Computing Machinery and Intelligence commentaire thématique

Bastien Guerry

26 janvier 2002

> « Il y avait donc toujours une méthode sous cette folie apparente. » [Hod88, p.46]
>
> Andrew Hodges

Table des matières

1 Introduction

L'article d'Alan Turing, *Computing Machinery and Intelligence*[1] est souvent cité comme étant l'article fondateur du projet de l'intelligence artificielle (IA). Et effectivement, c'est ici que la comparaison entre l'esprit et l'ordinateur passe du statut de simple métaphore à celui de paradigme : Turing construit un cadre conceptuel et expérimental à l'intérieur duquel la comparaison entre pensée et machine est susceptible de se faire *à un même niveau*. C'est l'unicité de ce niveau, établie par ce qui sera plus tard appelé le « test de Turing », qui est au centre de cet article : grâce à elle, Turing évite les pièges que nous tendent continuellement les termes équivoques de « pensée », « intelligence » et « machine ».

On a plus rarement souligné le paradoxe suivant[2] : c'est dans l'esprit du même chercheur que s'est élaborée une conception montrant les limites imposées à n'importe quelle *machine universelle*, et une conception montrant l'étendue des capacités dont on pourra doter ces machines universelles. Il n'y a là aucune contradiction : plus nous pénétrons dans le détail des deux articles, plus nous percevons l'unité de la méthode et des théories que Turing veut défendre. Plus nous maintiendrons la tension entre ces deux perspectives, plus nous serons capables de rendre justice à l'article qui nous occupe.

On connaît le sort des textes fondateurs : ils sont toujours cités, on se réfère constamment à leur autorité, ils forment un fonds commun dans lequel chacun puise ce qu'il entend. Mais à force de les mentionner, on oublie parfois de s'y ressourcer. Or, quand on entreprend l'effort d'analyser l'article de Turing, on rencontre au moins deux obstacles. D'une part, l'inertie historique de la métaphore de l'homme-machine tend à fausser la lecture. Tantôt elle nous incline à voir dans la position de Turing un simple avatar du mécanicisme, tantôt elle nous conduit à déformer l'idée que Turing devait se faire de l'intelligence. Or le concept de « machine » a bien évolué depuis La Mettrie (Turing est bien placé pour le savoir), ainsi que les différentes conceptions de l'intelligence. D'autre part, en baptisant Turing « père » de l'IA, on risque de juger son article à l'aune des succès et échecs de l'IA. Or, il n'est pas dit que l'IA se soit développée comme Turing l'imaginait : c'est certainement manquer la singularité du projet turingien que de l'assimiler rapidement aux diverses réalisations du domaine de l'IA. Notre analyse s'efforcera tant qu'elle pourra de contourner ces obstacles.

Il y a selon nous trois points essentiels pour lesquels une analyse de cet ar-

1. Références : [Tur50] pour le texte original; [Gir95] et [And83] pour les traductions françaises.

2. Signalé par Daniel Andler dans [And99, p.155].

ticle s'avère aujourd'hui indispensable. Concernant d'abord le « test de Turing » : la simplicité de cette expression nous cache la complexité du jeu de l'imitation. Nous tenterons de montrer que cette complexité n'est pas gratuite. Ensuite, nous pensons que Turing est trop vite attaqué sur ses positions béhavioristes et sur ses conceptions soi-disant naïves du langage. Enfin, nous analyserons dans le détail les sixième et huitième objections, car elle nous semblent relever toutes deux d'une même idée fondamentale, idée qui est toujours d'actualité pour défendre le point de vue de l'IA contre les fausses attaques.

2 Du jeu de l'imitation

2.1 Position du problème

L'expression de « test de Turing » ne se trouve pas dans l'article que nous étudions. Elle est beaucoup plus tardive[3]. Le succès de cette expression ne tient pas seulement à ce qu'elle rend hommage à l'inventeur du test, mais aussi à ce qu'elle parle facilement à l'imagination. Pourtant la représentation simple que suscite le terme de « test » nous cache la complexité du jeu de l'imitation. Elle nous fait croire qu'une machine est soumise à un test dont l'issue nous permettra de *mesurer* son intelligence, comme on soumet un enfant à un test spécifique pour mesurer son quotient intellectuel. Dans cette manière de voir, on fait comme si l'intelligence était quelque chose à prouver - ou pire : à trouver. Si la perspective de Turing correspondait à cette façon de poser le problème, on ne comprendrait pas pourquoi le jeu se joue entre trois joueurs.

On pourrait en effet concevoir un test plus simple, lequel ferait communiquer un interrogateur avec un seul agent : le but de l'interrogateur serait de deviner s'il a affaire à une machine ou à un être humain, et le but de la machine serait de se faire passer pour un être humain. En ajoutant des contraintes sur la durée du test et sur le quotient intellectuel de l'interrogateur, on mesurerait alors la capacité d'une machine à se faire passer pour un être humain dans un dialogue désincarné. Et il serait ensuite possible d'arguer que, puisque l'intelligence se manifeste essentiellement dans le comportement verbal d'un être humain, une machine passant le test avec succès peut légitimement être dite « intelligente ».

Cette manière de procéder soulèverait au moins deux objections. On dira d'abord qu'*imiter* n'est pas *être*, qu'il y a un seuil infranchissable entre la réalité et sa simulation. De quel droit déformerions-nous la signification des mots au point de

3. Jean Lassègue situe son apparition au milieu des années 70' : voir [Las96].

dire qu'une machine passant le test avec succès *est intelligente* ? On dira ensuite que ce test s'appuie sur une vision réductrice de l'intelligence : le plus important est de trouver une bonne définition de l'intelligence, puis une bonne mesure de cette capacité, ensuite seulement serait-il légitime de soumettre une machine à ce test.

Or la démarche de Turing est différente : il ne cherche pas à montrer que la machine fait preuve d'intelligence (cela supposerait donnée une bonne définition de l'intelligence), mais il monte un protocole expérimental à la suite duquel il n'y aura aucune raison de refuser d'utiliser le mot « intelligence » pour qualifier ce que fait la machine (ce qui ne remet pas en cause nos différents usages actuels du mot « intelligence »).

Dans ce qui suit, nous voudrions montrer que les complications introduites par Turing dans son exposition du jeu de l'imitation ne sont pas gratuites, et qu'elles représentent chacune une manière de répondre aux deux objections présentées ci-dessus.

2.2 Une femme entre deux jeux

Qu'est-ce que le jeu de l'imitation ? Le jeu de l'imitation se joue à trois. Dans un premier temps, les trois joueurs sont un homme A, une femme B et un interrogateur C. Le but du jeu pour l'interrogateur est de savoir qui est la femme. Le but du jeu pour l'homme est de faire croire qu'il est la femme. Le but du jeu pour la femme est aussi de faire croire qu'elle est la femme[4]. A et B ne peuvent pas communiquer entre eux, mais ils communiquent tous les deux avec C, par un canal ne laissant passer rien d'autre qu'une information écrite. Dans un deuxième temps, l'homme A est remplacé par une machine. La question décisive est : « l'interrogateur se trompera-t-il aussi souvent que lorsque le jeu se déroule entre un homme et une femme ? » [Gir95, p.136]

Levons tout de suite une ambiguïté. La question posée signifie bien : est-ce que la machine sera aussi capable que l'homme de se faire passer pour une *femme* aux yeux de l'interrogateur ? Ce n'est que par une généralisation de cette question qu'on en vient ensuite - mais la transition reste implicite dans l'article - à se demander : pour un prédicat quelconque P, autre que le prédicat *être-une-femme*,

4. Nous aurions envie de dire que le but du jeu pour la femme est de *convaincre* l'interrogateur qu'elle est *réellement* la femme, mais cette dernière formulation introduit des considérations psychologiques qui privilégient le point de vue de la femme (certainement agacée de voir sa féminité remise en cause); ces considérations doivent être écartées si nous voulons une description objective du jeu de l'imitation.

est-ce qu'une machine saura se comporter de manière à faire croire à C que le prédicat P lui revient réellement ? Nous soulignons ce point, car les analyses du jeu de l'imitation passent souvent très vite à sa version généralisée, n'expliquant pas ce que le choix du prédicat *être-une-femme* apporte à l'argument de Turing. La description ci-dessus nous oblige donc à rendre compte de deux traits caractéristiques : ce choix particulier de la féminité comme prédicat à imiter; la distinction de deux moments à l'intérieur du jeu.

Jean Lassègue propose une thèse intéressante pour expliquer le lien qu'il y a entre les deux moments du jeu[5]. Dans le premier, la différence homme-femme serait mise en scène comme étant la différence physique fondamentale au sein du genre humain. Un échec répété de l'interrogateur montrerait que cette différence n'a aucune pertinence dans la comparaison de l'intelligence des joueurs. Dans le second, l'opposition physique entre un être humain et une machine serait encore plus grande[6]. Mais ici encore, un échec répété de l'interrogateur montrerait que cette différence physique n'a aucune incidence sur l'évaluation des intelligences, de même que la différence des sexes n'en avait aucune dans le premier moment du jeu. Le mouvement général du jeu consisterait donc à s'abstraire progressivement de la spécificité des différents substrats physiques pour montrer que la pensée et l'intelligence en sont indépendantes.

Nous nous accordons à cette interprétation en ceci qu'elle rend bien l'esprit du jeu. Elle met en relief un lien logique (et pas seulement pédagogique) entre les deux moments du jeu. Mais elle ne rend pas tout à fait justice à la lettre du jeu. Car cette interprétation, supposant que le second moment du jeu est une comparaison entre une machine et un *être humain*, ne nous dit pas pourquoi Turing a choisit de remplacer l'homme A, plutôt que la femme B. Caprice ? Favoritisme ? La lecture que nous proposons va au contraire permettre de répondre à ces trois questions : pourquoi est-ce l'homme qui imite la femme (et non l'inverse) dans le premier moment du jeu ? Pourquoi l'homme est-il remplacé (plutôt que la femme) dans le passage au second moment ? A quoi sert le premier jeu, relativement à la généralisation qu'on en tire après ?

La remarque clef est d'ordre linguistique : le prédicat *être-un-homme* est d'un usage plus équivoque que le prédicat *être-une-femme*. « Equivoque » signifie ici que le prédicat *homme* peut qualifier en des sens différents le même individu : dire de quelqu'un qu'il est un homme nous informe soit au sujet de l'espèce à laquelle

5. Il considère même ces deux moments comme constituant deux jeux distincts : cf. [Las98, p.149].

6. Jean Lassègue parle d'une opposition entre « espèces ».

il appartient, soit au sujet de son identité sexuelle. A l'inverse, dire d'un prédicat qu'il est « univoque » signifiera qu'il peut se dire en un même sens d'individus différents. L'attribut *féminité* est univoque en ce qu'il peut qualifier un homme aussi bien qu'une femme. Ces remarques sont passablement triviales. Mais le but de Turing étant de poser la question de l'intelligence des machines « en des termes relativement non-ambigus » [And83, p39], il n'est pas inutile de s'interroger sur l'équivocité des prédicats invoqués dans le jeu de l'imitation.

Cette asymétrie entre l'usage des deux prédicats *homme* et *femme* a une conséquence directe : il est plus difficile de se faire passer pour une femme que de se faire passer pour un homme, car il est plus facile de jouer sur l'équivocité du prédicat *homme*. Il nous est maintenant possible de répondre à notre première question. Car s'il est plus difficile pour un homme d'imiter une femme que pour une femme d'imiter un homme[7], alors il est *a fortiori* encore plus difficile pour une machine de se faire passer pour une femme que pour un homme. Si Turing donne à l'homme le but d'imiter la femme, ce n'est pas en raison de quelque singularité sexuelle propre à Turing, mais en raison de la plus grande difficulté de ce but (relativement à celui, pour une femme, de faire croire qu'elle possède le prédicat *homme*), difficulté qui ne tient pas à des considérations psychologiques, mais à des considérations linguistiques sur les usages plus ou moins équivoques des prédicats *homme* et *femme*.

Que se passe-t-il avec l'intervention de la machine ? D'après ce que nous avons dit précédemment, une machine capable de rivaliser d'intelligence avec un homme dans la tentative de se faire passer pour une femme est plus intelligente qu'une machine rivalisant d'intelligence avec une femme dans la tentative de se faire passer pour un homme. Le choix du prédicat *femme* comme prédicat à imiter est donc justifié comme choix du prédicat le plus difficilement imitable. Ce qui répond bien à notre deuxième question.

Enfin, cette interprétation renforce la solidarité argumentative des deux moments du jeu : car si la machine est assez intelligente pour se faire passer pour une femme, et étant donné qu'il n'y a aucune particularité de l'intelligence des femmes qui puissent les empêcher de résoudre un problème quelconque, alors la machine est-elle *a fortiori* capable de résoudre n'importe quel problème. La mise en scène d'une femme dans le jeu de l'imitation n'est pas l'occasion de faire des conjectures sur les ambiguïtés sexuelles de Turing, mais l'occasion pour Turing d'introduire un prédicat à la fois univoque et général, de manière à placer le plus haut possible le défi que la machine doit relever.

7. Ceci, bien sûr, dans les conditions imposées par le jeu de l'imitation.

2.3 Pourquoi trois joueurs ?

Nous avons rendu compte du rôle joué par la femme dans le jeu de l'imitation, mais nous n'avons pas encore dit pourquoi le jeu de l'imitation a besoin de faire interagir *trois* joueurs.

Avant de nous plonger dans cette dimension du jeu, une simple remarque méthodologique. Ce n'est pas seulement l'expression de « test de Turing » qui nous donne une vision réductrice du jeu de l'imitation, mais encore la manière dont cette expression est souvent invoquée : on fait comme si l'on savait déjà de quoi il s'agissait, comme s'il n'était pas besoin de jouer pour savoir ce que le jeu met précisément en jeu. Or cette façon de procéder nous condamne à ne pas comprendre l'esprit du jeu de l'imitation, car le propre de la démarche de Turing dans tous ces travaux est de n'aborder la question *Comment ça marche ?* qu'au moyen d'une mise en œuvre *effective* de ce dont on cherche à comprendre le mécanisme. Ainsi, la question « Qu'est-ce qu'un calcul ? » est remplacée par la question « Comment fonctionne l'esprit qui calcule, la machine qui calcule ? ». Notre article transforme la question de savoir ce qu'*est* l'intelligence par le problème de savoir ce qui est mis en œuvre dans un comportement dit « intelligent ». De même, nous nous interdisons de comprendre ce qu'*est* le « test de Turing » si nous ne nous demandons pas d'abord « Comment marche ce jeu de l'imitation ? ». Un seul mot d'ordre méthodologique : jouons!

Pourquoi faut-il trois agents ? Pourquoi le jeu ne consisterait-il pas à mettre un interrogateur en face d'une seule entité, lui permettre de poser toutes les questions qu'il souhaite à cette entité, puis de lui demander s'il s'agit d'un homme ou d'une femme ? La situation du jeu à trois implique une compétition entre A et B. Cette compétition oblige A et B à imiter du mieux qu'ils le peuvent ce que c'est qu'être une femme. On remarque déjà ici que la femme est obligée d'objectiver pour elle-même les traits caractéristiques de la féminité, de manière à pouvoir convaincre l'interrogateur. Elle ne peut pas se contenter d'*être* une femme, elle doit aussi posséder une théorie de la féminité, construire un modèle stéréotypique de femme, afin d'imiter ce modèle en retour. A ce seul niveau, les frontières de la réalité et de la simulation sont amenées à se confondre : dans ce jeu, l'avantage n'est pas donné à l'agent qui est *réellement* une femme, mais à celui qui possède un modèle *plausible* de la féminité.

Franchissons encore un pas : pour gagner, la femme ne doit pas seulement avoir une théorie de la féminité, elle doit en plus tenter de s'imaginer quelle est la théorie de la féminité que l'interrogateur a en tête au moment où il lui pose une question. Autrement dit, elle doit former une théorie de la théorie de la féminité

qu'est susceptible de possèder l'interrogateur. Et elle doit se servir des conjectures qu'elle fait sur le modèle féminin que C possède pour rendre ses réponses les moins équivoques possible. Arrivé à ce point, on voit bien que ce n'est pas la réalité qui importe, mais les modèles de féminité qu'ont à l'esprit les trois agents. Le jeu ne consiste pas en une évaluation de la réalité, mais en une évaluation des modèles que l'interrogateur construit, de manière à répondre de manière pertinente aux questions qu'il pose. A la fin du jeu, la réalité reprendra le dessus, l'interrogateur aura perdu ou gagné. Mais *pendant le jeu*, seule importe la capacité des joueurs à construire des modèles plausibles de féminité et, pour A et B, à donner des réponses telles que le modèle de féminité que l'interrogateur pourra en inférer corresponde bien au modèle de féminité qu'il possède lui-même.

Mais ce n'est pas tout. Le femme sait que l'interrogateur peut comparer le modèle de féminité qui se reflète dans ses réponses au modèle de féminité qui se reflète dans les réponses de l'homme. Soit $T_X(P)$ la théorie T qu'a le joueur X du prédicat P. Soit F le prédicat *être-une-femme*. La femme sait non seulement que l'interrogateur évalue le rapport $\frac{T_B(F)}{T_C(F)}$, mais elle sait en plus qu'il évalue le rapport des deux rapports, qu'il teste l'égalité suivante :

$$\frac{T_A(F)}{T_C(F)} = \frac{T_B(F)}{T_C(F)}$$

La femme ne doit pas seulement chercher quel est le modèle de femme de l'interrogateur (tel, du moins, que ses questions le laisse supposer), mais elle doit en plus chercher à rendre le modèle de femme qu'elle exhibe plus plausible que n'importe quel autre modèle susceptible d'être proposé par l'homme. C'est ici que la différence avec une situation à deux joueurs devient essentielle. Nous sommes dans un jeu à information doublement partielle[8]. A et B connaissent les questions que leur pose l'interrogateur, mais ils ne connaissent pas $T_C(F)$. De plus, A et B connaissent chacun l'identité de l'autre, mais chacun ignore les réponses que l'autre soumet à C. Ce jeu de chassé-croisé entre ce qui est su et ignoré de part et d'autre permet une démultiplication indéfinie des stratégies que pourront mettre en œuvre les trois joueurs.

Prenons un exemple. Soit G le prédicat *savoir-jouer-au-Go*[9]. L'interrogateur mène deux parties simultanées contre les deux joueurs. Comment va-t-il s'y prendre

8. L'exclusion de l'argument de la perception extra-sensorielle rend l'information incomplète et l'exclusion du solipsisme exclus l'hypothèse de nullité de l'information : ce qui est transmis par l'écriture nous donne une information partielle mais *objective* concernant l'intelligence de l'agent.

9. Turing avait une prédilection marquée pour les échecs, mais il accordait aussi ses faveurs au jeu de Go.

pour savoir lequel est l'ordinateur ? Il va d'abord se faire une idée de ce que c'est, pour un homme, que de jouer au Go. Il aura aussi probablement quelques idées sur la manière dont les ordinateurs jouent au Go, sur ce qu'ils savent anticiper et ce qu'ils ne savent pas prévoir. L'ordinateur, de son côté, ne s'entêtera pas essentiellement à gagner, mais à jouer de telle manière qu'il paraisse être un homme. Il est donc possible de lui faire faire des erreurs dans certains joseki[10], ou de lui faire jouer des coups tellement bizarres qu'une appréhension seulement psychologique de ces coups les jugera audacieux ! L'homme, de son côté, ne cherchera pas à jouer tout simplement de son mieux, mais à trouver les coups décisifs, ceux que l'interrogateur ne pourrait pas attribuer à une machine. La recherche de ces coups implique que l'homme se fasse une idée de la manière dont l'ordinateur joue, en plus d'une idée de ce que l'interrogateur attend de lui. Ainsi, on voit comment les stratégies peuvent se compliquer au fur et à mesure que le jeu avance, selon la profondeur des théories de l'esprit qu'on suppose chez chaque joueur (« je sais qu'il sait que je sais... »), et selon la profondeur des inférences qu'il est capable d'en tirer (« je présume qu'il suppose que je crois... »).

Avec seulement deux agents, la dynamique des théories de l'esprit impliquées dans les différentes stratégies est nettement moins forte. Avec trois agents, la complexité explose. De même qu'il suffit d'introduire un troisième corps dans un système dynamique simple pour que « déterminisme » ne soit plus synonyme de « prédictibilité », de même l'introduction d'un troisième joueur rend-elle imprédictible le jugement de l'interrogateur. Remarquons que la décision de C quant à l'identité de chaque personne peut se formuler comme un problème d'arrêt : l'interrogateur pourra décider pour sa part que, si la différence entre $\frac{T_A(F)}{T_C(F)}$ et $\frac{T_B(F)}{T_C(F)}$ franchit un certain seuil S, alors il décidera d'attribuer telle identité à A et à B. Si ce seuil est franchi avant la fin d'une durée déterminée (cinq minutes, par exemple) on dira que l'interrogateur a gagné, sinon il aura perdu. Existe-t-il une machine de Turing telle que, si on lui donne en entrée les différentes théories que se font les trois joueurs d'un prédicat donné, elle nous permette de décider *en un temps fini* que l'interrogateur réussira à lever le défi ? La réponse de Turing est non, et formulée ainsi, elle acquiert un pouvoir de conviction énorme.

Que se passe-t-il lorsqu'on généralise le jeu à n'importe quel prédicat ? La pire stratégie pour chaque joueur tentant de faire croire qu'il possède un prédicat P (*être-une-femme*, *savoir-jouer-au-Go*), serait d'avoir une théorie essentialiste de ce prédicat. Si l'homme ou la femme se réfère à un modèle platonicien de femme

10. Joseki : correspond à des modèles de début de partie équitables pour les deux joueurs, séquences souvent jouées de manière automatiques. Il y a à peu près dix mille joseki répertoriés.

dans la première partie du jeu, alors ils se condamnent à ne pas pouvoir adapter leur modèle de façon à le rendre plausible aux yeux de l'interrogateur. Le modèle le meilleur sera celui qui parvient à se fondre dans la dynamique des conjectures plutôt que celui qui reste impassible dans son immobilité[11].

Considérons maintenant le prédicat *être-intelligent* dans sa généralité. D'après ce que nous venons de dire, la machine qui réussira le mieux à faire croire qu'elle possède ce prédicat sera celle qui aura une théorie la moins essentialiste de ce qu'est l'intelligence... autrement dit : celle qui saura exploiter l'équivocité dont nous faisons spontanément preuve dans notre usage du terme « intelligence »! Nous trouvons ici la réponse à la deuxième objection que soulève spontanément un simple « test de Turing » à deux joueurs : cette objection disqualifiait le test sous prétexte qu'il ne reposait sur aucune définition stable de l'intelligence. Mais la démarche de Turing ne consiste pas à supposer une définition (forcément réductrice) de l'intelligence pour ensuite dire que la classe ainsi définie comprend certaines machines; il part au contraire de l'usage normal que nous faisons du terme et construit une expérience dans laquelle nous ferons certainement usage de ce terme comme prédicat d'une machine. Et le fait que cette expérience implique trois joueurs plutôt que deux permet à l'expérience de faire de l'instabilité de notre usage du prédicat *être-intelligent* un moyen pour la machine de « s'approprier » ce prédicat à notre insu.

3 Du béhaviorisme et du langage

Dans tout ce qui a précédé, nous avons déjà introduit quelques considérations linguistiques. Nous voudrions maintenant examiner la question d'un point de vue plus général : quel idée Turing se faisait-il du langage ? Comment ses conceptions peuvent-elles nous permettre de mieux comprendre la portée de l'article ? Mais avant d'aborder la question du langage, nous nous demanderons en quel sens le point de vue de Turing se rapproche du béhaviorisme, car c'est souvent dans un même geste qu'on se moque de la naïveté de Turing à l'égard du langage et qu'on fustige ses « préjugés » béhavioristes.

11. Pourquoi Turing attribue-t-il alors une stratégie parfaitement naïve à la femme ? Tout simplement parce que, étant donné la difficulté de s'adapter aux différentes conjectures envisagées sans entrer en contradiction avec soi-même, il est certainement raisonnable pour elle de faire le pari que le vrai paraîtra plus vraisemblable. Mais ce n'est qu'un pari.

3.1 Point de vue traditionnel

Turing est l'héritier direct du positivisme logique des années trente. Géographiquement et idéologiquement, on ne peut le comprendre parfaitement si on ignore que ses vues s'inspirent largement de ce courant. Ce qu'il adopte en particulier, c'est une exigence de scientificité qui tient en trois préceptes : poser les problèmes dans un contexte où ils peuvent être falsifiés (exigence popperienne); évacuer tout psychologisme, notamment dans les questions de signification (exigence frégéenne, puis wittgensteinienne); comprendre la nature humaine en fonction de lois du comportement, plutôt qu'en fonction de principes non observables (exigence béhavioriste[12]). Il n'a de cesse de conserver ces exigences, qu'il s'occupe de mathématiques, de philosophie ou de morphogénèse.

Aussi trouve-t-on parfois des critiques de la portée du « test de Turing » s'appuyant sur l'attachement de Turing à ces exigences, notamment sur son attachement au béhaviorisme et aux vues réductrices de celui-ci sur le langage. La disqualification massive dont a été victime le béhaviorisme suffirait à rejeter l'opérationalisme de Turing, et la « découverte » des problèmes posés par la construction d'une théorie valable du langage contrasterait absolument avec la naïveté avec laquelle Turing envisage cette question.

Certes, l'optimisme de Turing quant à la manière dont pourront être levés les obstacles liés au langage prête aujourd'hui à sourire. Mais il faudrait distinguer ici entre deux catégories d'objections faites à Turing. La première est d'ordre pratique : elle émerge en même temps qu'on se heurte à des difficultés dans la mise en œuvre effective du programme dont il pose les fondements. Les échecs successifs de la construction d'une machine passant avec succès le test de Turing, dans les conditions linguistiques que celui-ci évoque dans l'article de 1950, montrerait combien la conception que Turing avait du langage et des difficultés de son implémentation sur une machine était rudimentaire. La deuxième est d'ordre théorique : elle refuse les bases argumentatives de Turing en réfutant le point de vue béhavioriste. L'argument rapidement reconstruit est : puisque Turing endosse le paradigme béhavioriste et puisque ce paradigme a été mis en défaut depuis bien longtemps, alors les suppositions de Turing n'ont plus de fondement solide, spécialement celles qui ont trait au langage.

12. Bien sûr, cette exigence n'est pas propre au béhaviorisme, mais c'est néanmoins ce courant de la psychologie qui l'incarne le mieux à l'époque de Turing.

3.2 Portée des objections d'ordre pratique

Pour répondre au premier type d'objections, on peut se montrer aussi généreux avec Turing qu'il se montre condescendant avec Charles Babbage et lady Lovelace, et dire qu'« [Il] n'étai[t] pas dans l'obligation d'avancer tout ce qu'il y avait à avancer. » [Gir95, p.160] Car nous devons soigneusement éviter toute illusion rétrospective, laquelle reviendrait à mesurer la consistance du projet de Turing à l'aune des résistances rencontrées en IA. Cette attitude est encore aujourd'hui largement partagée, et elle mène souvent à deux erreurs : d'une part, prenant l'ensemble du projet de l'IA comme un nouvel avatar des prétentions prométhéennes de l'homme, elle incline à ne voir dans l'IA qu'un vaste champ de ruines et de faux espoirs[13] ; d'autre part, elle suggère une forme de responsabilité de Turing par rapport aux différentes tentatives qui seront faites pour donner corps à son idée. Mais s'il est bien l'inspirateur essentiel de l'IA, celui dont les vues paradoxales et l'audace théorique seront une source intarissable d'énergie pour les chercheurs à venir, il n'en est pourtant pas le porteur immédiat : ses épaules son libres de la charge qui pèsera ensuite sur des gens comme H. Simon et A. Newell. Ce que fait Turing, c'est de justifier et de défendre le bien-fondé d'une *hypothèse*, non d'un fait. Et comme il le dit lui-même : « Pourvu que nous sachions clairement quels sont les faits prouvés et quelles sont les hypothèses, aucun mal ne peut en résulter. [Gir95, p.143] »

Enfin, l'optimisme de Turing peut aussi se comprendre d'un point de vue psychologique : d'abord il a assisté à l'émergence très rapide des inventions liées aux ordinateurs, puis il a lui-même contribué à la résolution de problèmes pratiques de programmation, le tout lui donnant certainement de bonnes raisons (subjectives) de croire ses prédictions raisonnables. L'idée prêtant aujourd'hui à rire selon laquelle « une soixantaine de personnes travaillant assidûment pendant cinquante ans pourraient accomplir le travail [Gir95, p.168] » devient peut-être moins délirante si l'on imagine que Turing se retient de dire : « une soixantaine de personnes *aussi douées que moi* »! Le fait que Turing soit tenu par le secret militaire concernant ses recherches en cryptographie explique aussi sa légère retenue chaque fois qu'il fait mention de son expérience technique : il donne bien des descriptions détaillées des ordinateurs dont il parle, mais lorsqu'il s'évoque en train de les manipuler, les détails se mettent à manquer. Mais ceci n'est pas à mettre au compte

13. C'est cette impression que peut laisser la lecture d'un livre comme celui de Hubert Dreyfus [Dre84]. Or c'est une faute logique que d'invalider la portée théorique des propositions de Turing en s'appuyant sur les obstacles techniques rencontrés par ses successeurs, faute que l'auteur est souvent en passe de commettre.

d'un quelconque « bluff » rhétorique, ni d'une impasse sur les difficultés techniques insolubles; seulement au fait que Turing ne peut pas tout dire. Si on ignore cette contrainte du secret, on risque de ne pas comprendre toutes les raisons subjectives qu'il a d'être optimiste quand à la faisabilité de son projet.

Dans tous les cas, on ne saurait attribuer sa naïveté (nous aimerions dire son « innocence ») à un manque de lucidité technique sur ce qui était en jeu. Tous ces arguments d'ordre pratique ont le même défaut d'être *a posteriori*, de sortir le texte de son contexte, de prendre prétexte de ce que Turing ne pouvait pas prévoir pour réduire ses vues à néant.

3.3 Portée des objections d'ordre théorique

Pour répondre au deuxième type d'objections, on peut invoquer deux arguments. Premièrement, les objections valant contre le béhaviorisme ne valent pas toutes contre l'opérationalisme. Deuxièmement, la conception que Turing a du langage ne se laisse pas nécessairement réduire aux thèses béhavioristes.

Béhaviorisme et opérationalisme

Par *opérationalisme* nous entendons la théorie qui considère un état mental comme une opération effectuée sur des informations. L'approche turinginenne de l'opérationalisme se rapproche du béhaviorisme en ce que, si chaque état mental est le résultat d'un calcul, alors pointe à l'horizon l'espoir d'une description possible des lois de comportement de la pensée. Mais cet opérationalisme ne se confond pas en tout point avec le béhaviorisme; sur de nombreuses questions, il s'accommode plus facilement des obstacles que rencontre le béhaviorisme.

Soit l'argument de la pauvreté du stimulus : d'après cet argument, le béhaviorisme est incapable d'expliquer comment sont acquises les compétences liées au langage, car les stimuli ne permettent pas à eux seuls d'expliquer la structuration complexe de ces compétences[14]. Les béhavioristes seraient en face du langage comme un biologiste qui voudrait rendre compte de la structure d'un organe en ne faisant aucune hypothèse sur sa préformation, en faisant tout émerger des seules interactions de l'organe avec l'environnement.

L'opérationalisme ne prête pas un flanc aussi fragile à l'argument de la pauvreté du stimulus : car plutôt que de considérer un individu comme une boîte noire, il entre à l'intérieur de cette boîte et se demande quelles sont les opérations que

14. Voir la critique de Skinner par Chomsky pour la forme canonique de cet argument.

l'on doit supposer pour comprendre les lois de son comportement. L'opérationaliste serait alors comme un biologiste qui dissèque l'organe et se pose la question : « comment agit et réagit tel élément de l'organe ? » Du point de vue de la méthode, la démarche générale reste à peu près la même, mais du point de vue des moyens mis en œuvre dans l'analyse des lois du comportement, la différence est de taille. Car si le béhaviorisme a une vision plutôt statique des lois du comportement, l'opérationalisme en a une vision très dynamique : le concept de boîte noire est conservé - en vertu de son importance méthodologique - mais on peut démultiplier indéfiniment les relations d'emboîtement.

Dans sa description des machines qui apprennent, Turing invoque constamment la notion béhavioriste d'apprentissage par renforcement. Mais il prend aussi un soin tout particulier à exposer les différents niveaux d'organisation de la machine : un niveau matériel (qu'il décrit précisément dans la troisième section de son article); un niveau logique[15]; un niveau sémantique (la représentation, sous forme symbolique, de propositions)[16] . Pour Turing, cette distinction en différents niveaux est ce qui rend la machine semblable à l'être humain du point de vue d'un apprentissage possible [Gir95, p.168]. Mais d'une manière générale, on peut suggérer que la différence entre ces niveaux d'organisation est l'élément conceptuel censé dissoudre l'argument par la pauvreté du stimulus. Car un stimulus n'est pauvre que relativement à la structure avec laquelle il réagit : en se donnant la possibilité de structures complexes, l'opérationalisme se donne ainsi les moyens de comprendre comment des compétences complexes (notamment le langage) peuvent être apprises.

L'idée fondamentale qui servira à formuler l'hypothèse opérationaliste était déjà dans l'article de 1936[17] : la distinction du ruban et de la table d'instructions d'une machine de Turing permet de dissocier son état structurel et son état logique. La possibilité de coder la table d'instructions d'une machine de manière à en faire l'état structurel d'une autre machine (opération analogue à la gödélisation des propositions de l'arithmétique dans le langage de l'arithmétique chez Gödel [NNGG89]) permet de prendre une opération comme objet d'une autre opération. C'est cette possibilité qui donne aux perspectives opérationalistes une complexité que n'atteint pas le béhaviorisme. Dès lors qu'une opération peut de-

15. Voir [Gir95, cf. note b p.171], où Turing précise que le niveau logique ne devra pas être appris.

16. On remarque ici que, selon la terminologie de J. R. Searle, la distinction de ces différents place Turing parmi les défenseurs d'une IA « faible », plutôt que « forte ». Voir [Gan90] pour une critique de cette bipartition artificielle.

17. [Tur36] pour le texte original; [Gir95] pour la traduction française.

venir objet d'une autre opération, les différentes lois de comportement d'un agent sont susceptibles de s'organiser hiérarchiquement, et cette structure hiérarchique va donner un rôle différent au couple stimulus-réponse selon le niveau d'organisation auquel on se trouve.

D'un manière générale, Turing ne soutient jamais que l'organisation d'un individu est déductible de ses seules interactions avec l'environnement : il accepte, pour l'homme comme pour la machine, l'idée d'une structure initiale [Gir95, p.168] déterminant la structuration progressive de l'individu. Mais il suppose aussi qu'on peut éduquer cet individu, lui donner une structure particulière en le soumettant à des stimuli appropriés aux réponses qu'on en veut obtenir. Et de la pauvreté explicative du couple *stimulus-réponse* chez le béhavioriste, on ne peut conclure à la pauvreté explicative du couple *input-output* de Turing, car la prise en compte des différents niveaux d'organisation donne une richesse fonctionnelle supplémentaire aux relations *input-output*.

Comprenons bien : nous ne voulons pas dire que Turing n'était pas béhavioriste, nous voulons seulement mettre en garde contre la facilité avec laquelle on réduit ses vues aux vues béhavioristes pour faire qu'un rejet de celles-ci soient un argument valable contre celles-là.

L'engagement pratique du langage

Etant donné que les critiques relatives au béhaviorisme de Turing sont souvent des critiques faites à sa conception du langage, il n'est pas inutile de se demander quelle conception linguistique soutient l'hypothèse et le projet de Turing.

Relativement à la question du langage, la critique spontanée du projet turingien consiste à dire qu'on a toujours pas réussi à programmer une machine qui puisse comprendre un texte et émettre des propositions sensées. D'une manière générale, cela revient à affirmer :

1. que l'on n'a pas levé les obstacles techniques relatifs à la programmation d'une telle machine;
2. que l'on n'a pas réussi à formaliser le langage naturel de manière à en implémenter un modèle efficace sur une machine;
3. qu'une machine ne produit jamais de « sens ».

Toutes ces affirmations sont, dans une certaine mesure, fausses. Mais surtout, elles ne constituent aucunement une critique valable des conceptions linguistiques de Turing. La première est une objection d'ordre pratique, elle n'invalide pas les théories de Turing sur le langage. La deuxième est d'ordre théorique, mais le

problème qu'elle soulève concerne le linguiste, non le théoricien se demandant s'il y a un sens à affirmer qu'une machine est intelligente. La dernière affirmation - celle qui retient l'attention de tous les philosophes - est d'ordre métaphysique. Turing n'aurait certainement pas reculé devant ce terme de « métaphysique », l'objet de son test étant justement de se placer hors de toute hypothèse spéculative sur la pensée (ni solipsisme, ni omniscience) et sur le langage (problème de la référence, problème de l'intentionalité).

Une critique plus réfléchie de la manière dont Turing met le langage au service de son argumentation serait d'accuser Turing de réduire la signification d'un mot (notamment celui, ici, d'intelligence) aux différents usages de ce mot. D'après cette perspective, voici quelle description rendrait compte de la démarche générale de Turing :

1. Le domaine d'application du prédicat *être-intelligent* se réduit à une classe de comportements observables;
2. La notion d'intelligence se réduit aux différents usages que nous faisons du mot « intelligence »;
3. Donc l'usage que nous ferons du mot « intelligence » pour qualifier le comportement observable de la machine pendant le jeu de l'imitation doit nous faire admettre que la machine *est* intelligente (puisque nous ne nous y prenons pas autrement pour dire que quelqu'un *est* intelligent).

Les critiques ne porteraient pas sur la conclusion, mais sur les deux prémisses du raisonnement. Pour la première, on soutiendra d'abord que cette définition est logiquement incohérente, car elle fait comme s'il n'y avait qu'une seule classe de comportements intelligents, alors qu'il y a plusieurs classes différentes d'intelligences[18]. On pourra ensuite arguer que cette définition de l'intelligence est empiriquement inadéquate, car ce que l'on entend habituellement par « intelligent » ne se rapporte pas à des comportements observables, mais à des qualités mentales de l'individu, des dispositions intrinsèques qui ne donnent signe extérieur de leur présence que de manière accidentelle.

Pour la deuxième prémisse, on pourra critiquer Turing en disant d'une part qu'il y a toujours plus de signification dans une notion que ce que ses usages ne le laissent supposer; et d'autre part, en admettant que cette prémisse puisse être théoriquement soutenue, elle pose néanmoins le problème pratique de savoir comment déterminer l'*ensemble* des usages d'une notion. A défaut de cette détermination

18. Voir par exemple le point de vue que soutient Howard Gardner dans [Gar97].

exacte, la définition n'est pas clause : or d'autres usages du mot « intelligent » seront toujours possibles qui devraient conséquemment modifier la signification de ce terme. Ce qui rend donc impossible la tâche de dire, de manière stable, qu'une machine est intelligente.

Nous défendrons Turing en montrant que la description faite ci-dessus de son raisonnement, quoique séduisante, reste schématique. Relativement à la première prémisse, rappelons que Turing ne définit pas l'intelligence. Son test n'est pas une manière de partir des préjugés que nous avons sur l'intelligence humaine pour montrer que ces préjugés qualifieraient adéquatement le comportement d'une machine. Le jeu de l'imitation est justement construit pour déjouer les présupposés que nous avons concernant la notion d'intelligence : il est destiné à nous placer dans une situation telle que nous serons obligés de faire comme si la machine était intelligente, donc dans une situation où l'usage de ce terme sera parfaitement justifié. Donc Turing n'est pas tenu d'avoir une définition logiquement cohérente de l'intelligence (première partie de la première critique), ni de déterminer effectivement quels sont tous les usages effectifs du terme « intelligent » (deuxième partie de la deuxième critique).

Pour ce qui est de la deuxième prémisse, nous dirons simplement que comprendre la signification d'un mot à partir de ses différents usages n'est pas *réduire* la sémantique à la pragmatique. Dire que nous n'avons pas d'autre information objective sur la signification d'un terme que celles qui nous sont données par ses différents usages, et dire que le sens d'un mot est tout entier décrit par ses multiples emplois sont deux choses différentes. Tout ce qui se rapporte à la profondeur subjective d'une notion relève d'une discussion métaphysique, du genre de celles que Turing veut explicitement éviter en répondant à l'argument issu de la conscience. Donc il n'y a aucun sens ni à tenir rigueur à Turing d'une inadéquation empirique de son approche de l'intelligence (deuxième partie de la première critique), ni à invoquer contre lui une quelconque profondeur sémantique d'un mot, profondeur qui échapperait irrémédiablement à la description de ses usages (première partie de la deuxième critique).

Turing aborde la question du langage de manière essentiellement diachronique. D'où son refus de répondre à la question de l'intelligence des machines en faisant un simple sondage d'opinion [Gir95, p.135], sondage qui ne donnerait qu'un instantané des différents préjugés que chacun nourrit concernant les machines et l'intelligence. Ce que propose Turing, c'est de construire un cadre expérimental dans lequel l'expression d'« intelligence des machines » ne sera plus

considérée comme irrémédiablement déviante[19]. Pour qu'une expression déviante devienne normale, il y a deux manières de s'y prendre : soit en stipulant qu'un mot acquiert une nouvelle signification (ce qui serait une interprétation faible de l'hypothèse de Turing); soit en supposant que la signification reste la même, mais que les anciens usages autorisent un nouvel usage différent, à la faveur d'un nouveau contexte de croyances. La phrase « La Terre est ronde » est déviante dans un contexte de croyances où il est admis qu'elle est plate, mais il n'est pas besoin de supposer que le mot « Terre » acquiert un nouveau sens dans la proposition où sa rotondité est défendue. Il suffit de voir qu'un changement dans le contexte scientifique permet une projection des anciens usages du mot « Terre » sur le nouvel usage, sans supposer un changement de signification. De même pour les changements de contexte technique : l'invention de l'écriture a rendu des phrases telles que « Je suis à mille kilomètres de vous » non déviantes, sans pour autant changer le sens des termes.

L'intérêt de ces considérations est de montrer que Turing ne veut en aucun cas stipuler arbitrairement qu'un nouveau sens du mot « intelligence » doit être accepté. Il ne fait que créer, avec son jeu de l'imitation, les conditions techniques dans lesquelles une projection des anciens usages du mot « intelligence » viendra justifier un nouvel usage. C'est dans cette perspective qu'il envisage « qu'à la fin du siècle l'usage, les mots et l'éducation de l'opinion générale auront tant changé que l'on pourra parler de machines pensantes sans s'attendre à être contredit. » [Gir95, p.148]

Nous croyons que cette interprétation s'accorde bien avec la lecture que nous faisons du jeu de l'imitation : il n'y a pas comparaison entre l'homme et la machine du point de vue d'une définition préalable de l'intelligence (laquelle serait alors à discuter), mais création d'un nouveau contexte expérimental dans lequel l'expression de « machine intelligente » n'est plus déviante. C'est en maintenant cette dimension linguistique de l'argument de Turing qu'on évite de caricaturer sa naïveté.

4 Du fonctionnement de deux arguments

Parmi les réactions philosophiques qu'a pu susciter l'article de Turing, un bon nombre peuvent être suspectées de ne pas avoir vraiment « joué le jeu ». Une fois qu'on prend le parti de voir comment fonctionne le jeu de l'imitation, une fois

19. Nous utilisons le terme « déviant » en nous inspirant de l'article de Hilary Putnam dans [And83] : déviant veut dire « logiquement bizarre ».

que l'on entre dans les rouages de l'argumentation, certaines objections s'évanouissent d'elles-mêmes. Ainsi des objections relatives à la mauvaise définition de l'intelligence ou celles stigmatisant le réductionisme linguistique de Turing. Pourtant, il en subsiste deux qui semblent si profondément ancrées dans la psychologie populaire qu'elles doivent avoir encore de beaux jours devant elle : la première soutient qu'une machine est incapable de *créer*, de produire du nouveau (objection de lady Lovelace); la deuxième affirme que, le comportement humain n'étant pas formalisable, on ne saurait réduire l'intelligence à du mécanique, ni *a fortiori* hisser le mécanique jusqu'à l'intelligence humaine (objection du comportement informalisable). Il nous semble que ces deux arguments doivent être traités ensemble, car ils relèvent tous deux du même sophisme consistant à croire que, « dès qu'un fait se présente à l'esprit, toutes les conséquences de ce fait jaillissent simultanément avec lui. » [Gir95, p.162] La tâche de débusquer ce sophisme est centrale, non seulement dans la réfutation de ces deux objections, mais dans toute l'œuvre de Turing.

4.1 L'idée fondamentale

L'idée fondamentale est toujours de voir comment fonctionnent *effectivement* les choses. Là où notre imagination nous invite à faire mille raccourcis, restons sur le long chemin de l'effectuation; là où un schématisme spontané nous incline à saisir un mécanisme dans sa représentation instantanée, laissons l'opération se déployer dans le détail de sa durée. Au cœur des deux réticences mentionnées ci-dessus, une vision trop rapide des machines de Turing nous fait passer sans heurt apparent du concept de machine au concept de mécanisme, puis de celui-ci à l'idée de détermination, puis de cette détermination à la conclusion d'une prédictibilité absolue de tout ce que la machine peut faire. Mais voyons ce qui cloche à chaque bond.

Le concept de machine de Turing, tel qu'il est exposé dans l'article de 1936, est un concept essentiellement mathématique et abstrait. Il n'est pas inventé dans l'idée d'une réalisation particulière, mais dans le but de résoudre par la négative le problème mathématique de l'*Entscheidung*. Est-ce à dire qu'il est complètement étranger aux extensions techniques auxquelles il a donné lieu ? Non, car l'exigence d'effectivité qui est au centre de la théorie des nombres calculables est aussi celle que l'on retrouve dans l'effectivité des calculs menés par une machine réelle. C'est même la rigueur de la formalisation abstraite de cette notion de calcul qui rend possible la réalisation d'une machine à calculer universelle. Le concept abstrait de machine de Turing rejoint bien son incarnation concrète dans un mécanisme,

mais l'idée générale de « mécanisme » voile un peu ce qu'une machine *universelle* a de singulier .

Quand nous entendons « mécanisme », nous concevons généralement un ensemble prédéterminé de rouages et de rapports, agencé de manière à servir d'instrument. Mais une machine universelle est bien plus qu'un instrument : c'est une machine à simuler *n'importe quel* instrument - pourvu que le fonctionnement de celui-ci soit descriptible sous forme de calcul. La passivité d'un instrument ordinaire est une inertie; la passivité d'une machine universelle est une potentialité. La forme d'un instrument ordinaire détermine une utilisation particulière; la forme d'une machine universelle ne détermine rien d'autre que le langage dans lequel celle-ci recevra les instructions. A associer trop rapidement le concept de machine de Turing et celui de mécanisme, on risque de croire que l'attitude d'instrumentalisation qui préside à leur usage est la même. Or c'est justement parce qu'une machine universelle n'est pas un simple instrument qu'il devient envisageable d'interagir avec elle de manière à la rendre « intelligente ».

On saisit du même coup les précautions qu'il faut prendre quand on passe de l'idée d'un mécanisme à celle de son absolue détermination. Certes, les machines de Turing ne sont jamais que dans un nombre dénombrable d'états différents, et le passage d'un état à un autre se fait selon une entière nécessité. Cependant, le dénombrable ne mène pas nécessairement au fini, et la nécessité n'est pas une « cécité ». Si une machine de Turing est déterminée, ce n'est pas à la manière d'un mécanisme aveugle et obligatoirement destiné à s'achever. La portée de l'idée de détermination est donc doublement limitée : d'un côté par l'indécidabilité du problème de l'arrêt pour un nombre infini de machines de Turing (limite externe); de l'autre par la possibilité d'interagir avec une machine de Turing de manière à en modifier les états et le fonctionnement (ce qui rend possible, pour la machine, une forme d'apprentissage). Ces deux limites remettent en cause la conception d'une simple passivité instrumentale des machines de Turing, et le passage du mécanique au déterminé n'est légitime que si l'on garde à l'esprit cette double limitation.

Enfin, l'association que nous sommes tentés de faire entre détermination et prédictibilité est illégitime. La détermination indique qu'aucun élément de hasard ne se glisse dans les mécanismes dans la machine - ce qui assure certes une forme de prédictibilité *du point de vue de Dieu*. Mais du point de vue d'un observateur limité dans le temps, la notion de prédictibilité est relative. Alors que la détermination d'une machine de Turing est parfaitement objective, la prédictibilité de son comportement dépend de l'observateur, du temps et de la vitesse à laquelle il fait ses prévisions. Chaque fois que Turing introduit des considérations de temps dans

son article, il y a en toile de fond cette distinction fondamentale entre détermination et prédictibilité. Confondre les deux notions en une seule vague intuition nous conduit à deux illusions psychologiques complémentaires : du fait de la détermination des machines, nous imaginons qu'elles sont totalement prévisibles; et parce que le comportement humain nous paraît imprévisible, nous imaginons qu'il est aussi indéterminé.

Ce sont ces deux illusions que Turing va s'employer à rejeter en analysant les sixième et huitième objections qu'il se fait à lui-même. La méthode sera la même dans les deux cas : là où notre esprit croit faire une prédiction réelle en se plaçant dans l'absolu (soit pour dire que le comportement d'une machine est entièrement prédictible, soit pour dire que celui d'un homme est radicalement imprévisible), Turing exige que nous *prenions le temps* de faire notre prédiction, de dérouler dans la durée les mécanismes supposés : en s'y prenant ainsi, la prédictibilité de la machine et l'imprévisibilité de l'homme ne seront plus deux absolus déterminés seulement subjectivement, mais deux manières d'appréhender un comportement, relativement à nos capacités déterminées de prévision.

4.2 L'objection de lady Lovelace

L'objection de lady Lovelace est d'autant plus pertinente que la machine dont elle commente le fonctionnement - la machine analytique de Charles Babbage - était bel et bien une machine universelle. De celle-ci, elle affirme qu'elle ne peut « donner naissance à quoi que ce soit. » [Gir95, p.160] Turing propose examine successivement deux variantes de cette affirmation.

La première est : « La machine ne fait jamais rien de nouveau ». A cette formulation de l'objection de lady Lovelace, Turing oppose le dicton : « Il n'y a rien de nouveau sous le soleil. » A quoi rime cette réponse ? Plutôt que de se placer sur un terrain technique et de montrer ce que la machine est capable d'inventer, Turing se place sur un terrain pour ainsi dire « métaphysique », répondant qu'il y aussi peu de certitude concernant l'originalité de nos pensées que concernant l'inventivité de la machine. Le contenu de cette réponse a des implications extrêmes et laisse supposer que Turing adhère à une forme radicale de déterminisme. Il est cependant plus raisonnable de conjecturer qu'il ne vise pas à convaincre son objecteur qu'un déterminisme absolu dirige le cours de l'univers, mais à lui montrer combien l'objection de l'impossible nouveauté est subjective. En donnant une tournure métaphysique à l'hypothèse du déterminisme, il souligne par contraste ce que les présupposés de l'objecteur ont d'incertain. Au moment où l'objecteur croit énoncer une trivialité parfaitement objective, Turing l'invite donc à reconnaître que la

notion de nouveauté n'a qu'un fondement psychologique. Et si cette notion n'a qu'un fondement psychologique, il est aussi absurde de s'en servir pour décrire ce que nous faisons réellement que pour prédire ce que la machine est effectivement capable de faire.

Il n'en reste pas moins que cette intervention de Turing n'est pas très convaincante. Aussi ressent-il le besoin d'une autre formulation de l'objection, dont il tentera une réfutation plus rigoureuse : « la machine ne peut jamais nous prendre par surprise. » Ici, Turing invoque de manière abrupte son expérience personnelle, la surprise qu'il a souvent éprouvée lorsqu'il faisait fonctionner ses machines. Cette réponse est très étonnante : ne vient-il pas de dissoudre l'argument de l'impossible nouveauté en montrant que l'impression de nouveauté n'avait qu'une origine psychologique ? N'en est-il pas de même pour l'impression de surprise ? La nouveauté et la surprise ne sont-elles pas les deux faces d'un même « acte de création mentale »[Gir95, p.161] ? A bien y réfléchir, il semblerait même que l'impression de nouveauté soit plus susceptible de s'appuyer sur des critères objectifs que l'effet de surprise...

La seule manière de caractériser l'effet de surprise de manière mesurable est de le concevoir comme résultant d'un décalage temporel : soit la machine va plus vite que nous pour parvenir à l'issue prévue d'un calcul; soit elle va moins vite que nous et produit un résultat différent de nos prédictions. Dans le premier cas, la surprise vient de ce que nous sommes pris de court dans nos prédictions, dans le deuxième cas elle vient de ce que nos prédictions sont prises en défaut. Chaque fois, elle naît d'un contraste entre le déterminisme objectif du comportement de la machine, et la prédictibilité subjective avec laquelle nous appréhendons ce comportement. La surprise ne dérive donc pas d'une simple impression psychologique, elle vient du déséquilibre entre les traits objectifs du comportement de la machine et la trace mentale de nos prédictions, ce déséquilibre étant lui-même objectivement descriptible. Si Turing préfère défendre l'idée qu'une machine peut être source de surprise plutôt que de soutenir l'hypothèse de son inventivité, c'est parce que nous ne pouvons donner qu'une définition négative de la nouveauté (comme ce qui n'était pas *a priori* déductible de notre connaissance de la machine), alors que nous pouvons donner une définition plus positive de la surprise (comme décalage mesurable entre nos prédictions et ce qui a eu lieu). L'impression de nouveauté n'admet pas de degré, de même que celle d'originalité; l'effet de surprise, lui, est plus ou moins fort, selon le degré d'implication du sujet dans les prédictions qu'il fait.

Observons à ce propos que ce n'est pas la prédiction en elle-même qui est subjective (on peut savoir exactement à quel moment l'évolution d'un système

dynamique devient imprévisible), mais plutôt les *conditions* dans lesquelles elle est faite. Comme le dit lui-même Turing, le fait qu'il se sente surpris doit nous pousser à lui reprocher la manière « rapide et bâclée »[Gir95, p.161] dont il fait ses calculs, non sa sincérité! Et dans la vie courante, les conditions dans lesquelles nous prédisons le comportement d'un agent, homme ou machine, sont toutes extrêmement subjectives.

Certes, l'effet de surprise ne nous oblige pas à admettre un libre arbitre en la machine, mais psychologiquement, il nous y incline par degré[20]. D'où les passages où Turing suggère qu'un élément aléatoire pourrait être introduit dans la machine, précisément pour déjouer les prédictions que nous pouvons faire à son sujet. A elle seule, cette inclination ne fait donc pas force de preuve; cependant, le jeu de l'imitation crée une situation artificielle dans laquelle le degré de surprise devient l'élément le plus déterminant pour décider de l'identité des joueurs. L'objection de lady Lovelace vaudrait peut-être dans l'absolu - encore que l'hypothèse du « Rien de nouveau sous le soleil » reste envisageable pour montrer qu'à l'égard de l'inventivité, nous ne sommes pas mieux placés que des machines; mais dans l'espace relatif du jeu de l'imitation, elle ne peut en aucun cas infirmer l'hypothèse de machines intelligentes.

4.3 L'objection de l'informalité du comportement

Admettons que lady Lovelace se soit laissée convaincre par les arguments de Turing. Elle accepterait de ne plus dire qu'une machine ne fait rien de nouveau, ni qu'elle est incapable de nous surprendre; elle reconnaîtrait de plus que, dans le jeu de l'imitation, cette capacité à nous surprendre nous porterait à croire qu'elle est intelligente. Mais lady Lovelace pourrait se retrancher derrière l'idée que, puisque le comportement humain est informalisable, il n'est pas décemment envisageable d'en construire un modèle implémentable sur une machine de analytique. Voilà la teneur de la huitième objection à laquelle répond Turing.

Ici encore, Turing répondra à l'argument en l'envisageant de deux manières différentes. Il suppose d'abord une définition *normative* du comportement formalisable : un comportement est formalisable si nous disposons de règles de conduite auxquelles nous pouvons nous référer pour décider, selon le contexte, de l'action à entreprendre. Pour donner plus de force à l'idée que le comportement humain ne peut pas s'appuyer en toutes situations sur de telles règles de conduites, Turing

20. N'est-ce pas au moment où nous ne comprenons plus ce que fait notre ordinateur que nous sommes le plus enclins à supposer en lui une forme de « volonté » - souvent opposée à la nôtre ?

choisit un exemple où le choix à faire est minimal : passer ou ne pas passer, selon que le feu est vert ou rouge. Que se passe-t-il lorsque les deux passent au vert en même temps ? Dans ce cas, nous ne disposons certainement pas de règle de conduite nous indiquant ce que nous devons faire : nous improvisons. Si cela est vrai pour une situation aussi simple, cela doit *a fortiori* être encore mieux vérifié dans les situations complexes de la vie courante.

Cet argument serait valable si les règles de conduite étaient établies objectivement, indépendamment des sujets qui s'y soumettent. Or le propre d'une norme est d'être conventionnelle. L'exemple du feu vert le montre bien : il n'y a aucune nécessité à passer lorsque le feu est vert et à s'arrêter lorsqu'il est rouge. Même collective, même publiquement déclarée, une norme n'en reste pas moins subjective, conformément à son degré de conventionalité. L'interprétation des normes est donc soumise aux aléas de la psychologie individuelle, et il devient normal que nous soyons toujours à la recherche de compromis improvisés avec nos règles de conduite, car celles-ci ne sont elles-mêmes que des compromis collectifs, des conventions improvisées pour assurer l'ordre social. C'est pourquoi l'impossibilité de trouver des règles de conduite rendant compte de tous les comportements humains ne dit rien quant à la nature machinique ou indéterminée de nos comportements : elle prouve seulement le caractère fragmentaire et subjectif de nos conventions. La nécessité d'avoir recours à des normes ne révèle aucune indétermination réelle; elle ne fait qu'indiquer l'ignorance dans laquelle nous sommes de nos déterminations profondes. Et l'insuffisance des normes ne montre pas que nous sommes réellement en train d'improviser, d'inventer de nouveaux comportements; elle souligne simplement le fait que les conventions ne suffisent jamais à résoudre le décalage entre notre indétermination psychologique et nos règles de conduite.

D'où la formulation d'une variation plus sérieuse de l'argument, supposant une définition *descriptive* du comportement formalisable. L'objection devient : il est impossible de définir toutes les lois de comportement qui règlent la vie d'un individu, alors que l'ensemble des lois de comportement d'une machine est quasiment connu d'avance. Cette objection est bien plus pertinente que la première : non seulement elle ne mêle aucune considération morale ou psychologique à la question; mais elle accepte en plus la corrélation que suppose Turing entre *être-une-machine* et *obéir-à-des-lois-de-comportement*. L'objection ne dit plus que nous improvisons réellement en de nombreuses situations, elle dit qu'à la différence des lois de comportement d'une machine, celles d'un individu ne se laissent pas si facilement embrasser. En assimilant cette extrême difficulté de la formalisation à une impossibilité, l'objection devient plus rude encore, et soutient par

principe que le comportement humain est informalisable. Cette formulation de l'objection nous entraîne sur le terrain de l'épistémologie : ce n'est plus la nature du comportement humain qui nous permet de décider s'il y a improvisation ou machinisme, mais la nature nécessairement limitée de notre savoir qui empêche de donner une consistance factuelle à l'hypothèse de Turing.

La réponse de celui-ci consiste encore une fois à s'extraire de la spéculation pour s'enfoncer dans l'expérimentation. Il lance un défi : trouver la totalité des lois de comportement d'un ordinateur, en fonction de notre seule connaissance des *input* et *output*. D'après ses pronostics, il faudrait certainement plus de mille ans afin de décrire la totalité des lois de comportement de la machine. Le but de ce défi est de montrer qu'il est illégitime de prendre la difficulté de la formalisation pour une impossibilité. L'objection du comportement informalisable ne saurait donc valoir par principe. Pire, elle oblige à considérer qu'il n'est pas nécessairement plus difficile de formaliser la totalité du comportement humain que de formaliser la totalité du comportement d'une machine[21] .

C'est toujours le même sophisme que Turing veut contrer : psychologiquement, nous exagérons l'imprédictibilité du comportement humain et nous minimisons celle de la machine. Le défi qu'il lance vise à induire chez le lecteur une expérience de pensée (en attendant de devenir une expérience effective) qui annulera progressivement cette impression de dissymétrie : plus nous mesurerons la difficulté de formaliser entièrement le comportement d'une machine à partir de la simple observation, moins nous semblera plausible le fait que le comportement humain soit par principe informalisable. Turing ne réfute pas directement l'hypothèse épistémologique affirmant qu'il est impossible d'atteindre les lois du comportement humain dans leur totalité, réfutation qui exigerait - réfutation qui nécessiterait certainement d'entrer dans une discussion hautement spéculative. Pour inciter à admettre l'idée de machines intelligentes, il a tout juste besoin de montrer que les raisons qui nous poussent à admettre cette hypothèse sont uniquement psychologiques. Et puisque tel est le cas, la meilleure stratégie est d'opérer directement sur les préjugés du lecteur, de les « tordre » dans le sens contraire de leur déséquilibre spontané. Car plus nous nous délivrerons de nos préjugés concernant

21. Le parallèle est frappant entre le défi que Turing évoque et ceux qu'il a constamment dû relever dans ses recherches en cryptographie : tout son travail consistait à anticiper sur les lois de comportement de la machine *Enigma*, avec à sa disposition des messages codés et - si ce n'est les messages clairs eux-mêmes - des hypothèses plausibles sur le texte clair. C'est aussi certainement à son expérience des « bombes » qu'il se réfère implicitement lorsqu'il explique que les machines l'ont souvent surpris. D'une manière générale, l'approche cryptographie nous semble être la perspective-clef pour mieux comprendre la démarche de Turing.

les machines, plus nous semblera plausible l'idée qu'elle sont intelligentes. Cette stricte « plausibilité » est la condition nécessaire et suffisante pour que le jeu de l'imitation en vienne bien à soutenir la thèse que Turing veut défendre.

5 Conclusion

Nous avons préféré partir de trois questions particulières pour analyser la démarche de Turing et évaluer la portée de ses thèses. La première nous a entraîné à l'intérieur de cette belle machine argumentative qu'est le jeu de l'imitation; nous y avons trouvé une complexité que ne laissait pas présager l'expression de « test de Turing », complexité qu'il faut bien clarifier pour ne pas se méprendre sur le projet de Turing. La deuxième nous a amenés à relativiser les critiques habituelles relatives au béhaviorisme de Turing et à la pauvreté de ses conceptions linguistiques. La troisième nous a confrontés à deux objections que Turing envisage, objections d'autant plus intéressantes qu'elles sont toujours actuelles; et nous pensons que les débats philosophiques en intelligence artificielle ont encore à méditer sur le « sophisme » que Turing dénonce.

Après ce parcours, nous saisissons mieux l'unité de la pensée de Turing. Au centre de son propos, l'idée qu'il y a toujours plus dans l'effectuation d'une opération que dans sa simple description, et l'idée que seule l'expérimentation peut décider du bien-fondé d'une hypothèse. Son extraordinaire intelligence des machines est solidaire de l'hypothèse de machines intelligentes, de même que sa compréhension des limites du formalisme lui donne les moyens d'en mesurer la réelle extension. Aussi soutient-t-il que, lorsque nous jouerons réellement au jeu de l'imitation (et il envisage que les conditions du jeu seront des conditions banales d'ici un demi-siècle[22]), nos opinions se modifieront à un tel point qu'il y aura un sens à parler de *machines intelligentes*. Ainsi, le défi de son article consiste à anticiper du mieux qu'il peut sur les conditions techniques et psychologiques dans lesquelles son hypothèse sera acceptée. C'est pourquoi cet article est un tel mélange d'audace théorique et d'anticipation technique. Turing ne propose pas un seulement un programme de recherche; il insuffle à l'IA tout son esprit. Car plus qu'aucun autre en IA, « il pousse à réfléchir, et à réfléchir à ses réflexions. » [HD87, p.77]

22. Signalons le site http ://www.alicebot.org, qui propose de faire passer à un programme le test de Turing. Dans une version plus populaire, moins professionnelle et plus drôle, voir aussi le module *doctor* sous Emacs...

Références

[And83] Alan Ross Anderson, editor. *Pensée et machine*. Editions du Champ Vallon, Seyssel, 1983.

[And99] Daniel Andler. *Science et philosophie*. Bibliothèque du CREA. CREA, 1999.

[Dre84] Hubert L. Dreyfus. *Intelligence artificielle: mythes et es*. Editions Flammarion, Paris, 1984.

[Gan90] Jean-Gabriel Ganascia. *L'âme-machine, les enjeux de l'intelligence artificielle*. Editions du Seuil, Paris, 1990.

[Gar97] Howard Gardner. *Les formes de l'intelligence*. Editions Odile Jacob, Paris, 1997.

[Gir95] Alan Turing; Jean-Yves Girard. *La machine de Turing*. Editions du Seuil, Paris, 1995.

[HD87] Douglas Hofstadter and Daniel C. Dennett. *Vues de l'esprit*. Inter-Editions, 1987.

[Hod88] Andrew Hodges. *Alan Turing ou l'énigme de l'intelligence*. Editions Payot, 1988.

[Las96] Jean Lassègue. What kind of turing test did turing have in mind? *Tekhnema*, (3), 1996.

[Las98] Jean Lassègue. *Turing*. Edition Les Belles Lettres, Paris, 1998.

[NNGG89] Ernest Nagel, James R. Newman, Kurt Gödel, and Jean-Yves Girard. *Le théorème de Gödel*. Editions du Seuil, Paris, 1989.

[Tur36] A. M. Turing. On computable numbers, with an application to the entscheidungsproblem. In *Proceedings of the Mathematical Society*, volume 42 of *2*, pages 230–265, 1936.

[Tur50] A. M. Turing. Computing machinery and intelligence. *Mind*, LIX(236), 1950.