Publié dans Travaux de logique 7, 73-89, 1992,
source qui doit être utilisée pour toute référence à ce travail

ARGUMENTS GÖDELIENS CONTRE LA PSYCHOLOGIE COMPUTATIONNELLE

Jacques Paul Dubucs

Le premier théorème d'incomplétude de Gödel établit que tout système formel S suffisamment riche contient, s'il est cohérent, un énoncé élémentaire G_S improuvable mais vrai («élémentaire» signifie ici: du type $\forall x \psi x$, ψ récursif). Cet énoncé affirme sa propre improuvabilité dans S. Comme il est, en fait, improuvable dans S, il est vrai.

Je viens de montrer que l'énoncé G_S est vrai, bien qu'il ne soit pas prouvable dans S. Je suis donc capable de reconnaître la vérité de plus d'énoncés arithmétiques que le système S ne peut en prouver (ou qu'une machine de Turing T(S) associée à S ne peut en énumérer). Le même raisonnement pourrait s'appliquer à tout système formel suffisamment riche, et à toutes les machines qui énumèrent les théorèmes de ces systèmes. Donc je ne suis pas une machine de Turing, et les psychologues qui affirment le contraire ont tort. Telle est, sommairement exposée, la conclusion que l'on a cru pouvoir tirer des résultats de Gödel: l'esprit humain ne «fonctionne» pas comme une machine de Turing.

1. Les arguments qui visent à établir la supériorité de l'esprit humain sur les machines peuvent pécher de deux façons: soit en attribuant à l'esprit la possibilité de performances dont il est incapable, soit en déniant aux machines la possibilité de performances dont elles sont capables. En un certain sens, l'argument anti-mécaniste esquissé ci-dessus tombe nécessairement dans l'un ou l'autre de ces travers.

Le raisonnement qui me convainc de la vérité de G_S est le suivant:

a) S est «suffisamment riche», c'est-à-dire que les relations arithmétiques récursives y sont fortement représentables: si ϕ est un prédicat récursif, alors pour chaque entier n l'énoncé $\phi(\overline{n})$ est prouvable si n satisfait ϕ et réfutable sinon.

b) G_S exprime «correctement» sa propre improuvabilité, c'est-à-dire qu'un entier qui ne satisfait pas le prédicat ψ est bien le numéro d'une preuve de G_S[1] .

Par conséquent si G_S était faux il existerait, d'après a), un entier n tel que $\neg\, \psi(\overline{n})$ soit prouvable: G_S serait donc réfutable. Mais, d'après b), G_S serait également prouvable. S serait donc incohérent. Donc G_S est vrai.

L'anti-mécaniste soutient que la performance cognitive dont je viens de faire preuve n'est pas à la portée de la machine T(S). Il a tort. Car ou bien il pense que je viens de prouver que G_S est vrai, et alors il surestime cette performance: j'ai simplement montré que *si* S est cohérent, alors G_S est vrai. Ou bien il reconnaît que j'ai seulement prouvé que si S est cohérent, alors G_S est vrai, et dans ce cas il sous-estime les capacités de T(S), puisque l'énoncé qui affirme que la cohérence de S entraîne la vérité de G_S *est* un théorème de S.

2. Il existe de l'argument anti-mécaniste une interprétation plus charitable que celle dont on vient de montrer l'absurdité.

Je ne peux, c'est entendu, *prouver* G_S que sous l'hypothèse de la cohérence de S. Mais si S est, en fait, cohérent, je peux prouver chaque instance de G_S *sans* recourir à l'hypothèse que S est cohérent. En effet je suis évidemment capable de reconnaître que la phrase «n n'est pas le numéro d'une preuve de G_S» est vraie, lorsqu'elle l'est. Et si S est cohérent cette phrase est vraie, et donc *reconnaissablement* vraie pour chaque entier n. En somme si S est cohérent je suis en mesure de prouver inconditionnellement $\psi(\overline{0})$, $\psi(\overline{1})$, $\psi(\overline{2})$ ...

Cette nouvelle performance, pas plus que la précédente, ne permet d'affirmer ma supériorité sur la machine T(S). Car le rai-

---

1 Rappelons que $\psi(x)$ est la formule ouverte $\neg\, Pr_S(x, \ulcorner G_S \urcorner)$ («x n'est pas le numéro d'une preuve dans S de la formule G_S»).

sonnement qui me convainc de la vérité de $\psi(\overline{n})$ consiste simplement à vérifier, au cas où n serait le numéro d'une preuve, que la dernière formule de cette preuve n'est pas G_S: rien ici qui ne soit à la portée d'une machine, puisque la notion de preuve dans S est effective. De fait toutes les instances $\psi(\overline{0})$, $\psi(\overline{1})$, $\psi(\overline{2})$... de G_S sont bel et bien prouvables dans S si S est cohérent.

Le point sur lequel mes capacités de justification semblent excéder celles d'une machine est plutôt le suivant. Si je possède, pour chaque objet d'un certain domaine, une raison probante d'affirmer que cet objet satisfait le prédicat ϕ, alors je suis autorisé, en vertu de l'interprétation familière du quantificateur universel, à affirmer que l'énoncé $\forall x \phi x$ est vrai dans le domaine considéré. Ayant prouvé que chaque entier satisfait le prédicat ψ, je suis donc justifié à affirmer que $\forall x \psi x$, c'est-à-dire G_S, est vrai dans le domaine des entiers naturels. G_S est bien un énoncé que S est incapable de prouver, mais dont je suis en mesure de reconnaître la vérité.

3. A l'argument anti-mécaniste ainsi reconstruit peut être opposée une objection bien connue: le fait que je suis en mesure de justifier G_S établit que je ne suis pas la machine T(S), mais il n'établit pas que je ne suis pas une machine. En effet, puisque le système S prouve $\mathrm{Coh}(S) \rightarrow G_S$, le système $S' = S \cup \{\mathrm{Coh}(S)\}$ prouve G_S[2]. Par conséquent une machine qui énumère les théorèmes de S' est, quant à elle, capable d'engendrer G_S. L'argument anti-mécaniste qui établit ma supériorité sur T(S) échoue donc à établir ma supériorité sur T(S').

C'est vrai. Mais l'anti-mécaniste peut répondre que la conclusion putative de son argument n'est pas la «supériorité» de l'esprit sur toutes les machines, c'est-à-dire une thèse du type

> (AM) Il existe un énoncé arithmétique que je suis capable de justifier, mais qu'aucune machine de Turing cohérente (c'est-à-dire incapable d'engendrer $\overline{0} = \overline{1}$) ne peut engendrer

2 Coh(S) est la formule $\forall x \neg \mathrm{Pr}_S(x, \ulcorner \overline{0}=\overline{1} \urcorner)$.

mais plutôt la fausseté de toute thèse du type

> M(T) La machine de Turing T peut, tout en restant cohérente, énumérer toutes les vérités arithmétiques que je suis capable de justifier.

En d'autres termes l'argument anti-mécaniste peut se présenter comme un «schéma de réfutation», destiné à montrer l'inadéquation de tout modèle mécanique de l'esprit qui pourrait être effectivement proposé. Entendue en ce sens, la réfutation de l'hypothèse selon laquelle T(S) serait un modèle de l'esprit n'est nullement astreinte à contredire l'hypothèse selon laquelle T(S') en serait également un. Et la possibilité de *modifier* la machine T(S) et de la transformer en une machine T(S') capable d'engendrer l'énoncé G_S ne saurait être invoquée contre l'anti-mécaniste: ce dernier est en droit d'exiger qu'on lui présente une machine qui, *tout en restant la même*, parviendrait à simuler le comportement cognitif sous-jacent à l'inférence qui conclut de $\psi(\bar{0})$, $\psi(\bar{1})$, $\psi(\bar{2})$, ... à $\forall x \psi x$.

Naturellement la correction éventuelle du schéma anti-mécaniste de réfutation n'invalide pas, en toute rigueur, la thèse mécaniste selon laquelle il existe une machine de Turing cohérente capable d'énumérer la totalité des vérités arithmétiques que je suis en mesure de justifier: la réfutation de chaque instance de cette thèse établit que le mécanisme est au mieux une doctrine psychologique «ω-incohérente», qui soutient que l'esprit est une machine quoiqu'il soit établi pour chaque machine qu'il n'est pas *cette* machine. Bien que la doctrine en question soit probablement digne d'attention[3] , je ne discuterai pas ici la portée résiduelle d'un tel affaiblissement du mécanisme, et je me limiterai dans la suite à l'examen de la correction du schéma de réfutation lui-même.

---

3 En un certain sens elle a été explorée par Benacerraf (1967) qui pour sa part infère des résultats d'incomplétude de Gödel que nous ne sommes pas des machines de Turing, ou que si nous en sommes nous ne pouvons en tout cas pas déterminer notre propre «table d'instructions».

4. Une grande partie de la littérature[4] consacrée aux implications anti-mécanistes éventuelles des résultats de Gödel témoigne d'une propension fâcheuse à envisager le problème sous l'aspect purement *extensionnel* de la classe des énoncés arithmétiques susceptibles d'être certifiés respectivement par une machine et par un agent humain: la question décisive semble être de savoir qui, de l'homme ou de la machine, triomphera de l'autre dans cette compétition. Il faut sans doute reconnaître que la pertinence des résultats de Gödel pour la psychologie dépend bien d'une hypothèse de cet ordre, puisque l'argument anti-mécaniste ne mériterait aucunement d'être discuté si l'on n'admettait pas d'emblée qu'un agent humain est en mesure de reconnaître comme vrais *au moins* autant d'énoncés arithmétiques qu'une machine de Turing associée à un système formel cohérent assez puissant n'est capable d'en engendrer. Mais une fois cette hypothèse «quantitative» accordée, c'est-à-dire étant admis que la psychologie susceptible d'être affecté par les résultats de Gödel est une discipline largement «idéale», qui postule chez les agents une forme d'«omniscience logique»[5] , l'évaluation raisonnée de l'argument anti-mécaniste doit tenir compte des aspects *qualitatifs* des modes de certification à l'oeuvre chez les humains et dans les machines.

Le raisonnement par «ω-induction» décrit plus haut est particulièrement instructif à cet égard. Ce qui semble me distinguer de la machine T(S), c'est la transition que j'effectue - mais pas elle - de «pour tout n: $\psi(\overline{n})$» à «$\forall x \psi x$». L'improuvabilité de $\forall x \psi x$ dans le système S, jointe à la prouvabilité dans S de chaque instance $\psi(\overline{n})$, s'explique tout naturellement par l'existence d'un modèle «non standard» de S, dont l'univers $\mathbf{N}^*$ contient, à côté de l'univers $\mathbf{N}$ des entiers naturels, d'autres objets que les entiers: interprétées dans $\mathbf{N}^*$, les prémisses de l'inférence sont vraies mais sa conclusion est fausse. Quant aux raisons pour lesquelles j'effectue moi-même cette inférence, elles vont de soi: engagé dans la recherche d'énoncés vrais dans l'univers $\mathbf{N}$ des entiers naturels, c'est *dans cet univers* que j'interprète les variables qui figurent dans les énoncés que

4 Cette littérature est pour l'essentiel issue de Lucas 1961.
5 Cf. sur ce point Dubucs 1992.

j'examine. Ayant justifié $\psi(\overline{n})$ pour tout n, je peux ajouter «$\forall x\psi x$» à la liste des énoncés justifiés, car pour moi c'est précisément *cela*, la signification de «$\forall x\psi x$». Dès lors la différence «quantitative» de mes capacités recognitionnelles et de celles de la machine T(S) n'est que l'indice d'une différence, beaucoup plus significative, entre la structure qualitative de ces capacités. A l'inverse de la machine T(S), dont le comportement démonstratif s'explique exhaustivement en termes compu-tationnels, c'est-à-dire par référence aux seules propriétés *formelles* des symboles arithmétiques, je suis pour ma part guidé par la référence «attendue» de ces symboles, c'est-à-dire capable d'invoquer le fait que certains de leurs regroupements reflètent éventuellement les propriétés de la structure mathématique **N** que je *vise*. C'est la raison pour laquelle il est impossible de soutenir, comme le défenseur du mécanisme est tenté de le faire, que le comportement de la machine T(S') est identique au mien. Erronée sur le plan «quantitatif» (contrairement à moi, T(S') ne parvient pas à certifier $G_{S'}$), cette affirmation est en outre insoutenable sur le plan «qualitatif»: l'aptitude de T(S') à inscrire G_S sur son ruban de sortie est le pur résultat de son architecture combinatoire. T(S'), et c'est là sa différence cruciale d'avec moi, est *intentionnellement inerte* [6].

5. Dans la dimension sous laquelle il peut se manifester à un observateur extérieur, le travail mathématique est une activité de traitement de symboles, dont le résultat tangible est la répartition de certaines concaténations de symboles, les «formules bien formées», en deux classes (disjointes dans le meilleur des cas): la classe des formules acceptées, et celle des formules rejetées. En

---

6 Il faut cependant reconnaître que la sensibilité des processus inférentiels au contenu sémantique des symboles manipulés ne découle pas, en toute rigueur, de l'hypothèse selon laquelle la vérité de G_S est reconnaissable. Il se pourrait que le raisonnement qui conduit de $\psi(\overline{0})$, $\psi(\overline{1})$, $\psi(\overline{2})$,... à $\forall x\psi x$ se fonde sur la mobilisation d'une règle formelle d'un type nouveau, la «règle ω», qui autorise précisément l'inférence en question (sur cette règle, cf. J.-Y. Girard 1987, chap. 6) Comme un raisonnement de ce type n'est pas «effectif» (il ne peut pas être simulé par une machine de Turing), ceci suffit à établir la partie négative de la thèse anti-mécaniste. La thèse de la «sémanticité», quant à elle, n'en découle pas.

vertu d'une hypothèse particulièrement séduisante par son économie et son uniformité, la partie «non visible» du travail mathématique est homogène à sa partie visible, c'est-à-dire consiste elle aussi en un traitement de symboles: le fonctionnement de l'appareil cognitif pourrait être intégralement décrit en termes de manipulations symboliques écartant toute référence à un *contenu sémantique* éventuellement attaché aux symboles traités. Cette hypothèse «computationnaliste» suggère une explication «causale» de l'activité mathématique. Car chaque réalisation d'une formule (considérée comme «type» syntaxique) est un objet physique (ou neuro-physiologique) par lequel peuvent évidemment transiter des déterminations causales: le fait que l'appareil cognitif soit à l'instant t dans une certaine relation *matérielle* R à une instance de la formule A peut être cause du fait qu'il est à l'instant t' dans la relation R à une instance de la formule B. Le computationnaliste dispose alors d'une explication naturaliste assez plausible de l'activité cognitive consistant à «accepter» ou à «justifier» certaines formules arithmétiques: ce que nous décrivons, dans un vocabulaire épistémique qui exige analyse, comme une «opération intellectuelle» conduisant à «justifier» la formule A n'est que le processus causal déterminé par lequel l'appareil cognitif se place dans une relation matérielle donnée à un *token* de «A» (la relation matérielle en question est tout à fait analogue à celle d'un processeur informatique à un opérande qui figure dans un registre déterminé de la mémoire d'un ordinateur). Compte tenu du caractère compositionnel du symbolisme mathématique, une explication de ce genre est assurément capable de rendre compte de l'essentiel des propriétés logiques satisfaites par l'activité de justification, et notamment du fait que l'ensemble des formules justifiées est stable par les règles d'inférence qui préservent la vérité: puisque l'effet causal des occurrences du symbole «A $\supset$ B» est une fonction des effets causaux des occurrences des symboles «A» et «B», une inférence comme le modus ponens, qui spécifie une relation logique entre formules, possède une contrepartie causale, c'est-à-dire que l'on peut formuler une loi de transition affirmant que si l'appareil cognitif est à l'instant t dans la relation pertinente à des *tokens* de «A» et de «A $\supset$ B» il doit être à l'instant t' dans la même relation

à un *token* de «B». En bref le computationnalisme se propose de rendre compte de l'activité mathématique en invoquant le fait que la structure combinatoire des formules arithmétiques est capable d'*encoder* la totalité des aspects de leur contenu sémantique qui interviennent dans les processus inférentiels. Et plus généralement il prétend expliquer les activités cognitives comme des processus de manipulations de symboles *cognitivement impénétrables*[7], c'est-à-dire de symboles à la référence desquels le sujet n'a jamais accès.

La plupart des philosophes hostiles au computationnalisme sont tentés d'admettre que les choses *pourraient* bien se passer comme le computationnaliste le prétend. Ils sont convaincus que l'intelligence humaine mobilise, au moins partiellement, la référence des symboles manipulés, mais ils se résignent généralement à reconnaître que les données de l'observation sont tout à fait incapables d'attester cette particularité du répertoire cognitif humain, et ils sont donc disposés à concéder - puisque l'explication computationnaliste du comportement cognitif des ordinateurs est évidemment la seule possible - que l'on pourrait concevoir des machines dont le comportement cognitif «manifeste» («le comportement d'entrée/sortie») serait indiscernable de celui d'un agent humain. Aussi concentrent-ils leurs critiques sur le célèbre test proposé par Turing (1950) pour remplacer la question vague «Les machines peuvent-elles penser?» (pour réussir le test, une machine doit pouvoir induire en erreur sur sa vraie nature un interrogateur qui ne la voit pas, mais se contente de lire ses réponses sur un téléscripteur). Ils objectent à ce test que la similitude du comportement cognitif manifeste ne suffit pas à garantir la similitude des processus sous-jacents à ce comportement: même si une machine répondait régulièrement par les mêmes «sorties» qu'un agent humain à une classe donnée d'«entrées», nous ne serions autorisés ni à en conclure que les processus cognitifs à l'oeuvre chez l'agent humain sont de nature computationnelle, ni à qualifier sérieusement de «compréhension» ou d'«intelligence» les capacités dont fait preuve la machine. L'argument favori des anti-computationnalistes consiste ici

7 L'expression est de Z.W. Pylyshyn, dont l'ouvrage Pylyshyn (1984) est une expression particulièrement élaborée des thèses computationnalistes.

à invoquer certaines *Gedankenexperimente* censées établir qu'une activité de manipulation symbolique, quelque «performante» qu'elle puisse paraître, ne saurait être qualifiée d'«intelligente» que si elle repose sur une «pénétration cognitive» des symboles manipulés. Ainsi, pour prendre un exemple inspiré de l'argument bien connu de la «chambre chinoise» de Searle (1980), il *se pourrait* qu'un locuteur chinois répondant en chinois à des messages chinois voie ses performances linguistiques égalées par un locuteur français ignorant le chinois, mais disposant d'un ensemble de consignes écrites en français du type «Si vous recevez le message «¶†®ƒ∂» (supposé être une phrase chinoise), envoyez le message «Ò‡◊©Ÿ™¤» (idem)»: dans une telle situation, souligne l'anti-computationnaliste, nous dirions du locuteur pour qui les idéogrammes chinois sont impénétrables, et qui détermine son comportement sur la seule base de leur *forme*, qu'il est parvenu à *simuler* la connaissance du chinois et non qu'il comprend «réellement» cette langue. Mais de telles paraboles, et c'est là leur fragilité, supposent en partie ce qui est en question, puisqu'elles se bornent à réaffirmer que nous ne sommes en aucun cas disposés à employer les mots «compréhension» ou «intelligence» là où les performances cognitives ne résultent pas de l'accès à un *sens*.

Un théoricien anti-mécaniste pourrait être tenté de soutenir qu'il détient avec l'argument issu des résultats de Gödel une arme d'une tout autre force contre la psychologie computationnaliste: puisque le traitement de la formule G_S semble bien *manifester* la différence entre un dispositif qui manipule les symboles arithmétiques d'après leurs seules propriétés combinatoires et un dispositif sensible à leur contenu sémantique, pourquoi ne pas *accepter* le verdict du test de Turing? Un «engin sémantique» pourrait être discriminé de n'importe quelle machine de Turing T (au programme supposé connu) au vu de ses réponses à une classe bien déterminée de questions arithmétiques: il est celui qui répond «NON» à au moins une question du type «Acceptez-vous A?», où A est l'un des axiomes du système formel S instancié par T, ou celui des deux qui répond «OUI» à la question «Acceptez-vous G_S?». A s'en tenir à l'argument exposé dans le paragraphe 2, une «expérience cruciale» de ce type n'a malheureusement aucune

chance d'être concluante, quand bien même on la laisserait se poursuivre indéfiniment. Car une chose est de dire que la classe des énoncés assertés par un agent qui «vise» le modèle standard doit contenir «$\forall x \psi x$» si elle contient «$\psi(\overline{n})$» pour tout n. Et autre chose est de dire qu'un tel agent est en mesure d'en arriver au point où il a effectivement établi «$\psi(\overline{n})$» pour tout n, même si l'on admet qu'il dispose d'un temps de calcul illimité, et même si l'on reconnaît qu'il est, pour tout n, en mesure d'établir effectivement «$\psi(\overline{n})$»: s'il commence ses calculs à l'instant t_0, il possède au plus n instances de $\forall x \psi x$ à l'instant t_n, et n'est donc en mesure de conclure $\forall x \psi x$ qu'à l'«instant» $t_{\omega+1}$ où il contemple et résume l'infinité dénombrable des résultats qu'il a obtenus jusqu'«ici».

Sauf à admettre - et qu'invoquerait-on pour justifier une telle hypothèse? - que les événements mentaux prennent place dans une temporalité dense et non pas discrète, les différences éventuelles entre le comportement cognitif des «engins sémantiques» et celui des «engins syntaxiques» ne peuvent donc être attestées par les données d'une observation indéfiniment prolongée, et l'invocation de l'«argument gödélien» reconstruit dans le paragraphe 2 ne met pas le théoricien anti-computationnaliste dans une situation sensiblement plus avantageuse que celle qu'il obtiendrait par un argument purement qualitatif comme celui de la «chambre chinoise»: le schéma de réfutation qui établit l'irrecevabilité de la thèse M(T) est vide de tout contenu empiriquement manifestable.

6. La thèse majeure de l'anti-mécaniste est qu'un agent qui «vise» le modèle attendu de l'arithmétique doit être capable, contrairement à la machine T(S), de reconnaître la vérité de G_S. Ce processus de reconnaissance comporte nécessairement une étape non mécanisable. Dans la version discutée jusqu'ici, cette étape non mécanisable est celle qui conduit à G_S à partir d'une infinité dénombrable d'énoncés dont chacun peut être engendré mécaniquement. Mais cette version de l'argument anti-mécaniste, on vient de le voir, ne permet aucune *manifestation* de sa propre correction: bien que la vérité de G_S (dans le modèle attendu) résulte

simplement de la vérité de chacune des instances numériques de G_S, la reconnaissance de la vérité de G_S ne peut pas, quant à elle, provenir de la reconnaissance «totalisée» de la vérité de chacune de ces instances.

Il existe toutefois une autre version de l'argument anti-mécaniste, pour ainsi dire duale de la première, qui semble exempte de ce défaut. Selon cette version, la disposition à «viser» le modèle attendu de l'arithmétique confère à qui la possède la capacité de reconnaître «intuitivement» la vérité éventuelle d'un certain énoncé dont G_S découle mécaniquement. Cet énoncé est celui qui exprime la cohérence de S. Si S est cohérent, Coh(S) est un énoncé dont toutes les instances $\neg\, \mathrm{Pr}_S(\overline{0}, \ulcorner \overline{0} = \overline{1} \urcorner)$, $\neg\, \mathrm{Pr}_S(\overline{1}, \ulcorner \overline{0} = \overline{1} \urcorner)$,... sont, à nouveau, vraies et reconnaissablement vraies, puisque c'est une tâche mécanique que de décider, un entier étant donné, s'il est ou non le numéro d'une preuve dans S et, dans l'affirmative, si la preuve dont il est le numéro possède « $\overline{0} = \overline{1}$ » pour formule finale. Mais le raisonnement susceptible de conduire à la reconnaissance de la vérité de Coh(S), contrairement au raisonnement analogue pour G_S, n'est pas une inférence fondée sur une infinité dénombrable de prémisses: c'est un raisonnement par induction (au sens ordinaire du terme) sur la longueur des preuves dans S. Il consiste à vérifier que les axiomes de S sont satisfaits dans le modèle attendu (donc que les énoncés dont la preuve dans S est de longueur 1 sont vrais dans ce modèle) et que les règles d'inférence de S préservent la vérité. Dans ces conditions tous les énoncés prouvables dans S sont vrais dans le modèle attendu, $\overline{0} = \overline{1}$ n'est donc pas prouvable dans S, Coh(S) est donc vrai, et l'on peut en déduire G_S. Un raisonnement de ce type, qui ne comporte aucun processus inférentiel «actuellement infini», est visiblement capable de conduire à une classe de réponses *effectivement discriminable* de celles de la machine T(S) dans le test construit dans le paragraphe précédent. La seule étape non mécanisable est la première, dans laquelle est évaluée la correction des axiomes de S dans le modèle visé: consistant à accepter ou à rejeter une classe d'énoncés *selon leur référence* (VRAI ou FAUX) dans l'interprétation attendue, elle n'est pas plus effectuable par une machine que ne le serait une

instruction du genre «Si le symbole observé est une réalisation du type T, faire A; sinon faire B».

7. J'examinerai pour terminer une objection possible à l'argument anti-mécaniste sous sa dernière version, celle qui suggère que la vérité de G_S peut être établie au travers de la reconnaissance intuitive de la correction éventuelle de S dans le modèle attendu de l'arithmétique. Cette objection sceptique vise à mettre en doute le gain en «manifestabilité» procuré par le passage de la première version de l'argument à la seconde.

Le test du paragraphe 5 n'est certainement pas capable de manifester la différence entre la machine T(S) et n'importe quel agent (cohérent et «logiquement omniscient») sensible au contenu sémantique du symbolisme arithmétique. Car il ne suffit pas à discriminer de celui de T(S) le comportement cognitif d'un agent qui «viserait» pour sa part un modèle non standard de l'arithmétique: un tel agent pourrait très bien reconnaître la correction des axiomes et des règles d'inférence de S sans pour autant asserter G_S, s'il s'avérait incapable de reconnaître sur cette base la cohérence de S. En effet l'affirmation selon laquelle aucune suite finie de formules de S n'est une preuve de $\overline{0} = \overline{1}$ est elle-même libellée dans un langage («suite finie») qui met implicitement en jeu la définition des entiers naturels, de sorte qu'un agent guidé par une conception «déviante» de la totalité de ces entiers pourrait estimer insuffisante la preuve par récurrence de l'improuvabilité de $\overline{0} = \overline{1}$: cette preuve, tout juste capable d'assurer que la longueur d'une démonstration dans S de $\overline{0} = \overline{1}$ ne peut être mesurée par aucun entier pouvant être obtenu par l'itération, un «nombre fini» (au sens standard) de fois, de l'opération «successeur» appliquée à zéro, laisse ouverte la possibilité de démonstrations de $\overline{0} = \overline{1}$ qui auraient une longueur finie (au sens non standard) ...

La forme finale de la thèse anti-mécaniste est donc la suivante: c'est la fausseté de chaque énoncé du type

> M'(T) La machine de Turing T est à la fois cohérente et capable d'énumérer tous les énoncés dont un agent *qui vise le*

modèle standard de l'arithmétique est capable de reconnaître la vérité.

Mais une thèse qui mentionne le «modèle standard de l'arithmétique» s'expose à une forme de scepticisme dont on peut trouver chez Wittgenstein et chez la plupart de ses disciples[8] une expression particulièrement intéressante et radicale, et en vertu de laquelle il n'y a tout simplement pas de sens à dire que G_S est vrai (le premier théorème d'incomplétude de Gödel devrait alors, en toute rigueur, être formulé de manière purement syntaxique: chaque système formel suffisamment riche contient, s'il est cohérent, un énoncé indécidable). Selon Shanker (1989: 221 sq), qui a récemment défendu ce point de vue de manière particulièrement éloquente, le défaut majeur de la formulation sémantique du premier théorème d'incomplétude (selon laquelle G_S est vraie (dans le modèle standard) mais improuvable dans S) tient à ce qu'elle rend «purement externe» la connection entre une proposition mathématique et sa preuve: il est «incohérent», écrit-il, d'affirmer que G_S est à la fois vraie et improuvable. Si on la comprend correctement, cette objection ne se limite nullement à refuser toute signification à un énoncé dont les conditions de vérité transcenderaient notre aptitude à reconnaître si elles sont ou non remplies. Car l'aptitude à «viser le modèle standard» suffit, en vertu de ce qui précède, pour reconnaître la vérité de G_S. L'objection, beaucoup plus radicale, requiert en outre que l'aptitude recognitionnelle en question puisse *publiquement* se manifester par la production d'une classe d'énoncés assez riche pour déterminer sans ambiguïté les propriétés du modèle considéré. Mais comme la mention du «modèle standard» est, nous l'avons vu, inéliminable de l'argument anti-mécaniste, et que par ailleurs le premier théorème d'incomplétude de Gödel établit précisément l'impossibilité de définir (à un isomorphisme près) ce modèle à l'aide d'une classe récursivement axiomatisable d'énoncés (du premier ordre), l'objection wittgensteinienne vide l'argument anti-mécaniste de tout contenu.

8 Cf. par exemple Goodstein 1980.

Il existe à cette objection de bonnes et de mauvaises parades. L'une des moins convaincantes, quoique des plus usitées, consiste à invoquer la possibilité de caractériser le modèle standard au sein d'un langage d'ordre supérieur: l'arithmétique de Peano du second ordre, dans lequel l'axiome de récurrence est écrit sous la forme

$$(*) \quad \forall\phi\{[\phi(0) \,\&\, \forall x(\phi(x) \rightarrow \phi(S(x)))] \rightarrow \forall x(\phi(x)\}$$

est une théorie catégorique, dont l'unique modèle est, à un isomorphisme près, le modèle standard[9]. Mais l'affirmation selon laquelle le modèle standard peut être décrit sans ambiguïté dans le langage de la logique du second ordre suppose que ce langage lui-même soit compris sans ambiguïté, c'est-à-dire que soit donnée au quantificateur universel de second ordre qui figure dans (*) sa signification «attendue», dans laquelle la variable ϕ parcourt la *totalité* des sous-ensembles de $\mathbf{N}$[10].

La réponse la plus raisonnable au scepticisme wittgensteinien consiste probablement à soutenir que notre impuissance à caractériser complètement des notions comme celle d'«entier» ou de «nombre fini» à l'aide d'une classe récursivement énumérable d'énoncés ne constitue pas un réel obstacle à l'intelligibilité et à l'absence d'ambiguïté de ces notions, et qu'en définitive l'interprétation «attendue» de l'arithmétique ne dépend pas du langage utilisé pour la décrire. Car les modèles «non standard» de l'arithmétique comportent tous des éléments qui ne peuvent pas être obtenus à partir de zéro par itération de l'opération «successeur», et la question de savoir si un objet peut ou non être engendré de cette manière ne soulève en pratique aucune espèce de difficulté: comme le remarque M. Dummett (1963: 193), «nous n'entretenons aucun doute sur le point de savoir si Jules César est ou non un nombre naturel». En d'autres termes, bien qu'il nous soit impossible d'obtenir, par le biais d'un accord sur une classe «présentable» d'énoncés, une *garantie publique* de l'identité de nos notions «privées» d'«entier» ou de «nombre fini», nous avons de bonnes raisons de supposer que nous par-

9 Cf. Montague 1965.
10 Cf. Shapiro 1985: 720-721.

venons à nous entendre lorsque nous disons d'une preuve dans un système formel qu'elle est un objet «fini» (même si nous ne spécifions pas de borne supérieure pour sa longueur). Or si une telle convergence est possible, qui ne se résume ni à une disposition communément partagée à asserter les énoncés constitutifs d'un certain corpus ni, *a fortiori*, à une propension commune à entretenir avec certaines *occurrences* de ces énoncés une relation matérielle déterminée, c'est qu'il existe au moins une notion dont la psychologie computationnaliste est incapable d'expliquer la maîtrise.

Appendice. On peut donner une forme plus précise à la question de l'incidence des résultats d'incomplétude de Gödel sur la psychologie spéculative des attitudes propositionnelles. Moyennant l'hypothèse d'omniscience logique mentionnée plus haut, un opérateur épistémique K supposé rendre compte des propriétés formelles du savoir satisfait certainement les axiomes suivants:

$\vdash_T K(A) \rightarrow A$	(ce qui est su est vrai)
$\vdash_T K(K(A) \rightarrow A)$	(la propriété précédente est elle-même connue)
$\vdash_T K(A \rightarrow B) \rightarrow (K(A) \rightarrow K(B))$	(le *modus ponens* est maîtrisé)
Si A est une tautologie, alors $\vdash_T K(A)$	(omniscience logique)

Nous pouvons ajouter sans contradiction au système T ainsi obtenu la conjonction Q de la classe finie des axiomes de l'arithmétique de Robinson. Mais si, soucieux de conformité à l'analyse computationnaliste des processus cognitifs, nous décidons de traiter le symbole K comme un prédicat métalinguistique attaché au *nom d'un énoncé*, et non plus comme un opérateur de type modal, associant une proposition à une *proposition*, nous obtenons le système T' défini par:

$\vdash_{T'} Q$

$\vdash_{T'} K(\ulcorner A \urcorner) \rightarrow A$

$\vdash_{T'} K(\ulcorner K(\ulcorner A \urcorner) \rightarrow A \urcorner)$

$\vdash_{T'} K(\ulcorner A \rightarrow B \urcorner) \rightarrow (K(\ulcorner A \urcorner) \rightarrow K(\ulcorner B \urcorner)$

Si A est une tautologie, alors $\vdash_{T'} K(\ulcorner A \urcorner)$

système dont Montague (1963) a montré l'incohérence (l'opérateur K ainsi défini possède toutes les propriétés de l'opérateur de prouvabilité de Gödel, augmentées du «principe de réflexion» $Pr(\ulcorner A \urcorner) \rightarrow A$).

Signalons enfin que cette impossibilité logique de reconstruire la psychologie des attitudes propositionnelles sur une base purement computationnaliste a été récemment étendue par Thomason (1980) aux attitudes «solipsistes» comme la croyance, qui n'entraînent pas la vérité de leur corrélat.

C.N.R.S. Institut d'Histoire et
de Philosophie des Sciences et des Techniques
Université de Paris-I Sorbonne
13, rue du Four; F - 75006 Paris

Références bibliographiques

ANDERSON, A.R. (ed.) (1964). *Minds and Machines*. Englewood Cliffs, N.J.: Prentice Hall (trad. fçse *Pensée et Machine*. Seyssel, 01420: Champ Vallon, 1983).

BENACERRAF, P. (1967). God, the devil, and Gödel. *The Monist, LI-1*, 9-32.

DUBUCS, J.-P. (1992). Omniscience logique et frictions cognitives. In: D. Andler *et al.* (éds), *Philosophie et Cognition*. Bruxelles: Mardaga.

DUMMETT, M. (1963). The philosophical significance of Gödel's theorem. In: *Truth and Other Enigmas*. London: Duckworth, 1978.

GIRARD, J.-Y. (1987). *Proof Theory and Logical Complexity I*. Napoli: Bibliopolis.

GOODSTEIN, R.L. (1963). The significance of incompletness theorems. *British Journal for the Philosophy of Science, XIV*, 208-220.

LUCAS, J.R. (1961). Minds, machines, and Gödel. *Philosophy, XXXVI* (repr. in: Anderson 1964, 43-59).

MONTAGUE, R. (1963). Syntactical treatment of modality, with corollaries on reflection principles and finite axiomatizability. In: R. Montague, *Formal Philosophy*. Yale U.P., 1974, 286-302.

MONTAGUE, R. (1965). Set theory and higher-order logic. In: J. Crossley & M. Dummett (eds.), *Formal Systems and Recursive Functions, Proceedings of the Eighth Logic Colloquium, Oxford, July 1963*. Amsterdam: North-Holland Pub., 131-148.

PYLYSHYN, Z.W. (1984). *Computation and Cognition*. M.I.T.: Bradford Books.

SEARLE, J. (1980). Minds, brains, and programs. *The Behavioral and Brain Sciences, III*: 417-457 (repr. in: J. Haugeland (ed.), *Mind Design*. M.I.T.: Bradford, 3e éd., 1985, 282-306).

SHANKER, S.G. (1989). Wittgenstein's remarks on the significance of Gödel's theorem. In: S.G. Shanker (ed.), *Gödel's Theorem in Focus*. London/New York: Routledge, 155-256.

SHAPIRO, S. (1985). Second-order languages and mathematical practice. *TheJournal of Symbolic Logic, L*, 714-742.

THOMASON, R.H. (1980). A note on syntactical treatments of modality. *Synthese, XLIV*, 391-395.

TURING, A.M. (1950). Computing machinery and intelligence. *Mind, LIX* (repr. in: Anderson 1964, 4-30).