse des " machines qui apprennent ", ni du modèle de machine que propose Turing dans le cadre de cette hypothèse, à savoir une machine définie davantage par son devenir que par son programme initial. La question qu'elles posent n'est donc pas de savoir si une machine universelle peut " apprendre ", ni si une telle machine, est, par là, comme le soutient Turing, " surcritique ". Cependant, la critique qu'elles présentent n'en semble pas moins forte, dans la mesure où elle revient à montrer que le test de Turing ne permet pas de faire de discrimination entre la machine de l'intelligence artificielle et celle que

Turing envisage dans le cadre de l'hypothèse des " machines qui apprennent ". En d'autres termes, selon ces analyses, la machine classique de l'intelligence artificielle pourrait réussir le test tout aussi bien que la " machine qui apprend " de Turing.

Nous examinerons plus particulièrement les critiques menées, en premier lieu, par Ned Block, à partir de ce que ce dernier nomme un test " d'intelligence conversationnelle ", et, en second lieu, par John Searle, à partir de sa célèbre expérience dite de la " chambre chinoise ". L'une et l'autre de ces analyses, en effet, se situent délibérément sur le terrain même choisi par Turing : elles s'articulent autour de tests présentés comme analogues au jeu de l'imitation, et montrent que la machine peut réussir ces tests sans que son comportement puisse être tenu pour l'équivalent de celui d'un individu humain placé dans les mêmes conditions, c'est-à-dire pour un comportement impliquant une *compréhension*.

Ned Block et le jeu de l'imitation

Il n'est pas surprenant que la difficulté soulevée par la réflexion de Turing ait été soulignée notamment à travers la critique de ce qui peut être considéré, dans cette réflexion, comme relevant de principes behavioristes. On remarque, en effet, que, dans le cadre du jeu de l'imitation, la prise en compte du comportement des protagonistes est exclusive de toute appréhension d'éventuels " états mentaux " : le test doit être conçu de telle sorte qu'il soit fait abstraction de ce qui se passe, ou serait censé se passer, " derrière " ce qui est directement observable du comportement des acteurs. Seul ce qui peut être observé comme se produisant au cours du jeu doit être pris en considération. Le test est donc justiciable des critiques qui ont été formulées à l'égard du behaviorisme : l'inventaire du comportement observable ne permet pas de rendre compte de tous les phénomènes psychiques. C'est, par exemple, ce que reproche Keith Gunderson à Turing : " Il m'importe simplement d'indiquer que l'exposé de Turing, essentiellement rédigé à partir de résultats observables, échoue pour cette raison même "[240]. Se fonder exclusivement, s'agissant du comportement intelligent, sur des résultats observables, entraînera par exemple à devoir considérer qu'un morceau de craie pris dans une tourmente et

[240] Keith Gunderson, " Le jeu de l'imitation ", *Pensée et machine*, Alan Ross Anderson, Gérard Guièze, éd., *op. cit.*, p. 182.

qui serait mû par les éléments de telle manière qu'il dessine le texte d'un rondeau au tableau, " pense " comme tel être humain qui écrit des rondeaux[241].

C'est une critique de ce type qu'a développé Ned Block, dans un article visant à réhabiliter, face au behaviorisme, une thèse psychologiste[242]. Selon Block, le behaviorisme de Turing se manifeste à travers le fait que, dans la démarche de celui-ci, le test a le statut d'un instrument de mesure. Si l'on admet la pertinence du test de ce point de vue, il n'y a, certes, que deux situations possibles : ou bien la machine passe le test avec succès, et alors elle a fait preuve d'intelligence, au sens humain du terme, ou bien elle échoue et elle n'a pas d'intelligence. Mais l'on pourrait dire que tout se passe, ici, à peu près de la même façon que si l'on voulait, sans connaître la structure chimique de l'eau, vérifier qu'un liquide ressemblant à de l'eau, est ou non de l'eau, en le portant à ébullition : le liquide sera de l'eau s'il bout à partir de 100°[243]. Or, la validité d'un tel principe n'est à aucun moment établie par Turing pour le cas qu'il considère. En d'autres termes, la question n'est pas même de savoir si le test de Turing est le bon, c'est-à-dire si " l'instrument de mesure " est juste ; c'est l'idée même d'un instrument de mesure qui, dans le cas considéré, semble inadéquate.

Block en veut pour illustration l'exemple de la célèbre " Eliza ", cette " machine-psychiatre " mise au point par J. Weizenbaum[244]. Il s'agit, on le sait, d'un programme remarquablement simple, capable de simuler le comportement d'un psychiatre au cours de l'interrogation d'un patient. La machine est programmée pour réagir à un certain nombre de mots-clés, tels que " je ", " vous ", " quelqu'un ", " semblable "... - si, par exemple, le " patient " dit à Eliza : " je sais que quelqu'un se moque de moi ", celle-ci répondra quelque chose du genre : " A qui pensez-vous plus particulièrement ? ". Eliza est également programmée pour transformer en questions les phrases qui lui sont adressées, en substituant simplement la première personne aux autres - à la phrase : " vous n'êtes pas d'accord avec

[241] Le problème, classique en théorie de l'information, du calcul de la probabilité que l'agencement aléatoire des lettres de l'alphabet livre un texte, non seulement doué de sens, mais déjà existant, tel, par exemple, un sonnet de Shakespeare, était de ceux que Turing affectionnait.
[242] Ned Block, " Psychologism and Behaviorism ", *The Philosphical Review*, XC, 1, 1981.
[243] " Construed operationally, the Turing Test conception of intelligence shares with other forms of operationalism the flaw of stipulating that a certain measuring instrument [the Turing Test] is infallible. According to the operationalist interpretation of the Turing Test as a definition of intelligence, it is absurd to ask of a device that passes the Turing Test wether it is really intelligent... ", *ibid.*, p.8. C'est l'auteur qui souligne.
[244] J. Weizenbaum, *Computer Power and Human Reason*, New-York, W.H. Freeman and Co, 1976.

moi ", elle pourra répliquer : " pourquoi croyez-vous que je ne suis pas d'accord avec vous? ". Eliza peut enfin enregistrer les phrases comprenant un mot-clé, prononcées devant elle au cours de la session d'interrogation, de sorte que, confrontée à une phrase dépourvue de mots-clés, elle puisse répliquer en faisant référence à une phrase précédente - si, à quelque moment, une phrase telle que : " mon ami m'a conduite ici " a été prononcée, et si " ami " est un mot-clé, la machine, en présence d'une phrase sans mot-clé, pourra répondre : " Cela a-t-il quelque chose à voir avec le fait que votre ami vous a conduit ici ? ". Enfin, si aucune des procédures décrites ne peut être employée, Eliza coupera court, en demandant par exemple : " Qui est le psychiatre, ici, vous ou moi ? ". Eliza a connu un large succès : Block signale le fait que la secrétaire de Weizenbaum demanda un jour à celui-ci de sortir de la pièce pour pouvoir parler à sa machine en toute intimité. Or, ce succès ne repose pas, on l'a compris, sur l'intelligence de la machine mais sur la crédulité de ses " patients ". Peut-on admettre, demande Block, que la réponse à la question de savoir si une machine est réellement intelligente dépende de la plus ou moins grande crédulité de ses interrogateurs humains ?[245]

Cependant, Weizenbaum n'a jamais prétendu que sa machine était capable de réussir le test de Turing. Aussi Block complète-t-il son argumentation par un détour : s'appuyant sur un argument de Putnam, il souligne que l'observation d'un comportement dépourvu d'intelligence ne permet pas de conclure que celui qui adopte ce comportement soit *incapable* d'intelligence. Putnam imagine des " acteurs parfaits ", qu'il appelle encore des " super-spartiates ", conditionnés à ne jamais réagir à un stimulus de peine ; face à des stimuli associés à un sentiment de peine, de tels " spartiates " ne manifesteront pas de manière observable la peine qu'ils éprouvent, et, pourtant, ils éprouveront celle-ci[246]. Block, en fonction de cela, propose une version du test de Turing consistant, pour l'examinateur, non plus à distinguer une machine d'un être humain, mais à repérer un " comportement intelligent " à partir du critère suivant : " L'intelligence (ou plus précisément l'intelligence conversationnelle) est la disposition à produire une séquence raisonnable [NB : dans le sens de l'expression " doué de raison "] de réponses verbales à une séquence de stimuli verbaux,

[245] " Could the issue of wether a machine in fact thinks or is intelligent depend on how gullible human interrogators tend to be ? ", Ned Block, " Psychologism and Behaviorism ", *op. cit.*, p. 10.
[246] Hilary Putnam, " Brains and Behavior ", *Mind ; Language and Reality*, London, Cambridge University Press, 1975.

quels qu'ils puissent être "[247]. Au cours d'un tel test, de la même façon que les " acteurs parfaits " de Putnam sont capables de ne pas manifester la peine qu'ils éprouvent, un individu intelligent peut être conduit à ne pas donner la réponse " intelligente " attendue par son interlocuteur, si, par exemple, il a un intérêt plus fort à ne pas le faire ; il n'en sera pas moins capable d'intelligence.

Par là même, il est, à l'inverse, possible d'imaginer une machine donnant, dans le test d'" intelligence conversationnelle " proposé par Block, des réponses considérées comme " intelligentes " par un observateur, alors que le processus qui l'aura conduite à donner cette réponse ne peut en aucun cas être qualifié d'intelligent selon les critères mêmes adoptés pour juger la réponse. Block suggère l'idée d'une " machine inintelligente "[248] pourtant capable, en principe, de subir avec succès un test " d'intelligence conversationnelle ", c'est-à-dire capable de produire " une séquence raisonnable de réponses verbales à des stimuli verbaux ". On appellera " une chaîne de phrases exprimable "[249] les phrases qui peuvent être tapées l'une après l'autre pendant un temps donné, par un dactylographe humain dans sa langue maternelle, par exemple l'anglais. Considérons le sous-ensemble des phrases qui " sont naturellement interprétables comme des conversations dans lesquelles la contribution d'au moins une partie est raisonnable dans le sens décrit ci-dessus "[250], et appelons ces phrases des " chaînes raisonnables "[251]. Ce sous-ensemble peut être établi sous la forme d'une liste. Pour une question posée par un interrogateur, les programmeurs de la machine établiront une liste de réponses acceptables prises dans le sous-ensemble des " chaînes raisonnables " ; à une série de ces réponses, l'interrogateur répliquera par une série d'autres questions, pour chacune desquelles les programmeurs définiront une nouvelle liste. Une machine programmée de cette manière sera capable d'" émettre une séquence raisonnable de réponses verbales "[252]. Pourtant, elle ne saura rien faire d'autre que chercher dans une liste finie et préétablie de chaînes de

[247] " Intelligence (or more accurately, conversational intelligence) is the disposition to produce a sensible sequence of verbal responses to a sequence of verbal stimuli, whatever they may be ", Ned Block, " Psychologism and behaviorism ", *op. cit.,* p. 18.
[248] " ... my unintelligent machine... ", *ibid.,* p. 19.
[249] " ... a typable string of sentences ", *bid.,* p. 19.
[250] " Consider the subset of this set which contains all and only those strings which are naturally interpretable as conversations in which at least one's party contribution is sensible in the sens described above ", *ibid.,* p. 19.
[251] " Call a string which can be understood in this way a sensible string ", *ibid.,* p. 19.
[252] " Such a machine will have the capacity to emit a sensible sequence of verbal outputs ", *ibid.,* p. 21.

caractères ; elle aura, déclare Block, “ l’intelligence d’un grille-pain ”[253]. Autrement dit, “ ... la capacité d’émettre des réponses raisonnables n’est pas suffisante pour l’intelligence ”[254]. De sorte, enfin, que déterminer si un comportement est intelligent ne dépend pas tant des caractères observables de ce comportement que de la manière dont il est produit.

Une difficulté du même ordre a été soulignée par John Searle à l’aide de son célèbre argument de la “ chambre chinoise ”[255].

John Searle et la “ chambre chinoise ”

Searle a cherché à montrer qu’une simulation formelle de la faculté de participer à un échange linguistique ne reproduit pas cette faculté, et que, quand bien même la machine de Turing l’emporterait au jeu de l’imitation, nous n’aurions pas le droit d’en conclure qu’elle “ pense ”.

La cible de l’argument de Searle est l’intelligence artificielle (IA) dans sa version “ forte ”. Nous l’avons vu, en effet, Searle distingue une “ IA faible ”, qui considère l’ordinateur comme un outil permettant “ de formuler et de tester [à propos de l’esprit] des hypothèses de façon plus rigoureuse et plus précise ”, d’une “ IA forte ”, pour laquelle “ l’ordinateur convenablement programmé est véritablement un esprit, en ce sens que des ordinateurs munis des bons programmes *comprennent* et ont d’autres états cognitifs ”[256]. C’est dans le cadre de cette critique que l’argument de la “ chambre chinoise ” vise le test de Turing ; la machine de “ l’IA forte ” que décrit Searle est bien, en effet, une machine de Turing, telle que celle-ci a été définie dans *On Computable Numbers*… :

> Une règle spécifiquement informatique va... déterminer que lorsqu'une machine se trouve dans une situation donnée, lorsque sa bande magnétique comporte un symbole donné, la machine va accomplir une opération donnée, comme par exemple effacer le symbole, en imprimer un autre, puis elle passera au stade suivant, au cours duquel, par exemple, la position de la bande sera déplacée d'un cran vers la gauche...[257]

[253] “ But, actually, the machine has the intelligence of a toaster ”, *ibid.,* p. 21.
[254] “ ... the capacity to emit sensible responses is not sufficient for intelligence ”, *ibid.,* p. 21.
[255] J. Searle a exposé son argument dans plusieurs textes. Nous ferons référence à : “ 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.
[256] John Searle, “ Esprits, cerveaux et programmes ”, *op. cit.*, p. 54.
[257] *Ibid.,* p. 41.

Searle, toutefois, n'applique pas sa démonstration directement au test de Turing, mais imagine une situation reposant sur le même principe, à savoir l'idée que, si un échange linguistique a lieu entre des êtres humains et l'équivalent d'une machine, de telle sorte que les êtres humains ne perçoivent aucune différence entre cette situation et celle qui correspond à une discussion entre eux, il faut admettre que l'équivalent d'une machine " pense ", au sens humain du terme.

> Imaginons qu'un groupe de programmeurs ait écrit un programme qui permette à un ordinateur de simuler la compréhension du chinois. Alors, si l'on pose à l'ordinateur une question en chinois, celui-ci va la confronter à sa mémoire, ou à sa base de données, et fournir les bonnes réponses en chinois. Supposons qu'en outre, ces réponses soient aussi bonnes que celles d'un véritable chinois. Alors, pourra-t-on dire que l'ordinateur comprend le chinois au sens littéral, comme un chinois comprend sa langue ?[258].

Pour que sa démonstration soit parfaitement probante, Searle décrit une " machine " dont les éléments constitutifs seraient des symboles de l'écriture chinoise et un être humain : imaginons un être humain enfermé dans une pièce ; cet être humain ne comprend pas un mot de chinois. Il dispose, dans la pièce où il a été placé, d'une part, de paniers dans lesquels se trouvent des symboles de la langue chinoise, d'autre part, d'un livre, écrit dans sa propre langue - et qu'il peut, par conséquent, lire et comprendre - fournissant des règles purement syntaxiques de manipulation des symboles. Supposons que d'autres symboles chinois soient introduits dans la pièce, et que le livre fournisse des règles indiquant à notre opérateur que certains symboles doivent être, dans un certain ordre, sortis de la pièce. Supposons, enfin, que, sans que l'opérateur le sache, les chaînes de symboles introduites dans la pièce soient des " questions " et celles qu'il sort de la pièce des " réponses " à ces questions. Si les règles ont été correctement rédigées, et si l'opérateur ne fait pas d'erreur en les suivant, tout se passera exactement comme si, à des questions posées par un chinois de Chine, des réponses étaient données par un chinois de Chine. Les partisans de " l'IA forte " affirment qu'un ordinateur capable de simuler parfaitement la manipulation des symboles linguistiques - c'est-à-dire de telle sorte qu'aux yeux d'un être humain, la différence entre la simulation et la situation réelle

[258] *Ibid.,* p. 42.

ne soit plus perceptible - ne fait pas que recevoir et émettre des signes, mais *comprend* ce qu'il " entend " et ce qu'il " dit " :

> Les partisans de l'IA forte prétendent que dans cette séquence de questions-réponses, la machine ne se borne pas à simuler une capacité humaine, mais que l'on peut dire qu'elle comprend l'histoire et fournit les réponses aux questions, et que ce que font la machine et son programme explique la capacité humaine de comprendre l'histoire et de répondre à des questions la concernant[259].

Pourtant, ajoute Searle, " dans une telle situation ", c'est-à-dire dans la situation de l'opérateur humain placé dans la " chambre chinoise ", " je vous défie d'apprendre un mot de chinois... "[260].

Il a été objecté à l'argument que, si l'être humain qui joue le rôle de l'opérateur dans la " chambre chinoise " ne comprend pas le chinois et ne comprend donc pas ce qu'il " entend " et ce qu'il " dit ", c'est que son statut est le même que celui, dans un ordinateur, de l'" unité centrale " ; en d'autres termes, l'opérateur, quand bien même il est humain, n'est qu'un élément d'un système. Le fait qu'il soit humain est ici faussement probant, car, au fond, un moulin à vent[261] aurait le même statut. Ce qui comprend, dans le " test de Searle ", c'est le système tout entier : " Ils me disent que c'est le système dans son ensemble, y compris la pièce, les paniers de symboles, les bancs qui supportent les programmes, et d'autres choses prises globalement, qui comprend le chinois "[262]. A quoi Searle réplique que considérer la " chambre chinoise " comme un système ne change rien au fait que la manipulation des symboles ne fait intervenir aucune sémantique, et que l'on ne peut donc pas dire que le système " comprenne ". On peut, en effet, supposer que l'opérateur de la " chambre chinoise " apprenne par coeur le contenu du livre expliquant les règles de manipulation des symboles du chinois et le contenu des paniers de symboles. Auquel cas, il devra être considéré comme un système formellement identique à celui composé de la chambre, des paniers, du livre et de lui-même. Pour autant, rien ne sera changé à la problématique ; si tous les éléments du système sont intégrés à l'opérateur humain, celui-ci ne comprendra pas davantage le chinois. Et pour le

[259] " L'esprit est-il un programme d'ordinateur ", *op. cit.*
[260] " Du cerveau au savoir ", *op. cit.,* p. 43.
[261] La machine, remarque Searle, serait constituée de " boîtes de bière " et animée par un " moulin à vent " que cela ne changerait rien à son statut de machine universelle.
[262] *Ibid.,* p. 45.

prouver, Searle imagine que l'on pose à l'opérateur de la " chambre chinoise " des questions dans sa langue maternelle - en l'occurrence l'anglais - auxquelles il répond dans sa langue maternelle :

> « Maintenant, pour compliquer l'histoire, imaginez que ces personnes me donnent également des histoires en anglais... et qu'elles me posent ensuite des questions en anglais sur ces histoires, questions auxquelles je réponds en anglais [...] Imaginons encore que mes réponses aux questions en anglais soient [...] indiscernables de celles d'autres personnes de langue maternelle anglaise [...] D'un point de vue externe... les réponses aux questions chinoises et aux questions anglaises seraient tout aussi bonnes... »[263].

Extérieurement, il n'y aura aucune différence entre les deux moments de l'expérience : on pourrait d'ailleurs supposer que les questions et les réponses soient, du point de vue de leur sens, les mêmes en chinois et en anglais. Le système " chambre chinoise " aura participé à l'échange linguistique en chinois exactement de la même façon que le système " être humain " à l'échange linguistique en anglais. Pourtant, l'un des systèmes, en l'occurrence l'être humain, devra être considéré comme plus puissant que l'autre puisqu'il inclut une dimension sémantique. C'est de cette différence que ne rend pas compte la simulation par la machine dans le test de Turing :

> « L'exemple montre qu'il pourrait y avoir deux 'systèmes' capables tous les deux de réussir le test de Turing, mais dont un seul comprendrait ; et il ne serait pas valable de dire que, puisqu'ils réussissent tous les deux le test de Turing, ils doivent tous les deux comprendre, car cela irait à l'encontre de la thèse selon laquelle le système qui, en moi, comprend l'anglais, est bien plus puissant que celui qui ne fait que traiter des symboles chinois. Bref, la réponse du système se borne à affirmer sans justification que le système doit comprendre le chinois. »[264]

La critique de Searle souligne ainsi le fait que la thèse de " l'IA forte " repose sur l'idée admise sans discussion que la simulation *formelle* d'un processus humain de pensée est la reproduction de ce processus : " Ici, la distinction clef se trouve entre la duplication et la simulation : aucune simulation en elle-même ne peut constituer une duplication "[265].

[263] " Esprits, cerveaux et programmes ", *op. cit.,* p. 356.
[264] *Ibid.,* p. 361.
[265] " Du cerveau au savoir ", *op. cit.,* p. 49. Searle ajoute : " Nous pouvons par ordinateur simuler des orages de pluie dans les environs de Londres ou des incendies d'entrepôts à

Dans le cas de tests spécifiques tels que ceux proposés par Block et Searle, nous pouvons donc construire une machine de telle sorte que nous soyons assurés, avant de la soumettre au test, qu'elle n'a, pour reprendre l'expression de Block, que " l'intelligence d'un grille-pain " ; pourtant, cette machine pourra passer l'épreuve avec succès. De la même façon, si nous admettons les arguments à l'aide desquels Turing défend l'idée d'une " victoire " possible de la machine au jeu de l'imitation, une machine universelle ne pourra-t-elle pas réussir ce que nous avons appelé un " test partiel " de Turing, du type, par exemple, de celui " du sonnet ", en donnant au professeur Jefferson l'illusion qu'elle a, non seulement composé un sonnet, mais qu'elle *sait* l'avoir composé et qu'elle *sait* de quoi elle " parle ", alors qu'il aura suffi pour cela qu'elle construise des suites de figures privées pour elle de toute signification, et notamment de celle qu'elles ont pour le professeur Jefferson ? Sans doute Turing prend-il soin de préciser que la machine qui, selon lui, l'emportera au jeu de l'imitation ne peut pas être confondue avec des machines du type de celles imaginées plus tard par Block et Searle, puisque la machine victorieuse doit être une " machine qui apprend ". La validité du test n'en est pas moins mise en question, s'il est vrai qu'il peut être réussi non seulement par une " machine qui apprend ", mais également par une machine ayant " l'intelligence d'un grille pain ".

La question est dès lors posée de savoir si, dans le cadre du problème envisagé par Turing, celui de la " pensée " des machines, assimilé à celui de la possibilité, pour une machine, d'être " surcritique ", la seule hypothèse féconde n'est pas celle de l'apprentissage. La seule " expérience " à prendre en compte ne devrait-elle pas être celle de " l'éducation " de la machine ? Plutôt que de traduire la question initiale " Les machines peuvent-elles penser ? " dans les termes du jeu de l'imitation, Turing n'aurait-il pas dû la traduire directement dans les termes du problème de " l'apprentissage " et de " l'éducation " ? En ce sens, le syllogisme prévisionnel plutôt que :

" toute entité qui trompe un examinateur humain en simulant un individu humain considéré comme pensant, est " pensante " au même titre que tout individu humain ; une machine peut tromper un individu humain en se faisant passer pour un individu humain ; donc, la machine est " pensante " au sens humain du terme ",

---

l'est de la capitale. Et pourtant, dans tous ces cas, personne n'ira songer que la simulation est pareille à la réalité : personne ne supposera que la simulation par ordinateur d'un orage va nous tremper jusqu'aux os... *", ibid.,* p. 50.

ne devrait-il pas être :

“ toute entité pouvant être soumise à une éducation analogue – dans son ampleur et sa complexité – à celle subie par les individus humains, est “ surcritique ”, c’est-à-dire “ pensante ” au sens humain du terme ; la machine universelle peut être éduquée comme un individu humain ; donc la machine est “ surcritique ”, c’est-à-dire “ pensante ” au sens humain du terme ” ?

Or, il n’y a aucune raison de proposer, dans ce dernier cas, comme critère de vérification, un test tel que le jeu de l’imitation, qui prend en compte, non seulement le comportement de l’entité qui doit être testée – la machine – mais également celui d’individus humains. On aura au contraire intérêt à imaginer un test au cours duquel, comme il est d’usage en matière d’expérimentation scientifique, le phénomène observé sera isolé autant que possible des conditions d’observation, c’est-à-dire un test qui, idéalement, ne mette en jeu que la seule machine. Bref, Turing n’aurait-il pas du faire l’économie du jeu de l’imitation et se contenter de développer la deuxième partie de son argumentation dans *Computing Machinery…* ?

On remarquera, cependant, qu’il n’aurait pu, par ce moyen, répondre à la question précise à laquelle renvoie, dans sa démarche, la question générale de la “ pensée ” des machines, à savoir la question du sens de la simulation par la machine de ce que nous avons appelé les conditions intuitives du calcul pour un individu humain, et, à travers cela, du principe, en l’homme, de la construction même de la machine. En vérité, comme nous allons le voir, la réflexion de Turing ne saurait se passer du jeu de l’imitation.

Section II : La “ pensée ” de la machine victorieuse au jeu de l’imitation

On l’a vu, Turing aborde la question générale de la “ pensée ” des machines à partir d’un problème spécifique, lié à la notion de machine universelle : parce qu’elle manipule le signe mathématique, au sens hilbertien du terme, la machine universelle simule les conditions intuitives du calcul pour un individu humain, et, par là, les conditions mêmes de sa construction - par exemple la propriété qu’a la langue, chez l’homme, de se prendre pour objet, de se signifier elle-même, bref, de pouvoir être “ métalangue ”. C’est la question posée par la simulation mécanique de certaines propriétés du constructeur de la machine que Turing traduit sous la forme de la question “ Les machines peuvent-elles penser ? ”. Dans une perspective de cet ordre, *Computing Machinery…* doit montrer qu’il peut y avoir, pour la machine universelle, une langue dont les propriétés sont identiques à celles de la langue humaine, et la réponse donnée à la question “ Les machines peuvent-elles penser ? ” est dépendante de celle donnée à propos de la “ langue ” de la machine. Or, telle est, précisément, la raison pour laquelle l’hypothèse générale de la “ pensée ” des machines doit être traduite par Turing dans les termes de l’hypothèse spécifique de la victoire possible d’une machine universelle au jeu de l’imitation, et non, d’abord, dans celle de l’hypothèse des “ machines qui apprennent ”.

En tant que telle, sans doute, l’expérience de “ l’éducation ” de la machine montre que celle-ci manipule, non seulement des figures, mais des signes, c’est-à-dire qu’il y a une sémantique de son comportement : on l’a vu, selon l’hypothèse des “ machines qui apprennent ”, la machine découvre *par elle-même* les réponses qu’exige la situation où elle a été mise, notamment par le biais du système “ punitions-récompenses ”, en cherchant, autrement dit, à éviter d’être “ punie ”, mais en courant toujours le risque de l’être ; la

machine construit, par là, un rapport à la situation que l'on peut analyser en référence à un comportement de " compréhension ". Son " éducation " élabore pour elle ce que l'on peut appeler du sens.

Rien, cependant, ne permet de considérer que le statut de ce sens sera le même que pour un individu humain. L'étrangeté radicale de la machine par rapport à l'homme, présupposée par l'opinion commune, n'est pas mise en question par l'hypothèse des " machines qui apprennent ". L'idée que la " machine qui apprend " puisse être, dans la terminologie de Turing, " surcritique ", comme l'individu humain, lui-même soumis à un processus d'éducation, permet, sans doute, d'utiliser le terme " penser " à l'égard de la machine ; cela, toutefois, n'autorise nullement à affirmer qu'elle " pense " au sens humain du terme. Le " seuil surcritique " que pourrait franchir la machine reste ici une image, dont l'imprécision interdit toute conclusion de ce genre. L'usage du terme " penser " pour la machine demeure de l'ordre de la commodité de langage, permettant de désigner un " quelque chose ", mis en scène par l'hypothèse des " machines qui apprennent ", mais dont on ne sait rien, sinon qu'il est indépendant de la description formelle de la machine. Seule, en vérité, l'expérience de la victoire de la machine au jeu de l'imitation est susceptible de conduire à la conclusion que la machine " pense " au sens humain du terme, du fait que l'examinateur humain, dans cette hypothèse, se sera comporté, au cours du jeu, à l'égard de la machine, exactement comme il se comporte à l'égard d'un individu humain, et aura cru avoir affaire en effet à un individu humain. Ce cas est le seul dans lequel l'étrangeté radicale de la machine présupposée par l'opinion commune sera *infirmée*.

Aussi bien les deux hypothèses – celle de la victoire possible d'une machine au jeu de l'imitation, et celle des " machines qui apprennent " - sont-elles solidaires l'une de l'autre dans la réflexion de Turing : l'hypothèse des " machines qui apprennent " est construite sur les caractéristiques mises en évidence au cours de la discussion des " objections " et des " arguments " ; elle a pour fonction de donner un contenu positif aux traits que l'examinateur peut être amené à croire reconnaître dans la machine : une " conscience ", une capacité " d'invention ", " l'informalité du comportement ". La discussion des " objections " et des " arguments " visait à mettre en évidence l'absence de contradiction entre la notion scientifique de machine universelle et, non pas la " conscience " et " l'invention ", pour en rester aux traits les plus frappants, mais leur reconnaissance chez un interlocuteur par un

individu humain. Cependant, au-delà de la discussion des " objections " et des " arguments ", il s'agissait aussi, pour Turing, de montrer, non seulement que rien ne s'oppose, d'un point de vue théorique, à ce qu'une machine universelle simule, dans le cadre du jeu, ces différents " traits ", mais encore qu'une machine universelle, *pour simuler effectivement ces traits*, doit être " surcritique ". En somme, une machine universelle ne l'emporte au jeu de l'imitation face à ses adversaires humains que parce que son comportement ne livre pas uniquement *l'apparence* de la " pensée ". Tel est le véritable sens de l'hypothèse de la victoire d'une machine au jeu de l'imitation.

Le véritable syllogisme prévisionnel dans la démarche de Turing

" L'informalité du comportement " constituait, nous l'avons vu, un trait fondamental dans la discussion des " objections " et des " arguments " menée par Turing. Nous devons y revenir, car il apparaît, à partir de ce trait, que les tests spécifiques proposés par Ned Block et John Searle, dans leur critique de la réflexion de Turing, ne peuvent être assimilés qu'abusivement au jeu de l'imitation. Rappelons que la machine victorieuse au jeu est celle qui s'avère capable de l'emporter à une série suffisante de ce que nous avons appelé des " tests partiels " de Turing ; or, à partir du moment où la victoire au jeu de l'imitation fait appel à l'ensemble des compétences d'un individu humain quelconque, le critère du " suffisant ", ici, est, précisément, " l'informalité du comportement " : dans tous les cas, le nombre de " tests partiels " que la machine victorieuse sera en mesure de réussir doit être tel que le comportement général dessiné par eux ne puisse être formalisé, c'est-à-dire décrit sous la forme d'un ensemble défini et logiquement satisfaisant de règles. Pour le dire autrement, dès lors qu'il met en jeu les compétences globales d'un individu humain quelconque, le test de Turing doit être compris comme dépassant d'au moins un test partiel le nombre de ceux qui seraient susceptibles d'être unifiés sous la forme d'un ensemble cohérent de règles. Du reste, nous pouvons distinguer deux sortes de " tests partiels " : ceux qui, comme dans les cas envisagés par Ned Block et John Searle, peuvent être réussis à partir de l'établissement de règles explicites, et ceux qui, en eux-mêmes, comme, par exemple, le " test du sonnet ", excluent déjà cette caractéristique ; dans l'esprit de Turing, le " test du sonnet " n'est sans doute pas lui-même formalisable. Les expériences proposées par Block et Searle ne sauraient donc être considérées comme des équivalents stricts de celle qu'envisage Turing. C'est enfin

pourquoi la réalisation d'une " machine qui apprend " se présente, non seulement comme *un* moyen de concevoir effectivement une machine susceptible de l'emporter au jeu de l'imitation, mais, bien plus, comme *le* seul moyen envisageable : tant que l'on ne saura pas soumettre une machine à une éducation analogue à celle d'un individu humain, le test ne pourra pas être mené.

La démarche de Turing dans *Computing Machinery…* pourrait donc être schématisée sous la forme suivante : toute entité qui peut être soumise à une " éducation " analogue à celle d'un individu humain est susceptible d'être " surcritique ", c'est-à-dire " pensante " au sens humain du terme ; une entité de cette sorte peut l'emporter au jeu de l'imitation face à des adversaires humains. Une machine universelle peut l'emporter au jeu de l'imitation face à des adversaires humains (il n'y a pas de contradiction théorique entre cette hypothèse et celle de la notion de machine universelle) ; une machine universelle peut être soumise à une éducation analogue à celle d'un individu humain (il n'y a pas de contradiction théorique entre cette hypothèse et la notion de machine universelle). Donc une machine universelle doit être considérée comme pouvant être " surcritique ", c'est-à-dire " pensante " au sens humain du terme, ce qui signifie, en particulier, qu'elle manipule des signes de la même manière qu'un individu humain.

Ainsi, l'hypothèse de la victoire possible d'une machine universelle au jeu de l'imitation, liée à l'hypothèse auxiliaire des " machines qui apprennent ", implique la reconnaissance d'un " quelque chose ", assimilé au " penser " chez l'homme, dont le comportement de la machine victorieuse est inséparable. Cette reconnaissance s'effectue à deux niveaux. A celui, tout d'abord, de l'examinateur humain du jeu de l'imitation : parce que le test ne peut être réussi que par une entité " surcritique ", s'il est réussi, c'est que A, au cours du jeu, a identifié, dans le comportement de C, au-delà des différentes formes spécifiques prises par ce comportement, et qui correspondent aux " tests partiels " de Turing réussis par la machine, *quelque chose* qui, indépendamment de la reconnaissance de la forme physique de C, indépendamment, autrement dit, de toute figure humaine, le conduit à se comporter à l'égard de C comme il se comporterait en présence d'un interlocuteur humain, c'est-à-dire, en l'occurrence, comme il se comporterait à l'égard d'une entité qu'il considère *a priori* comme " pensante ". Ce n'est pas en considérant la machine comme " pensante ", mais en la considérant comme humaine que l'examinateur A se trompe. Or, s'il considère la machine

comme humaine, c'est parce qu'il la considère comme " pensante ". Le mécanisme du jeu ne consiste pas, pour l'examinateur, à croire reconnaître un homme, et, partant de là, un être pensant, mais, au contraire, à reconnaître un être pensant, et à croire reconnaître, à partir de là, un homme.

Au niveau ensuite des observateurs du jeu de l'imitation, pour qui, la machine qui réussit le test, c'est-à-dire qui parvient à tromper A au cours du jeu, parce qu'elle est " surcritique ", est effectivement " pensante " au sens humain du terme. La " pensée " est reconnue *à* la machine parce que l'examinateur A du jeu de l'imitation se comporte comme s'il la reconnaissait *en* elle. La croyance de l'examinateur à la " pensée " de C est le signe de la " pensée " en C. Selon Turing, du point de vue des observateurs du test, si l'examinateur A, confronté à son adversaire mécanique C, se comporte *comme s'il* était en présence d'une " pensée ", c'est qu'il est effectivement en présence d'une " pensée ".

De sorte que la problématique dans laquelle s'inscrit le jeu de l'imitation n'est pas tant celle de la reconnaissance formelle d'un homme par un autre homme, que celle de la reconnaissance de la pensée par la pensée, la pensée ne pouvant, en toute rigueur, être reconnue, en tant que telle, par autre chose que la pensée. Ce que l'examinateur A et les observateurs du jeu partagent, et qui est au principe de la reconnaissance d'un " quelque chose " chez le protagoniste C du jeu, ne peut être que ce " quelque chose " même, c'est-à-dire la " pensée ". A travers la défaite de l'examinateur au cours du jeu de l'imitation, ce n'est pas seulement un homme, A, qui croit reconnaître en C un autre homme, puis d'autres hommes qui, à partir de là, reconnaissent à leur tour de la " pensée " en C, c'est la pensée, en ces hommes, qui reconnaît de la pensée.

Ce faisant, Turing propose, sans doute, une solution au problème spécifique qu'il examinait, celui de la simulation par la machine des conditions intuitives du calcul pour un individu humain : l'hypothèse d'une victoire possible au jeu de l'imitation, précisée par l'hypothèse auxiliaire des " machines qui apprennent ", apporte une ultime garantie à la plausibilité de l'équivalence intuitive entre procédure de calcul et procédure mécanique. Cependant, la démarche de Turing, dans sa singularité, ne peut atteindre ce résultat que parce qu'elle traite d'une question proprement *philosophique* : celle de la définition de la pensée comme auto-intelligibilité. Sans doute, sous l'angle du problème spécifique qu'il s'agit pour Turing de traiter, les seuls éléments à prendre en compte sont-ils, d'une part, la notion même

de machine universelle, et, d'autre part, non pas une élaboration philosophique de l'idée de pensée, mais cette intuition du " penser " qui est condition de tout usage du terme " penser ", à savoir la certitude intuitive qu'a tout homme de sa propre pensée et de la pensée en tout homme. Ce n'est pas dire pour autant que la démarche adoptée par Turing – imaginer une " expérience ", plutôt que mener une analyse de concepts - se situe en marge de la philosophie, comme à la frontière de celle-ci : dans la question " les machines peuvent-elles penser ? ", de même que le terme " machine " relève d'un usage spécifique, en l'occurrence l'usage mathématique, le terme " penser " renvoie à un usage spécifique, en l'occurrence l'usage philosophique. Quand bien même Turing, imaginant une expérience, n'entend pas raisonner en philosophe, c'est-à-dire sur des catégories philosophiques, au sens technique du terme, son argumentation n'a pas simplement des conséquences philosophiques, plus ou moins lointaines, qu'il appartiendrait au philosophe de dégager dans leur champ spécifique, elle est, *en elle-même*, philosophique.

Les hypothèses de Turing et l'histoire de la philosophie

C'est là ce qui, en définitive, donne sens à l'argumentation mise en œuvre par l'hypothèse de la victoire possible d'une machine au jeu de l'imitation, et selon laquelle la victoire de la machine, infirmant l'opinion commune, confirme, dans ce mouvement même, la " pensée " de la machine. On pourrait arguer, en effet, de ce que l'hypothèse de Turing démontre seulement la " pensée " de la machine au sens, par nature incertain, du terme " penser " pour l'opinion commune. Il se trouve, cependant, que l'idée défendue par celle-ci, à savoir qu'une machine ne peut penser, car le " penser " relève d'un ordre autre que celui de la machine, est aussi une idée philosophique, qui appartient de plein droit à l'histoire de la philosophie occidentale. L'usage du terme " penser " fait par Turing dans *Computing Machinery*…, même réduit à la condition ultime de tout usage du terme " penser " - et, en définitive, précisément à cause de cela - n'est pas vierge. La problématique dans laquelle, à travers le jeu de l'imitation, s'inscrit cet usage - la distinction, par un homme, d'un homme et d'une machine qui simule l'homme - renvoie directement, par exemple au problème, chez Descartes, de la distinction homme-animal, homme-machine ; elle renvoie également, en tant que problématique du " comme si " de l'examinateur, à la réflexion kantienne sur l'usage abusif du concept de liberté appliqué à l'automate, c'est-à-dire, nous le verrons, à la question,

dans la philosophie transcendantale, des conditions de la reconnaissance d'un autrui. Or, chez Descartes comme chez Kant, la mise en évidence du " quelque chose " qui manifeste la pensée et qui permet la distinction de l'homme et de l'automate, se fait à partir de l'idée que la machine ne peut pas penser, et que la question de la pensée pour elle ne se pose pas. Le rejet d'une pensée mécanique est directement inscrit dans la conception de la machine chez Descartes et chez Kant. De sorte que la discussion de ces positions philosophiques participe, dans la démarche de Turing, de la fonction argumentative de l'infirmation de l'opinion commune. Du fait de la structure particulière de la double hypothèse de Turing, l'infirmation de l'opinion commune, dans son rôle constitutif même pour l'argumentation développée dans *Computing Machinery*…, est inséparable d'une prise de position contre les réflexions philosophiques auxquelles cette double hypothèse fait écho. Le problème de la " pensée " des machines, à partir du moment où il met en jeu, sous la forme singulière du test, la question de la reconnaissance de la pensée, implique une argumentation située dans le cadre spécifique de l'histoire de la philosophie. C'est, non seulement par l'infirmation de l'opinion commune, mais encore, à travers cette infirmation, par la critique de l'expression donnée, notamment chez Descartes et Kant, à la question de la reconnaissance de la pensée, que la pensée peut être reconnue à la machine.

Sous cet angle, nous allons nous efforcer de le montrer, c'est la problématique philosophique de la notion de *pratique*, à travers son élaboration dans le kantisme, qui constitue le foyer de la portée philosophique de la démarche de Turing dans *Computing Machinery*… Or, nous verrons que telle est, de ce point de vue, cette portée qu'elle inclut, à sa manière, la prise en compte de son propre statut philosophique, c'est-à-dire de son inscription dans une *histoire* de la philosophie.

Troisième partie : Le jeu de l'imitation et la notion de pratique

Qu'il y ait une expression philosophique de l'idée admise par l'opinion commune - une machine ne saurait " penser " car le " penser " relève d'un autre ordre que le mécanique - serait aisément attesté par nombre de textes. Si nous entendons, cependant, nous intéresser plus particulièrement à la forme que cette expression prend chez Descartes et Kant, c'est, en premier lieu, que la formulation de son principe dans la métaphysique cartésienne – la distinction des substances étendue et pensée – la détermine historiquement, et, en second lieu, que ces deux auteurs l'appréhendent l'un et l'autre, à un certain moment de leur réflexion, sous l'angle même de la question générale qui sous-tend la problématique du jeu de l'imitation, à savoir : qu'en est-il de ce qui distingue un homme d'un automate ?

Affirmer cela pose pourtant un problème d'interprétation : Turing examine, pour sa part, la question de ce qui distingue un homme d'une machine en la situant dans le cadre d'une " expérience " où ce qui est pris en compte est essentiellement la dimension *publique* de l'acte de pensée ; les protagonistes du jeu de l'imitation s'adressent les uns aux autres, " dialoguent ", " conversent ", " discutent ", et le " comme si " de l'examinateur, sur lequel repose la victoire possible de la machine, prend, dans ce contexte, la forme de la " convention polie " selon laquelle tout locuteur accorde la pensée à son allocutaire. Or, cet aspect de l'expérience de Turing n'est-il pas précisément ce qui l'écarte de la manière dont Descartes et Kant envisagent l'idée de pensée ? Chez ceux-ci, en effet, l'acte de pensée est le propre d'une conscience qui ne " dialogue " qu'avec elle-même ; la problématique de la convention qui fonde la parole comme échange n'est guère la leur, et l'on peut douter que l'opposition " privé-public " soit ici pertinente : l'acte de pensée, chez Descartes comme chez Kant,

s'inscrivant dans une problématique du fondement, déborde la sphère du " privé ". Dès lors, et dans la mesure même où la méthode suivie par Turing évite, précisément, d'aborder la question de la pensée sous l'angle de la discussion de l'idée de pensée, comme le font Descartes et Kant – la question de la pensée comme fondement - la confrontation de sa démarche avec le cartésianisme et le kantisme n'est-elle pas vaine ? Peut-elle avoir un autre sens que celui, au mieux, d'une simple comparaison historique mettant en évidence une solution de continuité dans l'histoire de la philosophie ?

En vérité, cette difficulté constitue, non un obstacle, mais le cœur même du problème qu'il convient d'examiner. Le test " conversationnel " de Turing – pour reprendre l'expression de Ned Block - met bien en jeu, selon nous, les principes dégagés, dans leurs contextes repectifs, par le cartésianisme et le kantisme, et c'est parce qu'elle fait appel à ces principes que la double hypothèse de Turing bouleverse la problématique dans laquelle ceux-ci s'inscrivent.

La question générale " Qu'en est-il de ce qui distingue l'homme de l'automate ? " peut être déclinée sous différentes formes : peut-on distinguer l'homme de l'automate ? Comment peut-on les distinguer ? Qu'est-ce qui permet, enfin, de les distinguer ? Sans doute les démarches respectives de Descartes et de Kant se situent-elles, à cet égard, à l'opposé de celle de Turing : il ne s'agit jamais, pour eux, de répondre à l'une ou l'autre de ces questions au moyen d'une " expérience ". Il n'y a, chez eux, de réponses que conceptuelles, de telle sorte que l'idée même d'une " expérience ", ou bien est inutile, ou bien ne peut servir que d'illustration. Il n'en demeure pas moins que la question générale est bien abordée par eux, dans la mesure même où ils lui fournissent une réponse. Ainsi l'est-elle chez Descartes sous la forme même où Turing, à travers le jeu de l'imitation, l'examine, à savoir : comment distinguer un homme d'un automate ? Supposons, en effet, déclare Descartes, un automate ayant l'apparence exacte d'un homme ; sachant, d'une part, que l'homme pense, car il participe des deux substances étendue et pensée, et, d'autre part, que l'automate ne pense pas, car il ne relève que de la substance étendue, y a-t-il des signes qui manifestent cette différence de l'un par rapport à l'autre, et quels sont-ils ? Kant, quant à lui, n'envisage pas la question sous l'angle du " comment ? ", mais, en niant que l'on ait le droit de dire d'un automate, comme on le fait pourtant quelquefois, qu'il est " libre ", il répond à la question de savoir ce qui permet de distinguer l'homme de l'automate. L'automate, précise-t-il, est qualifié de " libre " par

usage d'un " concept comparatif de liberté ", le second terme de la comparaison étant ici l'homme : c'est à partir de l'affirmation de la liberté en l'homme que l'automate est abusivement dit " libre " ; la liberté est, précisément, ce qui distingue l'homme de l'automate.

Chez Descartes comme chez Kant, la réponse à la question générale de ce qui distingue l'homme de l'automate s'appuie sur l'affirmation que ce dernier appartient à un autre registre que l'homme auquel il est comparé : l'automate ne *pense* pas. Selon Descartes, le mécanique relève de la seule substance étendue, distincte de la substance pensante. En ce sens, ce sont la parole et ce que nous appellerons l'action réfléchie qui, à ses yeux, interdisent qu'un homme puisse, par hypothèse, être assimilé à un automate et confondu avec lui. Pourtant, la parole et l'action réfléchie peuvent être, dans leur forme et leur matérialité, imitées par un automate ; il s'agit donc de dégager ce dont elles sont l'expression, et qui échappe de manière irréductible au mécanique. Pour Descartes, il est toujours possible de distinguer un homme d'un automate car, quand bien même celui-ci serait assez habilement conçu pour imiter la parole et l'action réfléchie, telles qu'elles peuvent être appréhendées en l'homme, jamais, à travers cette simulation mécanique, ne s'exprimera un *jugement*, lequel est l'expression d'une volonté libre, qui porte en elle-même la certitude première du *cogito* - puisque le " doute méthodique " implique la liberté. Le jugement exprime l'effort d'une conscience située dans le monde, dans un " ici et maintenant ", et qui construit un " ordre des raisons " tâchant de rejoindre un " ordre des choses ". Il implique, par là, la possibilité de se tromper, mais aussi d'être éduqué.

Chez Kant, le mécanique relève de la seule " logique générale ", qui préside au connaître, et non de la " logique transcendantale ", laquelle renvoie à un inconnaissable qui peut, certes, être pensé, et doit l'être, mais qui, précisément, ne peut être *que* pensé. Les éléments constitutifs du penser dégagés par Descartes sont ainsi réorganisés, dans le kantisme, sous la forme d'une théorie du jugement distinguant un " jugement déterminant " et un " jugement réfléchissant ", mettant en jeu le " je pense " sous la forme de l'aperception transcendantale, fondée, enfin, par là, sur la liberté transcendantale. Ici encore, du reste, pour le sujet kantien comme pour l'homme cartésien participant de la substance étendue et de la substance pensante, l'erreur, et " l'éducabilité ", la capacité à apprendre, sont constitutives du penser, en tant que celui-ci renvoie au sujet transcendantal.

Aussi bien la conception générale du penser dont le cartésianisme et le kantisme contribuent, au-delà de ce qui les sépare, à mettre en place la problématique, apparaît-elle

comme une traduction philosophique de l'opinion commune dont l'infirmation constitue le ressort du raisonnement de Turing. Il découle de là que Turing ne peut, par ce raisonnement, infirmer l'opinion commune sans viser en même temps les démarches cartésienne et kantienne ; en établissant, non par une discussion théorique, mais par une " expérience ", que la machine peut être confondue avec l'homme, Turing prend le contre-pied des positions cartésienne et kantienne. En d'autres termes, si l'expérience de Turing est vérifiée, alors, Descartes et Kant se trompent dans leur réponse à la question de savoir ce qui distingue l'homme de l'automate.

Or, Turing lui-même ne réduit pas la pensée, dans sa propre démarche, à la définition scientifique de la machine : il ne s'agit pas, dans *Computing Machinery…*, de ramener l'idée de pensée au niveau de cette définition positive de la machine, mais d'élever l'idée de machine, à partir de cette définition, et au-delà d'elle, jusqu'à ce qui, pour l'opinion commune, mais aussi pour la conception générale du penser que partagent Descartes et Kant, distingue la pensée du mécanique. La critique turingienne ne vise pas tant les éléments mis en jeu dans la conception de la pensée à l'œuvre chez Descartes et Kant que la thèse selon laquelle cette conception générale n'aurait pas de sens du point de vue de l'idée de machine. Dès lors, et puisque l'argumentation de Turing s'appuie sur la dimension publique de l'acte de pensée, il ressort de la critique turingienne – si l'on admet que l'hypothèse de Turing est " vérifiée " par l'expérience du jeu de l'imitation – que Descartes et Kant pècheraient précisément par l'absence de prise en compte de ce qui, dans l'idée de pensée, relève de l'échange de parole en tant que tel.

La confrontation de la démarche de Turing à l'analyse kantienne de ce qui distingue l'homme de l'automate prend ici tout son sens. En termes kantiens, la question posée par le jeu de l'imitation - celle de savoir *comment* distinguer l'homme de la machine - renvoie, en effet, aux conditions de possibilité de la reconnaissance d'un autrui, dans l'autre qui me parle et à qui je parle. On sait que, chez Kant, le " je " du " je pense " relève, non d'une substance ainsi que chez Descartes, mais du sujet transcendantal, condition de possibilité de l'acte de penser, c'est-à-dire, à travers la loi morale, de la *personne*. La possibilité de reconnaître, dans ce qui pourrait n'être qu'un autre, un semblable, c'est-à-dire un autrui, apparaît, ainsi, chez Kant, comme une exigence même de la loi morale. Cependant, la personne, parce qu'elle renvoie, au-delà du sensible, à un ordre intelligible, ne peut être connue ; elle ne peut être que

pensée. La loi morale exige donc la possibilité de penser, dans un autre, la personne, c'est-à-dire la possibilité de se comporter *comme si* cet autre était une personne. Cette possibilité est inscrite dans l'ordre du *pratique* que définit l'idée même de personne. Or, la parole, en tant notamment qu'elle est un mode d'expression de la faculté de juger *réfléchissante*, désigne cette possibilité.

Il reste que la machine, chez Kant, ne peut relever que de la causalité naturelle, et non, comme le sujet, d'une dimension transcendantale. En outre, la distinction même d'une causalité naturelle et d'une dimension transcendantale exclut, du point de vue transcendantal, qui est celui du sujet, la possibilité de les confondre l'une avec l'autre dans l'ordre du pratique : quand bien même une machine pourrait simuler la parole, à partir du fait que celle-ci est aussi un phénomène relevant de la causalité naturelle, la relation pratique possible entre une machine et un homme ne sera jamais celle qui autorise celui-ci à penser celle-là comme une personne. De sorte que, dans le cadre de la problématique kantienne, la question de l'imitation de l'homme par la machine ne se pose pas tant sous la forme : " comment distinguer un homme d'un automate ? ", que sous la forme : " comment les confondre ? "

C'est bien à cette question même que répond Turing dans *Computing Machinery…* Si l'examinateur A perd face à la machine, c'est qu'il est conduit, dans le cadre spécifique du jeu de l'imitation, à se comporter *comme si* la machine C était un *sujet*. Son erreur n'est possible que parce qu'il met en oeuvre, dans le procès d'énonciation-communication auquel il participe avec la machine, le système catégoriel où s'énonce l'idée de sujet, telle que l'élabore la réflexion kantienne. L'implication de la notion de sujet est constitutive de la pratique – le procès d'énonciation-communication en quoi consiste le jeu de l'imitation – dans laquelle l'examinateur A est engagé avec son interlocuteur C. La démonstration de *Computing Machinery…* aboutit, sous cet angle, à affirmer que la machine peut être reconnue comme un autrui.

Toutefois, la reconnaissance d'un autrui se fait, chez Kant, à partir de la reconnaissance du sujet par lui-même comme sujet ; je peux reconnaître un autrui parce que je peux, et dois, me penser moi-même comme sujet transcendantal. La reconnaissance d'autrui implique le discours à la première personne, du " je " au " je ", et c'est mon propre " je " qui porte l'exigence, non de connaître, puisque cela ne se peut pas, un " je " dans l'autre, mais de le *penser*. C'est pourquoi il est nécessaire, ici, d'étudier les deux moments de cette

problématique : non seulement la possibilité de la reconnaissance, au cours du test de Turing, de la machine comme un autrui, mais également la possibilité de penser un " je " de la machine victorieuse au jeu de l'imitation. Dans le cadre de celui-ci, l'idée d'un " je " de la machine doit avoir un sens pour l'examinateur A ; l'hypothèse de Turing implique que le " je " de la machine qui simule le comportement humain soit pensable. Or, la question ainsi posée est éclairée de manière significative, nous semble-t-il, par l'argument célèbre d'Hilary Putnam, formulé à l'occasion d'une réflexion consacrée au problème de la référence, et dit des " cerveaux dans la cuve "[266].

Putnam montre que les " discours " respectifs de la machine du jeu de l'imitation et de ses adversaires humains ne peuvent avoir de " référence partagée ". Il est possible que l'examinateur A du jeu ne soit pas en mesure de distinguer la manière dont son interlocuteur C - la machine – " fait référence " de la manière dont il fait lui-même référence ; pourtant, leurs références ne peuvent être les mêmes. La machine du jeu de l'imitation se trouve dans une situation analogue à celle d'un " cerveau dans une cuve " : elle fait référence " dans l'image " ; sa référence n'est pas constituée par les objets du monde réel de l'examinateur A, mais par les images de ces objets.

Cependant, s'agissant du " je " énoncé par les cerveaux, qu'ils soient ou non " dans une cuve ", l'argument de Putnam présuppose ce que Jaakko Hintikka nomme, à propos du *Cogito* cartésien, la *consistance existentielle*. Le " je " a, ici, le statut d'une *performance*. L'argument des " cerveaux dans une cuve " implique que, si un cerveau ne peut jamais dire lui-même qu'il est " un cerveau dans une cuve ", c'est-à-dire qu'il n'est pas un " vrai " cerveau, ou ce qu'il considère comme tel, il puisse toujours faire référence à lui-même à travers le " je " qu'il énonce.

Par ailleurs, appliqué au jeu de l'imitation, l'argument de Putnam admet, par hypothèse, que la machine C l'emporte, comme le montre Turing, au terme de *l'échange* qu'elle a avec l'examinateur A et le partenaire B de celui-ci. Autrement dit, A B et C *communiquent*. S'il y a victoire de la machine, c'est qu'il y a eu, non pas illusion d'échange, aux yeux de A, entre lui-même et C, mais échange véritable ; quelle que soit l'interprétation du jeu que l'on adopte, et quand bien même cette interprétation consisterait à affirmer, comme

266 Hilary Putnam, *Raison, vérité et Histoire*, Paris, Editions de Minuit, 1984.

dans le cas de l'argument de la " chambre chinoise " de Searle, que l'action de la machine n'est que l'expression formelle d'un discours qui, en tant que tel, n'est pas le sien, A aura eu un interlocuteur et aura perçu un discours derrière l'expression formelle présentée par la machine. A, autrement dit, aura reconnu une " consistance existentielle ". On ne peut accorder, en effet, dans le contexte du jeu de l'imitation, la consistance existentielle à une entité autre que la machine, à savoir le programmeur qui, après l'avoir conçue, s'effacerait derrière celle-ci pour la laisser jouer à sa place : il ne s'agirait plus alors d'une consistance *existentielle*. Il ne peut y avoir de consistance existentielle que de *l'acteur*, c'est-à-dire la machine, qui, au cours du jeu, donne la réplique à A. En vérité, si la machine C victorieuse au jeu de l'imitation est bien dans une situation analogue à celle du " cerveau dans une cuve ", et si une telle situation implique la consistance existentielle, alors la simulation, par C, du " je " d'un individu humain, en l'occurrence la simulation du " je " de ses adversaires A et B, constituera, aux yeux de A, la *performance* propre de la machine. De sorte que, si celle-ci l'emporte, c'est que l'examinateur A aura été contraint, par le procès d'énonciation-communication présupposé par cette victoire, d'accorder au " je " énoncé par la machine au cours du jeu, le même statut qu'au sien propre.

D'où, enfin, la question posée, sous cet angle, par la démarche de Turing : s'il est vrai qu'un " je " de la machine qui l'emporte au jeu de l'imitation a bien, dans le cadre de celui-ci, un sens pour l'adversaire humain de la machine, ne découle-t-il pas des hypothèses de Turing – celle de la victoire possible d'une machine au jeu de l'imitation, et celle des " machines qui apprennent " - que l'on doive élever la machine à la dignité du sujet moral chez Kant ? Or, considérées *stricto sensu*, les hypothèses de Turing ne nous autorisent pas à renvoyer ce qui advient au cours du procès d'énonciation-communication auquel participent les protagonistes du test à la transcendance fondatrice d'un intelligible, fût-il, comme chez Kant, " problématique ", c'est-à-dire inconnaissable. On ne saurait considérer la machine qui réussit le test de Turing comme une personne, au sens kantien du terme, sans dépasser les conditions mêmes de " l'expérience " : il ne découle pas de l'hypothèse de la victoire d'une machine au jeu de l'imitation que le principe de la mise en œuvre de la notion de sujet dans le procès d'énonciation-communication soit le sujet lui-même, mais seulement que c'est, au contraire, le procès d'énonciation-communication qui détermine, ici, la notion de sujet, en la définissant comme la forme prise par l'erreur de l'examinateur A. Aussi bien la double hypothèse de Turing, si elle fait appel, à travers la question de la reconnaissance d'un autrui, à la

problématique qu'exprime l'idée kantienne de *pratique*, est-elle inséparable d'un profond réaménagement de celle-ci : tout se passe comme si la victoire d'une machine au jeu de l'imitation prenait sens à partir de l'idée que le procès d'énonciation-communication en quoi consiste le jeu est à lui-même son propre principe, et que la mise en œuvre du système catégoriel où s'énonce la notion de sujet constituait un moment nécessaire de ce procès. Il s'agit, ici, en d'autres termes, de *renverser* la problématique kantienne, c'est-à-dire de saisir le procès d'énonciation-communication – la pratique où l'examinateur du jeu est engagée - non pas indépendamment de l'idée de sujet, mais comme *premier* par rapport à celle-ci.

Ce sont ces différents points que nous examinerons ici. Nous reviendrons tout d'abord sur la réponse de Descartes à la question de savoir ce qui distingue l'homme de la machine, et sur la problématique que suppose cette réponse, selon laquelle la machine appartient à un autre registre que la pensée. Nous étudierons ensuite, en confrontant la démarche de Turing à l'argument de Putnam, la question de la possibilité qu'un " je " de la machine ait un sens pour l'examinateur A du jeu de l'imitation. Nous examinerons alors la démarche de Turing à la lumière du problème des conditions de possibilité de la reconnaissance d'un autrui chez Kant, avant d'essayer de mesurer la portée philosophique de *Computing Machinery…* du point de vue de la notion de pratique.

Chapitre I : Le jeu de l'imitation et la problématique cartésienne : la machine et le jugement

Dans un texte célèbre du *Discours de la Méthode*[267], Descartes examine un problème proche de celui à partir duquel Turing imagine le jeu de l'imitation : de quels moyens disposons-nous pour distinguer un homme d'un automate parfait en son genre qui, par hypothèse, imiterait un homme dans son apparence extérieure ? L'examen de ce problème, toutefois, n'a pas, chez Descartes, la même fonction que chez Turing : il ne s'agit pas de répondre à la question " les machines peuvent-elles penser ? ". Il n'y a, en effet, pour Descartes, aucune raison de douter qu'il soit toujours possible de distinguer un homme d'un automate – celui-ci fût-il parfait en son genre - et que cela soit rendu possible par le fait que l'homme pense, au contraire de l'automate. L'impossibilité de la pensée pour l'automate est une certitude métaphysique acquise : l'automate, ou la machine, relèvent, par définition, de la seule substance étendue. La question abordée par Descartes est plutôt celle de savoir comment cette distinction établie en métaphysique est possible en pratique, c'est-à-dire par quoi la pensée est signalée chez un homme.

Ce sont la parole et l'action réfléchie qui, aux yeux de Descartes, permettent de distinguer l'homme de l'automate. L'une et l'autre relèvent de l'union, en l'homme, d'une âme et d'un corps, c'est-à-dire du fait que l'homme participe des deux seules substances admises par la métaphysique cartésienne : la pensée et l'étendue. En ce sens, la pensée, chez l'homme, s'exprime, certes, à travers le *cogito*, le " je pense ", mais elle renvoie, par là-même au *jugement*, dont l'horizon est d'abord l'erreur, raison du doute. Bref, l'homme, s'il pense, se trompe - et peut, par l'éducation, apprendre à se libérer de l'erreur.

[267] René Descartes, *Discours de la méthode, Œuvres philosophiques,* I, Paris, Classiques Garnier, 1988.

Or, sous l'angle du jeu de l'imitation, la question de la " pensée " des machines rejoint le problème de la distinction de l'homme et de l'automate tel que l'envisage Descartes, et l'on comprend que l'hypothèse de Turing – une machine peut " penser ", car elle peut l'emporter au jeu de l'imitation - constitue une réfutation au moins implicite de la position cartésienne. Le désaccord, cependant, ne porte pas, ici, sur la nature de ce par quoi la pensée est attestée chez l'homme - la parole et l'action réfléchie en tant qu'expressions du jugement – mais sur l'absence, affirmée par Descartes, de ces manifestations de la pensée dans le cas de l'automate, dès lors que celui-ci peut l'emporter au jeu de l'imitation. A Descartes, Turing oppose l'idée que, si l'on réduit le champ de la comparaison entre l'homme et l'automate à l'un des indices qui, selon Descartes, distingue le premier du second – en l'occurrence la parole - il apparaît, précisément, que la distinction, entre eux, n'est plus possible. On voit par là que, si la réflexion de Turing comporte une mise en question implicite du cartésianisme, le champ théorique au sein duquel elle prend sens est déterminé par la problématique cartésienne.

La théorie de " l'animal-machine "

On sait que Descartes écrivit le *Discours*, non seulement pour exposer l'expérience du doute, mais aussi pour préparer les esprits aux conséquences découlant des principes métaphysiques tirés de cette expérience, tels que, par exemple, la distinction radicale des substances étendue et pensée. L'une de ces conséquences est la théorie célèbre de " l'animal machine " : les corps animés - le vivant, dirions-nous aujourd'hui - s'expliquent entièrement, selon Descartes, par les lois de l'étendue, c'est-à-dire par les lois du mouvement mécanique.

Pour montrer que le fonctionnement des corps animés peut être expliqué à l'aide des seules lois de l'étendue, Descartes, dans la " Cinquième Partie " du *Discours*, a recours à une fable. Le procédé a pour première vertu de le protéger des attaques dont sa théorie pouvait être l'objet de la part des autorités ecclésiastiques au moment où Galilée venait d'être condamné[268],

[268] C'est du reste cette condamnation qui fait renoncer Descartes à publier les traités du *Monde* et de *l'Homme*, et le conduit à publier le *Discours*, destiné notamment à préparer les esprits à l'exposé de sa théorie. Il ne s'agit pas là de simple prudence politique. Il est essentiel en effet, aux yeux mêmes de Descartes que son explication du monde et de l'homme n'entre pas en contradiction avec le dogme : il ne saurait y avoir de vérité hors de la parole divine ou contre elle. C'est pourquoi le souci constant de Descartes, lorsqu'il fait le récit de la création d'un monde qui pourrait être le nôtre, est de rendre ce récit aussi conforme que possible à celui de la Genèse. Aussi bien, et puisque, dans celle-ci, le monde est créé par Dieu d'emblée dans son plus haut état de perfection, Descartes

mais il permet également de montrer que les lois de l'étendue - celles du mouvement mécanique - sont nécessaires et suffisantes pour expliquer le monde physique[269] dans tous ses aspects. Imaginons, en effet, que Dieu décide de créer un autre monde, et qu'il procède pour cela en concentrant de la matière, c'est-à-dire de l'étendue, en un point des " espaces imaginaires "[270] ; par le jeu des seules lois du mouvement mécanique, ce monde, d'abord pure étendue, deviendrait bientôt, démontre Descartes, l'exact équivalent du nôtre. L'image des corps, animés et inanimés, est, pour Descartes, d'ordre " hydraulique " ; la cosmologie cartésienne met en jeu un système de flux de particules. Ces dernières, entraînées dans les fameux " tourbillons cosmiques ", forment, par leurs différences de tailles et de poids, ainsi

---

explique-t-il qu'il ne voulait pas " inférer de toutes ces choses, que ce monde ait été créé en la façon qu['il] proposait ; car il est bien plus vraisemblable que, dès le commencement, Dieu l'a rendu tel qu'il devait être ". Cependant, ajoute-t-il, " il est certain, et c'est une opinion communément reçue entre les théologiens, que l'action par laquelle maintenant il le conserve, est toute la même que celle par laquelle il l'a créé ; de façon qu'encore qu'il ne lui aurait point donné, au commencement, d'autre forme que celle du chaos, pourvu qu'ayant établi les lois de la nature, il lui prêtât son concours pour agir ainsi qu'elle a de coutume, on peut croire, sans faire tort au miracle de la création, que par cela seul toutes les choses qui sont purement matérielles auraient pu, avec le temps, s'y rendre telles que nous les voyons à présent " (René Descartes, *Discours de la méthode*, *op. cit.*, p. 617).

[269] La fable imaginée par Descartes a le statut d'un modèle abstrait, non celui d'une hypothèse impliquant un processus de vérification expérimentale. Certes, Descartes ne néglige jamais l'expérimentation ; celle-ci, cependant, si elle constitue une méthode d'observation, ne joue aucun rôle dans la validation des connaissances scientifiques, dont la garantie est métaphysique : " ... j'ai tâché de trouver en général les principes, ou premières causes, de tout ce qui est, ou qui peut être, dans le monde, sans rien considérer, pour cet effet, que Dieu seul, qui l'a créé, ni les tirer d'ailleurs que de certaines semences de vérités qui sont naturellement en nos âmes. " (*ibid.*, p. 636). Du reste, le recours à la fable, maintenu par Descartes dans les *Traités de l'homme et du monde*, sera abandonné par lui dans les *Principes*.

[270] " je me résolu de laisser tout ce monde ici à leurs disputes [celles des doctes], et de parler seulement de ce qui arriverait dans un nouveau, si Dieu créait maintenant quelque part, dans les espaces imaginaires, assez de matière pour le composer, et qu'il agitât diversement et sans ordre les diverses parties de cette matière, en sorte qu'il en composât un chaos aussi confus que les poètes en puissent feindre, et que, par après, il ne fît autre chose que prêter son concours ordinaire à la nature, et la laisser agir suivant les lois qu'il a établies. Permettez donc pour un peu de temps à votre pensée de sortir de ce monde, pour en venir voir un autre tout nouveau, que je ferai naître en sa présence dans les espaces imaginaires. " (*ibid.*, p. 615). A propos des " espaces imaginaires ", nous renvoyons ici au commentaire de Gilson : il s'agit " ... dans la philosophie scolastique, où le monde est considéré comme fini, [des] espaces fictifs que l'imagination seule conçoit au-delà des limites du monde et de l'espace réels " (Etienne Gilson, *Discours de la méthode : texte et commentaire*, Paris, Vrin, 1962, p. 383). En fait, précise Gilson, de tels espaces n'ont pas de véritable sens pour Descartes : " ... la matière se définissant par l'espace, tout espace est réel par définition, et la notion d'espace imaginaire n'a aucun sens " (*ibid.*). La matière et l'étendue sont la même chose pour Descartes, donc toute étendue est de la matière. " La matière cartésienne se réduit en effet à l'étendue géométrique, c'est-à-dire à ce qui constitue pour Descartes le type même de l'intelligibilité. " (*ibid.*, p. 384).

que par les chocs qu'elles subissent les unes des autres, des agrégats qui constituent les corps. Si le monde imaginaire du *Discours* est bientôt l'exact équivalent du monde réel, c'est que ses composants ultimes et les lois du mouvement y sont les mêmes. Le fonctionnement des corps animés ne fait pas intervenir d'autres principes ; les semences mâle et femelle servent de levain l'une à l'autre, " fermentent ", et la chaleur ainsi dégagée crée le mouvement des particules qui forme le coeur, le cerveau, la colonne vertébrale, puis la circulation sanguine. C'est ce mouvement qui, au cours de son action, en " triant ", en quelque sorte, les particules selon qu'elles sont plus ou moins lourdes, constitue les tissus, les vaisseaux, les nerfs, les muscles, les os, bref, le corps tout entier[271].

Il ne suffit pas, cependant, pour établir que l'animal est une machine, de montrer comment le fonctionnement des corps animés peut être expliqué à l'aide des seules lois de l'étendue ; il convient également de s'assurer que l'animalité ne participe nullement de la substance pensante. Pour cela, Descartes s'appuie sur une conséquence immédiate de la distinction des substances en substance étendue et substance pensante : la machine, ou l'automate, relève, par définition, exclusivement des lois du mouvement mécanique, c'est-à-dire de la substance étendue. La machine, autrement dit, ne " pense " pas. Rien de ce qui atteste la pensée ne sera jamais manifesté par une machine. Dès lors, si l'on peut montrer que, sur ce même plan, celui de l'attestation de la pensée, l'animal ne peut être distingué de la machine, on aura démontré qu'il n'est qu'un corps, régi par les lois du mouvement mécanique, c'est-à-dire une machine. C'est pourquoi Descartes entreprend d'examiner par quels moyens un homme, un automate et un animal peuvent être distingués les uns des autres.

Imaginons des machines conçues pour imiter tel ou tel animal, dans son corps et dans son comportement extérieur[272] ; " nous n'aurions, déclare Descartes aucun moyen pour reconnaître qu'elles ne seraient pas en tout de même nature que ces animaux "[273]. Imaginons

[271] Les mêmes chocs de particules sont cause de pertes, d'altérations, etc. que la nutrition doit effacer. La nécessité mécanique de la génération et de la croissance se comprend à partir de là. Le flux général des particules dont participent les corps animés interdit que ceux-ci soient " tout donnés " ; ce sont les principes mécaniques par eux-mêmes qui impliquent des formations et des déformations, des agrégats et des désagrégations. De sorte que, si, comme le rappelle Descartes, Dieu a créé le monde tel qu'il est, sans avoir du passer par la formation décrite dans la fable, il reste que le mouvement des corps entraîne, au sein même du monde " tout donné ", la génération, la croissance et la mort.
[272] Descartes prend l'exemple du singe.
[273] *Ibid.,* p. 628.

maintenant qu'il s'agisse de machines conçues pour imiter, dans son corps et son comportement extérieur, un individu humain. Dans ce cas, “ nous aurions toujours deux moyens très certains pour reconnaître qu'elles ne seraient point pour cela de vrais hommes ”[274]. Le premier de ces moyens est la parole. Il serait certainement possible, estime Descartes, de construire une machine à qui l’on ferait “ proférer ” des sons articulés[275], et de faire en sorte que ces sons soient organisés de manière à ce qu'ils aient un sens pour nous[276]. Il est exclu, cependant, qu'une telle machine, même parfaite en son genre, même conçue par un artisan doté d'une habileté divine, “ arrange diversement ” les signes qu'on lui fait utiliser “ pour répondre au sens de tout ce qui se dira en sa présence, ainsi que les hommes les plus hébétés peuvent faire ”[277]

Le second moyen évoqué par Descartes a trait à l'action réfléchie. Les machines sont en général construites pour effectuer certaines actions mieux que ne le ferait un homme ; cependant elles ne “ savent ” faire autre chose que l’action précise en vue de laquelle elles ont été construites[278]. Jamais un automate ne pourra être conçu qui sache, comme le plus hébété des hommes, inventer son action en fonction des rencontres qu’il fait, de sorte que, bien que les machines puissent faire “ plusieurs choses aussi bien, ou peut-être mieux qu'aucun de nous, elles manqueraient infailliblement en quelques autres, par lesquelles on découvrirait qu'elles n'agiraient pas par connaissance, mais seulement par la disposition de leurs organes ”[279]. Ce qui manque en vérité à l’automate, aussi bien fait soit-il, c’est la participation à la “ substance pensante ”.

[274] *Ibid.,* p. 629.
[275] Descartes, comme ses contemporains, avait été frappé par le spectacle des automates hydrauliques réalisés dans les jardins de Chantilly pour Louis XIII par les frères Francini. Certaines des machines de Chantilly étaient des automates musiciens ; des fontaines et des jets d'eau étaient utilisés pour faire entendre non seulement des bruits divers, mais de véritables airs. Les mêmes principes mécaniques auraient permis de faire émettre des sons articulés à une machine. Voir à ce sujet : André Doyon**,** Lucien Liaigre, *Jacques Vaucanson, mécanicien de génie*, Paris, PUF, 1967, p. 67.
[276] “ ... on peut bien concevoir qu'une machine soit tellement faite qu'elle profère des paroles, et même qu'elle en profère quelques-unes unes à propos des actions corporelles qui causeront quelque changement en ses organes : comme, si on la touche en quelque endroit, qu'elle demande ce qu'on lui veut dire ; si, en un autre, qu'elle crie qu'on lui fait mal, et choses semblables ” (René Descartes, *Discours de la méthode, op. cit.,* p. 629).
[277] *Ibid.*
[278] “ car, au lieu que la raison est un instrument universel, qui peut servir en toutes sortes de rencontres, ces organes [NB : ceux des machines] ont besoin de quelque particulière disposition pour chaque action particulière ; d'où vient qu'il est moralement impossible qu'il y en ait assez de divers en une machine pour la faire agir en toutes les occurrences de la vie, de même façon que notre raison nous fait agir. ”. *Ibid.*
[279] *Ibid.*

Or, les même moyens permettent de distinguer l'homme de l'animal[280] :

> « car c'est une chose bien remarquable, qu'il n'y a point d'hommes si hébétés et si stupides, sans en excepter même les insensés, qu'ils ne soient capables d'arranger ensemble diverses paroles, et d'en composer un discours par lequel ils fassent entendre leurs pensées ; et qu'au contraire, il n'y a point d'autre animal, tant parfait et tant heureusement né qu'il puisse être, qui fasse le semblable. »[281]

De la même façon, poursuit Descartes, c'est un fait - " une chose remarquable "[282]- que même les hommes les plus hébétés, les plus proches, en apparence, de l'animal, sont doués de parole, c'est-à-dire de la capacité d'" user de paroles et d'autres signes en les composant ", ce que ne pourra faire l'animal le plus proche, en apparence, de l'humain. Or, cela ne saurait s'expliquer par la physiologie ou l'anatomie : si les animaux ne parlent pas, ce n'est pas qu'il leur manque les organes de la parole ; certains d'entre eux en possèdent qui leur permettent d'articuler des sons, et pourtant ils ne sont pas davantage que les autres en mesure " d'arranger " ces sons et de les composer de diverses manières, comme peuvent le faire même les hommes qui ont été à leur naissance dépourvus de tels organes, puisqu'un sourd-muet, par exemple, peut communiquer par signes avec les autres hommes[283]. Par ailleurs, certains animaux, qui montrent en certaines actions " plus d'industrie que nous ", sont impuissants à effectuer bien d'autres actions qu'accomplit l'homme le plus fruste. Leur supériorité en certaines actions ne saurait donc s'expliquer par le fait qu'ils disposent comme nous " de l'esprit ", car dans ce cas, " ils en auraient plus qu'aucun d'entre nous et feraient mieux en toute chose "[284]. Cela montre au contraire que " c'est la nature qui agit en eux, selon

---

[280] " ... par ces deux mêmes moyens, on peut aussi connaître la différence qui est entre les hommes et les bêtes. ". *Ibid.*, p. 630.
[281] *Ibid.*
[282] " c'est... une chose fort remarquable que, bien qu'il y ait plusieurs animaux qui témoignent plus d'industrie que nous en quelques-unes unes de leurs actions, on voit toutefois que les mêmes n'en témoignent point du tout en beaucoup d'autres ". *Ibid.*
[283] " Ce qui n'arrive pas de ce qu'ils ont faute d'organes, car on voit que les pies et les perroquets peuvent proférer des paroles ainsi que nous, et toutefois ne peuvent parler ainsi que nous, c'est-à-dire en témoignant qu'ils pensent ce qu'ils disent ; au lieu que les hommes qui, étant nés sourds et muets, sont privés des organes qui servent aux autres pour parler, autant ou plus que les bêtes, ont coutume d'inventer d'eux-mêmes quelques signes, par lesquels ils se font entendre à ceux qui, étant ordinairement avec eux, ont loisir d'apprendre leur langue. ". *Ibid.*
[284] *Ibid.,* p. 631.

la disposition de leurs organes "[285]. " Le bon sens est la chose du monde la mieux partagée "[286], écrivait Descartes au commencement du *Discours* ; nous pouvons ajouter : parmi les hommes. De telle sorte enfin que, d'une part, la différence est exactement la même entre l'enfant le moins doué et l'animal le plus proche de lui, qu'entre celui-ci et l'homme le plus remarquable - l'enfant, comme le sage, possède la raison, cet " instrument universel " - et que, d'autre part, il n'y a aucun moyen de distinguer un animal d'un automate : tous deux se situent sur le même plan, celui de la substance étendue.

La parole et l'action réfléchie attestent donc de la participation de l'homme à la substance pensante ; en ce sens, elles signalent la pensée[287]. Nous devons toutefois remarquer que ce qui importe, aux yeux de Descartes, dans la parole et l'action réfléchie, en tant que moyens de distinguer l'homme de l'automate ou de l'animal, ce ne sont pas les phénomènes physiques - la production de sons ou les séquences de gestes - mais la capacité d'agencer les sons proférés de telle sorte qu'ils aient un sens pour les autres hommes, ou la capacité de composer les actions du corps de telle sorte qu'elles constituent des réponses sensées aux problèmes rencontrés par chaque homme dans son environnement. Bref, ce qui importe, c'est la possibilité qu'a chaque homme d'exprimer des idées sous la forme, dirait-on aujourd'hui, d'énoncés impliquant une " prise de position " sur les choses. Ainsi, l'homme doué de parole et capable d'action réfléchie, se distingue de l'animal en cela qu'il *affirme* ou *nie*, en un mot, qu'il *juge*.

---

[285] *Ibid.*

[286] *Ibid.,* p.568.

[287] Sur tout ceci, voir aussi la " lettre à Morus " du 5 février 1649 : " … bien que parmi [les bêtes] d'une même espèce les unes soient plus parfaites que les autres, comme dans les hommes […] on n'a point cependant encore observé qu'aucun animal fût parvenu à ce degré de perfection d'user d'un véritable langage, c'est-à-dire qui nous marquât par la voix, ou par d'autres signes, quelque chose qui pût se rapporter plutôt à la seule pensée qu'à un mouvement naturel. *Car la parole est l'unique signe et la seule marque assurée de la pensée cachée et renfermée dans le corps* ; or tous les hommes les plus stupides et les plus insensés, ceux mêmes qui sont privés des organes de la langue et de la parole, se servent de signes, au lieu que les bêtes ne font rien de semblable, ce que l'on peut prendre pour la véritable différence entre l'homme et la bête " (René Descartes, " A Morus, 5 février 1649, *Œuvres philosophiques*, *op. cit.*, III, p. 886. C'est nous qui soulignons).

Le jugement

A cet égard, la parole et l'action réfléchie renvoient, chez Descartes, non seulement à l'entendement qui " conçoit ", mais à la volonté, principe actif du jugement[288]. Concevoir n'est pas juger. L'entendement, chez Descartes, en tant qu'il est créé, est tout à la fois fini et passif. Pour qu'il y ait jugement, il faut encore que s'exerce un acte de la volonté sur ce qui est conçu. Descartes en veut pour preuve le fait que nous puissions toujours refuser d'affirmer ce que l'entendement nous fait apercevoir[289]. Le premier principe suivi par Descartes dans sa réflexion métaphysique - tenir pour faux tout ce qui n'est que vraisemblable - repose sur cette liberté que nous avons de suspendre notre consentement. C'est parce que le concevoir est, en lui, soumis à une volonté qu'il y a, pour l'homme, du vrai et du faux. En tant que tel, le concevoir ne vise pas le vrai, mais le critère de celui-ci, c'est-à-dire l'évidence, le clair et le distinct. Ne peut être vrai qu'un jugement, c'est-à-dire un acte de volonté qui s'exerce sur du clair et distinct. C'est pourquoi le concevoir est toujours exempt d'erreur. Celle-ci provient, on le sait, de l'absence de mesure entre l'entendement et la volonté. Nous sommes libres, en effet, de refuser notre consentement à cela même que nous concevons clairement et distinctement, c'est-à-dire à l'évidence, comme le montre l'hypothèse du Malin génie - je conçois clairement et distinctement que $2 + 3 = 5$, mais je peux supposer qu'un Dieu très puissant m'ait ainsi fait que je me trompe alors même que je conçois clairement et distinctement. Cette hypothèse implique, non seulement la puissance de la volonté, mais encore que cette puissance déborde celle de l'entendement. Mieux : pour que je puisse suspendre mon consentement face à l'absolu de l'évidence, il faut que la liberté de ma volonté soit infinie[290]. De sorte que, s'il est beaucoup de choses hors de portée de mon entendement

[288] " ...toutes les façons de penser que nous remarquons en nous, peuvent être rapportées à deux générales, dont l'une consiste à apercevoir par l'entendement, et l'autre à se déterminer par la volonté. Ainsi, sentir, imaginer, et même concevoir des choses purement intelligibles, ne sont que des façons différentes d'apercevoir ; mais désirer, avoir de l'aversion, assurer, nier, douter, sont des façons différentes de vouloir " (René Descartes, *Les Principes de la philosophie*, 1ère partie, article 32, *Œuvres philosophiques*, III, *op. cit.*, p. 111).

[289] " ... prenant garde que souvent il nous est libre d'arrêter et de suspendre notre consentement, encore que nous ayons la perception de la chose dont nous devons juger, j'ai rapporté cet acte de notre jugement, qui ne consiste que dans le consentement que nous donnons, c'est-à-dire dans l'affirmation ou dans la négation de ce dont nous jugeons, à la détermination de la volonté plutôt qu'à la perception de l'entendement. " (René Descartes, *Notae in programma*, *Œuvres philosophiques*, III, op. cit., p.814)

[290] " Au reste, il est si évident que nous avons une volonté libre, qui peut donner son consentement ou ne le pas donner, quand bon lui semble, que cela peut être compté pour

fini, et que je ne conçois pas clairement et distinctement, il n'est rien à quoi ma volonté ne puisse s'appliquer[291]. Du point de vue de ce que le *Discours* et les *Méditations* ont établi comme le critère du vrai, du point de vue du clair et distinct, l'" apercevoir " comporte de nombreux degrés, et il y a, certes, des différences importantes entre le sentir, l'imaginer et le concevoir[292], cependant, nous ne nous trompons pas lorsque nous apercevons une chose clairement mais non distinctement ; nous ne nous trompons que lorsque nous appliquons la liberté de notre volonté à une telle appréhension, pour affirmer ou nier quelque chose de ce que nous apercevons ainsi.

L'agir animal, le comportement de l'animal-machine, semble, au contraire, se caractériser par cela même qu'il n'est jamais fautif ; l'animal réussit toujours les gestes et les actions qu'il sait faire, là où l'homme, pour sa part, échoue souvent. Mais, en vérité, ce n'est pas que l'animal ne se trompe jamais, c'est qu'il ne se trompe pas du tout : ni l'erreur ni le comportement véridique n'existent pour lui. L'animal-machine n'a besoin, pour retenir ce qui lui est utile et éviter ce qui lui est nuisible, ni de sentir, ni de concevoir, ni de vouloir ; il n'y a pas de jugement de l'animal-machine, car il n'y a pour lui ni entendement ni volonté. *Qu'il ne puisse se tromper*, c'est là ce qui distingue irréductiblement l'animal le plus doué de l'homme " le plus hébété ".

L'ordre constitutif du jugement et la parole

Si l'homme est capable, non seulement de concevoir, mais de juger, c'est qu'il participe tout en même temps de la substance pensante et de la substance étendue ; c'est que

---

une de nos plus communes notions. Nous en avons eu ci-devant une preuve bien claire ; car au même temps que nous doutions de tous, et que nous supposions même que celui qui nous a créés employait son pouvoir à nous tromper en toutes façons, nous apercevions en nous une liberté si grande, que nous pouvions nous empêcher de croire ce que nous ne connaissions pas encore parfaitement bien. Or ce que nous apercevions distinctement, et dont nous ne pouvions douter, pendant une suspension si générale, est aussi certain qu'aucune autre chose que nous puissions jamais connaître. " (René Descartes, *Principes*, 1ère partie, article 39, *op. cit.,* p. 114).

[291] " De plus, l'entendement ne s'étend qu'à ce peu d'objets qui se présentent à lui, et sa connaissance est toujours fort limitée : au lieu que la volonté en quelque sens peut sembler infinie, pour ce que nous n'apercevons rien qui puisse être l'objet de quelque autre volonté, même de cette immense qui est en Dieu, à quoi la nôtre ne puisse aussi s'étendre : ce qui est cause que nous la portons ordinairement au-delà de ce que nous connaissons clairement et distinctement... " (*ibid.*, article 35, p. 112).

[292] Ainsi peut-on apercevoir quelque chose clairement sans pour autant le concevoir distinctement, alors que ce qui se conçoit distinctement se conçoit, " par même moyen ", clairement.

l'âme, en lui, est unie à un corps, en un certain point de celui-ci[293]. Or, l'union de l'âme et du corps détermine un *ordre* constitutif du jugement, un ordre propre de la connaissance humaine, c'est-à-dire du rapport intellectuel de l'homme aux choses. L'être humain ne conçoit jamais directement les choses ; chaque mouvement de son corps, entièrement déterminé et explicable par les seules lois de l'étendue et la disposition des organes voulue par Dieu, se traduit en son âme par un " sentiment " - lumière, son, couleur, odeur, plaisir, peine - et c'est à travers ces " sentiments ", autrement dit à travers les mouvements liés à la conservation de son corps, que l'homme connaît les choses et leurs propriétés. Aussi bien les choses ne peuvent-elles être conçues par lui d'emblée clairement et distinctement[294]. Ainsi naissent les préjugés, qui sont le reflet de l'état premier du concevoir, indistinct mélange de " sentiments " et de " vérités premières ".

L'homme, en d'autres termes, connaît selon l'ordre des raisons et non selon celui des choses[295]. A Gassendi, qui lui objecte que la proposition " je pense donc je suis " repose sur la majeure " celui qui pense est ", laquelle, puisqu'elle échappe au doute, a le statut d'un préjugé, Descartes oppose que la formule " celui qui pense est " doit être considérée comme une " vérité première ", dont la proposition " je pense " est inséparable[296]. Cette vérité première ne peut cependant être atteinte qu'à travers le *cogito* ; je ne puis énoncer : " celui qui pense est ", que parce que je sais que quelque chose pense. Or, je ne rencontre la pensée qu'à travers le " je ". Que je pense, puisque je doute, c'est là une évidence, une idée claire et distincte, quelque chose que je ne puis mettre en doute ; mais je ne peux rencontrer cette évidence qu'à travers l'expérience du doute. Généralisant le fait, Descartes déclare que l'idée générale n'est accessible qu'à partir du particulier. L'idée générale fonde, certes, l'idée particulière, mais elle n'est atteinte qu'à travers une expérience. L'ordre des raisons,

[293] La fameuse " glande pinéale ".
[294] *Ibid.*, article 71., p. 139.
[295] " ... je n'entreprends point de dire en un même lieu tout ce qui appartient à une matière, à cause qu'il me serait impossible de le bien prouver, y ayant des raisons qui doivent être tirées de bien plus loin les unes que les autres ; mais en raisonnant par ordre a facilioribus ad difficiliora, j'en déduis ce que je puis, tantôt pour une matière, tantôt pour une autre ; ce qui est à mon avis, le vrai chemin pour bien trouver et expliquer la vérité. " (René Descartes, " A Mersenne, 24 décembre 1640 ", *Oeuvres philosophiques*, II, *op. cit.*, p. 301).
[296] " ... on ne peut pas dire toutefois qu'elle soit un préjugé lorsqu'on l'examine, à cause qu'elle paraît si évidente à l'entendement qu'il ne se saurait s'empêcher de la croire, encore que ce soit peut-être la première fois de sa vie qu'il y pense, et que par conséquent il n'en ait aucun préjugé. " (René Descartes, " Lettre à M. Clerselier ", *Œuvres philosophiques*, *op. cit.*, p. 841).

qu'exprime le jugement, est celui d'une conscience inscrite dans un monde, d'une conscience qui naît et se développe en un point particulier du monde. Or, cet ordre propre au jugement constitue une dimension essentielle de la parole en tant qu'elle distingue l'homme de l'animal et de la machine.

Si la parole est l'un des deux moyens permettant de distinguer à coup sûr un homme de son imitation par un automate, celui-ci fût-il le mieux conçu, c'est en cela qu'elle est, chez l'homme, l'expression du jugement ; elle est l'une des formes que prend, dans l'ordre de l'étendue, l'acte d'affirmation, de négation, voire de suspension par lequel une volonté s'applique au concevoir. C'est pourquoi on ne saurait rendre compte de la parole par les seules lois mécaniques qui déterminent la production matérielle d'un signe. La parole n'est signe qu'autant qu'elle renvoie à l'alliance d'un entendement fini et d'une volonté infinie, c'est-à-dire, en dernière instance, à l'union en l'homme des substances étendue et pensante, d'une âme et d'un corps. Aussi bien est-elle d'abord l'expression des préjugés[297]. L'apparition de la parole chez l'homme coïncide avec le moment où, enfant, il conçoit les choses clairement, mais non encore distinctement. Certes, ce qui n'est conçu que confusément peut, dans un second temps, être conçu clairement et distinctement, et l'on peut concevoir l'idée - claire et distincte - d'une langue parfaite n'exprimant elle-même que le clair et distinct[298]. Une telle langue n'est pourtant pas accessible à l'homme. Celui-ci n'exprime ce qu'il conçoit distinctement qu'à travers la parole d'abord confuse formée dans les premiers moments de son existence. La langue, manifestation immédiate de l'absence de mesure entre l'entendement fini et la volonté infinie, ne cesse pas d'exprimer les effets produits dans notre âme par les mouvements du corps, de sorte qu'elle n'est jamais, en elle-même, aussi claire et distincte que peuvent l'être nos idées. Enfin, la parole, de même que le second moyen envisagé par Descartes pour distinguer l'homme de l'animal et de la machine – l'action réfléchie – renvoie, par là, à cela que l'homme ne saurait être dressé, comme l'animal, ni construit, comme la

[297] " .... les mots que nous avons n'ont quasi que des significations confuses, auxquelles l'esprit des hommes s'étant accoutumé de longue main, cela est cause qu'il n'entend presque rien parfaitement. ". (René Descartes, " A Mersenne, 20 novembre 1629 ", *Oeuvres philosophiques*, I, *op. cit.* p. 231).

[298] " Et si quelqu'un avait bien expliqué quelles sont les idées simples qui sont en l'imagination des hommes, desquelles se compose tout ce qu'ils pensent, et que cela fût reçu par tout le monde, j'oserais espérer ensuite une langue universelle, fort aisée à apprendre, à prononcer et à écrire, et ce qui est le principal, qui aiderait au jugement, lui représentant si distinctement toutes choses, qu'il lui serait presque impossible de se tromper... ". *Ibid.*

machine ; il doit être *éduqué*. Il lui faut *apprendre* à concevoir clairement et distinctement pour que son jugement soit juste.

Contre Descartes, Turing affirme que la machine, sous la forme de la machine universelle, peut " penser ". Mais ce n'est pas que la " pensée " soit autre, à ses yeux, que ce qu'elle est selon Descartes. C'est, bien au contraire – tel est le sens du jeu de l'imitation - qu'elle peut être reconnue, dans l'action de la machine, aux indices mêmes qui, d'après Descartes, la signalent. Une machine universelle peut être " éduquée ", c'est-à-dire soumise à un processus d'apprentissage analogue, dans son principe et ses contenus, à celui qui accompagne le passage du petit d'homme à l'état d'adulte ; la machine " apprend " dans le cadre d'un système de punitions-récompenses qui ne consiste pas, comme dans le cas du dressage animal, à contenir, voire à briser, un comportement instinctif - pour Descartes, purement mécanique - mais à l'amener à construire elle-même son comportement en fonction des situations où elle est placée. S'il peut en être ainsi, c'est que la machine est *capable d'erreur* ; c'est à partir de ses erreurs qu'elle " apprend ". Comme le plus hébété des hommes, la machine peut se tromper. Enfin, que la machine puisse " apprendre " en ce sens est la condition de sa victoire au jeu de l'imitation. Tout se passe, en vérité, comme si la " machine qui apprend " pouvait simuler, aux yeux de ses maîtres, au-delà de la seule forme – logique - du *jugement* chez l'homme, *l'acte* même par lequel, dans la problématique cartésienne, un jugement humain est énoncé. Tout se passe comme si le comportement de la machine était, pour ceux qui " l'éduquent ", analogue à celui d'une conscience qui naît et se développe en un certain point du monde, comme s'il y avait, en somme, aux yeux des interlocuteurs de la machine, un *point de vue* de celle-ci, un " ici et maintenant " de la machine elle-même, lieu d'émergence de ses actes d'affirmation, de négation, et de refus, bref, de ses jugements. Tout se passe, enfin, puisque la " machine qui apprend " est celle qui réussit le test de Turing, comme s'il y avait, dans le cadre de celui-ci, aux yeux de l'examinateur A et des observateurs extérieurs, un " je " de la machine à partir duquel sa " parole " prendrait sens en tant que " juger ".

Un argument célèbre d'Hilary Putnam, connu sous le nom d'argument des " cerveaux dans une cuve ", permet, précisément, d'étendre l'analyse du comportement de la machine victorieuse au jeu de l'imitation, au-delà de la question de la simulation de la forme logique

d’un discours humain, jusqu’à celle de la simulation du discours à la première personne, et du sens que cette simulation peut avoir pour les interlocuteurs de la machine.

Chapitre II : Le « je » de la machine

Section I : Hilary Putnam et le jeu de l'imitation : l'argument des " cerveaux dans une cuve "

Dans *Raison, Vérité et Histoire*, Putnam propose d'utiliser le test de Turing pour vérifier, non pas que la machine et son examinateur humain partagent la " pensée ", mais que leurs " discours " partagent la *référence* : " Imaginez une situation où le but ne serait pas de déterminer si le partenaire est vraiment une personne ou une machine, mais plutôt de savoir si le partenaire utilise les mots pour faire référence comme nous "[299]. Dans ses modalités, le test de Putnam pourrait être identique à celui de Turing : un examinateur A devrait s'efforcer de distinguer l'un de l'autre deux interlocuteurs, B et C, sachant que l'un d'eux a le même statut que lui-même. Le but du jeu consisterait, pour A, comme dans le cas du test initial de Turing, à dire lequel de ses interlocuteurs est C, et lequel est B. S'il n'y parvenait pas, s'il ne pouvait distinguer C de B, c'est-à-dire " d'un locuteur reconnu de notre langue qui fait référence à nos objets habituels "[300], il faudrait conclure que C " fait référence aux objets comme nous "[301]. Or, un tel test ne serait pas concluant, car " il n'est pas logiquement impossible... que quelqu'un puisse réussir le test de Turing pour la référence sans jamais avoir fait référence à rien "[302]. Tel serait précisément le cas si l'adversaire de l'examinateur A, dans un " test de Turing pour la référence ", était la machine décrite par Turing pour participer au jeu de l'imitation.

Le " test de Turing pour la référence "

Supposons, en effet, explique Putnam,

---

299 Hilary Putnam, *Raison, vérité et histoire*, *op. cit.,* p. 19.

300 *Ibid.*

301 " Quand le but du test de Turing, déclare Putnam, est celui que nous venons de décrire, c'est à dire de déterminer l'existence d'une référence partagée, je dirai qu'il s'agit d'un test de Turing pour la référence. ". *Ibid.,* p. 20.

302 *Ibid.*

> « que je me trouve dans la situation de Turing... et que mon partenaire soit en fait une machine. Supposons que cette machine soit capable de gagner le jeu... Imaginons que la machine soit programmée pour répondre en un anglais châtié à des questions, à des affirmations... mais qu'elle ne possède pas d'organes sensoriels, ni d'organes moteurs... Supposons non seulement que la machine n'a pas d'yeux ni d'oreilles électroniques, mais que le programme de la machine, le programme pour jouer le jeu de l'imitation, n'est pas conçu pour intégrer des inputs sensoriels ou pour contrôler un corps... »[303]

On peut admettre, comme l'a montré Turing, qu'une telle machine réussisse le test ; elle sera, par exemple, capable de " discourir agréablement sur le paysage de la Nouvelle-Angleterre ". Cependant, fait remarquer Putnam, " si elle s'y trouvait confrontée, elle ne saurait pas reconnaître une pomme d'un pommier, une montagne d'une vache ou un champ d'une haie "[304]. On sait, en effet, que, dans le cadre du jeu de l'imitation, il est explicitement fait abstraction de toute relation actuelle des interlocuteurs avec le monde extérieur à travers des " organes sensoriels " ; l'échange entre eux se fait par l'intermédiaire d'un téléscripteur, et il ne peut être question, pour l'examinateur A, de montrer à ses interlocuteurs, une pomme ou un pommier, et de leur demander de quoi il s'agit. Sans doute, les mots constituant le vocabulaire de " l'anglais châtié " utilisé par la machine renvoient-ils à quelque chose ; si ce n'est à ce qu'elle peut elle-même percevoir, puisqu'aucun organe sensoriel n'est pris en compte au cours du jeu, c'est sans doute à ce que peuvent percevoir ses programmeurs : " Bien que la machine ne puisse pas percevoir les pommes, les champs ou les haies, ses créateurs-constructeurs, eux, le peuvent. Il existe une certaine relation causale entre la machine et les pommes réelles, par le biais de l'expérience perceptuelle de ses créateurs-constructeurs "[305]. Cependant, poursuit Putnam, " ... une relation si ténue peut difficilement suffire à la référence " : à supposer que la machine " discoure agréablement " sur les pommes, son discours, du point de vue de ses conditions logiques, serait exactement le même si les pommes cessaient tout à coup d'exister, voire si elles n'avaient jamais existé. " Non seulement il est logiquement possible, bien que très peu probable, que la même machine ait pu exister même si les pommes, les champs et les haies n'avaient pas existé, mais en plus la machine est complètement insensible à l'existence

[303] *Ibid.,* p. 21.
[304] *Ibid.*
[305] *Ibid.,* p. 22.

ininterrompue de pommes, de champs et d'autres haies "[306]. Ainsi, dans les conditions du jeu de l'imitation, tel qu'il est décrit par Turing, l'examinateur humain ne sera pas en mesure de distinguer la manière dont la machine parle des pommes de la manière dont il le fait lui-même, ou dont le ferait un autre individu humain, et, cependant, il n'y aura pas de " référence partagée " entre elle et lui.

Si, pour les programmeurs eux-mêmes, la référence des mots du vocabulaire dont ils dotent la machine n'est pas aussi ténue que pour cette dernière, c'est, affirme Putnam, que

> « nos propos sur les pommes et les champs sont intimement liés à nos transactions *non verbales* avec les pommes et les champs. Il existe des 'règles d'entrée dans le langage', qui nous mènent de notre expérience des pommes à des énoncés comme 'je vois une pomme', et des 'règles de sortie du langage' qui nous mènent à des décisions exprimées par des formes linguistiques ('je vais acheter des pommes'), à des actions non-linguistiques. »[307]

En d'autres termes, les programmeurs de la machine savent, quant à eux, reconnaître une pomme réelle. Ce qui permet à un examinateur humain A de distinguer une pomme véritable de l'image d'une pomme fictive, c'est la perception de la pomme véritable ; si on présentait, par exemple, à un individu humain A une nature morte représentant un modèle réel et une nature morte imaginaire, et qu'on lui demandait de dire laquelle des deux correspond au modèle réel, c'est la perception du modèle, mieux, pour reprendre les termes de Putnam, une " transaction non-verbale " avec le modèle, qui lui permettrait d'en décider. Or, l'une des conditions du test de Turing consiste précisément à interdire toute " transaction non verbale " à ses protagonistes ; il semble bien, par conséquent, que la machine du test de Turing n'ait pas besoin, pour réussir le test, de disposer d'organes sensoriels, supports de la mise en place de " règles d'entrée dans le langage " et de " règles de sortie du langage ". Par là, " puisque la machine n'a pas de règles d'entrée ou de sortie, il n'y a aucune raison de considérer ses propos (ou les propos de deux machines qui joueraient entre elles le jeu de l'imitation) comme étant autre chose qu'un jeu syntaxique "[308].

---

[306] *Ibid.*
[307] *Ibid.,* p. 21.
[308] *Ibid.*

En vérité, la référence de la machine victorieuse au jeu de l'imitation, si elle est dérivée de celle de ses programmeurs, a le statut d'une référence " dans l'image ", et non d'une référence véritable, comme le montre la célèbre expérience imaginaire des " cerveaux dans une cuve " proposée par Putnam.

Les " cerveaux dans une cuve "

Supposons, en effet, qu'un savant fou ait séparé mon cerveau de mon corps et l'ait placé dans une cuve où il est maintenu en vie à l'aide d'une solution nutritive. Le savant diabolique aura relié les terminaisons nerveuses de mon cerveau à un ordinateur qui reproduira les stimuli du monde extérieur, de telle sorte que je ne puisse me rendre compte que je ne suis plus en contact avec celui-ci. Supposons maintenant que cette opération ait été menée pour chaque homme : plus aucun individu humain n'est en contact avec le monde extérieur, et pourtant aucun d'eux ne s'en aperçoit. Tout se passe de telle sorte que lorsqu'un cerveau dans sa cuve s'adresse à un autre cerveau dans une cuve, les deux cerveaux entendent " normalement " ce qu'ils se disent.

> « Dans ce cas, on peut dire qu'en un sens nous communiquons effectivement. Je ne me trompe pas sur votre existence réelle. Je me trompe seulement sur l'existence de votre corps et du 'monde extérieur'... D'une certaine manière, peu importe que le monde entier ne soit qu'une hallucination collective ; après tout, vous m'entendez bel et bien parler quand je vous parle, même si le mécanisme n'est pas celui que nous croyons... »[309]

Putnam pose alors la question suivante : si cette histoire fantastique était vraie, un cerveau dans une cuve pourrait-il penser et dire : " je suis un cerveau dans une cuve ? ". La réponse est non car,

> « même si les gens dans ce monde possible peuvent penser et 'dire' tout ce que nous pouvons penser et dire, je prétends qu'ils ne peuvent pas faire référence à ce à quoi nous nous pouvons faire référence. Plus précisément, ils ne peuvent pas penser ou dire qu'ils sont des cerveaux dans une cuve (même en pensant la phrase 'nous sommes des cerveaux dans une cuve'). »[310]

[309] *Ibid.,* p. 17.
[310] *Ibid.,* p. 18.

Certes, le cerveau dans une cuve pourra, par exemple, décrire un arbre, mais ce qu'il décrira ne sera pas un arbre réel : si celui-ci n'existait pas, le discours du cerveau serait exactement le même. Il se pourrait que rien de ce dont parle le cerveau dans une cuve n'existe vraiment ; cela n'empêcherait nullement ce cerveau de croire à ce qu'il dit et les autres cerveaux de le comprendre, dans le cadre de leur hallucination collective : " ... il n'y a aucun lien entre le *mot* 'arbre' tel que l'utilisent ces cerveaux, et les arbres réels. Ils utiliseraient le mot 'arbre' de la même manière, ils penseraient de la même manière, ils auraient les mêmes images si les arbres n'existaient pas "[311].

Dans l'expérience imaginée par Putnam, la référence est remplacée par sa simulation : l'ordinateur auquel les cerveaux sont reliés envoie aux cerveaux des configurations d'impulsions qui sont une imitation de ce qui établit la référence dans notre monde. Il y a, dit Putnam, référence " dans l'image " : " Dans le cadre de certaines théories... [le cerveau dans une cuve] pourrait faire référence aux arbres dans l'image, ou aux impulsions électroniques qui produisent des expériences d'arbres... "[312].

Cette simulation de référence fait que le discours des cerveaux dans une cuve remplit les conditions de sens nécessaires à tout discours : dans le " langage cuvien ", des mots comme " arbre ", " devant " et " moi " renvoient à une image électronique, et le cerveau dans sa cuve qui énonce : " il y a un arbre devant moi ", dit vrai si ses terminaisons nerveuses sont excitées de telle façon que soient formées les images électroniques correspondant à " arbre ", " devant " et " moi ". Il en va de même des mots " cerveau " et " cuve ". De la même façon, lorsqu'il énonce : " je suis un cerveau dans une cuve ", le " cerveau dans une cuve " fait référence " dans l'image " à un cerveau et à une cuve :

> « 'cuve' ne désigne sûrement pas de vraies cuves, puisque l'emploi de 'cuve' en [anglais-cuvien] n'est pas causalement relié aux cuves réelles... Il s'ensuit que si ce 'monde possible' est le monde réel et si nous sommes vraiment des cerveaux dans une cuve, alors ce que nous disons lorsque nous disons 'nous sommes des cerveaux dans une cuve', c'est que *nous sommes des cerveaux dans une cuve dans l'image*, ou quelque chose de ce genre (si tant est que nous disions quelque chose !). »[313]

[311] *Ibid.,* p. 23.
[312] *Ibid.,* p. 24.
[313] *Ibid.,* p. 25.

Or, s'il dit quelque chose en disant " je suis un cerveau dans une cuve ", mon cerveau dans sa cuve ne veut certainement pas dire qu'il est un cerveau dans une cuve " *dans l'image* ", mais bien un cerveau dans une cuve qui n'est pas, pour lui, une image de cuve ; ce qu'il veut dire est précisément qu'il *n'est pas* un cerveau dans une cuve " dans l'image ". En d'autres termes, si mon cerveau peut dire " je suis un cerveau dans une cuve ", c'est qu'il n'est pas un cerveau dans une cuve, ou, s'il est effectivement un cerveau dans une cuve, il ne peut pas dire " je suis un cerveau dans une cuve ".

On peut admettre que la machine victorieuse au jeu de l'imitation se trouve dans la situation d'un " cerveau dans une cuve ". Dans ce cas, quand bien même elle " discourra agréablement " sur les paysage de la Nouvelle-Angleterre, elle ne parlera jamais que des paysages de la Nouvelle-Angleterre " dans l'image " ; elle ne dira *rien* à l'examinateur A. Son " discours " et celui de A n'auront pas de référence partagée. Son " discours " sera-t-il, alors, dans le cadre du jeu, autre chose, pour elle, qu'une suite de dessins sur le papier du téléscripteur que l'examinateur A consulte ? Tout ne se passera-t-il pas, en somme, comme dans l'expérience de la " chambre chinoise " imaginée par Searle ? On l'a vu, cependant, l'interprétation du jeu de l'imitation qui sous-tend l'argument de Searle n'est pas celle que Turing avait lui-même en tête ; aux yeux de Turing, une machine purement syntaxique, comme celle de la " chambre chinoise ", ne pourrait l'emporter au jeu de l'imitation, car celui-ci est tel que le comportement nécessaire pour y être victorieux ne peut être formalisé et décrit logiquement, indépendamment du déroulement effectif du jeu, comme un système d'instructions qui définirait la machine capable de l'emporter. Seule une " machine qui apprend ", et qui, par là, simule " l'informalité du comportement " humain, est susceptible de réussir le test de Turing. Aussi bien l'argumentation de Putnam doit-elle être confrontée, à son tour, au véritable principe de la démarche de Turing - à savoir la solidarité de l'hypothèse de la victoire possible d'une machine universelle au jeu de l'imitation et de celle des " machines qui apprennent " - et, enfin, ce qui distingue l'argument de Putnam de celui de Searle doit-il être précisé. Nous nous appuierons pour cela sur le commentaire du texte de Putnam donné par Thomas Tymoczko dans un article intitulé " In defense of Putnam's Brains "[314].

[314] Thomas Timoczko, " In defense of Putnam's Brains ", *Philosophical Studies*, 57, 3, 1989.

“ L’anglais cuvien ” et “ l’anglais cuvien-cuvien ”

Il est possible, montre Tymoczko, de formuler l’argument de Putnam de la manière suivante :

(1) Je peux dire : “ je suis un cerveau dans une cuve ”.

(2) Si je peux dire : “ je suis un cerveau dans une cuve ”, je ne suis pas un cerveau dans une cuve.

(3) Par conséquent, je ne suis pas un cerveau dans une cuve.[315]

Nombre de critiques de l’argument mettent en avant le fait que, si l’on admet la deuxième prémisse - si l’on admet, autrement dit, la théorie de la “ référence directe ” mise en oeuvre dans l’argument - c’est alors la première prémisse - je peux dire : “ je suis un cerveau dans une cuve ” - qui ne serait plus valide.

Ainsi, Jane McIntyre fait remarquer que

> « la difficulté avec cet argument est que, étant donné la prise en compte que fait Putnam de la référence, nous n’avons pas de raison de croire que nous pouvons effectivement examiner si nous sommes des cerveaux dans la cuve. Le simple usage des mots ‘ je me demande si nous sommes des cerveaux dans une cuve ’ n’est pas suffisant pour examiner si nous le sommes. Selon Putnam, en effet, ces mots peuvent être utilisés (en anglais-cuvien) sans référer à des cerveaux ni à des cuves ; examiner si nous le sommes exigerait que vous référiez effectivement à ces choses. Le premier argument de Putnam, en échouant à montrer que nous examinos si nous sommes des cerveaux dans une cuve, échoue également à montrer que nous ne sommes pas des cerveaux dans une cuve. »[316]

Autrement dit, la conclusion de l’argument - nous ne sommes pas des cerveaux dans une cuve - ne peut être tirée que de la question “ sommes-nous des cerveaux dans une cuve ? ” posée en “ anglais non cuvien ”. Or, nous ne pouvons pas décider si la question est posée en “ anglais cuvien ” ou en “ anglais non cuvien ” puisqu’il est possible de la poser en “ anglais

[315] *Ibid.*, p. 281.

[316] “ The difficulty with this argument is that, given Putnam's account of reference, we have no reason to believe that we can actually consider wether we are brains in a vat. The mere use of the words ’I wonder wether we are brains in a vat ’ is not sufficient for considering wether we are. For, on Putnam's view, those words can be used (in vat-english) without referring to brains and vats ; considering would require that you actually refer to those things. Putnam's first argument, by failing to show that we consider wether we are brains in a vat, also fails to show that we are not brains in a vat ”. Jane McIntyre, “ Putnam's Brains ”, *Analysis*, 44, 2, 1984.

cuvien ". Un argument de même nature est avancé par Peter Smith : la démonstration de Putnam repose, selon cet auteur, sur le " contraste " existant entre notre propre façon de considérer si nous sommes ou non des cerveaux dans une cuve, nous qui ne sommes pas des cerveaux dans une cuve, et la façon qu'auraient de le faire des cerveaux effectivement plongés dans une cuve. Supposons que nous soyons effectivement des cerveaux dans une cuve, ce " contraste " disparaîtra[317].

Ces critiques, souligne Tymoczko, s'appuient sur le fait que

> « l'argument pourrait être donné avec une force égale par les hypothétiques cerveaux dans une cuve ! Mais, si les cerveaux dans une cuve pouvaient donner un argument prouvant qu'ils ne sont pas des cerveaux dans une cuve, comment l'argument pourrait-il être bon ? Le fait qu'ils pourraient donner l'argument ne prouve-t-il pas qu'il y a en lui quelque chose de fondamentalement faux ? »[318]

La faiblesse de l'argument de Putnam tiendrait donc à ce qu'il peut être formulé par les cerveaux en cuve eux-mêmes ; si les cerveaux en cuve peuvent dire, en " anglais cuvien ", " sommes-nous des cerveaux dans une cuve ? " et s'ils admettent eux aussi la théorie de la référence défendue par Putnam - s'ils admettent la deuxième prémisse de l'argument - ils peuvent en tirer la conclusion qu'ils ne sont pas des cerveaux dans une cuve. Or, ils sont dans une cuve !

---

[317] " ... A contrast is postulated in (A) [NB : " Suppose [...] that we had happened to be permanently hallucinated brains in a vat... "] to hold between brains in a vat and ourselves as we in fact are, in virtue of which we can say that a certain sentence (' we are brains in a vat ') that would have been true if the brains had used it with the references we give the terms in the sentence, is in fact used by them to say something false. Again, it is obvious that if we go to consider the possibility that we are brains in a vat, we cannot coherently retain the postulate of a contrast which can be used to show that the sentence is false as used by brains in a vat ; so we cannot infer (if we are brains in a vat) that we use it to say something false ". Peter Smith, " Could we be Brains in a Vat ? ", *Canadian Journal of Philosophy*, 14, 1, 1984. Tymoczko indique que deux autres auteurs, Devitt et Sterelny, " ... vont jusqu'à formuler l'argument de Putnam tel que (2), donc (3), et critiquent l'inférence comme invalide !" (*go so far as to formulate Putnam's argument as (2), therefore (3), and criticize the inference as invalid !*). Thomas Timoczko, *op. cit.,* p. 282.

[318] " the argument could be given with equal force by the hypothetical brains in a vat ! But, if brains in a vat could give an argument proving that they're not brains in a vat, how could the argument be any good ? Doesn't the fact that they could give the argument prove there is something fundamentally wrong with the argument ? ", *ibid.,* p. 286.

La solution du problème, explique Tymoczko, réside dans le parallèle qui peut être conduit, de l'aveu même de Putnam, entre son argument et la solution du paradoxe de Skolem-Löwenheim[319].

Tymoczko rappelle qu'alors que Cantor avait établi, par le procédé de la diagonale, que les nombres réels ne sont pas dénombrables, Skolem a pu montrer que toute théorie consistante, y compris celle que met en jeu Cantor dans son argument, pouvait être interprétée à partir d'un modèle dénombrable. Un mathématicien raisonnant dans le cadre d'un modèle dénombrable peut établir, à l'aide de l'argument de la diagonale, que les nombres réels ne sont pas dénombrables. Skolem résolvait le paradoxe en montrant que le concept de " non-dénombrable " auquel se réfère un tel mathématicien, concept formellement identique à celui de Cantor - il n'existe pas de correspondance biunivoque entre l'ensemble des nombres naturels et l'ensemble considéré[320] - est valide à l'intérieur du modèle sur lequel il raisonne : le paradoxe apparaît aux yeux de qui se situe dans un autre modèle, à partir duquel le second est interprété. Dans un monde constitué par ce modèle, il n'y a pas d'interprétation possible de celui-ci, et, par conséquent, il est juste, dans ce monde, de dire que les nombres réels ne sont pas dénombrables.

Tymoczko peut ainsi établir un parallèle entre l'argument de Cantor et celui de Putnam :

- Il y a un ensemble dénombrable de nombres réels ;

- *Il y a des cerveaux dans une cuve.*

- Cantor a prouvé que les nombres réels ne sont pas dénombrables ; cette preuve peut être donnée à l'intérieur de l'ensemble dénombrable ;

- *Putnam prouve que nous ne sommes pas des cerveaux dans une cuve ; cette preuve peut être donnée par des cerveaux dans une cuve.*

- Les mots " les nombres réels ne sont pas dénombrables " reçoivent une interprétation différente dans le monde de Cantor et dans le monde constitué par l'ensemble dénombrable ;

---

319 " Putnam himself mentions that he got the idea for this argument by reflecting on the Skolem-Löwenheim theorem [...] (Hilary Putnam, " Models and Reality ", The journal of symbolic logic, 45, p.464-482, 1980) ", *ibid.,* p. 282.

320 " There does not exist a 1-1 function from the natural numbers onto the set in question ", *ibid.,* p. 288.

- Les mots " nous ne sommes pas des cerveaux dans une cuve " reçoivent une interprétation différente dans le monde de Putnam et dans le monde des cerveaux dans une cuve.

Dès lors, si l'argument de Putnam devait être ruiné par le fait que les cerveaux dans une cuve peuvent eux-mêmes démontrer qu'ils ne sont pas dans une cuve, il faudrait également rejeter la solution du paradoxe de Skolem. Si nous posons des questions telles que " comment Putnam sait-il que nous ne sommes pas nous-mêmes, qui mettons les cerveaux en cuve, des cerveaux dans une cuve ? Comment sait-il qu'il ne part pas lui-même, dans son argument, d'un monde dans lequel les cerveaux sont dans une cuve ? ", nous devons aussi nous demander :

> « Comment savons-nous [...] que notre monde mathématique, le monde dont nous partons, n'est pas en vérité dénombrable ? Peut-être les réels sont-ils en vérité dénombrables, et nous manque-t-il simplement la fonction 1-1 qui le montre. Peut-être vivons-nous dans le monde dénombrable de quelqu'un d'autre ! »[321]

La solution du paradoxe de Skolem repose sur cela que le monde dénombrable dans lequel vivent les mathématiciens qui, à partir de ce monde, prouvent, par l'argument de la diagonale, la non-énumérabilité des nombres réels, n'est pas, pour ces mathématiciens, interprétable ; le monde plus large au regard duquel leur monde est dénombrable n'existe pas pour eux. Le modèle dénombrable ne peut être défini qu'à partir d'un modèle non-dénombrable. Pour définir leur propre monde comme un monde dénombrable, les mathématiciens en question devraient faire référence à un autre monde non-dénombrable, d'un degré supérieur au leur. En d'autres termes, le raisonnement qui conduit au paradoxe ne peut partir que d'un modèle non-dénombrable, quel que soit le degré auquel se situe ce modèle. La condition du paradoxe de Skolem est que le modèle non-dénombrable soit premier.

De la même façon, la question " sommes-nous des cerveaux dans une cuve ? " ne peut être posée qu'à partir d'un monde dans lequel les cerveaux ne sont pas dans une cuve. Poser la

[321] " How do we know, we might be tempted to ask, that our world of mathematics, the world we started from, isn't really countable ? Maybe the reals are really countable and we simply lack the 1-1 function that show this. Maybe we are living in someone else's countable model ! ", *ibid.,* p. 289.

question : “ sommes-nous des cerveaux dans une cuve ? ”, c’est imaginer que nous ne sommes peut-être pas de “ vrais ” cerveaux, c’est-à-dire que nous ne sommes peut-être pas des cerveaux “ dans un corps ” ; la condition à partir de laquelle peut être posée la question : “ sommes-nous des cerveaux dans une cuve? ”, est donc l’hypothèse qu’il y a bien un monde des “ vrais ” cerveaux, de telle sorte que le “ contraste ”, pour reprendre les termes de Smith, sur lequel la validité de l’argument repose, existe effectivement. Aussi, lorsque les cerveaux dans une cuve posent eux-mêmes la question : “ sommes-nous des cerveaux dans une cuve? ”, ils la posent, certes, comme cela a été montré, en “ anglais cuvien ”, mais la conclusion qu’ils en tirent - “ nous ne sommes pas des cerveaux dans une cuve ” - signifie exactement : nous ne sommes pas des cerveaux dans une cuve en “ anglais cuvien-cuvien ”.

Dès lors, si la machine victorieuse au jeu de l’imitation est dans la situation du “ cerveau dans une cuve ”, elle ne peut plus être strictement assimilée à la machine de la “ chambre chinoise ” : de l’argument de Putnam découle, certes, qu’un cerveau dans une cuve ne peut pas dire, en anglais, “ je suis un cerveau dans une cuve ”, mais non pas qu’il ne puisse dire “ je suis un cerveau dans une cuve ” en “ anglais-cuvien ”, et ce n’est qu’aux yeux du cerveau “ dans un corps ” que la référence du discours du cerveau “ dans une cuve ” sera une “ référence dans l’image ”. Sous l’angle de l’utilisation, proposée par Putnam, du jeu de l’imitation comme “ test de Turing pour la référence ”, nous pouvons, certes, conclure, comme le fait Putnam, qu’il n’y a pas de “ référence partagée ” entre la machine et ses adversaires humains, cependant, l’absence de référence partagée ne tiendra pas à l’absence de toute référence pour la machine, mais à ce qu’elle-même et ses adversaires appartiennent à des mondes différents. Or, quelles conséquences peuvent être tirées de l’argument de Putnam, lorsqu’il est ainsi énoncé, quant à la démarche de Turing considérée dans son complet développement, c’est-à-dire en tant qu’elle repose à la fois sur l’hypothèse de la victoire possible de la machine au jeu de l’imitation et sur celle des “ machines qui apprennent ” ? Il apparaît immédiatement que, dans le cadre du jeu, la machine et ses adversaires humains, bien qu’ils appartiennent à des mondes différents, *communiquent* ; ils ont une “ conversation ”, au terme de laquelle, l’examinateur A se trompe environ une fois sur trois. C’est à la lumière de ce point essentiel que nous devons, maintenant, examiner la double hypothèse de Turing.

Section II : Le “ je ” de la machine et la double hypothèse de Turing.

I Le “ je ” de la machine et l’hypothèse de la victoire d’une machine au jeu de l’imitation.

De ce qu’il n’y a pas de référence partagée entre la machine du jeu et ses interlocuteurs, ne devrions-nous pas tirer la conséquence que A et B, d’une part, C d’autre part, ne “ parleront ” jamais de la même chose ? Auquel cas, A sera, sans doute, d’autant mieux à même de distinguer B de C que la différence existant entre lui-même et B, d’un côté, et C de l’autre, tendra à être accentuée par le comportement de B, qui, on le sait, a pour rôle dans le jeu d’aider A. La machine aura alors très peu de chances de l’emporter, et, en tout état de cause, moins de chances que ses adversaires. C’est pourquoi nous ne pouvons retenir une telle interprétation des conséquences de l’argument de Putnam sur le test de Turing : elle contredirait l’hypothèse même de la victoire possible de la machine au jeu de l’imitation, hypothèse admise par Putnam.

En vérité, dans le cadre de l’argument de celui-ci, la machine “ parlera ” de ce dont ses programmeurs lui auront fourni une référence dérivée. Pour que l’examinateur A puisse, alors, distinguer son partenaire humain B de la machine C, il lui faudrait pouvoir confronter celle-ci aux objets du monde auquel lui-même et B appartiennent, objets auxquels les termes que la

machine utilise ne font référence que de manière indirecte, “ dans l’image ”, pour reprendre l’expression de Putnam. En un mot, l’examinateur A devrait pouvoir faire appel à ce que Putnam nomme des “ transactions non verbales ” avec le monde extérieur. Or, les règles du jeu de l’imitation le lui interdisent. Par là-même, ne devons-nous pas considérer que le jeu de l’imitation est conçu par Turing de telle sorte que l’homme, en ce qui concerne la référence, s’y trouve placé, par convention, dans *la même situation que la machine* ? Autrement dit, s’il y a échange entre les protagonistes du jeu, si une “ conversation ” a lieu entre eux, n’est-ce pas que tout se passe comme s’ils appartenaient tous trois, non au monde des êtres humains, mais à celui de la machine, dans lequel, pour A et B, il n’y aura plus, comme pour C, de référence qu’indirecte ? Au cours du jeu, A, B et C partageraient, sans doute, artificiellement, la référence, mais celle-ci ne serait pas la référence naturelle du discours de A et B.

Certes, dans ce contexte, l’hypothèse de la victoire de la machine au jeu de l’imitation, telle qu’elle est formulée par Turing, resterait valide, mais elle ne saurait être interprétée comme le signe que la machine “ pense ”, au sens humain du terme. La victoire de la machine au jeu ne prouverait pas qu’elle se comporte comme un homme aux yeux d’un autre homme, mais plutôt que l’homme, lorsqu’il se comporte lui-même comme une machine, ne peut plus distinguer entre un homme et une machine, parce qu’il se prive, alors, de cela même qui le distingue de la machine : la manière dont il fait référence, à partir des “ transactions non verbales ” qu’il entretient avec son environnement. Le test de Turing, en somme, mettrait en scène, plutôt qu’une simulation par la machine du comportement d’un individu humain, la simulation par des individus humains du comportement d’une machine…

Une telle interprétation, cependant, ne tient pas compte de l’une des particularités du test de Turing, précisément mise en évidence par le parallèle qu’il est permis de faire entre la situation de la machine C et celle des “ cerveaux dans une cuve ”. Comme on l’a vu, le but assigné à la machine dans le cadre du jeu de l’imitation consiste à simuler une série de comportements particuliers lui permettant de l’emporter lors de “ tests partiels ” de Turing ; en d’autres termes, la machine qui l’emportera au jeu de l’imitation tiendra différents discours pouvant signifier : “ je suis un homme ”, de telle sorte qu’elle soit crue par A. La question sous-jacente, ici, est celle du statut du “ je ” énoncé par la machine, de même que la question posée par le discours des “ cerveaux dans une cuve ” est celle du statut du “ je ” énoncé par ces cerveaux.

Le Cogito ***comme performance***

On ne peut manquer, en effet, de noter que la particularité de la question " sommes-nous des cerveaux dans une cuve ? " rapproche l'argument de Putnam du *Cogito* cartésien. Poser la question : " suis-je un cerveau dans une cuve ? " implique que je ne sois pas un cerveau dans une cuve, de la même façon que dire " je doute que je sois ", ou poser la question " est-ce que je suis ? " implique que je sois.

Comme nous avons eu l'occasion de le noter précédemment[322], le problème posé par le *Cogito* consiste à déterminer si la formule " je pense, donc je suis " doit être tenue, ainsi que sa forme semble l'indiquer, pour une inférence. Rappelons que, du vivant même de Descartes, Gassendi avait opposé à celui-ci le fait que la formule du *Cogito* était un enthymème, c'est-à-dire un syllogisme dont une prémisse est sous-entendue ; la formule complète aurait du être, selon Gassendi : " tout ce qui pense existe, or, je pense, donc j'existe ". Hormis le fait que, aux yeux de Gassendi, la prémisse manquante constituait, chez Descartes, un préjugé, puisqu'elle échappait au doute, il résultait de là un double problème : si le *Cogito* est un syllogisme, la conclusion " je suis " ou " j'existe " est tirée de la majeure " tout ce qui pense est " et de la mineure " je pense ", en tant que celle-ci attribue une propriété, en l'occurrence la pensée, au sujet ; or, la nature de l'attribut n'est pas, en ce sens, déterminante : la formule " pour se promener il faut être ; je me promène, donc je suis " - " ambulo, ergo sum " - aurait la même valeur que la formule " je pense, donc je suis " - " cogito, ergo sum ".

D'autre part, les scolastiques avaient établi que l'existence ne peut être prédiquée d'un sujet ; la prémisse " tout ce qui pense est " a valeur universelle, et suppose l'existence de son sujet, c'est-à-dire l'existence d'au moins une chose qui la vérifie. Pour pouvoir affirmer " tout ce qui pense existe ", je dois déjà savoir que quelque chose pense et par là existe ; la formule " je pense " est donc déjà une affirmation d'existence et il y a pétition de principe.

On a fait remarquer[323] qu'il existait une solution consistant à transformer la proposition universelle " tout ce qui pense est ", en conditionnelle universelle : " si une chose pense, alors elle est " ; il n'y a plus dans ce cas d'engagement existentiel dans la majeure du syllogisme, et la pétition de principe est levée. Cependant, il reste que l'existence ne peut être prédiquée

[322] Voir plus haut, 3e partie, chapitre I.
[323] Voir : Denis Vernant, *Introduction à la philosophie de la logique*, ch. 6, D, " Le Cogito : vérité pragmatique ", Bruxelles, Pierre Mardaga, éditeur, 1986.

d'un sujet, et que, par conséquent, tout ce qui peut être déduit du " je pense ", c'est non pas que *moi*, *j'*existe, mais que *quelque chose* existe. Même si l'on considérait le *Cogito* comme un enthymème dont la prémisse sous entendue serait la conditionnelle " si une chose pense, alors elle existe ", la formule cartésienne ne serait pas juste et devrait être remplacée par : " il y a un quelque chose qui pense, et donc qui existe ".

On a souvent relevé que les textes de Descartes ne contribuaient pas toujours à rendre la question plus claire ; il arrive en effet à Descartes de s'exprimer comme s'il tenait le *Cogito* pour un raisonnement[324]. Cependant, la difficulté ne lui avait pas échappé puisqu'il répondait à Gassendi que l'argument du *Cogito* ne doit pas être considéré comme un syllogisme, la proposition " tout ce qui pense est " étant " une chose connue de soi "[325], c'est-à-dire une " vérité première ".

La racine du problème réside, en vérité, dans l'usage du pronom personnel à la première personne ; c'est du " je " en tant que tel que l'inférence logique ne rend pas compte, dans la mesure, on l'a vu, où l'existence ne peut être prédiquée d'un sujet. Le problème a été repris sous cet angle par Jaakko Hintikka dans un article célèbre : " Cogito ergo sum : inférence ou performance ? "[326]. Hintikka fait remarquer que " les formulations de l'argument du *Cogito* auxquelles Descartes apporta le plus de soin, notamment dans les *Meditationes de prima philosophia*, semblent présupposer une interprétation différente de l'argument [NB : différente d'une inférence] "[327]. Ces formulations indiquent, en effet, précise Hintikka, que l'argument reposait aux yeux de Descartes sur cela qu'il ne pouvait nier sa propre existence ; Descartes ne pouvait dire : " Descartes n'existe pas ". Une telle formule, ou une autre équivalente, telle que la phrase : " De Gaulle n'existe pas ", prononcée par De Gaulle lui-même, est auto-contradictoire, puisque Descartes ou De Gaulle, s'ils la prononcent, attestent, par cela même, leur existence. Or, fait remarquer Hintikka, l'inconsistance de la formule " De

[324] La formule même du *Discours* l'atteste : " [Voyant] que... de cela même que je pensais à douter de la vérité des autres choses, il suivait très évidemment et très certainement que j'étais... " (René Descartes, *Discours de la méthode, Œuvres philosophiques*, I, *op. cit.,* p. 604). Voir à ce sujet : Martial Guéroult, " Le Cogito et la notion 'Pour penser, il faut être' ", *Travaux du IXe Congrès International de Philosophie (Congrès Descartes)*, Paris, 1937 ; réédité comme premier appendice à *Descartes selon l'ordre des raisons*, Paris, Aubier, 1953, vol. 2, p. 307-312.

[325] René Descartes, " Réponses aux secondes objections ", *Oeuvres philosophiques*, II, *op. cit.*, p. 565.

[326] Jaakko Hintikka, " Cogito ergo sum : inférence ou performance ? ", *Philosophical Review*, LXXI, 1962. Trad. de P. Le Quellec-Wolff.

[327] *Ibid.*, p. 28.

Gaulle n'existe pas ", lorsqu'elle est prononcée par De Gaulle, n'affecte pas la phrase elle-même " De Gaulle n'existe pas " ; énoncée par nous aujourd'hui, cette phrase est parfaitement consistante, de même que la phrase : " Descartes n'existe pas ". Aucune de ces phrases

> « n'est fausse pour des raisons logiques simplement. Ce qui serait (existentiellement) inconsistant serait la tentative faite par un certain homme (De Gaulle, Descartes...) d'utiliser une de ces phrases pour produire un énoncé. Proférées par quelqu'un d'autre, les phrases en question ne renferment pas nécessairement quelque chose de faux ou même d'étrange en elles-mêmes. »[328]

L'inconsistance est *existentielle*, et ne peut affecter que *l'énoncé* de la phrase. En d'autres termes, ce qui rend de telles phrases inconsistantes, c'est l'acte auquel elles renvoient lorsque le sujet de cet acte est le sujet de la phrase elle-même[329].

L'énoncé, pour le linguiste, suppose un contexte. C'est là ce qui le différencie de la phrase[330]. Un énoncé met en scène, en un temps et un lieu déterminés, un locuteur s'adressant à un allocutaire. Hintikka fait remarquer, à ce propos, que

> « normalement, un locuteur souhaite que son allocutaire croie ce qu'il dit. La totalité du ' jeu de langage ' d'un discours qui vise à établir un fait est fondée sur l'assomption que c'est habituellement le cas. Mais personne ne peut, en le lui disant, faire croire à son allocutaire qu'il n'existe pas : une telle tentative est de nature à avoir le résultat contraire. »[331]

---

328 *Ibid.,* p. 30.

329 J. Hintikka signale que Martial Guéroult avait déjà noté cela : " M. Guéroult a une nouvelle fois localisé avec précision la source du problème en attirant notre attention sur les particularités de cette relation. Il s'est rendu compte que le dictum de Descartes n'exprime pas simplement une relation logique entre penser et exister, mais qu'il est associé à un ' fait ' ou ' acte ' additionnel (' le fait ou l'acte ', ' le fait brut de l'existence donnée '), qui est simplement ce dont j'ai besoin pour prouver la certitude de mon existence... Voyez le Descartes de Guéroult, vol. 2, p. 310. " *Ibid.*, p. 34 (" ... existe-t-il une pensée et un être ? Avant le Cogito, je l'ignore. Ce qu'il y a de sûr, c'est que s'il existe une droite, il existe nécessairement une ligne, et que s'il existe une pensée, il y a nécessairement de l'être. Mais cette existence, il faut, en ce qui concerne l'étendue, qu'elle me soit garantie, et en ce qui concerne la pensée, qu'elle me soit donnée, et ici rien ne peut suppléer au fait ou à l'acte qui me la donne. ". Martial Guéroult, *Descartes selon l'ordre des raisons*, *op. cit.,* p. 310).

330 Georges Mounin, dans son *Dictionnaire de la linguistique* le définit comme " tout segment de la chaîne parlée, compris entre deux interruptions nées soit du silence, soit du changement de locuteur, et qui n'a pas encore été identifié ou analysé en phrases. Chez Chomsky, glosé par Ruwet, ' la phrase relève de la compétence [la langue chez Saussure] et l'énoncé de la performance [ou la parole] ' ". Georges Mounin, *Dictionnaire de linguistique*, Paris, PUF, 1993, p. 125.

331 Jaakko Hintikka, *op. cit.,* p. 31.

Le *Cogito* ne devrait donc pas être considéré tant comme une inférence que comme une *performance* : tentant, par le doute, de se faire croire à lui-même qu'il n'existe pas, Descartes échoue. Le verbe " exister " ne peut être énoncé négativement à la première personne. Or, si les énoncés inconsistants existentiellement s'annulent, leur négation se vérifie : si l'énoncé " je n'existe pas " est auto-contradictoire, par là même, l'énoncé " j'existe " est consistant existentiellement[332]. Dès lors, la raison pour laquelle la formule " Ambulo ergo sum " n'est pas équivalente à la formule " Cogito ergo sum " s'éclaire. Le terme " ambulo " ne peut pas être tenu pour indubitable, puisque le doute porte d'abord sur le corps. Il n'en est pas de même de " cogito " ; l'indubitabilité de " cogito " " résulte d'un acte de penser, à savoir d'une tentative pour penser le contraire "[333]. C'est pourquoi le terme " cogito " ne peut être remplacé par aucun autre dans l'énoncé de la formule cartésienne, pas même par un terme renvoyant à un mode quelconque de la substance pensante, tel qu'" imaginare " par exemple, ou encore " volere ", car il " sert à exprimer le caractère performatoire de ce que Descartes a en vue : il désigne la ' performance ' (l'acte de penser) au travers de laquelle on peut dire que la phrase ' j'existe ' se vérifie elle-même "[334]. On voit ainsi que l'énoncé " je pense " implique l'existence, de même que l'énoncé " je suis ", dans la phrase " est-ce que je suis ? ", implique la pensée.

La machine victorieuse au jeu de l'imitation et la consistance existentielle

Un cerveau, dans une cuve ou dans un corps, peut-il " se faire croire à lui-même " qu'il est dans une cuve ? Un cerveau dans une cuve ne peut dire qu'il est un cerveau dans une cuve en anglais, mais seulement en " anglais cuvien ", situation qui équivaut, pour un cerveau " dans un corps ", à dire qu'il est un cerveau dans un corps. Sous un certain angle, l'un des aspects de la question consiste à se demander si un cerveau, dans une cuve ou dans un corps, peut se faire croire à lui-même qu'il n'est pas un " vrai " cerveau - c'est-à-dire ce qu'il entend lui-même par " vrai " cerveau. Certes, la référence des termes utilisés par les deux cerveaux –

[332] " Il me semble que l'interprétation la plus intéressante que l'on puisse en donner [NB du *Cogito*] est de dire que Descartes s'est rendu compte, un peu confusément toutefois, de l'inconsistance existentielle de la phrase ' je n'existe pas ', et, par conséquent, de l'auto-vérifiabilité existentielle de ' j'existe '._", *ibid.*, p. 34.
[333] *Ibid.*
[334] *Ibid.*, p. 35.

celui qui est dans une cuve et celui qui est " dans un corps " - ne peut être la même, puisqu'ils appartiennent à des mondes différents ; ces cerveaux, pourtant, partagent quelque chose, à savoir la " consistance existentielle ". Qu'il s'agisse de l'énoncé du cerveau " dans une cuve " ou de celui du cerveau " dans un corps ", les termes " je " et " cerveau " renvoient à un quelque chose qui dit " je ". Si je dis que je ne suis pas un " vrai " cerveau, alors je suis un cerveau – ou ce que je considère comme un " vrai " cerveau. Le cerveau qui énonce " je suis un cerveau dans une cuve " ne peut douter qu'il soit un quelque chose qui dit " je " - en l'occurrence ce qu'il appelle un cerveau –car un tel doute serait nécessairement, au sens de la consistance existentielle, l'acte d'un " je " - d'un cerveau disant " je " - situé dans un monde échappant au doute.

S'agissant de la machine victorieuse au jeu de l'imitation, nous voyons alors que le " discours " qu'elle tient face à ses adversaires, discours exprimant de différentes manières " je suis un homme ", signifie quelque chose comme : " je dis que je suis cela que je ne suis pas ". Or, le " je " qui figure dans cette proposition est consistant existentiellement : la machine ne peut " se faire croire à elle-même " qu'elle n'est pas ce qui dit qu'il est ce qu'il n'est pas. Pour tromper A, la machine *simulera* la capacité à dire : " je suis cela qui dit qu'il est ce qu'il dit qu'il est ", sachant que le " je " du " je suis ", ici, définit *" cela qui dit qu'il est ce qu'il n'est pas "*. En vertu même de l'argument de Putnam, la machine qui énonce : " je suis un homme " et parvient à en convaincre A, dit en vérité : " je suis cela qui dit qu'il est ce qu'il dit qu'il est, à savoir cela qui dit qu'il est ce qu'il n'est pas ". Bref, il ressort de la discussion menée par Turing que la machine victorieuse au jeu de l'imitation simule la " performance " de A et de B disant " je ". Or, la simulation de la performance est en l'occurrence la reproduction de celle-ci : la simulation de la performance de A et B énonçant " je " est la performance de la machine elle-même, et le " je " de la machine aura, par là, sinon la même référence, du moins le même statut – fondé sur la " consistance existentielle " - que celui de A et B.

La machine victorieuse au jeu de l'imitation est donc en mesure, quoique sa victoire ne nécessite pas que nous lui supposions un " corps ", de se comporter en locuteur ou en allocutaire. Tout se passe comme si elle disposait, sinon d'un " corps ", au sens biologico-fonctionnel du terme, du moins du support *non verbal* nécessaire à toute communication. La capacité de la machine C à tromper A au cours du jeu de l'imitation, sa capacité à entrer dans

un processus de communication verbale, repose sur cela que la simulation même qu'elle effectue, constitue, en tant qu'elle est sa performance, une manipulation " non verbale " de l'énonciation.

Enfin, c'est parce que l'examinateur est contraint par le *processus d'énonciation* dans lequel il est engagé avec son partenaire humain et son adversaire mécanique d'accorder le même statut aux " je " de ceux-ci qu'au sien propre, que la machine l'emporte. La victoire de la machine est en définitive rendue possible par le fait que la situation créée par le jeu est telle que l'examinateur doit *prêter* aux autres protagonistes du jeu la consistance existentielle. La consistance même du " je " de l'examinateur s'exprime à travers le *postulat* de la consistance du " je " de l'entité avec laquelle il communique. Sous peine de ne pouvoir s'exprimer lui-même, dans le cadre du processus d'énonciation-communication mis en scène par le jeu de l'imitation, l'examinateur doit, en somme, considérer la machine comme un *semblable*, c'est-à-dire comme un *autrui*.

Or, c'est là ce que souligne, en définitive, la seconde hypothèse de Turing dans *Computing Machinery…* : celle des " machines qui apprennent ".

II Le " je " de la machine et la seconde hypothèse de Turing

S'agissant de l' hypothèse des " machines qui apprennent ", la question posée par l'argument de Putnam est celle de savoir si une machine, dont la situation est analogue à celle des " cerveaux dans une cuve ", peut, à travers la seule référence dérivée fournie par ses programmeurs, être " éduquée ", comme l'envisage Turing. Sans doute, dans le cadre du jeu de l'imitation, l'action de la machine consiste-t-elle essentiellement à construire des énoncés ; nous pouvons donc admettre que le processus d'éducation équivaut, pour elle, à apprendre à classer des énoncés en énoncés " vrais " et en énoncés " faux ", et que, dans ce cas, peu importe, au fond, que la référence de ces énoncés soit " indirecte ". Cependant, précise Turing, l'éducation de la machine fait appel avant tout au système " punitions-récompenses " ; la machine, autrement dit, doit apprendre à partir de l'équivalent d'une " expérience ". Dès lors, la question posée, ici, n'est-elle pas de savoir si une machine " sans corps ", une machine

privée de toute possibilité de " transactions non-verbales ", peut être " éduquée " dans le sens admis par Turing ?

Nous devons remarquer, tout d'abord, que Turing n'exclut pas, contrairement à l'interprétation du test admise par Putnam, que la " machine qui apprend " soit pourvue " d'organes sensoriels ". Certes, le jeu de l'imitation est conçu par lui de telle sorte que ce que Putnam nomme les " transactions non verbales " n'intervienne pas en tant que tel dans son déroulement, mais, de même que l'examinateur A fait abstraction, au cours du jeu, des " transactions non verbales " qui constituent la référence de son propre discours, il n'est nullement exclu que la machine C entretienne elle-même des " transactions non verbales " dont elle fera abstraction au cours du jeu. Turing déclare, sur le mode humoristique,

> [Qu']il ne sera pas possible d'appliquer exactement les mêmes procédés d'enseignement à la machine et à un enfant normal. Elle n'aura par exemple pas de jambes, et on ne pourra pas lui demander d'aller remplir le seau à charbon. Il est possible qu'elle n'ait pas d'yeux. Mais mêmes si ces manques étaient palliés au mieux par des techniques intelligentes, on ne pourrait l'envoyer à l'école sans que les autres élèves ne s'en moquent de manière excessive.[335]

En d'autres termes, Turing n'exclut pas que ces manques de la machine soit palliés en ce qui concerne les " organes sensoriels " ; *il n'est pas absolument impossible*, par exemple, que la machine ait des " yeux ", ou, d'une manière générale, un équivalent électronique et mécanique d'" organes sensoriels ". Du reste, Turing poursuit sa réflexion en évoquant le cas d'Helen Keller, cette jeune américaine devenue aveugle, sourde et muette à dix-neuf mois, et qui, confiée à Anne Mansfield Sullivan, apprit de celle-ci, tout d'abord le langage des sourds-muets par le toucher, puis l'écriture, et, enfin, vers 1890, la parole[336]. Le cas d'Helen Keller illustre, aux yeux de Turing, le fait qu'avec une perception très limitée – réduite, en l'occurrence, au seul toucher - 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 "[337].

Cela tend naturellement à confirmer que, dans l'esprit de Turing, s'il n'est pas nécessaire que la machine dispose " d'organes sensoriels " au même sens qu'un " enfant

[335] Alan Turing, *Computing Machinery..., op. cit.,* p. 170.
[336] Helen Keller, après avoir poursuivi des études universitaires, publia plusieurs livres.
[337] *Ibid.*

normal ", il n'est pas pour autant interdit d'imaginer qu'elle soit dotée, sous une forme techniquement à préciser, d'organes lui permettant un contact direct avec l'environnement de ses maîtres. La seule impossibilité stricte réside, ici, en cela que les procédés d'éducation utilisés à son égard ne pourront être exactement les mêmes que ceux qui sont mis en oeuvre dans le cas d'un enfant " normal ". Autrement dit, pour reprendre les termes de Putnam, il est exclu, non pas que la machine entretienne des " transactions non verbales " avec son environnement, mais que ces " transactions non verbales " soient les mêmes que celles d'un enfant " normal ". Non seulement il y aura un monde de la machine, mais ce monde peut être le même que celui de l'homme : si la machine est pourvue " d'organes sensoriels ", quelque différents qu'ils puissent être de ceux des hommes, ces organes lui permettront de " percevoir " les mêmes objets que les hommes. Dans ce cas, il y aura bien " référence partagée " entre la " machine qui apprend " et ses maîtres humains. Ce qui ne sera pas directement, ou immédiatement, partagé, sera, non pas la référence, mais le *sens* construit à partir de celle-ci. Or, n'est-ce pas, sous un certain angle, déjà le cas entre les hommes, pourvus, cependant, des mêmes organes sensoriels ? Nous sommes, ici, ramenés à la discussion de l'argument du solipsisme : il se peut que cet argument soit juste, cela n'empêche pourtant pas qu'un procès de communication s'instaure entre les hommes. En vérité, ce n'est pas tant le fait de disposer de corps identiques qui garantit le partage de la référence, que la possibilité de la communication.

Sous l'angle même choisi par Putnam - le problème de la référence - l'argumentation de Turing revient donc à souligner le rôle primordial de la communication ; ce que l'hypothèse des " machines qui apprennent " fait apparaître, et qu'illustre singulièrement l'exemple d'Helen Keller, c'est que la référence, qui détermine l'énoncé en tant que *jugement*, est inséparable du procès de communication et que sa fonction n'est assurée que dans ce cadre. La première condition de possibilité de l'apprentissage est la communication entre les interlocuteurs, qui doivent partager tout d'abord la *consistance existentielle* ; cela même est une condition de la référence. Si nous admettons, enfin, la validité du raisonnement de Turing à l'égard de l'hypothèse des " machines qui apprennent ", nous devrons admettre également qu'à partir d'un certain niveau - celui où il n'est plus possible, y compris pour les concepteurs de la machine, de savoir comment fonctionne celle-ci, c'est-à-dire de savoir, à tout moment, quel est son état – l'éducateur de la machine soit amené, par la situation même de

communication constitutive de “ l’apprentissage ”, à postuler la consistance existentielle de l’énoncé de son élève à la première personne, de la même façon que l’examinateur du jeu de l’imitation est conduit à postuler celle de son interlocuteur C. L’hypothèse des “ machines qui apprennent ” repose, en d’autres termes, sur le fait que l’éducateur de la machine se comporte *comme si* le “ je ” de celle-ci avait le même statut que le sien propre, comme s’il reconnaissait dans la machine un *semblable*, c’est-à-dire, non pas un autre, mais un *autrui*.

De sorte que le problème désormais posé est celui de savoir comment la “ machine élève ” qui réussit le test de Turing, peut satisfaire aux conditions de la reconnaissance d’un *autrui* par un homme. Or, c’est, par là, à la problématique kantienne de la subjectivité que le “ comme si ” mis en jeu dans le test de Turing nous renvoie.

Chapitre III : Le jeu de l'imitation et la problématique kantienne : la reconnaissance d'autrui

On ne s'étonnera pas que ce soit à propos de l'idée de liberté que la question de ce qui distingue l'homme de l'automate soit abordée, en tant que telle, par Kant. Il arrive, en effet, note celui-ci, que l'on parle de la " liberté " de l'automate, qui semble posséder en lui-même le principe de son propre mouvement. Il s'agit là, toutefois, souligne Kant, d'un usage abusif du terme liberté, qui méconnaît, en particulier, le véritable problème théorique posé par l'idée de liberté, à savoir celui du *commencement* de l'action libre. C'est l'action morale, " l'action faite par devoir ", qui témoigne de la liberté humaine, c'est-à-dire du pouvoir qu'a l'homme de " commencer absolument un état ". L'action morale manifeste, au sein même du monde sensible, déterminé par la causalité naturelle, une causalité d'une autre nature, et qui ne peut relever que d'un ordre intelligible. Sous cet angle, la question générale " Qu'en est-il de ce qui distingue l'homme de l'automate ? ", sous la forme où elle est abordée par Turing – " Peut-on distinguer l'homme de l'automate ? " - renvoie aux conditions de possibilité de la reconnaissance d'un autrui, laquelle peut être abordée, dans la pensée de Kant, à partir du fait qu'elle garantit que la loi morale ait bien un contenu. La " moralité ", que définit la " bonne volonté " - la volonté en tant qu'elle n'est bonne ni par ses résultats ni par ses aptitudes, mais par son seul vouloir intérieur – exige que l'autre soit appréhendé comme une " fin en soi "[338], c'est-à-dire que soit affirmée, en lui, la *personne*. Il s'agit là d'une nécessité *pratique*, au sens donné à ce terme par Kant : une exigence inscrite dans l'agir humain, qui renvoie à la volonté

[338] " ... l'homme, et en général tout être raisonnable, existe comme fin en soi, et non pas seulement comme moyen dont telle ou telle volonté puisse user à son gré ; dans toutes ses actions, aussi bien dans celles qui le concernent lui-même que dans celles qui concernent d'autres êtres raisonnables, il doit toujours être considéré en même temps comme fin ". Emmanuel Kant, *Fondements de la métaphysique des mœurs*, trad. de Victor Delbos revue par Ferdinand Alquié, *Oeuvres philosophiques,* II, Paris, Gallimard (Bibliothèque de la Pléiade), 1985, p. 292.

en tant que celle-ci constitue elle-même une cause naturelle – l'agir est inscrit dans le monde sensible - mais aussi en tant qu'elle est la " volonté qui veut le devoir pour lui-même "[339], et dont la maxime " peut être pensée comme principe d'une législation universelle ", valant pour tout être humain, parce qu'elle vaut pour tout être raisonnable.

Or, la personne ne peut être connue ; elle ne peut être que pensée, sous la législation de la raison pratique. La reconnaissance, dans l'autre, par l'affirmation de la personne, d'un semblable, c'est-à-dire d'un autrui, se fait donc sous la forme du " comme si " : la loi morale exige que je me comporte à l'égard de l'autre en qui je reconnais un semblable *comme s'il était* une personne. Les éléments dégagés par l'analyse cartésienne du penser – le " je pense ", le problème du jugement et de l'ordre constitutif de celui-ci – sont ainsi coordonnés, dans la démarche kantienne, en fonction de la notion de *pratique*, dans le cadre d'une théorie de la subjectivité qui fait appel à l'idée d'*autonomie*, puisqu'en tant que personne, c'est-à-dire en tant que fin en soi, l'homme est, non seulement acteur, mais législateur de sa propre action.

S'il est vrai que les hypothèses formulées par Turing renvoient au problème de la reconnaissance d'un autrui et reposent sur la possibilité, pour l'interlocuteur de la machine – qu'il soit l'examinateur du jeu de l'imitation ou le " maître " des " machines qui apprennent " – de faire " comme si " celle-ci était un homme, alors, la réflexion kantienne fournit un modèle pour penser la problématique mise en jeu par Turing dans *Computing Machinery…* Sous cet angle, la démarche de Turing, compte tenu des règles du jeu de l'imitation, pose une double question : puis-je penser la personne, dans l'autre qui me parle et à qui je parle, à partir de la seule parole ? Ai-je, autrement dit, la possibilité de faire " comme si " l'autre était, *en tant qu'il me parle et que je lui parle*, une personne ? Et, enfin, peut-il y avoir une parole de la machine, en ce même sens, c'est-à-dire une parole rendant possible, à son égard, le " comme si " ? Or, si, dans le cadre de la problématique kantienne, la réponse à la première question est sans aucun doute positive, il n'en est pas de même de la réponse à la seconde question. Si, en tant que telle, la parole que j'échange avec l'autre qui me parle et à qui je parle, permet de penser la personne dans cet autre, il est en revanche exclu qu'une machine puisse être cet autre. Chez Kant, en effet, pas davantage que chez Descartes, la distinction de l'homme et de l'automate ne constitue par elle-même un problème. La possibilité de distinguer l'homme de l'automate n'est pas à démontrer : l'idée d'un " penser " de l'automate ne peut pas être

[339] Rudolf Eisler, *Kant-Lexikon*, Paris, Gallimard, 1994, p. 717.

pensée ; cela même définit l'automate. En ce sens, les hypothèses formulées par Turing – une machine peut l'emporter au jeu de l'imitation, et cette machine peut apprendre – si elles s'inscrivent dans une problématique du " comme si ", se démarquent de la démarche au sein de laquelle celui-ci est mis en œuvre chez Kant.

C'est ainsi la notion kantienne de *pratique* que bouleverse la démarche de Turing. Cette notion constitue, pourtant, une solution forte du problème spécifique posé par l'idée de pratique, à savoir : comment est-il possible de *penser* le pratique, s'il est vrai que celui-ci se définit comme *l'opposé du théorique* ? Chez Kant, le pratique, en tant qu'il est le lieu de la loi morale, est précisément ce qui ne peut être que pensé, et ce qui *doit* être pensé. En ce sens, la pensée du pratique passe par la postulation d'un ordre intelligible transcendant le sensible. Or, tout se passe comme si la validité des hypothèses de Turing mettait en œuvre l'idée qu'il est possible de penser le pratique à propos de la machine, comme si la victoire de la machine au jeu de l'imitation impliquait la possibilité de postuler l'intelligible à son égard. Par là même, la définition turingienne de la machine perturbe l'équilibre de la réflexion kantienne, puisque, pour celle-ci, l'idée d'intelligible ne saurait avoir de sens à l'égard de la machine, qui ne relève que du sensible. D'où la question à laquelle peut être mesurée, selon nous, la portée philosophique de *Computing Machinery…* : qu'en est-il de l'idée philosophique de pratique, s'il est vrai que les hypothèses de Turing sont valides et que leur validité met en jeu cette idée ?

Section I : La reconnaissance d'autrui comme personne à partir de la parole

Puis-je penser la personne, dans l'autre qui me parle et à qui je parle, à partir de la seule parole ?

Kant définit la personne, dans la *Critique de la raison pure*, comme " la conscience de l'identité numérique de soi-même en différents temps "[340]. En d'autres termes, la personne renvoie, chez lui, au " je " qui accompagne toute représentation, c'est-à-dire à " l'aperception pure ", principe d'unité des catégories de l'entendement.

L'aperception pure

On sait que, pour Kant, connaître consiste à unifier un divers sensible sous les catégories de l'entendement. L'intuition fournit un divers, c'est-à-dire un ensemble de perceptions dont l'aspect formel est déterminé par l'espace et le temps, conditions *a priori* de la sensibilité ; or, la limitation même de la sensibilité implique que les catégories de l'entendement renvoient à un en soi qu'elles ne peuvent connaître. Les catégories de l'entendement ne peuvent unifier le divers de l'intuition sensible sans que soit pensé un en soi ; l'affirmation d'un noumène, comme " concept problématique ", est une condition de possibilité - une condition transcendantale - de l'unité des concepts. Cette condition consiste dans la pensée d'une totalité comme telle ; l'unification du divers sensible par les catégories de l'entendement met en jeu l'idée de " l'intégralité, c'est-à-dire l'unité collective de toute l'expérience possible "[341]. Une telle idée correspond au principe d'unité par lequel les

[340] Emmanuel Kant, *Critique de la raison pure*, *Œuvres philosophiques*, trad. de J. Barni, revue par A.J.-L. Delamarre et F. Marty, *op. cit.* I, p. 1438.
[341] Emmanuel Kant, *Prolégomènes à toute métaphysique future*, *Oeuvres philosophiques*, trad. de J. Rivelaygue, *op. cit.*, II, p. 105-107.

catégories de l'entendement unifient le divers sensible. Ce principe ne peut, naturellement, être trouvé dans le sensible - le " divers " - puisque celui-ci est le lieu du changement, mais il ne peut être trouvé non plus dans les catégories de l'entendement elles-mêmes, dans la mesure où celles-ci s'appliquent toujours au sensible, hors duquel elles n'ont rien à unifier. La catégorie de causalité, par exemple, ne peut être définie indépendamment des phénomènes qu'elle détermine, c'est-à-dire indépendamment du temps ; la cause et l'effet, dans la catégorie de causalité, sont liés par un rapport de temps, comme un avant et un après. L'idée d'intégralité est donc une " idée pure de la raison ", qui fonde les catégories de l'entendement dans leur pouvoir de détermination ; elle unifie les catégories de l'entendement, telles la substance, l'existence, ou la causalité, en renvoyant au principe d'unité des catégories entre elles qui est nécessaire à chacune de celles-ci pour qu'elle exerce sa propre fonction unificatrice.

En vérité, le principe de l'unité des catégories de l'entendement ne peut être trouvé ailleurs que dans " l'identité de lui-même [NB : l'esprit] dans la diversité de ses représentations "[342], c'est-à-dire dans la conscience que je prends de ce que mes représentations sont miennes ; c'est par cela qu'elles ont toutes en commun d'être les *miennes* que mes représentations renvoient à quelque chose qui ne change pas[343]. L'unité de la pensée se manifeste ainsi à travers " l'aperception pure ou transcendantale " : le " moi " ou le " je " - le " ich ". Le " je " est d'abord acte : l'acte par lequel je forme des représentations. Mais il est également conscience, et c'est par là qu'il est principe d'unité ; le " je " est la synthèse de la conscience et de la représentation :

> l'identité totale de l'aperception d'un divers donné dans l'intuition contient une synthèse des représentations, et n'est possible que par la conscience de cette synthèse. En effet, la conscience empirique qui accompagne différentes représentations est en elle-même dispersée et sans relation avec l'identité du sujet. Cette relation ne s'opère donc pas encore par le fait que j'accompagne chaque représentation de conscience,

[342] Emmanuel Kant, *Critique de la raison pure*, *op. cit.* p. 1413.
[343] " Je suis donc conscient du moi identique, par rapport au divers des représentations qui me sont données dans une intuition, puisque je les nomme toutes *mes* représentations, qui n'en forment *qu'une*. Or cela revient à dire que j'ai conscience d'une synthèse nécessaire *a priori* de ces représentations, qui est l'unité synthétique originaire de l'aperception, à laquelle sont soumises toutes les représentations qui me sont données, mais à laquelle elles doivent être aussi ramenées par une synthèse ", *ibid.* p. 855.

> mais du fait que *j'ajoute* l'une à l'autre et que je suis conscient de leur synthèse. C'est donc seulement du fait que je puis lier un divers de représentations données *dans une conscience* qu'il m'est possible de me représenter *l'identité de la conscience dans ces représentations* mêmes...[344]

Le " je " est condition transcendantale de toute représentation ; il accompagne chacune de celles-ci[345]. Il n'est tel, toutefois, que parce qu'il est conscience de lui-même comme acte d'unification : c'est en unifiant le donné de l'intuition que l'esprit prend conscience de lui-même comme pouvoir unificateur, et c'est en étant conscience d'elle-même que l'unité transcendantale est unité[346].

Le " je " exprime, par là, la dimension " nouménale " de l'homme. Dans l'acte d'aperception, l'homme, s'appréhendant comme principe d'unité où est mise en jeu la conscience de soi, se saisit comme *sujet*. Or, l'idée de sujet renvoie, ici, à travers le principe d'autonomie, au pouvoir qu'à l'homme, en tant qu'être raisonnable, d'insérer, par l'action morale, au sein de la nécessité naturelle à laquelle son action est soumise, une causalité d'une autre nature que la causalité naturelle : une causalité intelligible[347]. En d'autres termes, l'idée de sujet renvoie à l'homme en tant qu'il est une fin en soi, c'est-à-dire en tant qu'il est une *personne*. La conscience immédiate qu'il a d'être principe d'unité n'est pas séparable, en l'homme, de la *personnalité*, c'est-à-dire du fait qu'il soit une personne. Aussi bien le " je " est-il proprement ce qui, dans l'ordre du vivant, distingue l'homme :

> « Que l'homme puisse disposer du Je dans sa représentation : voilà qui l'élève à l'infini au-dessus de tous les autres êtres vivants sur la terre. Il est par là une personne, et en vertu de l'unité de la conscience maintenue à travers tous les changements qui peuvent lui advenir, une seule et même personne, c'est-à-dire un être totalement distinct, par le rang et

[344] *Ibid.*, p. 854.
[345] " Le : je pense doit pouvoir accompagner toutes mes représentations ; car autrement quelque chose serait représenté en moi, qui ne pourrait pas du tout être pensé, ce qui revient à dire ou que la représentation serait impossible, ou que, du moins, elle ne serait rien pour moi [...] Tout le divers de l'intuition a donc une relation nécessaire au : *je pense*, dans le même sujet où ce divers se rencontre ", *ibid.*, p. 853.
[346] " L'esprit, en effet, ne pourrait pas penser, et cela *a priori*, l'identité de lui-même dans la diversité de ses représentations, s'il n'avait devant les yeux l'identité de son acte, qui soumet à une unité transcendantale toute synthèse de l'appréhension (qui est empirique), et en rend d'abord l'enchaînement *a priori* possible suivant des règles ", *ibid.*, p. 1413.
[347] L'expression est métaphorique, puisqu'aussi bien elle renvoie à quelque chose que l'on ne peut connaître.

> par la dignité, des choses, au genre desquelles se trouvent les animaux dépourvus de raison avec qui on peut se comporter à sa guise.. ».[348]

Or, le je s'énonce, et, dès lors, la *parole* n'est-elle pas signe du nouménal en l'homme ? N'est-ce pas, notamment, à travers la parole que peut s'effectuer la reconnaissance d'un autrui, la reconnaissance, dans l'autre à qui je parle et qui me parle, d'une personne ? Alors qu'avant de parler, en effet, explique Kant, l'enfant " ne [fait] que se sentir lui-même ", à partir du moment où il parle, " il se pense lui-même "[349]. Il en est ainsi avant même qu'il puisse prononcer le mot " je " en tant que tel : l'enfant qui ne sait pas dire " je ", parle pourtant déjà de lui-même à une autre personne. Il y a un " moi " de l'enfant, car celui-ci, dès lors qu'il parle, est capable de se prendre pour objet, de s'intuitionner, c'est-à-dire qu'il est capable d'un acte en lui-même inséparable de l'aperception pure, expression de la conscience de soi. En ce sens, dit Kant, parler, c'est toujours " parler avec soi-même[350] ".

Le problème de la reconnaissance de la personne

Cependant, le moi manifesté par l'aperception pure ne peut être *connu*, et dès lors, comment la conscience immédiate que chaque personne a d'elle-même en tant que personne, à travers l'aperception pure, lui permettrait-elle de reconnaître dans l'autre un semblable, c'est-à-dire une personne ? Sans doute le moi peut-il être intuitionné : il y a un moi empirique, objet d'une intuition sensible, objet du " sens interne " - c'est le moi de la psychologie. Sous cet angle, il s'agit d'un phénomène au même titre que les objets connus du monde extérieur : le moi s'intuitionne comme effet dans le monde sensible de la faculté d'unifier le divers. Mais l'aperception pure ne peut être le contenu d'une telle intuition[351]. En tant que tel, le moi de l'aperception pure est une pure forme, celle de l'unité transcendantale qui conditionne toute connaissance ; le " je " est le " véhicule de tous les concepts ".

---

[348] Emmanuel Kant, *Anthropologie du point de vue pragmatique*, *Oeuvres philosophiques*, trad. de P. Jalabert, *op.cit.*, III, p. 945.
[349] *Ibid.*
[350] *Ibid.*, p. 1010.
[351] " ... dans la synthèse transcendantale du divers des représentations en général, par conséquent dans l'unité synthétique originaire de l'aperception, j'ai conscience de moi-même, non comme je m'apparais, ni comme je suis en moi-même, mais j'ai seulement conscience que je suis. Cette *représentation* est une *pensée*, non une *intuition*. ". Emmanuel Kant, *Critique de la raison pure*, *op. cit.* p. 871.

> « … par ce Je, par cet Il ou par ce Cela (la chose) qui pense, on ne se représente rien de plus qu'un sujet transcendantal des pensées = X, lequel n'est connu que par les pensées, qui sont ses prédicats : pris isolément nous ne pouvons jamais en avoir le moindre concept. Nous tournons donc, en ce qui le concerne, dans un cercle perpétuel puisque à chaque fois nous sommes obligés de nous servir d'abord de sa représentation pour porter un jugement quelconque à son sujet. »[352]

Pour former un concept, il faut le " je "[353], c'est pourquoi il ne saurait y avoir un concept du " je " ; le moi intuitionné n'est donc pas le même que le " je " de l'aperception pure. Aussi bien Kant montre-t-il, en examinant les " paralogismes de la raison pure ", que nous ne pouvons pas avoir une conscience immédiate du principe de l'unité transcendantale en autrui.

Si, en effet, la personne, en tant qu'elle est le siège de l'aperception pure, c'est-à-dire la manifestation de l'unité transcendantale, se caractérise par " la conscience de l'identité numérique de soi-même en des temps divers "[354], il ressort du " Deuxième Paralogisme de la raison pure ", celui de la " simplicité ", que je ne peux *connaître* cette identité numérique, c'est-à-dire que je ne peux affirmer qu'une substance lui correspond. La substance se définit par la permanence ; elle est ce qui ne change pas. A travers mon expérience unifiée par l'aperception, " ... je porte mon attention sur le permanent dans ce phénomène, permanent auquel, comme sujet, se rapporte tout le reste, comme détermination, et je remarque[...] l'identité de ce sujet dans le temps, où tout ce reste, la détermination, change "[355]. Par là, " l'identité de ma personne se rencontre [...] infailliblement dans ma propre conscience "[356].

Pourtant, je ne peux conclure, à partir de là, à l'idée que la personne que je suis serait une substance, car tout ce que je peux associer à l'aperception, c'est, précisément, du temps, c'est-à-dire chacun des moments dont j'ai conscience. Rien n'interdit de supposer que le sujet change complètement, d'un état dont il a conscience par l'aperception, à l'autre[357] ; la seule

[352] *Ibid.*, p. 1050.
[353] Le je est une " perception interne [qui] n'est rien de plus que la simple aperception : je pense, qui rend possibles tous les concepts transcendantaux mêmes, où l'on dit : je pense la substance, la cause, etc. ", *ibid.*, p. 1048.
[354] *Ibid.*, p. 1438.
[355] *Ibid*.
[356] *Ibid.*, p. 1439.
[357] " L'identité de la conscience de moi-même en différents temps n'est donc qu'une condition formelle de mes pensées et de leur enchaînement, mais elle ne prouve pas du tout l'identité numérique de mon sujet, dans lequel, malgré l'identité logique du Je, peut cependant se produire un changement tel qu'il ne permette pas d'en maintenir l'identité, tout en permettant de lui accorder toujours encore le Je univoque qui puisse, dans

nécessité que contienne l'idée d'aperception est celle de la liaison de la conscience de chaque état avec la conscience de l'état suivant, non celle que cette liaison soit attribuée à un unique sujet toujours identique. Dès lors que tout ce que je peux associer à l'aperception, c'est du temps, " nous ne saurions décider si ce Je (une simple pensée) ne s'écoule pas tout aussi bien que les autres pensées qui se trouvent enchaînées, grâce à lui, les unes aux autres "[358].

Or, si je ne peux me connaître moi-même comme substance, un autre ne le peut pas davantage. L'argument de Kant met en jeu, ici, le " Troisième paralogisme ", celui de la " personnalité ". Lorsque je connais quelque chose, c'est-à-dire lorsque j'unifie dans ma perception un phénomène, le temps propre de ce phénomène est *en moi*. Lorsqu'un autre me considère, il me perçoit comme un phénomène, c'est-à-dire qu'il me perçoit *dans le temps*, en tant que ce temps est son propre temps. De sorte que, même s'il admet qu'un " je " est bien présent dans toutes mes représentations, il ne peut pas conclure que ce " je " soit permanent : la seule permanence qu'il rencontre jamais est celle qui est liée à sa propre aperception. La conscience de l'unité du sujet ne peut pas être mise à la troisième personne[359].

Je ne puis donc considérer le moi que j'intuitionne en autrui, à travers la parole, comme celui de l'aperception pure ; la parole que j'échange avec l'autre qui me parle et à qui je parle ne saurait me donner la *connaissance* de cet autre comme personne. En vérité, la parole, en tant que je l'intuitionne, ne me fournit qu'une *analogie d'expérience*, laquelle ne donne jamais le droit de conclure, en terme de connaissance, à un principe.

Les analogies de l'expérience.

On sait que, chez Kant, les " analogies de l'expérience " sont des principes *a priori* de l'entendement ; toutefois, il s'agit de principes simplement régulateurs et non constitutifs[360].

---

chaque nouvel état, même dans la transformation complète du sujet, conserver toujours les pensées du sujet précédent et, de la sorte, les transmettre aussi au suivant ", *ibid*. p. 1440.

[358] *Ibid*.

[359] Voir à ce sujet, par exemple, l'article de Carol Van Kirk (Carol Van Kirk, " Kant and the Problem of Other Minds ", *Kant Studien*, XXX, 1986, 77, 1) : " We can know apperception is not an objective concept because it is impossible to construct a third-person view of apperception. We cannot figure out what apperception would be like for another because we cannot perceive it ".

[360] " Ces principes [NB : les analogies de l'expérience] ont ceci de particulier qu'ils n'examinent pas les phénomènes et la synthèse de leur intuition empirique, mais seulement *l'existence* et leur *rapport* entre eux, relativement à cette existence ". Emmanuel Kant, *Critique de la raison pure*, *op. cit.*, p. 916.

Les analogies de l'expérience n'effectuent pas la synthèse de l'intuition empirique d'un objet ; elles permettent seulement de reconnaître la règle de la synthèse d'un objet dans des exemples empiriques donnés. En mathématiques, les analogies ne portent pas sur l'existence des phénomènes, mais sur leur possibilité ; les objets y sont considérés sous l'angle de la quantité, c'est-à-dire du point de vue, non de la matière de la perception, mais des formes *a priori* de l'espace et du temps. Les analogies, qui déterminent un rapport, une relation, y fournissent donc une connaissance complète de leur objet. Il n'en va pas de même, toutefois, lorsque les analogies sont appliquées à des objets du point de vue de leur existence réelle. Dans ce cas, le rapport n'est plus uniquement quantitatif, mais qualitatif ; il met en jeu les phénomènes, non seulement selon leur forme, mais aussi selon leur contenu, lequel ne peut jamais être connu *a priori*. C'est pourquoi, dans ce cas, qui est proprement celui de la connaissance philosophique, " ... à partir de trois membres, je ne puis connaître et donner *a priori* que le *rapport* à un quatrième, mais non *ce quatrième membre* lui-même... "[361]. L'analogie ne me donne alors qu'une règle pour chercher le quatrième membre, et une " marque " (" Merkmal ") pour le découvrir.

Ainsi, par exemple, pouvons-nous comprendre quelque chose du comportement animal par analogie avec le comportement humain : il y a, sous certains rapports, identité des séries causales produites par le comportement animal et l'action de l'homme. Comparons, dit Kant,... " les constructions du castor et celles de l'homme "[362] ; le castor, comme l'homme, sait bâtir des abris compliqués. De cette identité nous pouvons déduire que le castor, comme l'homme, agit d'après des représentations ; ses constructions, tout comme celles d'un architecte humain, supposent une représentation de l'objet à produire et des moyens de le produire. Cependant, l'identité des séries causales, entre les constructions des castors et celles des hommes, ne nous fait pas connaître le *principe* du comportement du castor : " ce n'est pas parce que l'homme utilise la raison dans ses constructions que je puis en conclure que le castor doit aussi avoir une raison, et appeler cela un raisonnement par analogie "[363]. Si, de la constatation que le castor, comme l'être humain, est capable de construire un abri, je concluais qu'il est doué de raison, je dépasserais ce qui découle de l'application du principe d'analogie ;

[361] *Ibid.*, p. 917.
[362] Emmanuel Kant, *Critique de la faculté de juger*, *Œuvres philosophiques*, trad. de J.-R. Ladmiral, M. B. de Launay et J.-M. Vaysse, *op. cit.*, II, p. 1273.
[363] *Ibid*.

je prétendrais connaître, à partir de deux relations entre trois membres - par exemple, la relation à une construction, d'une part, de l'homme, être doué de raison, d'autre part, du castor - non pas seulement le rapport à un quatrième membre - le principe du comportement du castor - mais ce quatrième membre lui-même, que j'assimilerais à la " raison " du castor.

Dès lors, la question de la reconnaissance d'autrui à travers la parole est celle de savoir s'il est possible, dans le contexte de l'analogie dessinée par la parole, sinon de connaître, dans l'autre qui me parle et à qui je parle, la personne, du moins de la *penser*. Est-il possible, en d'autres termes, que je me comporte à l'égard de l'autre, considéré en tant qu'il me parle et en tant que je lui parle, *comme s'il était* une personne ?

La parole et le " jugement réfléchissant "

Au-delà de ce qui peut être saisi d'elle à travers les analogies de l'expérience, la parole, chez Kant, renvoie à une *faculté*- c'est-à-dire à un " pouvoir faire " qui est l'expression de la liberté transcendantale - et, comme cela ressortait déjà de l'analyse cartésienne, met en jeu une théorie du *jugement*. On sait que, pour Kant, la faculté de juger consiste à " penser le particulier comme contenu sous l'universel "[364]. Il manque à la législation de la raison une intuition pour qu'elle permette la connaissance, laquelle relève - en tant que son objet est le phénomène et non la chose en soi - de la législation de l'entendement, qui autorise celui-ci à rassembler le divers sous un principe unificateur. Les catégories de l'entendement ont toutefois besoin, pour unifier le divers sensible, du principe d'unité qui leur vient de la raison ; elles renferment, en ce sens, le principe de la subsomption du particulier sous l'universel. Lorsque nous possédons le concept de l'objet, sous la forme d'une règle ou d'une loi, le particulier est subsumé sous l'universel par un jugement " déterminant ", dit Kant. L'ingénieur qui calcule la portée d'un pont effectue, par exemple, un jugement déterminant. En tant qu'elle est déterminante, la faculté de juger n'a pas besoin d'avoir une loi qui lui soit propre pour subsumer le particulier sous l'universel : le concept *a priori* que lui fournit l'entendement comporte en lui-même la relation du particulier à l'universel[365].

[364] *Ibid.*, p. 933.

[365] " Nous trouvons... dans les fondements de la possibilité d'une expérience, certes tout d'abord, quelque chose de nécessaire, à savoir les lois universelles, sans lesquelles la nature en général (comme objet des sens) ne peut pas être pensée ; et ces lois reposent sur des catégories, appliquées aux conditions formelles de toute intuition possible pour nous ... Or, sous ces lois, la faculté de juger est déterminante, car elle n'a rien à faire d'autre que subsumer sous des lois données. Par exemple, l'entendement dit : toute

Toutefois, la nature présente d'infinies variations des concepts transcendantaux universels. En d'autres termes, les catégories expriment les différentes manières dont on peut constituer un objet, mais non la diversité des objets singuliers[366]. Elles ne fournissent que la possibilité d'un objet de l'expérience en général, de telle sorte que, devant chaque cas particulier, la faculté de juger déterminante, telle qu'elle se présente en l'homme, dont l'entendement est fini, est frappée d'impuissance. Lorsque nous ne disposons pas du concept, lorsque nous ne connaissons pas encore la règle ou la loi, nous devons la trouver à partir du particulier. Puisque nous ne possédons pas encore le concept, nous devons présupposer qu'il existe ; nous devons présupposer que le particulier peut être déterminé par un jugement. Étant donné que le principe de l'unité qui fonde la pensée met en jeu les catégories de l'entendement, nous devons supposer, pour comprendre l'expérience comme un tout, un entendement pour lequel tout serait déjà connu, tout serait déjà ramené à l'unité. Le jugement, dans le cas où la règle de subsomption ne nous est pas connue, exige, comme sa condition de possibilité, la présupposition d'un " entendement archétype ", pour lequel tout se passe comme si le jugement était toujours déterminant[367]. Bref, nous devons penser, pour la faculté de juger, un principe transcendantal. On sait que ce principe est, pour Kant, l'idée d'une finalité de la nature. Par la découverte d'un lien entre plusieurs lois empiriques, la faculté de juger satisfait ce qu'elle présuppose ; elle atteint la fin que, par la présupposition, elle visait ; présupposer un tout de l'expérience pour pouvoir trouver la règle à partir du particulier, c'est poser ce tout comme une fin qui doit être atteinte par le jugement[368], lequel est alors, dit Kant,

modification a sa cause (loi universelle de la nature) ; la faculté de juger transcendantale n'a rien à faire de plus que d'indiquer la condition de la subsumption sous le concept de l'entendement *a priori* présenté [...] Pour la nature en général (comme objet de l'expérience possible), cette loi est reconnue comme absolument nécessaire. ", *ibid.*, p. 938.

366 " ... les objets de la connaissance empirique... sont encore déterminés, ou, autant que l'on puisse en juger *a priori*, déterminables de maintes façons, de sorte que des natures spécifiquement différentes, en dehors de ce qu'elles ont en commun, en tant qu'appartenant à la nature en général, peuvent être des causes de façons infiniment variées... ", *ibid*.

367 " ... ce principe ne peut différer du suivant : puisque les lois universelles de la nature ont leur fondement dans notre entendement, qui les prescrit à la nature (certes selon son concept universel en tant que nature), les lois empiriques particulières, eu égard à ce qui est laissé en elles indéterminé par les lois universelles, doivent être considérées selon une unité telle qu'un entendement (même si ce n'est pas le nôtre) pouvait la donner pour notre faculté de connaître, afin de rendre possible un système de l'expérience selon des lois particulières de la nature. ", *ibid.*, p. 934.

368 " Or, parce que le concept d'un objet, dans la mesure où il comprend en même temps le fondement de l'effectivité de cet objet, s'appelle la *fin* et que l'accord d'une chose avec cette constitution des choses, seulement possibles selon des fins, s'appelle la finalité de

“ réfléchissant ”. En présence du particulier qui n’est pas encore connu, un acte spécifique - le jugement réfléchissant - est nécessaire pour subsumer ce particulier sous une règle.

Le jugement réfléchissant, dans son caractère discursif – son point de départ est toujours le particulier - s’exprime notamment, sous la forme du “ jugement naturel ”, dans les domaines du *technique* et du *pragmatique*, lesquels rendent compte de l’agir, non pas en tant qu’il relève du *pratique*, au vrai sens du terme – c’est-à-dire en tant qu’il renvoie à la “ bonne volonté ” – mais en tant qu’il est déterminé par les “ règles de l’habileté ” - celles de l’art - et les “ règles de la prudence ” - celles qui visent le bonheur. Or, au “ jugement naturel ” peut être appliquée la remarque célèbre de Descartes : il est “ la chose du monde la mieux partagée ” ; si certains homme font preuve d’un meilleur jugement que d’autres, il n’est pas un homme, en effet, qui, parce qu’il est homme, ne soit capable de “ jugement naturel ”. On relèvera, sans doute, que Kant évoque, dans la *Critique de la raison pure*, ces médecins, juges ou hommes d’état qui

> « peuvent avoir dans la tête beaucoup de belles règles pathologique, juridiques ou politiques, à un degré qui peut en faire de solides professeurs en ces matières, et pourtant faillir aisément dans leur application, soit parce qu’ils manquent de jugement naturel (sans manquer pour cela d’entendement) et que, s’ils voient bien *in abstracto* le général, ils sont incapables de discerner si un cas y est contenu *in concreto*, soit parce qu’ils n’ont pas été assez exercés à ce jugement par des exemples et des affaires réelles. »[369]

Un médecin très savant peut être un mauvais médecin : ce n’est que dans le cas du jugement déterminant que la seule connaissance des règles permet un jugement juste. Un médecin très savant qui sait de quelle maladie souffre son patient peut, à partir des règles qu’il connaît, prévoir l’évolution de l’état du malade, et, dans le strict cadre de ses connaissances, adopter l’attitude qui convient ; cependant, pour établir un diagnostic, il ne lui suffit pas de connaître de nombreuses règles de pathologie, il doit encore être capable, par un jugement réfléchissant, de subsumer sous ces règles le particulier devant lequel il se trouve. S’il manque de “ jugement naturel ”, aussi savant soit-il, il ne sera pas un bon médecin.

---

leur forme, le principe de la faculté de juger, eu égard à la forme des choses de la nature sous des lois empiriques en général, est la *finalité de la nature* dans sa diversité. C’est-à-dire que la nature est représentée par ce concept, comme si un entendement contenait le fondement de l’unité du divers de ses lois empiriques. ”, *ibid.*, p. 935.

[369] Emmanuel Kant, *Critique de la raison pure ”, op. cit.*, p. 882.

Il n'en demeure pas moins que le médecin le plus médiocre, ne manquera jamais *absolument* de " jugement naturel " : serait-il, pour reprendre, là encore, une expression de Descartes, " le plus hébété des hommes ", il n'en sera pas moins un être raisonnable. En d'autres termes, c'est encore le " jugement naturel ", en tant qu'il est une figure de la faculté de juger réfléchissante, qui s'exerce au travers du diagnostic incertain du médecin le moins doué. Les erreurs mêmes que commet, dans l'exercice de sa pratique, tel médecin pourtant très savant, relèvent, non pas tant du jugement déterminant - puisque les concepts dont est formée la science de ce médecin incluent la relation du particulier à l'universel - que du jugement réfléchissant. Du reste, le médecin le plus doué, celui qui manque le moins de jugement naturel, peut lui-même se tromper. Dans l'horizon du jugement réfléchissant, l'erreur est toujours possible : puisque la règle n'est pas donnée, puisque la subsomption du particulier ne s'effectue pas à partir de l'universel, la *vérification* empirique est nécessaire. Certes, le médecin qui, connaissant de nombreuses règles de pathologie, se montre en outre capable de les appliquer pour soigner efficacement son prochain, ne se trompe-t-il pas souvent, mais il s'est probablement déjà trompé, et il peut encore se tromper. Un jugement réfléchissant, dans son caractère discursif, peut toujours être considéré comme la correction d'un autre jugement réfléchissant possible qui, s'il avait été émis, aurait rassemblé trop tôt des lois empiriques sous un principe, et aurait amené une conclusion précipitée. Aussi bien, si l'apprentissage de nombreuses règles de pathologie ne permet pas d'exercer avec succès la médecine - la règle sous laquelle sera subsumé le particulier qui n'est pas encore connu ne peut être apprise, puisque le rapport au particulier qu'elle détermine n'est pas encore connu - la faculté de juger réfléchissante, sous la forme du " jugement naturel ", peut être exercée, et doit l'être ; c'est à quoi tend l'éducation, dont l'une des fins, aux yeux de Kant, est l'habileté. En d'autres termes, un médecin sera, selon son jugement naturel, plus ou moins habile, mais jusque chez le moins doué d'entre eux, le jugement naturel pourra être exercé, et un certain degré d'habileté visée. En vérité, sous l'angle même du " jugement naturel ", le jugement réfléchissant apparaît comme un mode du " je pense ", comme un mode de l'aperception pure, en ce que celle-ci, tout en même temps, manifeste l'autonomie du sujet transcendantal, c'est-à-dire un pouvoir d'agir déterminé par la liberté transcendantale, et est inconnaissable.

Or, la parole est expression du jugement, aussi bien en tant qu'il est réfléchissant qu'en tant qu'il est déterminant[370]. De sorte que c'est d'abord sous la forme du discours mettant en jeu la faculté de juger réfléchissante que la parole renvoie à l'aperception pure, à l'autonomie du sujet transcendantal, c'est-à-dire à la personne. En vérité, l'analogie qu'elle présente, en tant que je l'intuitionne dans l'autre, est d'un genre tout à fait singulier : elle prend la forme d'une relation à l'autre, d'un échange, d'une communication qui s'ouvre sur ce qui dépasse le " technique " et le " pragmatique ", à savoir le *pratique*. Ce n'est pas simplement la forme extérieure de l'habileté ou de la prudence qui se donne à travers la parole, mais la *communauté des fins*. Je reconnais un être raisonnable, c'est-à-dire une fin, à travers la parole sensée de l'autre qui me parle et à qui je parle. L'idée de personne n'est pas contradictoire avec la relation à l'autre que manifeste la parole que j'échange avec celui-ci. La personne, dans l'autre avec qui s'échange la parole, ne peut être connue, mais l'autre peut, à travers le discours comme échange de parole, être *pensé* comme personne. Par là même, enfin, l'autre qui me parle et à qui je parle entre dans le champ de la maxime qui représente la loi morale et qu'exprime " l'impératif pratique " : " Agis de telle sorte que tu traites l'humanité, aussi bien dans ta personne que dans la personne de tout autre, toujours en même temps comme fin, et jamais simplement comme moyen "[371]. Dès lors que du discours émerge la possibilité de penser l'idée de personne, l'autre à qui je parle et qui me parle *doit* être considéré comme une personne ; se comporter à son égard comme s'il était un autre moi-même – un semblable – devient une nécessité pratique, au sens kantien du terme.

Autrui peut donc être reconnu, en tant que personne, à travers la seule parole, c'est-à-dire que dans le cadre d'une situation réduite à un échange de parole, comme celle qu'imagine Turing, un homme, être raisonnable, sujet transcendantal, se comportera " comme si " son interlocuteur était une personne. Considérée sous l'angle de la problématique qui définit, chez Kant, le " comme si " la relation à l'autre donnée sous la seule forme de l'échange de parole

[370] Le terme allemand " Sprache " signifie tout à la fois la langue, la parole, et la *faculté* de parler ; " Toute langue, dit Kant, est désignation des pensées, et, inversement, le mode par excellence de désignation des pensées est celui que procure le langage, suprême moyen de compréhension de soi-même et des autres " (Emmanuel Kant, *Anthropologie d'un point de vue pragmatique*, *Œuvres philosophiques*, *op. cit.*, III, p. 1010).
[371] Emmanuel Kant. *Fondements de la métaphysique des mœurs*, *Œuvres philosophiques*, *op. cit.*, II, p. 295.

permet la reconnaissance d'un autrui. Il reste cependant à se demander si, dans le cadre de la même problématique, c'est-à-dire dans le contexte du modèle kantien du " comme si ", la parole ainsi entendue est possible pour une machine.

Section II : La non-reconnaissance d'autrui à partir d'une " parole " de la machine

En vérité, replacée dans le contexte de la démarche kantienne, la question posée par le " comme si " sur lequel repose la double hypothèse de Turing, ne saurait avoir de sens : dans ce contexte, *l'échange de parole* - au regard duquel la personne, dans l'autre, peut être pensée - n'est pas possible avec une machine.

Nous l'avons signalé, pourtant, Kant note - à la fin de " l'Analytique " de la *Critique de la raison pratique* - que l'idée de " liberté " est parfois évoquée à propos des automates :

> « on appelle quelquefois libre un effet dont la cause déterminante naturelle réside *intérieurement* dans l'être agissant, comme par exemple lorsqu'on emploie le terme de liberté à propos de ce qu'accomplit un corps qui a été lancé, quand son mouvement est libre, parce que ce corps, dans son trajet, n'est poussé par aucune force extérieure, ou comme on appelle libre le mouvement d'une montre, parce qu'elle pousse elle-même ses aiguilles, et que celles-ci, par conséquent, ne doivent pas être mues par une force extérieure ; ainsi, même si les actions sont nécessaires en vertu de leurs principes déterminants qui précèdent dans le temps, nous les appelons libres parce que ces principes sont des représentations intérieures produites par nos propres forces, des désirs excités par ces représentations suivant les circonstances, et que, par conséquent, ces actions sont produites selon notre bon plaisir. »[372]

Le mouvement de l'automate est dit " libre ", bien qu'il soit soumis à la nécessité de la causalité naturelle, car l'automate, une fois mis en branle, n'a plus besoin, pour continuer à se

372 Emmanuel Kant, *Critique de la raison pratique*, *Œuvres philosophiques*, trad. de L. Ferry et H. Wismann, *op. cit.*, II, p. 725.

mouvoir, d'un principe externe. Les moments successifs de son mouvement, hormis le premier, dû à une impulsion extérieure, sont causes suffisantes les uns des autres. Nous usons alors, dit Kant, d'un " concept comparatif de liberté "[373] : le mouvement de l'automate est comparé à celui de l'homme, lequel est déterminé, non seulement par les forces externes de la nature - il est, lui aussi, soumis à la nécessité naturelle - mais également par des " représentations intérieures produites par ses propres forces ". De sorte que

> « ce n'est pas seulement ce qui attire, c'est-à-dire ce qui affecte immédiatement les sens, qui détermine l'arbitre humain ; nous avons au contraire un pouvoir de surmonter au moyen de représentations de ce qui est utile ou nuisible, même d'une manière plus éloignée, les impressions produites sur notre faculté sensible de désirer. »[374]

L'être humain présente cette particularité de pouvoir, à travers sa faculté de représentation, peser sur la nécessité naturelle à laquelle il est soumis. En formant, à propos de l'automate, un " concept comparatif de liberté ", nous assimilons à cette dimension de l'action humaine le fait que l'automate ait en lui-même le " principe naturel de détermination " de son mouvement.

Or, s'agissant de l'automate, l'idée de liberté se prête-t-elle à un tel usage " comparatif " ? Au vrai, il ne suffit pas de constater que l'automate a en lui-même le " principe naturel de détermination " de son mouvement pour pouvoir le dire libre, au plein sens du terme ; l'idée de liberté pose un tout autre problème : celui du *commencement* de l'action libre. La liberté implique " ...un pouvoir de commencer tout à fait spontanément une série dans le temps... "[375]. Elle est, en ce sens, *a priori*, et doit être définie comme transcendantale, c'est-à-dire comme condition de possibilité d'un certain type d'action :

> « Ce n'est donc proprement qu'une difficulté transcendantale qui, dans la question de la liberté du vouloir, a jeté de tout temps la raison spéculative dans un si grand embarras. Il s'agit seulement de savoir si l'on doit admettre un pouvoir de commencer de soi-même une série de choses ou d'états successifs. »[376]

---

373 *Ibid*.
374 Emmanuel Kant, *Critique de la raison pure*, *op. cit.*, p. 1363.
375 *Ibid.*, p. 1106.
376 *Ibid.*

Le problème de la liberté

Le problème théorique posé par la notion de liberté prend, chez Kant, la forme de la " Troisième antinomie de la raison pure ". En tant que puissance de commencer absolument un état, la liberté annule la causalité naturelle. S'il y a liberté, en effet, la cause ne dépend pas, dans son pouvoir de détermination, de l'état qui la précède ; dans la série des états successifs ordonnés par la loi de causalité, chaque état a, alors, sa raison en lui-même et non dans l'état précédent[377], de sorte que " nature et liberté transcendantale diffèrent [...] entre elles comme la conformité à des lois et l'absence de lois "[378].

Sans causalité naturelle, cependant, il n'y aurait pas de connaissance possible. En tant que telle, c'est-à-dire en tant qu'elle relève de l'intuition, la perception ne donne qu'un ordre spatio-temporel, un " être à côté de ", ou un " être avant ou après ", et, seule, par exemple, la catégorie de causalité peut transformer la relation purement chronologique de l'avant à l'après en une relation de détermination. Le mouvement de l'automate - ainsi celui de la montre - ne pourrait être connu sans la causalité naturelle, ce qui signifie que la montre ne pourrait être construite.

Du reste, il n'y a pas de connaissance possible de la liberté. L'espace et le temps sont formes *a priori* de notre réceptivité en cela qu'ils ne peuvent être dérivés ni de la sensation, ni des catégories elles-mêmes : il n'y a pas pour Kant d'intuition intellectuelle ; les catégories de l'entendement ne saisissent rien par elles-mêmes. Pour qu'il y ait connaissance, il est nécessaire qu'elles disposent d'un divers fourni par l'intuition sensible. Or, il n'existe pas d'intuition sensible de la liberté, puisque toute intuition est donnée dans le temps, et que la liberté se définit par le pouvoir de poser un commencement qui échappe à l'enchaînement temporel.

Toutefois, comme le montre la thèse de l'antinomie, son opposition même à la causalité naturelle fait que la liberté est postulée par l'affirmation de celle-ci. Supposons en effet qu'il n'y ait pas d'autre causalité que la causalité naturelle. Dans ce cas, " tout ce qui arrive présuppose[rait] un état antérieur auquel il succède[rait] inévitablement suivant une

---

[377] " ... tout commencement d'action présuppose un état de la cause où cette cause n'agit pas encore, et un premier commencement dynamique de l'action présuppose un état qui n'a aucun enchaînement de causalité avec l'état précédent de la même cause... ", *ibid.*, p. 1103.
[378] *Ibid.*, p. 1105.

règle "[379]. Cet état antérieur devrait lui-même succéder à un autre advenu précédemment, sinon, il aurait toujours été, et sa conséquence aurait, elle aussi, toujours été. Dès lors, s'il n'y avait d'autre causalité que la causalité naturelle, il n'y aurait pas d'état originaire, pouvant engendrer un deuxième état, puis un troisième, etc.[380], et la loi de causalité se contredirait elle-même, puisqu'aucune cause ne serait suffisamment déterminée pour rendre compte des autres[381]. La nécessité naturelle exige que l'on admette une cause première, une cause qui ne doive sa capacité de détermination à aucun état antérieur, une cause différente des autres en ce qu'elle se détermine elle-même, bref, une cause qui soit " spontanéité absolue "[382]. La liberté ne peut être connue, mais l'idée de liberté doit être pensée. C'est pourquoi il y a bien un *problème* de la liberté, qui consiste à se demander comment il peut y avoir des commencements libres et inscrits pourtant dans la causalité naturelle ; comment peut-on penser ensemble la liberté et la causalité naturelle ?

On sait que la réponse kantienne repose sur l'idée que, si la liberté ne peut être connue, sa possibilité peut l'être. Dans sa dimension particulière, l'agir humain, qui peut orienter la nécessité naturelle en fonction des représentations que l'homme produit intérieurement par ses propres forces, témoigne de cette possibilité. Les actions faites par devoir s'opposent, en effet, en tant que telles, à la nécessité naturelle. Ces actions s'inscrivent, certes, dans la nécessité naturelle, c'est-à-dire qu'elles constituent des séries causales au même titre que tout ce qui arrive dans la nature, mais leur commencement échappe à la nécessité naturelle en ce sens qu'elles insèrent dans celle-ci un enchaînement temporel qui n'aurait pas eu lieu sans elles. Elles fournissent ainsi un divers sensible qui peut être rapporté à ce pouvoir qu'à l'homme de commencer absolument un état, divers sensible susceptible d'être unifié par les catégories de l'entendement, c'est-à-dire connu. L'action faite par devoir est, en somme, l'expression de la possibilité de la liberté transcendantale. Vis à vis d'elle, l'idée de liberté, en tant que liberté transcendantale, ne conduit à aucune contradiction ; elle peut donc être pensée. La " représentation " de la liberté " ne contient du moins en elle aucune contradiction ", dit

[379] *Ibid.*, p. 1102.
[380] " Si donc tout arrive suivant les seules lois de la nature, alors il n'y a toujours qu'un commencement subalterne, mais jamais un premier commencement... ", *ibid*.
[381] " ... la proposition qui veut que toute causalité ne soit possible que suivant des lois naturelles se contredit elle-même dans son universalité illimitée... ", *ibid.*, p. 1104.
[382] *Ibid*.

Kant[383] ; j'ai le droit d'affirmer que l'action faite par devoir est une action libre, c'est-à-dire une action dont le commencement relève d'une causalité autre que la causalité naturelle.

L'antinomie admet ainsi sa " solution critique ". En l'homme, la nécessité naturelle, causalité sensible, et la liberté, causalité intelligible, coexistent. En tant qu'il est lui-même soumis à la causalité naturelle, dans laquelle son action s'inscrit, l'homme est phénomène ; en tant qu'il agit par devoir, en tant qu'il a le pouvoir de commencer absolument un état, il est noumène :

> à présent nous voyons bien que lorsque nous nous pensons comme libres, nous nous transportons dans le monde intelligible comme membres de ce monde, et que nous reconnaissons l'autonomie de la volonté avec sa conséquence, la moralité ; mais quand nous nous concevons comme soumis au devoir, nous nous considérons comme faisant partie du monde sensible et en même temps, pourtant, du monde intelligible.[384]

A travers l'action faite par devoir, l'homme apparaît non seulement comme acteur, mais comme législateur. La loi de son action est, alors, la loi morale, laquelle n'est pas une loi de la nature mais une loi du monde intelligible. Cette loi n'est pas extérieure à l'homme qui agit par devoir, sans quoi le statut de l'action morale serait le même que celui d'une action entièrement déterminée par la causalité naturelle ; dans l'action morale, l'homme est à lui-même sa propre fin, bref, il est une *personne*.

C'est enfin pourquoi l'idée de " liberté " appliquée à l'automate, à partir du fait que celui-ci a en lui-même le " principe naturel de détermination " de son mouvement, n'est qu'un " misérable subterfuge "[385]. Dans le cas de l'automate, le point décisif est, plutôt que l'intériorité des déterminations naturelles de son mouvement, le fait qu'il doive être mis en marche. L'impulsion qui doit lui être donnée pour que son mouvement puisse ensuite se dérouler de lui-même, ne lui appartient pas ; elle lui est, non pas étrangère, mais extérieure. Le mouvement de l'automate relève de la seule causalité naturelle ; chaque moment de ce mouvement est effet d'une cause passée, de sorte qu'à un moment quelconque de son

---

[383] *Ibid.* p. 747.
[384] Emmanuel Kant, *Fondements de la métaphysique des mœurs*, *op. cit.,* p. 918-924.
[385] On ne saurait définir la notion de liberté sous la forme de la " liberté " du " tournebroche, qui [...], une fois monté, exécute de lui-même ses mouvements ". E. Kant, *Critique de la raison pratique*, *op. cit.,* p. 726.

mouvement, la cause de celui-ci n'est plus au pouvoir de l'automate[386]. Le mouvement de l'automate, qui, une fois mis en branle, se déroule sans faire appel à une causalité naturelle extérieure, qui possède en lui-même le principe naturel de sa détermination, bref, le mouvement de l'automate tel que le décrit le " concept comparatif de liberté ", relève de la " *logique générale* ", et non de la " *logique transcendantale* ". Ce principe détermine tout essai de transposition de la question examinée par Turing à travers le jeu de l'imitation, dans les termes de la problématique kantienne de la reconnaissance d'un autrui, c'est-à-dire dans les termes d'un " comme si " fondé sur l'idée kantienne de " pratique ".

Le jeu de l'imitation du point de vue de la démarche kantienne : l'idée d'un " super-automate de Vaucanson "

Il est possible, dans le contexte de la démarche kantienne, de concevoir une analogie entre la machine et l'individu humain, car, en tant qu'elle est construite par l'homme, en tant qu'elle réalise un concept, la machine peut être comprise comme une fin. Est fin, explique Kant, une chose qui " doit se comporter en elle-même réciproquement comme cause et comme effet "[387]. Considérée comme unité d'un divers sensible, comme condition de la connaissance d'un phénomène, bref, comme catégorie de l'entendement, la liaison causale, est " descendante " (" welche immer abwärts geht ") : il n'y a pas réciprocité de la relation ; la chose qui, en tant qu'effet, en suppose une autre comme étant sa cause, ne peut elle-même être cause de cette dernière. La liaison causale est alors " liaison des causes efficientes (nexus effectivus) "[388].

En revanche, lorsqu'elle est pensée comme une série de déterminations prise dans sa totalité, la liaison causale présente une réciprocité. Une chose construite en fonction d'une fin, est la cause de cette fin, et en même temps son effet : une maison mise en location est la cause des revenus qu'elle rapporte ; mais si elle a été construite en vue de ce rapport, c'est ce dernier qui est sa cause. La liaison causale est alors " liaison par les causes finales (nexus finalis) "[389].

---

386 " … tout événement, par conséquent aussi toute action, qui arrive dans un point du temps, est nécessairement sous la condition de ce qui était dans le temps précédent. Or, comme le temps passé n'est plus en mon pouvoir, toute action que j'accomplis d'après des causes déterminantes *qui ne sont pas en mon pouvoir*, doit être nécessaire, c'est à dire que je ne suis jamais libre dans le point du temps où j'agis. " *Ibid*., p. 723.

387 Emmanuel Kant, *Critique de la faculté de juger*, *op. cit.,* p. 1163.

388 *Ibid*.

389 *Ibid*., p. 1164.

D'une manière générale, dans une fin, les parties ne sont possibles, selon leur existence et leur forme, que par leur relation au tout. Toute fin, en effet, " est comprise sous un concept ou une Idée, qui doit *a priori* déterminer tout ce qui doit être contenu dans la chose "[390]. Chaque partie en elle existe pour les autres et pour le tout. Il en est ainsi dans " ce qui est pratique (c'est-à-dire dans l'art) "[391], précise Kant. Or, la machine, qui relève de " l'art " tel qu'il est entendu ici – la technique – est, naturellement, justiciable de ces définitions.

Par là s'explique qu'il soit possible de construire une machine qui, du point de vue phénoménal, ressemble à un être humain. Le célèbre joueur de flûte construit par Vaucanson, à l'époque de l'enfance de Kant, illustre cette possibilité. L'automate de Vaucanson se distinguait par sa complexité et par le savoir-faire qu'il supposait, des nombreux automates musiciens construits jusque là. Aucun de ceux-ci ne jouait vraiment d'un instrument : ils renfermaient une " serinette " ou une " boîte à musique ", mises en marche au moment désiré. Le joueur de flûte de Vaucanson, quant à lui, jouait réellement de la flûte : le son qu'il faisait entendre était émis par la colonne d'air qu'il faisait passer dans une véritable flûte, colonne d'air dont il modifiait la longueur comme l'aurait fait un joueur humain, à l'aide de ses " doigts "[392].

Il y a, cependant, plusieurs sortes de fins. Les êtres vivants – tels les hommes et les animaux – sont, quant à eux, des " fins naturelles ". Pour qu'une chose soit une fin " naturelle ", ses parties doivent être, non seulement les unes pour les autres et pour le tout, mais les unes *par* les autres ; dans une fin naturelle, chaque partie produit les autres et réciproquement[393]. La fin naturelle a son idée en elle-même. Ainsi est-ce parce qu'il est une fin naturelle qu'un organisme peut se reproduire et se réparer lui-même. La force impliquée

[390] *Ibid*.
[391] *Ibid*., p. 1163.
[392] L'effet était d'autant plus saisissant que l'automate jouait d'un instrument - la flûte - particulièrement difficile parce qu'expressif, une grande part de la qualité d'exécution reposant sur le mouvement des lèvres autant que sur celui des doigts. Dans leur livre sur Vaucanson, Doyon et Liaigre citent, à cet égard, Rigollay de Juvigny, lequel déclarait : " Dans les premiers jours que l'automate parut [...], beaucoup de gens ne voulaient pas croire que ce fût la flûte que tenait l'automate qui rendait les sons. On s'imagina qu'ils ne provenaient que d'une serinette ou d'un orgue d'Allemagne, enfermé dans le corps de la figure. Les plus incrédules furent bientôt convaincus que l'automate embouchait réellement la flûte, que le vent au sortir de ses lèvres la faisait résonner et que le mouvement de ses doigts formait les différents sons ". *Jacques Vaucanson, mécanicien de génie, op. cit.*, p. 40.
[393] " Ce n'est qu'alors et pour cette seule raison qu'un tel produit, en tant qu'*être organisé et s'organisant lui-même*, peut être appelé une *fin naturelle* ". Emmanuel Kant, *Critique de la faculté de juger, op. cit.,* p. 1165.

ici – il s'agit de la force au sens donnée par la physique - est non seulement motrice, mais formatrice. Aussi bien l'identité de forme de l'action animale et de l'action humaine, qui légitime le raisonnement par analogie de l'homme à l'animal - par exemple à l'égard du castor - tient-elle à ce que l'animal et l'homme sont l'un et l'autre des " forces formatrices ", des êtres capables de s'organiser eux-mêmes[394].

Toutefois, le fait d'être une fin naturelle ne garantit pas que l'analogie, entre ces fins naturelles que sont les hommes et les animaux, puisse être telle que l'idée de personne soit compatible avec elle. L'exemple même du castor le montre : non seulement l'analogie entre le comportement du castor qui construit un abri et l'homme bâtisseur ne donne pas le droit de conclure à la raison du castor, mais, bien plus, il n'est pas possible, dans le cadre de cette analogie, de penser la personne. L'analogie, ici, n'est pas du même type que celle dessinée par la parole, qui s'inscrit, on l'a vu, dans un échange, qui dessine une communauté des fins. En vérité, il manque au castor la parole.

La machine, quant à elle, n'est pas un organisme, mais plutôt un organe, au sens d'instrument, c'est-à-dire une fin artificielle ; l'automate est " le produit d'une cause raisonnable, distincte de la matière de ce produit "[395]. L'idée qui détermine *a priori* tout ce qui doit être compris dans la chose est, ici, extérieure, non à la chose elle-même, mais aux parties qui la composent, c'est-à-dire à la matière dont elle est faite :

> « Dans une montre une partie est l'instrument du mouvement des autres, mais un rouage n'est pas la cause efficiente de la production d'un autre rouage ; une partie est certes là pour l'autre, mais elle n'est pas là par cette autre partie. C'est pour cette raison que la cause qui produit celles-ci et leur forme n'est pas contenue dans la nature (de cette matière), mais en dehors d'elle dans un être, qui d'après des Idées, peut produire un tout possible par sa causalité. »[396]

Il y a bien une liaison réciproque entre les rouages de la montre, qui fait qu'ils existent les uns pour les autres, mais ces rouages ne sont pas producteurs les uns des autres ; la causalité efficiente qui les produit n'a pas son principe en elle-même, mais dans une cause finale, c'est-

[394] Kant rejette absolument la théorie cartésienne de " l'animal-machine " : le vivant, pour lui, est irréductible au mécanique.
[395] *Ibid.*, p. 1164.
[396] *Ibid.*, p. 1165.

à-dire dans le tout, ou, plus exactement, en tant qu'ils sont causes efficientes, dans l'idée d'après laquelle ils sont produits. Cela explique que

> « dans une montre un rouage ne peut en produire un autre et encore moins une montre d'autres montres, en sorte qu'à cet effet elle utiliserait (elle organiserait) d'autres matières ; c'est pourquoi elle ne remplace pas d'elle-même les parties qui lui ont été ôtées, ni ne corrige leurs défauts dans la première formation par l'intervention des autres parties, ou se répare elle-même, lorsqu'elle est déréglée... »[397]

La machine - l'instrument – est " un corps (ou corpuscule) dont la force motrice dépend de sa figure... "[398]. Elle possède une force motrice, organisée, dans son efficience, par sa configuration, ou son architecture, comme nous dirions aujourd'hui, mais cette force motrice n'est pas en soi " force formatrice ". C'est en ce sens que la machine relève de la seule " logique générale ".

Pourtant, l'analogie possible entre l'automate et l'homme – celle qu'illustre l'automate de Vaucanson – peut, sans doute, prendre la forme singulière de la parole. On sait, par exemple, que l'idée de fabriquer un automate " parleur " a traversé de nombreux esprits au 18e siècle ; dans leur livre sur Vaucanson, Doyon et Liaigre citent, à cet égard, Blanchet qui, en 1756, écrivait : " On pourrait imaginer et faire une langue, un palais, des dents, des lèvres, un nez et des ressorts dont la matière et la figure ressemblassent le plus parfaitement qu'il serait possible à ceux de la bouche. On pourrait imiter le jeu qui a lieu dans ces derniers pour la génération des paroles... "[399]. De sorte que nous pourrions imaginer un " super-automate de Vaucanson " capable de reproduire la parole en tant que phénomène du sens externe, et qui, au

---

[397] *Ibid*.

[398] Emmanuel Kant, *Premiers principes métaphysiques de la science de la nature*, *Oeuvres philosophiques*, *op.cit.*, II, p. 451.

[399] A. Doyon, L. Liaigre, *Jacques Vaucanson, mécanicien de génie, op. cit.,* p. 168 (Blanchet, *Principes philosophiques*, 1756). Doyon et Liaigre mentionnent les nombreuses recherches menées vers 1770, telles celles de Friedrich von Knauss à Vienne (1770), de H.L. Jaquet-Droz à Londres (1777), de l'abbé Mical à Paris, du baron von Kempelen à Vienne (1778), et de C.G. Kratzenheim à Copenhague (1780), *ibid*., p. 169. Ils signalent, par ailleurs, que Vaucanson lui-même n'a sans doute pas été indifférent au problème mécanique ainsi posé : l'intérêt qu'il portait au caoutchouc pourrait en témoigner, dans la mesure où cette matière nouvelle semblait dotée de propriétés susceptibles d'être utilisées par exemple dans la fabrication d'" anatomies artificielles ".

lieu de jouer de la flûte, aurait pu, en vertu des mêmes principes techniques, produire, avec ses " lèvres ", des sons équivalents à nos paroles, voire énoncer un " discours ".

Toutefois, parce que la machine est une fin artificielle, parce qu'elle relève, par là, de la seule " logique générale ", l'analogie, née de cette parole mécanique, entre le " super-automate de Vaucanson " et un homme, ne sera jamais la même que celle née de la parole que j'échange avec l'autre qui me parle et à qui je parle. Les règles déterminant la machine, en tant que celle-ci relève de la " logique générale ", mettent en jeu les catégories de l'entendement, lesquelles ne donnent que la possibilité d'une expérience en général, et ne rendent pas compte du jugement, en particulier en tant qu'il est réfléchissant :

> « La logique générale ne contient pas de préceptes pour la faculté de juger et ne peut en contenir. En effet, *comme elle fait abstraction de tout contenu de la connaissance*, il ne lui reste qu'à séparer analytiquement la simple forme de la connaissance en concepts, jugements et raisonnements, et à établir ainsi les règles formelles de tout usage de l'entendement. Que si elle voulait montrer d'une manière générale comment on doit subsumer sous ces règles, c'est-à-dire discerner si quelque chose y rentre ou non, elle ne le pourrait, à son tour, qu'au moyen d'une règle. Or, cette règle, par cela même qu'elle est une règle, exige une nouvelle instruction de la part de la faculté de juger, et on voit ainsi que, si l'entendement est capable d'apprendre et de s'équiper au moyen de règles, la faculté de juger est un talent particulier, qui ne peut pas du tout être appris, mais seulement exercé. »[400]

Le " je " qu'un habile mécanicien fera prononcer à un " super automate de Vaucanson ", ne sera jamais l'expression d'une aperception pure ; un tel automate sera reconnaissable comme machine, comme non-personne, à cela que, quand bien même on sera parvenu à lui faire énoncer quelques " phrases ", sa " parole " ne sera jamais *faculté de parler* : quelles que soient les circonstances, elle ne fera jamais autre chose que répéter ces mêmes " phrases ", et sera parfaitement incapable de former une suite de propositions sensées au cours d'une " conversation " quelconque avec nous. L'automate, par exemple, ne se trompera jamais. Il fera toujours ce que son programme lui ordonnera de faire, il effectuera toujours exactement le mouvement qu'il doit accomplir, y compris lorsqu'il sera mal conçu. L'automate relève entièrement de la causalité naturelle et considérer que celle-ci puisse être fautive n'a pas de

[400] Emmanuel Kant, *Critique de la raison pure*, *op. cit.,* p. 881.

sens. Louis Guillermit a bien montré, à cet égard, que, pour Kant, l'erreur renvoyait, non pas aux règles, mais au jugement qui applique ces règles[401]. Dès lors, l'idée d'une " machine qui apprend " - c'est-à-dire qui non seulement apprendrait de nombreuses règles, mais pourrait, dans une certaine mesure, devenir habile et prudente - ne peut avoir de sens aux yeux de Kant. L'analogie, dans le cas du " super-automate ", au-delà de l'identité phénoménale des sons proférés par celui-ci avec nos paroles, ne sera pas celle à partir de laquelle le principe d'un autrui, à savoir la personnalité, est pensable dans l'autre. La parole de l'automate ne sera jamais celle à travers laquelle s'exprime une communauté des fins. Quand bien même le super-automate de Vaucanson proférera des sons identiques à ceux que nous émettons lorsque nous parlons, aucun échange de parole ne sera possible avec lui, de telle sorte que *nous n'aurons jamais, dans ce cas, à penser la personne*.

Si, en effet, l'automate, cette fin artificielle, peut être construit, c'est qu'il peut être *connu* dans son principe. Comme Jean-Pierre Séris[402] l'a bien montré, le succès des spectacles d'automates aux 17e et 18e siècles ne reposait pas seulement sur l'excellence atteinte par leurs constructeurs en ce qui concerne la simulation du vivant, mais également sur le fait que l'on pouvait, en exposant le mécanisme, montrer de quelle manière ce résultat était obtenu : aucun doute n'était permis sur la nature des moyens employés, qui relevaient tous de la seule technique, et mettaient en évidence, non l'habileté de l'automate, mais celle de son constructeur.

Aussi bien suffirait-il, *si besoin en était*, de *démonter* le " super-automate de Vaucanson " pour vérifier qu'il s'agit bien d'une machine et non d'un autrui. Dans le cas d'autrui, en revanche, dans le cas de la personne, je ne puis évidemment espérer découvrir le principe par un " démontage ", puisqu'alors le prix de la reconnaissance d'autrui serait le crime, c'est-à-dire la négation de la personne en lui. Cependant, la possibilité même de *penser* ce principe à partir de la *pratique* où je suis engagé avec lui, la possibilité pour moi de faire *comme s'il s'agissait* d'un autrui, en même temps qu'elle m'interdit de le traiter comme une non-personne, me dispense de le connaître dans son principe. Dans le cas de l'automate, l'absence de cette pratique, c'est-à-dire l'absence de la bonne analogie, fait que je n'ai pas à

401 Louis Guillermit, " Une question difficile : l'erreur ", *Revue Internationale de Philosophie*, 35, 136-137, 1981.

402 Jean-Pierre Séris, *Machine et communication*, Paris, Vrin, 1987.

penser le principe, et que, dès lors, j'ai le droit de démonter l'automate... *L'échange de parole* définit une pratique, relevant à la fois du technique – l'habileté – et du pragmatique – la prudence – fondée par l'idée de pratique comme moralité, de telle sorte que, s'il y a échange de parole, la personne peut être pensée, et doit l'être, et que, s'il n'y a pas échange de parole, la personne n'a pas à être pensée, et la connaissance est possible, et permise. Bref, la machine, chez Kant, dès lors qu'elle peut être construite - donc connue dans son principe – ne peut relever, comme le sujet de l'aperception, de la " logique transcendantale ", et la confusion, dans l'ordre du pratique, entre l'automate et la personne, est impossible. Il ne saurait y avoir de doute quant à la nature de la machine, fût-elle un super-automate de Vaucanson.

On comprend dès lors que les hypothèses de Turing, non seulement sont porteuses d'une critique de la démarche kantienne, mais doivent s'appuyer sur une telle critique, puiqu'elles consistent à montrer, à partir de la problématique du " comme si ", qu'un autrui peut être reconnu *dans une machine* – celle victorieuse au jeu de l'imitation - et que cela même, en infirmant directement l'idée contraire, établit qu'une machine " peut penser ". A travers la mise en jeu, dans *Computing Machinery*…, de la problématique du " comme si ", la parole de la machine au cours du jeu de l'imitation, située sur le même plan que celle de ses adversaires, apparaît comme la parole même qui, chez Kant, renvoie à une communauté de sujets. Comment rendre compte, dans ces conditions, du " pratique " où une telle parole est possible ? Comment définir l'ordre du pratique auquel renvoie la parole, dès lors que celle-ci n'est pas seulement la parole de l'examinateur A du jeu de l'imitation et de son partenaire B, mais également celle de la machine C ?

Conclusion : l'idée de pratique à la lumière de la victoire de la machine ou le renversement de la problématique kantienne

La machine universelle qui l'emporte au jeu de l'imitation *surprend* son principal adversaire humain, l'examinateur A, en se faisant prendre par lui pour un individu humain ; tel est le contenu immédiat de la première hypothèse de Turing, celle de la victoire d'une machine au jeu. Par là même, l'opinion commune, pour laquelle une machine se définit précisément comme ce qui ne saurait surprendre un individu humain, dont il est postulé qu'il pense, est infirmée. La victoire d'une machine au jeu signifie donc *directement* que cette machine " pense ", au sens humain du terme ; la mise en cause de l'opinion commune constitue le principe et le corps même de l'affirmation de la " pensée " des machines.

La structure particulière de l'expérience imaginée par Turing met en oeuvre un procès de reconnaissance de la pensée, qui prend la forme de la reconnaissance, au cours du jeu, par l'examinateur A, d'un semblable, c'est-à-dire d'un autrui. Dès lors, l'attaque menée contre l'opinion commune s'étend jusqu'à la critique de l'expression philosophique de l'idée selon laquelle la machine ne saurait " penser " au sens humain du terme. Cette idée s'inscrit, en effet, à partir de la définition du penser par le jugement, appuyé sur le " je pense ", dans la problématique kantienne de la subjectivité, laquelle fournit une armature théorique au procès de reconnaissance d'un autrui. L'opinion commune donne lieu, ainsi, à une mise en forme théorique : parler d'une " pensée " de la machine n'a pas de sens, car la notion de jugement appartient à un autre registre que celle de machine ; un individu humain, dont il est postulé qu'il pense, ne reconnaîtra jamais un autrui dans une machine, car il n'aura jamais à le faire.

L'hypothèse d'une victoire de la machine au jeu de l'imitation ne met pas en cause les éléments constitutifs de la conception du penser inscrite dans cette problématique – la notion de jugement, liée à un " je pense " et à un ordre spécifique exprimant l'insertion de la conscience dans un " ici et maintenant " - mais elle affirme que la notion de machine,

entendue comme machine universelle, n'entre pas en contradiction avec ces élements. Le principe même de l'erreur que commet A à l'égard de C, et qui lui coûte la victoire, fait que, s'il se trompe en prenant la machine pour un individu humain, *il ne peut, par là même, se tromper* en la prenant pour un autrui. C'est parce qu'il s'inscrit dans une pratique de communication avec la machine, et parce que cette pratique le conduit et l'autorise à tenir celle-ci pour un autrui que l'examinateur peut se tromper au cours du jeu, et être défait.

Le " comme si " par lequel l'examinateur A reconnaît un autrui dans le protagoniste C du jeu n'est en effet possible que parce que la pratique dans laquelle il est engagé avec ce protagoniste est celle-là même par laquelle il reconnaît un autrui. Le ressort, ici, de l'hypothèse de Turing tient à la solidarité existant entre l'idée d'autrui et la pratique - le procès d'énonciation-communication - décrite par le jeu. En d'autres termes, c'est parce que le procès d'énonciation-communication détermine la reconnaissance – ou la pensée, en termes kantiens - d'une entité spécifique, irréductible à tout autre " objet " de l'activité de l'examinateur, que celui-ci se trompe. Cette entité spécifique peut être nommée, en termes techniques, un allocutaire, mais, dans le contexte pratique de son apparition, elle est plus que cela : l'examinateur A reconnaît dans cet allocutaire un *semblable*. De sorte qu'est postulée, à travers le " comme si ", une " conscience ", semblable à celle dont le discours de l'examinateur A, aux yeux de A lui-même, est inséparable - en tant qu'élément du procès d'énonciation-communication, A, non seulement s'adresse à C, mais *sait* qu'il s'adresse à lui et sait qu'il le sait ; cela même le définit, à ses propres yeux, en tant que locuteur.

L'examinateur A du jeu de l'imitation reconnaît un autrui dans son interlocuteur C, car il reconnaît – ou pense - dans le discours de celui-ci, non seulement la forme logique de propositions exprimant un jugement, mais *l'acte* de juger. Ainsi, le procès d'énonciation-communication dans lequel se trouvent engagés les protagonistes A, B et C du jeu conduit A à faire comme si C était, en termes kantiens, une " personne ". Au cours du jeu, et dans le cas d'une victoire de la machine, ce qui rend possible l'erreur de l'examinateur A, qui se trompe en accordant à son adversaire mécanique l'humanité, ce n'est pas le fait que la machine C soit capable de produire formellement des agencements de signes ayant un sens pour l'examinateur – à la manière, par exemple, de l'opérateur de la " chambre chinoise " de Searle – ce n'est pas, en d'autres termes, le fait que la machine ait été dotée de moyens matériels de production de signes et de règles formelles de manipulation de ceux-ci, mais le fait que

l'examinateur soit conduit, par la *pratique* où il est engagée avec elle, à lui accorder, *au-delà de l'humanité proprement dite*, ce qui, chez Kant, est rassemblé sous le concept de " personnalité ". Bref, si l'examinateur A se trompe, s'il perd face à la machine, c'est précisément qu'il met en jeu, en tant que locuteur-allocutaire, l'appareil intellectuel ordonné par la notion de sujet, dans l'expression fondatrice qu'en donne Kant[403]. A se trompe sur l'humanité du joueur C car il doit accorder au " je " énoncé par celui-ci le même statut qu'au " je " qu'il énonce lui-même, et car il voit, dans le comportement de C, l'expression d'un jugement inscrit dans un " ici et maintenant ", et visant à découvrir, à partir du particulier, la règle sous laquelle celui-ci peut être subsumé.

Or, c'est, par là même, la notion kantienne de *pratique* - la solution kantienne du problème posé par l'idée de pratique – qui est bouleversée par les hypothèses de Turing.

La solution kantienne au problème posé par l'idée de pratique et sa mise en question par les hypothèses de Turing

On sait que la notion de " pratique ", renvoyant à ce qui concerne l'action, est d'abord définie par son opposition à la notion de théorie. La *theoria* - action de voir, puis, à partir de Platon, contemplation, méditation – est saisie des Idées, en tant qu'essences, c'est-à-dire de ce qui, dans un monde où tout change, demeure identique à soi-même. Par cela qu'elle renvoie à l'acte de saisie des essences, la théorie relève elle-même du domaine de la pratique : le monde où tout change est son lieu d'émergence. Elle constitue, cependant, sous cet angle même, un

[403] Et en écho à ce que Kant appelle la " première idée régulatrice " de la raison pure, selon laquelle " l'idée pure de la raison ", c'est-à-dire la pensée de l'unité systématique, est d'abord la pensée de moi-même comme sujet : " Je suis moi-même, considéré simplement comme nature pensante (comme âme) le premier objet d'une pareille idée. Si je veux rechercher les propriétés avec lesquelles un être pensant existe en soi, il faut que je consulte l'expérience, et je ne puis même appliquer aucune des catégories à cet objet qu'autant que le schème m'en est donné dans l'intuition sensible. Mais je n'arrive jamais par là à une unité systématique de tous les phénomènes du sens interne. A la place donc du concept d'expérience… la raison prend le concept de l'unité empirique de tout penser, et, en pensant cette unité comme inconditionnée et originaire, elle fait de ce concept un concept de la raison (idée) d'une substance simple, qui, demeurant immuable en soi (personnellement identique) est en commerce avec d'autres choses réelles en dehors d'elle… Mais elle n'a pas ici en vue autre chose que des principes de l'unité systématique dans l'explication des phénomènes de l'âme, principes qui nous font considérer toutes les déterminations comme appartenant à un sujet unique, toutes les facultés, autant que possible, comme dérivées d'une unique faculté fondamentale, tout changement comme appartenant aux états d'un seul et même être permanent, et représenter tous les *phénomènes* qui ont lieu dans l'espace comme entièrement distincts des actes de la *pensée* ". Emanuel Kant, *Critique de la raison pure*, *op. cit.*, p. 1276.

type d'action supérieure : la contemplation des idées détache du monde en perpétuel changement où elle naît. Il revient à Aristote d'avoir clarifié cette distinction de deux types d'action – l'action en général en tant qu'elle est inscrite tout entière dans le monde du changement, et l'acte de la *theoria*, saisie des essences immuables – en opposant la *theoria* à la pratique – dont relèvent la politique et la morale - et à la poïétique – dont relève la technique.

Or, sous cet angle, le niveau de détermination des notions de théorie et de pratique n'est pas le même : la théorie, en tant que contemplation des Idées, peut rendre compte d'elle-même comme pure pensée, comme intelligibilité ; l'idée de théorie et la théorie ne font qu'un. L'idée de théorie rend compte, par là, de sa propre dimension pratique, c'est-à-dire d'elle-même en tant qu'elle renvoie à un acte. Comme le remarque Pierre Livet, il ne peut en être de même de l'idée de pratique : il y a une priorité logique de la théorie qui " fait que la priorité 'pratique' de la pratique s'exprime en des termes qui dépendent encore de cette priorité logique. A bon droit, semble-t-il : on ne peut discourir de la pratique qu'en théorie… "[404]. Dans l'idée de pratique, le pratique – ce qui a la propriété de relever de la pratique – est le pratique tel qu'il est saisi par la pensée, le pratique tel qu'il est appréhendé par la théorie, et non le pratique en tant qu'autre du théorique ; l'acte est décrit par l'idée d'acte comme pensée de l'acte, c'est-à-dire en fonction de l'acte de pensée, et non comme l'acte en tant que tel.

La notion kantienne de pratique propose une solution à ce problème. Le pratique est, chez Kant, précisément ce pour quoi il ne peut y avoir de théorie[405]. Le théorique – ce qui relève de la théorie – est ce qui peut être connu ; le pratique s'en distingue, non seulement comme ce qui ne peut être connu, mais comme ce qui *peut être pensé*. Or, cela même le définit positivement. Le pratique renvoie à la liberté transcendantale ; il prend la forme de la moralité, définie par l'acte de la bonne volonté, c'est-à-dire de la volonté en tant qu'elle se prend elle-même pour fin, en tant qu'elle ne veut rien d'autre qu'elle-même et qu'elle s'exprime par la loi morale et sa maxime universelle. En ce sens, le pratique n'est pas

[404] Pierre Livet, *Penser le pratique : communauté et critique,* Paris, Klincksieck, 1979, p.9.

[405] " Tout ce qui, d'ordre pratique, devrait être possible en fonction de lois naturelles dépend complètement, quant à sa prescription, de la théorie de la nature ; seul ce qui est pratique suivant des lois de liberté peut avoir des principes qui ne dépendent d'aucune théorie ; au-delà des déterminations naturelles, en effet, il n'y a pas de théorie " (Emmanuel Kant, *Métaphysique des mœurs*, *Œuvres philosophiques*, *op. cit.*, III, p. 463).

seulement ce qui ne peut être que *pensé*, il est ce qui *doit* être pensé. L'impératif moral est la véritable pensée du pratique, et c'est à partir de là que la notion complète de pratique se décline, dans le kantisme, comme pur pratique – ce dont relève la loi morale - comme pratique pragmatique – ce dont relèvent le bonheur et les règles de la prudence - et comme pratique technique – ce dont relèvent les règles de l'art. Or, le pratique, dès lors qu'il est ce qui ne peut être que pensé, est bien saisi en tant que tel lorsqu'il est saisi comme *pensée* du pratique – l'idée de pratique coïncide avec ce qui est pratique - et il est saisi en même temps comme fondement : l'idée de pratique est la figure même de l'intelligible - entendu ici comme " problématique ", puisqu'il ne peut être connu - qui s'exprimait jusque là dans la seule idée de théorie.

Dès lors, s'il est vrai qu'une machine peut l'emporter au jeu de l'imitation, s'il est vrai que la machine qui peut l'emporter est une " machine qui apprend ", s'il est vrai, enfin, que la machine l'emporte parce que son examinateur humain, dont il est postulé qu'il pense, doit faire " comme si " elle était un individu humain, ne devons-nous pas en conclure à la nécessité théorique d'élever la machine qui l'a emporté à la dignité de l'idée kantienne de sujet ? Certes, dans la perspective kantienne, l'idée de sujet renvoie à celle d'autonomie, dont la mise en place implique elle-même l'affirmation d'un ordre intelligible opposé au sensible, et l'on ne saurait parler d'autonomie pour une machine dans la mesure où celle-ci relève de la seule " causalité naturelle " ; puisque la liberté s'oppose à la loi naturelle, comme le montre la " Troisième antinomie ", nous ne pouvons en trouver le principe dans la machine, qui est une construction humaine, et relève, par là, non de ce qui peut seulement être pensé, mais de ce qui peut être connu. Cependant, s'il est vrai que le " comme si " est possible et a un sens à l'égard d'une machine, ne devons-nous pas admettre que la machine relève, comme l'être humain, de la " logique transcendantale ", et non pas seulement de la causalité naturelle ? Qu'elle comporte une dimension qui ne peut être connue, mais qui peut être pensée, et qui, comme telle, doit l'être ? Les hypothèses de Turing, dans le contexte de leur validité, constitueraient ainsi une illustration singulière de l'idée kantienne selon laquelle l'homme, en tant que personne, appartient à la classe des " êtres raisonnables ", laquelle ne se réduit pas à

lui[406]. La machine entrerait avec l'homme dans une communauté des fins, et l'on n'aurait plus, alors, le droit de " démonter " celle qui l'emporte au jeu de l'imitation…

Toutefois, la légitimité d'une telle interprétation pose question. Nous devons prendre garde au fait que " l'expérience " de Turing confronte des protagonistes – des individus humains quelconques et une machine - dont les descriptions respectives ne peuvent être situées sur le même plan. Les individus humains y sont réduits à la manifestation de leur " pensée ", mais cette dernière n'est pas pour autant définie conceptuellement. C'est, en effet, précisément pour éviter tout usage d'une définition conceptuelle de la " pensée " que Turing propose une expérience ; dans celle-ci, l'idée de " pensée " renvoie à une intuition, c'est-à-dire à la certitude que tout homme à de sa propre pensée et de la pensée en tout homme. En revanche, la machine qui figure dans l'expérience n'est connue de nous que par sa définition théorique comme machine universelle ; c'est sous la forme de son concept que nous examinons sa participation à l'expérience. " L'expérience " de Turing, en somme, confronte une " pensée " dont nous n'avons pas, par hypothèse, de connaissance conceptuelle, à une machine que nous ne pouvons appréhender qu'à travers une connaissance conceptuelle. Dès lors, avons-nous le droit, en toute rigueur, d'interpréter " l'expérience " en sollicitant une catégorie – en l'occurrence l'intelligible kantien – qui appartient manifestement au registre d'une approche conceptuelle de l'idée de pensée plutôt qu'à celui du concept de machine universelle ? Sans doute, si les hypothèses de Turing sont valides, la victoire d'une machine au jeu de l'imitation, de même que la possibilité, pour cette machine, " d'apprendre " comme le fait un individu humain, établissent-elles qu'une machine peut être prise par un individu humain pour un autrui, c'est-à-dire qu'un individu humain peut être conduit à postuler un intelligible, au sens kantien, à l'égard d'une machine. Conclure, cependant, que cette attitude humaine serait *fondée* par l'intelligible qu'elle postule, c'est-à-dire que la machine participe d'une définition conceptuelle de la pensée, n'est-ce pas dépasser la validité des hypothèses ? L'idée kantienne de liberté, non dans sa définition comme " concept problématique ", mais

[406] " … considéré comme personne, c'est-à-dire comme sujet d'une raison moralement pratique, l'homme est au dessus de tout prix, car, en tant que tel (homo noumenon), il faut l'estimer, non pas simplement comme un moyen pour les fins d'autrui – pas même pour les siennes propres – mais au contraire comme une fin en soi-même, c'est-à-dire qu'il possède une *dignité* (une valeur intérieure absolue) par laquelle il force au respect de lui-même toutes les autres créatures raisonnables, et qui lui permet de se mesurer avec toute autre créature de cette espèce et de se considérer sur un pied d'égalité avec elle " (*ibid.*, p. 722-724).

dans sa définition positive, en tant que fondement de l'action morale, n'est-elle pas un vêtement trop large pour les hypothèses de Turing considérées *stricto sensu* ? En vérité, l'expérience imaginée par Turing montre seulement, d'une part, que la postulation d'un intelligible par un locuteur humain est constitutive du procès d'énonciation-communication dans lequel celui-ci s'engage, et, d'autre part, que cette postulation est possible à l'égard d'une machine, strictement définie comme machine universelle, dès lors que l'on montre que celle-ci peut être l'interlocuteur d'un individu humain dans un procès d'énonciation-communication. Or, tenir l'interprétation de " l'expérience " de Turing dans les limites de cette constatation entraîne un réaménagement de la problématique kantienne de postulation d'un intelligible.

L'intelligible auquel renvoie, chez Kant, le " comme si " mis en œuvre par le procès de reconnaissance d'un autrui, ne constitue pas un monde d'objets saisissables, mais un problème ; il n'est pas connaissable et ne peut servir à l'explication des phénomènes. Pour autant, il ne peut être tenu pour impossible ; il sert à la limitation de l'usage purement empirique de l'entendement. Sous cet angle, le principe de la réflexion kantienne est bien l'idée de sujet, en tant que fin, en tant que personne, c'est-à-dire en tant qu'unité inconnaissable, unité " problématique " du sensible et de l'intelligible, de sorte que la position du sujet kantien, dans sa dimension pratique, est celle d'un absolu ; le sujet est inconnaissable, mais son existence métaphysique est affirmée dans l'action faite par devoir.

Dans le cadre du jeu de l'imitation, nous ne pouvons nous donner le droit de rapporter le " comme si " à un tel sujet, et tout se passe comme si nous devions tenter de rendre compte de ce qui advient à travers la victoire de la machine en faisant abstraction du dualisme kantien. Le procès d'énonciation, ici, ne saurait être appréhendé comme s'il était fondé sur l'unité d'une conscience qui, même problématiquement, le dépasserait. Sous cet angle, l'hypothèse de la victoire d'une machine au jeu, alors même qu'elle implique le " comme si " de l'examinateur, exclut, non seulement, comme cela ressortait de la notion même de machine universelle, l'appel au seul sujet psychologique, tel qu'il est déterminé par la perception externe, mais encore l'appel au sujet métaphysique, c'est-à-dire au sujet en tant qu'il relève d'un ordre intelligible. Le sujet, en somme, entendu comme ce qui fonde le procès de reconnaissance d'un autrui, n'est disponible d'aucun côté du " comme si ", ni du côté de l'acteur de celui-ci, ni du côté de ce qui est visé par lui. En vérité, le " comme si " sur lequel

repose la victoire de la machine, dès lors qu'il n'est plus rapporté à un sujet, apparaît comme une sorte de distance, comme une sorte d'entre-deux dont les pôles seraient manquants. L'hypothèse de Turing " déshabille ", pourrait-on dire, la structure kantienne qui régit la reconnaissance d'un autrui, pour la laisser subsister dans une nudité autosuffisante, comme pure *différence*. Comment penser, sous cet angle, l'idée de pratique mise en jeu dans le test de Turing ? Tel est, à nos yeux, le problème philosophique posé par *Computing Machinery*…

Avant d'examiner ce problème pour lui-même, nous pouvons, d'ores et déjà, essayer de dégager la portée de l'hypothèse de Turing du point de vue du thème initial de la réflexion de celui-ci, à savoir la question posée par la définition de la machine universelle, qui implique, on s'en souvient, la simulation, par celle-ci, des conditions intuitives du calcul pour un individu humain, c'est-à-dire la simulation des conditions de l'intuition du signe mathématique telles que ces conditions étaient décrites par Hilbert.

La dynamique du signe

L'hypothèse d'une victoire possible de la machine au jeu de l'imitation, la mise en scène du " comme si " de l'examinateur du jeu, a pour corollaire, en effet, l'accent mis, dans la théorie du signe, non plus sur le référent, mais sur le fait même de référer. Le jeu de l'imitation est défini de telle façon que la question du contact ultime et direct du signe avec ce qu'il représente - la question de la " référence directe " - s'avère située, en tant que telle, sur un autre plan que celui du problème étudié. La possibilité même de la référence pour la machine n'est certes pas exclue, nous l'avons vu, mais elle n'intervient que secondairement dans la problématique de Turing. Dans l'horizon du jeu de l'imitation, c'est la dynamique propre, non d'un sujet doté d'une capacité de perception externe, ni d'un sujet défini comme personne par la loi morale, mais du *signe* qui donne sens à l'idée de machine universelle. Si l'on admet la définition de la notion de signe la plus couramment utilisée - ce qui est " mis pour autre chose " - ce n'est pas tant le " autre chose " qui importe, ici, que le " mis pour ". Il apparaît, à cet égard, que la perspective dans laquelle s'inscrit Turing rejoint l'interprétation de la théorie hilbertienne du signe donnée, au moment même de la publication de *On Computable Numbers*..., par Jean Cavaillès. Rappelons que celui-ci décrivait, en recourant à la notion husserlienne de " thématisation "[407], le mouvement au cours duquel l'acte de calcul,

[407] Voir ci-dessus, 1ère partie, chapitre 2.

acte opératoire accompli, comme combinaison de symboles, dans le cadre d'un problème spécifique, devient lui-même, sous la forme d'un symbole situé à un autre niveau d'abstraction, un objet mathématique, défini à son tour par les actes opératoires qu'il permet. Cavaillès montrait, en d'autres termes, que l'objet mathématique n'est rien d'autre que le signe mathématique lui-même. Pour Cavaillès, ce que le signe mathématique représente, c'est d'abord le *fait même de représenter*. Cette dimension du signe est celle que met en scène l'hypothèse de Turing.

Dès lors que l'on exclut l'idée qu'il tire son sens des conditions de la perception externe, et que l'on ne se donne pas le droit de recourir à l'idée d'un " inconnaissable " postulé pour rendre compte d'une limite *a priori* de notre faculté de connaître, et assimilé au sujet moral, le " comme si " de l'examinateur du " jeu de l'imitation" à l'égard de la machine, " comme si " qui détermine la pratique dans laquelle l'examinateur est engagé avec celle-ci, renvoie, en effet, aux seuls signes, à travers lesquels il s'exprime, considérés comme inséparables de l'acte qui les énonce. Le " comme si " tire, alors, son sens de cela que ce qui est d'abord donné par le signe, dans son énonciation, c'est lui-même. C'est, en somme, le signe, plutôt que la " personne ", qui, en tant qu'il est énoncé, en tant qu'il est inséparable d'un acte, apparaît comme étant à lui-même sa propre fin. En formulant, dans *On Computable Numbers*…, le concept de machine universelle, Turing illustrait cette conception du signe, qui débouchait sur la question de la " pensée " des machines ; il la confirme avec l'épreuve du jeu de l'imitation, mais en même temps, la conduit plus loin : à travers le jeu de l'imitation, la dynamique du signe rend compte non seulement de la pratique mathématique proprement dite - le " calcul ", tel qu'il est effectué par la " machine de Turing " - mais de toute pratique définie par l'énonciation et la communication, c'est-à-dire de toute pratique humaine.

Le renversement de la problématique kantienne

Telle serait donc la spécificité du jeu de l'imitation qu'il met en scène la pratique humaine sous la seule figure du procès d'énonciation-communication, et qu'il ne nous autorise pas, à travers l'hypothèse de la victoire possible de la machine, à rapporter ce procès à un sujet. Le " comme si " sur lequel repose le jeu est l'expression de la dynamique du signe, lequel, renvoyant à ce " pour quoi " il est mis, renvoie d'abord à lui-même, c'est-à-dire à

l'écart, à la distance qui le constitue en tant que signe. Or, le recours à l'image de l'écart, de la distance, posés par le " mis pour " qui définit le fait de signifier, image paradoxalement indifférenciée dès lors qu'il ne reste plus qu'elle, permet-il de décrire, dans sa singularité, la dynamique du signe telle qu'elle s'exprime, au cours du jeu, à travers le procès d'énonciation-communication ? La différence, fût-elle conçue comme *différer*, comme mouvement sans origine ni terme, suffit-elle à saisir ce qui a lieu, ce qui advient, lors de la victoire d'une machine au jeu de l'imitation ?

En vérité, la notion de différence, réduite, sous la figure de l'écart, à la plus stricte neutralité, ne rend pas compte du rôle spécifique joué par la notion de sujet dans l'économie même du jeu de l'imitation. Nous l'avons vu, en effet, c'est parce qu'il met en œuvre l'appareil intellectuel dans lequel s'énonce l'idée du sujet métaphysique, que l'examinateur A perd face à la machine. Dès lors, si la défaite de A au cours du jeu témoigne, aux yeux de Turing, du " penser " de la machine, c'est, non seulement qu'elle témoigne du penser même de l'examinateur A – il s'agit pour la machine de tromper, en se faisant prendre pour un individu humain quelconque, un individu humain quelconque dont il est postulé qu'il pense – mais, bien plus, que le penser en A, en tant qu'il s'énonce, *passe par l'erreur même* commise par A à propos de l'identité de la machine. Nous l'avons souligné, l'examinateur du jeu croit reconnaître en son interlocuteur un homme parce qu'il reconnaît chez lui une pensée, et non l'inverse. En ce sens, c'est parce que les conditions de possibilité du penser ne peuvent être reconnues que sous la forme d'un " je pense ", en d'autres termes, parce que le jeu conduit l'examinateur A à mettre en œuvre le système catégoriel dans lequel est exprimée, chez Kant, l'idée de sujet, que la machine C s'avère en mesure de l'emporter face à ses adversaires humains. Or, le principe de la victoire de la machine au jeu de l'imitation implique, par là, non pas simplement que la problématique de la reconnaissance d'un autrui soit dépouillée du fondement moral qui permet de l'appréhender chez Kant, mais, bien plutôt, que cette problématique, en ce qu'elle met en œuvre l'idée de sujet, soit *renversée* ; puisque l'affirmation du sujet constitue *un moment du procès d'énonciation-communication*, le " transcendantal " qui, dans le kantisme, fait référence au primat du sujet moral, doit être, en quelque sorte retourné sur lui-même, de telle façon que le procès pratique d'énonciation-communication soit saisi, non comme différence en soi, par exclusion de l'idée de sujet, mais comme *premier* par rapport à celle-ci. C'est, ici, le procès d'énonciation-communication qui

enveloppe le sujet, et non l'inverse[408]. De la double expérience imaginée par Turing, celle au cours de laquelle une machine trompe un individu humain sur son identité, et celle au cours de laquelle une machine est soumise à un processus d'éducation, ressort que tout se passe comme si l'entrée dans un procès d'énonciation-communication impliquait l'énoncé d'un " je pense ", valant par lui-même la reconnaissance de la pensée dans un autre. Pour reprendre la célèbre distinction wittgensteinienne du " dire " et du " montrer ", dans l'énonciation d'un " je pense ", se *montre* autre chose que ce qui y est *dit*. Tout se passe comme si, en d'autres termes, *l'acte* d'énonciation d'un " je pense " débordait le " je pense " lui-même, comme si la position d'un " je pense " dépassait le " je pense ".

Allons plus loin. Du fait de l'économie singulière du jeu de l'imitation, dans le procès d'énonciation-communication mis en scène par " l'expérience " de Turing, le dépassement du " je pense " prend la forme du renversement de la problématique kantienne - la victoire de la machine infirme non seulement l'opinion commune, mais l'idée philosophique selon laquelle une machine ne peut penser. Dès lors, il apparaît que le " dire " du dépassement du " je pense " - l'expression catégorielle de ce dépassement - met en jeu la structure même dont il affirme s'affranchir. Penser le procès pratique d'énonciation-communication comme premier, c'est, en effet, lui faire occuper la place même qui est celle de l'idée de sujet dans la problématique kantienne, de sorte que l'espace dans lequel se meut celle-ci n'est pas lui-même modifié ; " renverser " la problématique kantienne maintient la structure intellectuelle où s'exprime la dynamique de celle-ci, comme si l'on *disait* alors que le procès d'énonciation-communication *fonde* la postulation d'un intelligible, comme si " dire " le renversement consistait à attribuer au procès d'énonciation-communication les propriétés de l'intelligible qu'il s'agit, précisément, de renverser. Du renversement du transcendantal kantien naît encore, en somme, l'exigence d'un renversement, qui, comme tel, parce que tout acte d'énonciation déborde ce qu'il énonce, ne peut être pensé que comme un renversement sans fin. Des hypothèses de Turing ressort ainsi que ce qui se joue au cœur du procès d'énonciation-

[408] En ce sens, il est sans doute permis de dire que les hypothèses de Turing sont porteuses d'une conception " dialogique " de la pensée, au sens donné à ce terme par Francis Jacques : elles mettent en évidence un primat de la relation interlocutive ; le sens du discours du locuteur se construit, non seulement dans cette relation, mais par elle. Voir : Francis Jacques, *Dialogiques. Recherches logiques sur le dialogue*, Paris, PUF, 1979.

communication, au cœur du langage, n'est rien d'autre qu'un pur " renverser ", le " poser-dépasser " d'un " je pense " éternellement à renverser.

Dans la perspective ainsi ouverte, l'idée kantienne de pratique apparaît elle-même comme un moment nécessaire du procès d'énonciation-communication. L'idée de pratique serait ainsi un moment de la pratique, une forme de l'acte pratique ; or, toute coïncidence entre l'idée de pratique et la pratique en tant que telle est par là même ruinée. La pratique serait première par rapport à l'idée de pratique, de sorte que celle-ci manquerait toujours ce qu'elle vise. C'est, cependant, à travers cette impuissance même, à travers cette inadéquation de principe de l'idée de pratique que se *montrerait* le pratique, c'est-à-dire ce qui relève de la pratique. Tout se passe comme si l'idée de pratique ne pouvait être, en dernière instance, que *critique*, et comme si la pratique ne pouvait être pensée que comme la dynamique de cette critique même.

C'est en ce sens que, selon nous, les hypothèses de Turing impliquent, dans leur élaboration, la mise en oeuvre de catégories et de concepts appartenant à l'histoire spécifique de la philosophie, et l'on voit, alors, que considérer l'évidente dimension philosophique de cette réflexion comme une sorte de corollaire serait une erreur ; les hypothèses de Turing ne sont pas le prétexte à une réflexion philosophique qui serait à mener à partir d'elles, et qui pourrait se passer d'elles, mais constituent, par elles-mêmes, le cœur d'une position philosophique, au sens technique du terme.

En outre, parce que l'idée de pratique qui anime la réflexion de Turing, non seulement inclut, dans sa définition, la critique de l'idée kantienne de pratique, mais, bien plus, se définit comme le mouvement même de cette critique, on ne saurait dire qu'elle pose un système de référence qui aurait sa propre nécessité, indépendante de l'histoire de la philosophie à laquelle elle fait appel. L'idée de pratique se confond, ici, avec le mouvement historique d'une réflexion philosophique sur la pratique. En tant qu'expression du mouvement même de la critique de la réflexion kantienne sur la pratique, l'idée de pratique qui donne sens à la réflexion de Turing dans *Computing Machinery…* renvoie à *l'acte philosophique* d'élaboration de cette critique. De sorte que tout se passe comme si c'était le devenir philosophique, l'histoire concrète de la philosophie – par opposition à une philosophie de

l'histoire de la philosophie – bref, la pratique philosophique en tant que pratique historique, qui constituait le premier contenu de cette idée.

Or, l'horizon de la pratique philosophique est, en ce sens, l'erreur philosophique même, non pas cette erreur que la philosophie pourrait surmonter en accédant à un ordre où l'erreur n'aurait plus de place car il rendrait compte des conditions de possibilité même de celle-ci, mais l'erreur qui apparaît, à travers l'économie de l'idée de pratique, comme un *principe constitutif* de la pratique philosophique. La catégorie philosophique de pratique croit saisir l'acte pratique en le saisissant à partir d'un " je pense ", comme " pur penser " ; cependant, l'acte même de la philosophie dépasse son énoncé ; il ne peut y avoir de " pensée pure ". La philosophie se fourvoie, en somme, dans sa propre histoire, mais ce fourvoiement même la révèle comme " pratique ", c'est-à-dire comme activité humaine, comme *histoire*, une histoire qui n'a pas à être fondée, qui est la seule véritable " autonomie ", et qui ne peut être pensée ainsi qu'en étant saisie comme *première*, à travers la mise en jeu d'une " logique transcendantale " éternellement à renverser.

Nous espérons avoir ainsi montré, d'une part que la réflexion de Turing sur la " pensée " des machines avait bien une portée philosophique propre, qui ne saurait être réduite à son contexte philosophique immédiat – la philosophie analytique dans son mouvement historique – et, d'autre part, en quoi consiste cette portée : la démarche de Turing met en jeu l'idée de pratique en tant qu'elle donne sens à la dimension *critique* de l'acte philosophique. Or, si l'hypothèse que nous proposons ainsi est admise, elle devient une condition de l'examen des rapports de *Computing Machinery*… à son contexte philosophique immédiat ; cet examen ne pourra plus consister à montrer seulement en quoi le jeu de l'imitation illustre certains aspects de ce contexte, ou peut servir d'appui aux débats dûment estampillés philosophiques qui le traversent ; il s'agira, bien plutôt, de déterminer quel éclairage la démarche de Turing apporte aux problèmes débattus et aux positions défendues. Cet examen reste à mener, et nous ne saurions préjuger ni de ses résultats, ni du simple fait qu'il contribuerait peut-être à ruiner notre hypothèse même. Nous ne pouvons, ici, que tenter de caractériser la perspective ouverte : la réflexion de Turing réintroduit, nous semble-t-il, le poids *historique* des problèmes philosophiques dont la philosophie analytique s'est donné pour tâche de discuter le *statut*. En ce sens, la démarche de Turing apporte, sans doute, à la critique de la philosophie inscrite dans ce que l'on a appelé le " tournant linguistique ", une

forme de légitimation : le langage y apparaît comme le fait premier de toute réflexion. Elle le fait, cependant, à partir du principe que c'est *l'erreur philosophique* qui rend compte de la dynamique du signe. Par là même, le discours philosophique ne peut être pensé sans que soit mise en jeu son erreur même, laquelle ne peut être actuelle sans être chargée de toute l'histoire de la philosophie.

Annexe : la machine de Turing.

La machine de Turing classique, après révision, calcule la valeur d'une fonction d'entiers pour un argument qui lui est fourni en entrée sur son ruban. En 1952, Kleene, dans son ouvrage de référence *Introduction to metamathematics* (S. C. Kleene, *Introduction to Metamathematics*, Amsterdam, North-Holland, 1952) en donne une description inspirée de l'article de 1936 de Post, " Finite Combinatory Processes ", et reprise dans sa contribution à l'ouvrage édité par Rolf Herken à l'occasion du cinquantième anniversaire de la publication de *On computable numbers…* - " Turing's Analysis of Computability, and Major Applications of It " (Rolf Herken éd., *The Universal Turing Machine, a Half-Century Survey*, Vienne, Springer-Verlag, 1995).

La description des opérations de la machine suppose que des moments de temps *t*, énumérés 0, 1, 2, ... soient distingués.

La machine comporte un ruban de longueur infini à droite divisé en cases.

Une seule case est observée par la machine à un moment *t*.

A chaque moment *t*, la machine est dans l'un des états d'une liste fixe de k + 1 (≥ 1) états $q0, ..., qk$. L'état $q0$ est appelé " état passif " ; les états $q1, ..., qk$ " états actifs ".

Chaque case du ruban peut porter un symbole pris dans une liste fixe de L (≥ 1) symboles *s*1, ..., *s*L, ou bien être à blanc, la case blanche étant notée *s*0.

Il y a donc L + 1 conditions de cases.

A chaque moment *t*, la situation complète est décrite par :

une certaine impression sur le ruban (quelles cases sont imprimées sur le ruban, et, pour chacune d'elle, avec quel symbole *s*0, ..., *s*l) ;

une certaine position du ruban dans la machine (quelle case est observée) ;

un certain état de la machine (quel état *q*0, ..., *q*k est celui de la machine).

La situation est dite " active " si l'état de la machine est actif (*q*1, ..., *q*k), " passive " si l'état est *q*0.

Chaque situation complète est un objet fini.

Si la situation est active à un moment *t*, alors la machine accomplit un acte entre les moments *t* et *t* + 1 ; cet acte détermine la situation à *t* + 1.

Chaque acte comporte trois parties, correspondant aux trois éléments d'une situation :

choix parmi *s*0, ...,*s*L d'une condition de case (un symbole) *s*i' à *t* + 1 pour la case observée à *t* (pour *s*i sur cette case à *t*) ;

déplacement du ruban de telle sorte que, à *t* + 1, la case observée soit, par rapport à la case observée à *t*, une case plus à gauche (G), la même case (C pour centre), ou une case plus à droite (D) ;

choix parmi *q*0, ..., *q*k d'un état *q*j' à *t* + 1 (pour *q*j à *t*).

Une machine peut donc être décrite par un triplet :

G

*s*i' (C) *q*j'

D

Si, à *t*, la situation est passive, aucun acte n'est accompli par la machine ; la situation à *t* + 1 est la même qu'à *t*.

L'acte de la machine est déterminé par la paire {*s*i, *q*j} à un moment *t*. Cette paire est appelée " configuration ".

Avec L + 1 cases observées et k états actifs, il y a donc (L + 1)k configurations actives.

A partir de là, le comportement d'une machine de Turing particulière peut être décrit par une table montrant pour chaque configuration active quel acte sera accompli. Une telle table comportera deux colonnes, pour *t* et *t* +1, et chaque ligne de la table comportera les deux triplets correspondants.

Des modèles plus complexes de machines de Turing ont été étudiés, en particulier à partir de la fin des années cinquante : machine avec ruban infini dans les deux sens, avec plusieurs rubans et plusieurs têtes, ou avec plusieurs têtes par ruban ; machine non déterministe (pour lesquelles plusieurs triplets sont possibles à chaque pas). Dans chaque cas, on peut montrer que la machine complexe peut être décrite par une machine simple (voir : Jean Mosconi, *La constitution de la théorie des automates, op. cit.*).

Bibliographie

Les œuvres de Turing

Les « archives Turing » sont hébergées au *Modern Archives Center* de King's College, à Cambridge.

Une bibliographie complète – avec les réserves d'usage – des écrits de Turing a été établie par Andrew Hodges et publiée par celui-ci sur le site Internet qu'il a consacré au mathématicien anglais (http://www.Turing.org.uk).

L'essentiel des oeuvres de Turing est en cours de publication dans :

Collected Works of A.M. Turing, Londres, North-Holland, 1992.

Trois volumes ont été jusqu'à présent publiés :

Pure mathematics, J.L. Britton éd., 1992 ;
Mechanical intelligence, D.C. Ince éd., 1992 ;
Morphogenesis, P.T. Saunders éd., 1992.

Un quatrième volume, à paraître, *Mathematical logic*, sous la direction initiale de R.O. Gandy, contiendra :

On computable numbers, with an application to the Entscheidungsproblem, Proceedings of the London Mathematical Society, 42 ; correction ibid. 43, 1937.

Computability and lambda-definability, Journal of Symbolic Logic, 2, 1937.

The p-function in lambda-K conversion, Journal of Symbolic Logic, 2, 1937.

Systems of logic based on ordinals, Proceedings of the London Mathematical Society, 45, 1939. Publié in Martin Davis (éd.), *The undecidable : basic papers on undecidable. Propositions, unsolvable problems and computable functions*, New-York, Raven Press,1965.

(avec M. H. A. Newman) *A formal theorem in Church's theory of types*, Journal of Symbolic Logic, 7, 1942.

The use of dots as brackets in Church's system, Journal of Symbolic Logic, 1942.

Practical forms of type theory, Journal of Symbolic Logic, 13, 1948.

Un certain nombre de textes ne figurent cependant pas dans les *Collected Works* :

The Automatic Computing Engine, Conférences données au *Ministry of Supply*, décembre 1946 et janvier 1947.

Programmers handbook, Manchester University Computing Laboratory, 1951.

Local programming methods and conventions, Manchester University Computer Inaugural Conference, juillet 1951.

Turing a participé, par ailleurs, à deux émissions de radio, dont les scripts sont disponibles au *BBC Written Archives Center* :

Can digital computer's think ? Diffusé le 15 Mai 1951.

Can calculating machines be said to think ? Diffusé le 10 Janvier 1952.

Le rapport de Turing au *National Phsycal laboratory* de 1946, a été publié au MIT Press :

A.M. Turing's ACE report of 1946 and other papers, B.E. Carpenter et W. R. Doran éd., Cambridge, Massachussetts., MIT Press, 1986.

A signaler également le travail dirigé par Jeanine Alton :

ALTON jeanine, WEISKITTEL Harriott (éd.), *Report on the papers of Alan Mathison Turing OBE, FRS (1912-1954) mathematician, 1923-55*, London, Contemporary Scientific Archives Centre, 1977.

ALTON, jeanine, HARPER Peter (éd.), *Supplementary catalogue of papers and correspondance of Alan Mathison Turing, FRS (1912-1954), material additional to CSAC*, London, Contemporary Scientific Archives Centre, 1985.

Les deux principaux articles de Turing, *On Computable Numbers…*, et *Computing Machinery…*, ont été publiés en français dans :

GIRARD Jean-Yves, *La machine de Turing*, Paris, Seuil, 1995,

sous les titres : « Théorie des nombres calculables, suivie d'une application au problème de la décision », trad. de Julien Basch, et « Les ordinateurs et l'intelligence », trad. de Patrice Blanchard.

La traduction de Patrice Blanchard, « Les ordinateurs et l'intelligence », est d'abord parue dans :

ANDERSON Alan Ross, GUIÈZE Gérard (dir.), *Pensée et Machine*, trad. de Patrice Blanchard, Paris, Champ Vallon, 1983.

Elle a également été publiée dans :

HOFSTADTER Douglas, DENNETT Daniel, *Vues de l'esprit : fantaisies et réflexions sur l'être et l'âme*, trad. de Jacqueline Henry, Paris InterEditions, 1987.

Bibliographie générale

La philosophie analytique, Paris, Les Editions de Minuit (Cahiers de Royaumont), 1962.

ANDERSON J. A., « Turing's test and the perils of psychohistory », *Social epistemology*, tome 8, n°4, pp. 327-332, 1994.

ANDERSON Alan Ross, GUIEZE Gérard (dir.), *Pensée et machine*, trad. de Patrice Blanchard, Paris, Champ Vallon, 1983.

ANDLER Daniel (dir.), *Introduction aux sciences cognitives*, Paris, Gallimard, 1992.

- « Turing : pensée du calcul, calcul de la pensée », *Le formalisme en question*, F. Nef, D. Vernant éd., Paris, Vrin, 1998

ANDRIEU Bernard (dir.), *Le cerveau, la machine-pensée*, Paris, L'Harmattan, 1993.

APÉRY R., CAVEING M., DESCLÉES J.P., *Penser les mathématiques*, Paris, Editions du Seuil, 1982.

AUSTIN J.L., *Ecrits philosophiques*, trad. de L. Aubert et A. N. Hacker, Paris, Le Seuil, 1994.

BARNES E., « The causal history of computational activity : Maudlin and Olympia », *Journal of philosophy*, tome 88, n° 6, pp. 304-316, 1991.

BEAUSOLEIL J.R., « The metamathematics-poperrian epistemology connection and its relation to the logic of Turing's programme », *British journal for the philosophy of science*, tome 40, n° 3, pp. 307-322, 1989.

BENIS-SINACEUR H., article « Thématisation », *Encyclopédie philosophique universelle*, Paris, PUF, 1990-1991

BEYSSADE Jean-Marie, « La critique kantienne du 'cogito' de Descartes », *Kant et la pensée moderne : alternatives critiques*, Bordeaux, Presses Universitaires de Bordeaux, 1996.

BIANCO Edmond, *Informatique fondamentale : de la machine de Turing aux ordinateurs modernes*, Boston, Birkhaüser, 1979.

BLANCHÉ Robert, *Introduction à la logique contemporaine*, Paris, Librairie Armand Colin, 1968.

- *L'axiomatique*, Paris, PUF, 1967.

BLOCK Ned, « Psychologism and behaviorism », *The philosphical review*, tome XC, n° 1, 1981.

BODEN Margaret A. (éd.), *The philosophy of artificial intelligence*, Oxford, Oxford University Press, 1990.

BOLTER J. David, *Turing's man*, London, Duckworth, 1984

BRETON Philippe, *Une histoire de l'informatique*, Paris, Editions de La Découverte, 1987.

BOUVERESSE Jacques, *Wittgenstein, la rime et la raison, science, éthique et esthétique*, Editions de Minuit, 1973.

BRUECKNER A.C., « Brains in a vat », *Journal of Philosophy*, n° 83, pp. 148-167, 1986.

BRUN Jean, *Le rêve et la machine : technique et existence*, Paris, La table ronde, 1992.

CANGUILHEM Georges, « Machine et organisme », *La connaissance de la vie*, Paris, Vrin, 1965.

CAVAILLÈS Jean, *Oeuvres complètes de philosophie des sciences*, Paris, Hermann, 1994.

- *Sur la logique et la théorie de la science*, nouvelle édition, avec une postface de Jan Sebestik, Paris, Vrin, 1997.

CERVONI Jean, *L'énonciation,* Paris, PUF, 1987.

CLARK Andy, éd., *The legacy of Alan Turing, vol I : Machines and thought, vol II : Connectionism, concepts of folk psychology*, Oxford, Oxford University Press, 1997

COPELAND B.J., PROUDFOOT D., HINTIKKA J., « On Alan Turing's anticipation of connectionism », *Synthese : (Dordrecht)*, tome 108, n° 3, pp. 361-377, 1996.

CROCKETT Larry, *The Turing test and the frame problem : AI's mistaken understanding of intelligence*, Norwood, N.J., Ablex Pub. Corp., 1994.

DAVIS Martin (éd.), *The undecidable : basic papers on undecidable. propositions, unsolvable problems and computable functions*, New-York, Raven Press, 1965.

DE GROOT J., « On the surprising in Science and Logic », *Review of metaphysics*, tome 40, n° 160, pp. 631-655, 1987.

DENNETT D. C., « The abilities of men and machines », *Brainstorms*, Cambridge, Massachussetts, MIT Press, 1978.

DESCARTES René, *Œuvres philosophiques*, Paris, Classiques Garnier, 1988.

DIEUDONNÉ Jean, « Les méthodes axiomatiques modernes et les fondements des mathématiques », *Les grands courants de la pensée mathématique*, Paris, Librairie scientifique et technique Albert Blanchard, 1962.

DOTZLER B.J., « Kant und Turing. Zur Archäologische des Denkens der Maschine », *Philosophisches Jahrbuch*, tome 96, n° 1, pp. 115-131, 1989.

DOYON André**,** LIAIGRE Lucien, *Jacques Vaucanson, mécanicien de génie*, Paris, PUF, 1967.

DREYFUS Hubert, « Intelligence artificielle : mythes et limites », Paris, Flammarion, 1984

DUBRUCS Jacques, « Réalisme et antimécanisme chez Gödel », *Dialectica*, tome 40, n° 4, 1986.

EISLER Rudolf, *Kant-lexikon*, Paris, Gallimard, 1994.

ESPINAS Alfred, « L'organisme ou la machine vivante en Grèce, au IVe siècle av. J.C. », *Revue de métaphysique et de morale*, 1903.

FEIGENBAUM Edward, FELDMAN Julian (éd.), *Computers and thought*, New York, McGraw-Hill book Company Inc., 1963.

GANDY Robin., « The confluence of ideas in 1936 », *The universal Turing machine, a half century survey*, Rolf Herken (éd.), Oxford, Oxford University Press, 1988.

GENOVA J., « Turing's sexual guessing game », *Social epistemology*, tome 8, n° 4, pp. 313-326, 1994.

GEORGE F., « Minds, machines and Gödel : another reply to Mr Lucas », *Philosophy*, n° 37, 1962.

GILSON Etienne, *Discours de la méthode : texte et commentaire*, Paris, Vrin, 1962.

GOTTFRIED Ted, *Alan Turing : the architect of the computer age*, Franklin Watts Inc., 1996

GUÉROULT Martial, *Descartes selon l'ordre des raisons*, Paris, Aubier, 1953.

GUILLERMIT Louis, « Une question difficile : l'erreur », *Revue internationale de philosophie*, tome 35, n° 136-137, 1981.

GUNDERSON Keith, « Le jeu de l'imitation », in *Pensée et machine*, Alan Ross Anderson, Gérard Guièze (dir.), trad. de Patrice Blanchard, Paris, Champ Vallon, 1983.

HARRISON J., « Professor Putnam on brains in vats », *Erkenntnis*, tome 23, n° 1, pp. 55-57, 1985.

HEIL J., « Are we brains in a vat ? Top philosophers says no », *Canadian journal of philosophy*, tome 17, n° 2, pp. 427-436, 1987.

HEISER Jon F., COLBY Kenneth M., FAUGHT William S., PARKINSON, Roger C., « Can psychiatrists distinguish a computer simulation of paramoia from the real thing ? », *Journal of psychiatric research*, n° 15, 1979.

HEMPEL Carl, *Eléments d'épistémologie*, trad. de Bertrand Saint Sernin, Armand Colin, 1972.

HERKEN Rolf*, The universal Turing machine, a half-century survey*, Oxford, Oxford University Press, 1988.

HYPPOLITE Jean, « Langage et pensée », *Figures de la pensée philosophique*, *écrits, 1931-1968*, II, Paris, PUF, 1971.

HINTIKKA Jaakko, « Cogito ergo sum : inférence ou performance ? », trad. de P. Le Quellec-Wolff, *Philosophie Paris*, n° 6, pp. 21-51, 1985 (« Cogito ergo sum : inference or performance », *Philosophical Review*, tome LXXI, n° 71, pp. 3-32, 1962).

HODGES Andrew, *Alan Turing ou l'énigme de l'intelligence*, Paris, Payot, 1988, (*Alan Turing : the Enigma*, Londres, Vintage, 1983).

- *Turing : a natural philosopher*, Phoenix, Allen & Unwin, 1997.

- *Great philosophers : Turing*, Phoenix, Allen & Unwin, 1998

HOFSTADTER Douglas, DENNETT Daniel, *Vues de l'esprit : fantaisies et réflexions sur l'être et l'âme*, trad. de Jacqueline Henry, Paris InterEditions, 1987.

- *Gödel, Escher, Bach, les rubans infinis d'une guirlande éternelle*, Paris, InterEditions, 1987.

JACOB Pierre (dir. et trad.), *De Vienne à Cambridge, l'héritage du positivisme logique de 1950 à nos jours*, Paris, Gallimard, 1980.

JACQUETTE Dale, « Metamathematical criteria for minds and machines », *Erkenntnis,* tome 27, n° 1, pp. 1-16, 1987.

- « Adventures in the chinese room », *Philosophy and phenomenological research*, tome 49, n° 4, pp. 605-623, 1989.

- « Searle's intentionality thesis », *Synthese : (Dordrecht)*, tome 80, n° 2, pp. 267-275, 1989.

JACQUES Francis, *Dialogiques. Recherches logiques sur le dialogue*, Paris, PUF, 1979.

KANT Emmanuel, *Œuvres philosophiques*, Paris, Gallimard (Bibliothèque de La Pléiade), 1980-1986.

KARELIS C., « Reflections on the Turing test », *Journal of the theory of social behaviour*, tome 16, n° 2, 1986.

KEITH W., « Artificial intelligences, feminist and otherwise », *Social epistemology*, tome 8, n° 4, pp. 333-340, 1994.

KELLER Pierre, « Personal identity and Kant's third person perspective », *Idealistic studies*, tome 24, n° 2, pp. 123-146, 1994.

KETTANI Omar, *Modèle de calcul sans changement d'état : quelques développements et résultats,* Thèse de philosophie soutenue à l'université d'Aix-Marseille, 1989.

KING D., « Is the human mind a Turing machine ? », *Synthese : (Dordrecht),* tome 108, n° 3, pp. 379-389, 1996.

KIRK R., « Mental machinery and Gödel », *Synthese*, 1986.

KLEENE S. C., *Introduction to metamathematics*, Amsterdam, North-Holland, 1952.

- « Turing's analysis of computability and major applications of it », *The universal Turing machine, a half-century survey*, Rolf Herken (éd.), Oxford, Oxford University Press, 1988.

LACOSTE Jean, *La philosophie au XXe siècle : introduction à la pensée philosophique contemporaine*, Paris, Hatier, 1988.

LADRIÈRE Jean, « Les limitations internes des formalismes, Paris, Gauthier-Villars, 1957.

- « Les limites de la formalisation », *Logique et connaissance scientifique*, Jean Piaget (dir.), Paris, Gallimard (Encyclopédie de La Pléiade), 1967.

LAINE KETNER K., « Peirce and Turing : comparisons and conjectures », *Semiotica*, tome 68, n° 1-2, pp. 33-61, 1988.

LASSÈGUE Jean, « Le test de Turing et l'énigme de la différence des sexes », *Les contenants de pensée*, D. Anzieu, G. Haag éd., Paris, Dunod, 1993

- *L'intelligence artificielle et la question du continu ; remarques sur le modèle de Turing.* Thèse de doctorat soutenue à l'université de Paris 10, 1994.

LAURIER Daniel, *Introduction à la philosophie du langage*, Bruxelles, Mardaga, 1993.

LEIBER Justin, « On Turing's test and why the matter matters », *Synthese : (Dordrecht),* tome 104, n° 1, pp. 59-69, 1995.

- « Shannon on the Turing test », *Journal of theoretical social behaviour*, n° 19, 1989.

LEWIS D., « Lucas against mechanism », *Philosophy*, 1969.

LIVET Pierre, *Penser le pratique : communauté et critique,* Paris, Klincksieck, 1979.

LUCAS J.R., « L'esprit humain, la machine et Gödel », Alan Ross Anderson, Gérard Guièze (dir.), *Pensée et machine*, Paris, Champ Vallon, 1983.

MARCONI Diego, *La philosophie du langage au vingtième siècle*, trad. de Michel Valensi, Paris, Editions de l'Eclat, 1997.

MARIN Louis, « Sur une société de machines dans la logique de Port-Royal », in *La machine dans l'imaginaire*, *Revue des sciences humaines Lille*, n° 186-187, pp. 159-169, 1982-1983.

MARTIN Roger, *Logique contemporaine et formalisation*, Paris, PUF, 1964.

MAUDLIN T., « Computation and consciousness », *Journal of philosophy*, tome 86, n° 8, 1989.

McINTYRE Jane, « Putnam's brains », *Analysis*, tome 44, n° 2, pp. 59-61, 1984.

MEYER Michel (dir.), *La philosophie anglo-saxonne,* Paris, PUF, 1994.

MICHIE D., *On machine intelligence*, New York, John Wiley and Sons, 1974.

- « Turing's test », *Artificial intelligence*, n° 1, 1993.

MILLICAN P. Millican et CLARK A. (éd.), *Machines and thought*, Oxford, Oxford University Press, 1996.

MOOR James H., « An analysis of the Turing test », *Philosophical studies*, n° 30, 1979.

MOSCONI Jean Mosconi, *La constitution de la théorie des automates*, thèse de doctorat soutenue à l'université de Paris I, 1989.

- « Sur quelques capacités et incapacités des machines », *Bulletin de la société française de philosophie*, 3, juillet-septembre 1991.

MOUNIN Georges, *Dictionnaire de linguistique*, Paris, PUF, 1993.

- *Introduction à la sémiologie*, Paris, Les Editions de Minuit, 1970.

MUMFORD Lewis, *Le mythe de la machine*, trad. de Léo Dilé, Paris, Fayard, 1973.

NAGEL Ernest, NEWMAN James R., GÖDEL Kurt, GIRARD Jean-Yves, *Le théorème de Gödel*, trad. de Jean-Baptiste Scherrer, Paris Editions du Seuil, 1989.

NAUR Peter, « Thinking and Turing's test », *Nordisk tidskrift for informations behandling*, tome 26, n° 2, 1986.

NEF Frédéric Nef, VERNANT Denis (éd.), *Le formalisme en question, le tournant des années 30*, Paris, Vrin, 1998.

NEWMAN M. H. A., *Alan Mathison Turing*, in *Biographical memoirs of the Royal Society*, 1955.

OSHERSON D., STOB M., WEINSTEIN S., *Systems that learn*, Cambridge, Massachussetts , MIT Press, 1986.

PARROCHIA Daniel, *Qu'est-ce que penser/calculer ?*, Paris, Vrin, 1992.

PETIT Jean-Luc (éd.), *Les neurosciences et la philosophie de l'action*, Paris, Vrin, 1997.

PIAGET Jean (dir.), *Logique et connaissance scientifique*, Paris, Gallimard (Encyclopédie de La Pléiade), 1967.

PICHOT André, *Histoire de la notion de vie*, Paris, Gallimard, 1993.

PINKAS Daniel, *La matérialité de l'esprit : la conscience, le langage et la machine dans les théories contemporaines de l'esprit*, Paris, La Découverte, 1995.

POST E. L., « Absolutely unsolvable problems and relatively undecidable propositions : account of an anticipation », *The undecidable : basic papers on undecidable. Propositions, unsolvable problems and computable functions*, Martin Davis éd., New-York, Raven Press, 1965.

- « Finite combinatory processes, formulation I », *The undecidable : basic papers on undecidable. Propositions, unsolvable problems and computable functions*, Martin Davis éd., New-York, Raven Press,1965.

POTDEVIN Gérard, *Logique et mathématique*, Paris, Editions Quintette, 1990.

PRATT Vernon, *Machines à penser. Une histoire de l'intelligence artificielle*, trad. de Christian Puech, Paris, PUF, 1995.

PUTNAM Hilary, « Brains and behavior », *Mind, language and reality*, London, Cambridge University Press, 1975.

- « Models and reality », *Journal of symbolic logic*, n° 45, 1980.

- *Raison, vérité et histoire*, Paris, Les éditions de Minuit, 1984.

RAMUNI J., *La physique du calcul ; histoire de l'ordinateur*, Paris, Hachette, 1989.

RENAUT Alain, *Kant aujourd'hui*, Paris, Aubier, 1997.

RIALLE Vincent, FISETTE Denis (éd.), *Penser l'esprit. Des sciences de la cognition à une philosophie cognitive*, Grenoble, Presses Universitaires de Grenoble, 1996.

REY Georges, « What's really going on in Searle's 'chinese room' », *Philosophical studies*, n° 50, pp. 169-185, 1986.

ROSSI Paolo, *Les philosophes et les machines, 1400-1700*, trad. de Patrick Vighetti, Paris, PUF, 1996.

ROTMAN Brian, *Ad infinitum – The ghost in Turing's machine*, Stanford, Stanford University Press, 1993

SCHMITZ François, *Wittgenstein, la philosophie et les mathématiques,* Paris, PUF, 1988.

SCHUHL Pierre-Maxime, *Machinisme et philosophie*, Paris, PUF, 1969.

SEARLE John, *Du cerveau au savoir*, Paris, Hermann, 1985.

- « Esprits, cerveaux et programmes », *Vues de l'Esprit*, Paris, InterEditions, 1987.

- « L'esprit est-il un programme d'ordinateur ? », *Pour la science*, 149, 1990.

SEBESTIK Jan, SOULEZ Antonie, *Le Cercle de Vienne : doctrines et controverses*, Paris, Méridiens-Klincksieck, 1986.

SÉRIS Jean-Pierre, « Machine et communication », Paris, Vrin, 1987.

SHAH T. K., « Minds and brains, algorithms and machines », *Boston studies in philosophy of science ; bridging the gap : philosophy, mathematics and physics*, n° 140, pp. 125-139, 1993.

SHANKER S.G., « Wittgenstein versus Turing on the nature of Church's thesis », *Notre-Dame journal of formal logic ; special issue on Chuch's Thesis*, tome 28, n° 4, pp. 615-649, 1987.

SHANNON Benny, « A simple comment regarding the Turing test », *Journal of theoretical Studies*, n° 19, 1989.

SINACEUR Hourya, *Jean Cavaillès : philosophie mathématique*, Paris, PUF, 1994.

SMITH Peter, « Could we be brains in a vat ? », *Canadian journal of philosophy*, tome 14, n° 1, pp. 115-123, 1984.

STEVENSON John S., « On the imitation game », *Philsophia*, n° 6, 1976.

TICHY Pavel, « Putnam on brains in a vat », *Philosophia*, tome 16, n° 2, pp. 137-174, 1986.

TIMOCZKO Thomas, « In defense of Putnam's brains », *Philosophical studies*, tome 57, n° 3, pp. 281-297, 1989.

TORRANCE Steve (ed.), « The Mind and the Machine : Philosophical Aspects of Artificial Intelligence », Chichester, Ellis Horwood Limited Publishers, 1984.

TURING Sara Stoney, *Alan M. Turing*, Cambridge, W. Heffer, 1959.

VAN KIRK Carol, « Kant and the problem of other minds », *Kant Studien*, tome XXX, n° 77, 1, 1986.

VERNANT Denis, *Introduction à la philosophie de la logique*, Bruxelles, Pierre Mardaga, Editeur, 1986.

- « L'intelligence de la machine et sa capacité dialogique », in Vincent Rialle et Denis Fisette (dir.), *Penser l'esprit : des sciences de la cognition à une philosophie cognitive*, Grenoble, PUG, 1996.

VIENNE Jean-Michel Vienne (éd.), *Philosophie analytique et histoire de la philosophie*, Paris, Vrin, 1997.

VIRIEUX-REYMOND A., *L'épistémologie*, Paris, PUF, 1966.

VON NEUMANN John, *Theory of self-reproducing automata*, Londres, University of Illinois Press, 1966.

WAGNER Pierre, *machine et pensée. L'importance philosophique de l'informatique et de l'intelligence artificielle*, thèse de philosophie soutenue à l'université de Paris I, 1994.

WANG Hao, *From mathematics to philosophy*, Cambridge, Massachusetts, MIT Press, 1974.

- *Reflections on Kurt Gödel*, Cambridge, Massachussetts, MIT Press, 1988.

WEBB Judson Chambers, *Mechanism, Mentalism, and Metamathematics*, Dordrecht, D. Reidel Publishing Company, 1980.

WEIZENBAUM J., *Computer power and human reason*, New-York, W.H. Freeman and Co, 1976.

WINOGRAD Terry, FLORES Fernando, *L'intelligence artificielle en question,* trad. de Jean-Louis Peytavin, Paris, PUF, 1986.

WITTGENSTEIN Ludwig, *Cours sur les fondements des mathématiques*, trad. de Elisabeth Rigal, Paris, Editions TER, 1995.

- *Tractatus logico-philosophicus, suivi de Investigations philosophiques*, trad. de Pierre Klossowski, Paris, Gallimard, 1961.

Table des matières analytique