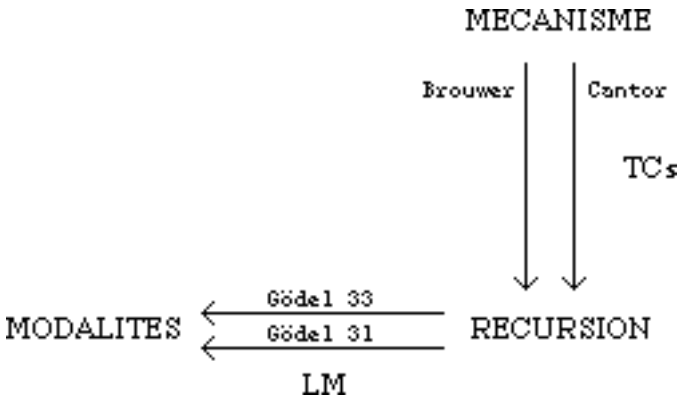
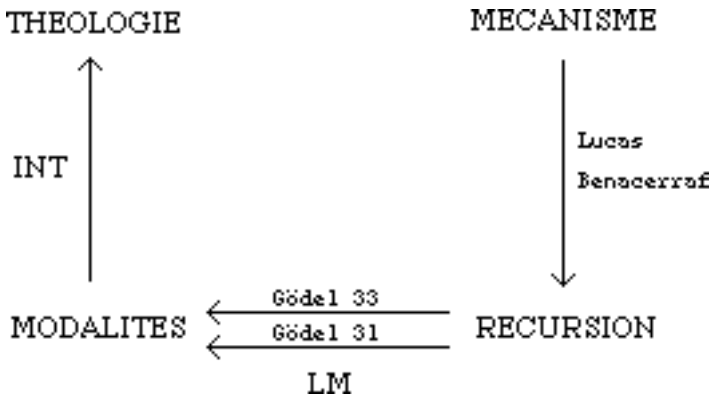
## 2.3 Des lois mécanistes de l'esprit

*RESUME DE TOUT 2.3*

-*types* essentiellement,




ainsi que



- *but et méthodologie*

Au lieu d'aborder de plein front l'analyse de la duplication, je vais étudier les utilisations du théorème de Gödel en philosophie de l'esprit en indiquant au fur et à mesure les relations avec les paradoxes de la duplication. Le but est d'étudier ce qu'une machine peut connaître ($\Box$) sur elle-même, et ce qu'il est possible pour elle de communiquer de façon convaincante-positiviste à son sujet ou au sujet de son éventuel duplicata. Pour cette *communication*, il ne peut pas y avoir de différences entre soi et le dupliqué, puisqu'une tierce personne ne peut pas a priori distinguer une personne "originale" de son duplicata. J'analyserai en profondeur la réfutation du mécanisme due à Lucas et une reconstruction de l'argument de Lucas due à Benacerraf. Les relations entre l'argumentation de Lucas et la duplication seront évidentes grâce à la définition de l'identité personnelle basée sur 2-REC.

L'autoréférence du mécanisme *indexical* a eté ramenée dans la section précédente à une structure de contrôle définie par diagonalisation. Cela permet d'étudier les discours indexicaux limites des machines. La différence entre le communicable et le connaissable sera alors extraite de cette analyse. Plus précisément la méthodologie va consister à analyser l'argumentation de Lucas, d'abord +/- historiquement (2.3.1), ensuite avec une théorie de la connaissance (S4, ce que j'appellerai aussi *l'avis des philosophes*, c'est 2.3.2), ensuite, à extraire une logique de la communication des machines auto-référentiellement correctes (2.3.3), et utiliser cette logique avec le stratagème (2.3.4) pour extraire une théorie de la connaissance introspective possible des machines afin de demander leur propre *avis* dans les voisinages de l'infini (2.3.5). L'erreur principale de Lucas pourra être interprétée par une permutation non valide du quantificateur existentiel et de l'opérateur de communication convaincante, dans la définition du mécanisme. En effet, on sera naturellement amené à la distinction

$$\Box\exists n \text{fonc}(n) \neq \exists n \Box \text{fonc}(n).$$

Il s'agit donc essentiellement de la même erreur que celle de Kalmar dans sa réfutation de la thèse de Church.

Il restera à expliciter comment des machines peuvent connaître plus qu'elles ne peuvent communiquer. Je le fais en introduisant la notion d'inférence inductive (2.3.5).

Ensuite, avec le matériel accumulé, on étudie la philosophie des machines dans les voisinages de l'infini. Une logique de la connaissance-temporelle est extraite. Elle éclaire les situations paradoxales de la duplication. Avec l'hypothèse mécaniste, les théorèmes d'incomplétude

font de la découverte de la machine universelle une véritable (re)découverte du sujet (ou *des* sujets) en mathématique.

Toutefois il reste assez bien de travail pour résoudre conceptuellement l'entièreté des paradoxes présentés en 1.3. En particulier la situation de l'annihilation retardée sort du cadre de 2.3.

C'est pourquoi dans la troisième partie, le paradoxe de l'annihilation retardée est généralisé avec le paradoxe du graphe filmé et le paradoxe du dovetelleur universel. Je montre qu'en plus d'extraire une temporalité, c'est toute une phénoménologie de la physique qui devient, au moins partiellement mais obligatoirement, isolable à partir de l'hypothèse mécaniste. Le caractère nécessairement non constructif de l'argument peut alors être exploité pour rendre compte de l'apparence *nécessairement empirique* des "lois de la nature" (de la *physicalité* locale).

## 2.3.1 Le théorème de Gödel et le mécanisme

*Brièvement*

*Ici je regarde les arguments traditionnels selon lesquels le mécanisme peut être réfuté par les théorèmes d'incomplétude (TI)*

*TI ==> ¬MEC (TI-)*

*et je montre, avec Webb, Benacerraf, ..., qu'on a plutôt l'inverse*

*TI ==> +MEC (TI+)*

---

Les relations entre les théorèmes d'incomplétude (TI) de Gödel et la question du mécanisme en philosophie de l'esprit sont denses et tumultueuses.

Elles sont denses en regard du nombre de publications.

Elles sont tumultueuses en regard de la diversité des opinions et du nombre de changements d'avis de nombreux chercheurs qui réfléchissent sur ces relations[1].

Ces relations reposent en partie sur l'identification de base entre les systèmes formels et les machines, et sur la thèse de Church qui identifie les dynamiques formelles avec le *Turing* (LISP, jeu de la vie, etc.) *émulable*.

On peut distinguer essentiellement trois types d'argumentation :

1) **TI$^-$** : les résultats d'incomplétude sont en faveur de la fausseté du mécanisme. Les théorèmes de Gödel montreraient, dans quelques sens, que nous ne sommes pas des machines (¬MEC-DIG-IND), (Par exemple Lucas 1961, Kreisel 1970, Penrose 1988, Neumaier 1987)

2) **TI°** : les résultats d'incomplétude ne sont pas pertinents pour la philosophie mécaniste de l'esprit. En général, cette position est par défaut celle de ceux qui tentent de réfuter *à la base* les arguments de type TI$^-$, soit en déniant l'identification entre machines et systèmes formels, comme celle que je décris dans la section précédente[2], soit en niant la thèse de Church.

Certains auteurs cependant, admettent cette identification, et réfutent TI$^-$ en approfondissant le théorème de Gödel, et en le réutilisant, en quelque sorte contre les arguments de type TI$^-$. Webb parle à ce sujet de l'aspect

1 Comme Post 1922, Putnam 1960, 1988, Webb 1968, 1980, Benacerraf 1967, Lewis 1969, 1979.

2 Voir aussi Yu 1992 pour une présentation assez claire de cette identification.

*double-edged* des arguments Gödéliens. Une analyse scrupuleuse des arguments de type $TI^-$ semblent alors les transformer en type $TI^+$ :

3) **$TI^+$** : les résultats d'incomplétude sont en faveur de MEC. De même que c'est l'*impossibilité* de distinguer mécaniquement entre le total et le partiel calculable qui rend la thèse de Church plausible (voir 2.2), les résultats d'incomplétude pourraient rendre plausible la philosophie mécaniste elle-même. Les résultats d'incomplétude constituent alors un squelette sur lequelle une *théorie* mécaniste de l'esprit et de la conscience peut venir se greffer.

La ressemblance de LWV, accompagnée de sa justification par MDI, avec l'énoncé du second théorème d'incomplétude, ainsi que le fait que ce dernier s'applique d'office aux machines (avec l'identification de base), sont autant d'indices pour $TI^+$.

Il n'est pas question de décrire ici exhaustivement[3] l'histoire des relations entre les théorèmes d'incomplétude de Gödel et le mécanisme. Je vais de façon synthétique décrire quelques morceaux choisis afin de mieux apprécier l'entourage historique de la réfutation du mécanisme par Lucas. L'approche est *en spirale*, autant pour des raisons pédagogiques, que par un souci de cohérence chronologique.

*1°) Post 1921 (1941, 1965)*

En 1921 (soit 10 ans avant que Gödel ne sorte son théorème d'incomplétude) Post avait déjà anticipé l'essentiel de l'histoire des relations "Gödeliennes" entre l'esprit et les machines à partir d'une réflexion sur ses *systèmes normaux complets* dont l'équivalence avec les machines universelles de Turing sera démontrée plus tard. Post démontre alors les résultats d'insolubilité (pas de système normal capable de résoudre tous les problèmes que l'on peut énoncer concernant les systèmes normaux). Il démontre[4] (de façon informelle et sémantique) des résultats d'incomplétude (qui le conduiront plus tard à définir les ensembles créatifs et à donner son "théorème de Gödel" *miniature*, (Post 1944), voir 2.1) Les réflexions de Post concernent les logiques qu'il appelle *finitaires*. (Celles-ci correspondent aux systèmes formelles, aux théories axiomatisables).

Il conçoit "l'identification de base", ainsi que le thèse de Post-Turing, le mécanisme digital et l'existence (dans cette philosophie) de propositions absolument indécidables. IL anticipe $TI^-$, $TI^\circ$, $TI^+$, ainsi qu'un renversement

---

3 Voir aussi Smart 1961 Lewis 1969, 1979, ou Hutton 1976, Coder 1969, Dennett 1972s, Dodd 1991.

4 D'une certaine façon la section 2.1 décrit, en termes anachroniques, ce que Post avait déjà réalisé dans les années 20 (Post 1921).

dans les fondements des mathématiques. Par exemple, soumettant, en 1941[5], son anticipation datant de 1921-22, Post écrit :

> *But perhaps the greatest service the present account could render would stem from its stressing of its final conclusion that mathematical thinking is, and must be, essentially creative. It is to the writer's continuing amazement that ten years after gödel 's remarquable achievement current views on the nature of mathematics are thereby affected only to the point of seeing the need of many formal systems, instead of a universal one. Rather has it seemed to us to be inevitable that these developpments will result in a reversal of the entire axiomatic trend of the late nineteenth and early twentieth centuries, with a return to meaning and truth. Postulational thinking will then remain as but one phase of mathematical thinking.* (Post 1921-22-41-65, Post souligne)

Toute cette section va tenter de préciser davantage cette citation[6] ainsi que les citations de Post qui vont suivre.

Au cours de la rédaction de cette anticipation, Post propose l'argument TI⁻ : l'ensemble des propositions prouvables par une machine étant récursivement énumérable, il ne peut pas exister de système complet (finitairement présenté) pour la logique symbolique. Se référant explicitement à "l'évolution créative" de Bergson, il conclut (dans un premier temps) :

> *The Logical Process is Essentially Creative*

et Post précise plus loin (c'est lui qui souligne) :

> *It makes of the mathematician much more than a kind of clever being who can do quickly what a machine could do ultimately. We see that a machine would never give a complete logic ; for once the machine is made we could prove a theorem it does not prove.*

C'est essentiellement le raisonnement que Lucas 1961 développera plus tard.

Dans l'appendice cependant, Post anticipe la thèse de Church (la version "Post-Turing") et sa relation avec la psychologie :

> *..., that development owes its significance entirely to the universal character of our characterization of an arbitrary generated set ... Establishing this universality is not a matter for a mathematical proof, but of psychological analysis of the mental processes involved in combinatory mathematical processes.*

---

[5] Laquelle sera refusée. Il faut attendre Davis 1965 pour que ce *joyau* paraisse.

[6] Cette idée est proche de Dummet 1963 qui utilise le théorème de Gödel en faveur de la philosophie intuitionniste qui elle aussi retourne vers la vérité et la signification (voir aussi Ladrière 1951) mais à la différence de Dummet, je pense que la prise en compte de l'intuionnisme ne doit pas se faire au dépend de la logique classique, et à la différence de Post je ne crois pas que ce renversement (dû à l'incomplétude) signera la fin de la pensée "postulationnelle" (hypothético-déductive, même s'il marque la fin du mythe de la pensée *purement* hypothético-déductive.

à partir de quoi il suggère l'existence de propositions absolument indécidables :

> *The unsolvability of the finiteness problem* (le problème de la décision) *for all normal systems, and the essential incompleteness of all symbolic logics, are evidences of limitations in man's mathematical powers, creative though theses be. They suggest that in the realms of proof, as in the realms of process, a problem may be posed whose dificulties we can never overcome; that is that we may be able to find a definite proposition which can never be proved or disproved.*

Suit alors, sous formes de notes, des réflexions sur la nature de la preuve, des relations entre la créativité et les ordinaux. Il en arrive finalement à corriger explicitement sa précédente argumentation de type $TI^-$, et à anticiper les arguments de type $TI^+$ :

> *The following suggestions came up.*
> *(a) The conclusion that man is not a machine is invalid. All we can say is that man cannot construct a machine which can do all the thinking he can. To illustrate this point we may note that a kind of machine-man could be constructed who would prove a similar theorem for his mental acts.*
> *(b) The creative germ, seems not to be capable of being purely presented but can be stated as consisting in constructing ever higher types. These are as transfinite ordinals and the creative process consists in continually transcending them by seeing previously unseen laws which gieve a sequence of such numbers. Now it seems that this complete seing is a complicated process mostly subconscious. But it is not given till it is made completely conscious. But then it ought to be constructable purely mechanically.*

Post termine quelques notes par "*all this before reading Brouwer*". On reconnaît par ailleurs l'inspiration Bergsonienne autant dans cette article de Post que dans la philosophie de Brouwer, lequel s'est aussi référé à Bergson. Brouwer est cependant très éloigné du mécanisme[7].

*2°) Turing 1936, 1939, 1948, 1950*

Si on fait abstraction des machines universelles naturelles, comme la donnée de *n* corps dans un système newtonien ($n>2$), le système génétique, ou le cerveau humain[8], on peut considérer que Babbage est le premier inventeur de la machine universelle, et Turing le second. Mais Turing la découvre *théoriquement*, attelé au problème de la décision, et reconnaît son

---

7 La notion de sujet créateur, ou plutôt sa formalisation à travers le schéma de Kripke, rend la thèse de Church inconsistante (Kreisel 1970). Dans la mesure où l'intuitionisme de Brouwer est pensé comme non formalisable et indépendant du langage (sur quoi Brouwer insiste assez bien), je pense qu'il est difficile de conclure.

8 Le cerveau est au moins une machine universelle, la thèse mécaniste affirme qu'il est au plus une telle machine

universalité. Il tentera de construire une réelle machine électronique qui soit universelle (l'ACE), et abordera l'hypothèse mécaniste béhavioriste de façon claire et directe. C'est sans doute l'influence du positivisme ambiant (il a même assisté à des exposés de Wittgenstein) qu'il se limitera à MEC-BEH.

Etant l'auteur des premiers théorèmes de limitation et d'insolubilité, et ayant été le premier a énoncé ces résultats directement en terme de machines, Turing était bien placé pour aborder la question de la pertinence de ces théorèmes pour réfuter ou encourager le mécanisme.

Sans entrer dans les détails je propose les remarques suivantes :

a) Dans son papier de 1936 sur le problème de la décision (posé par Hilbert) Turing présente sa machine comme émulant directement un calculateur humain (usant d'une feuille de papier!). Il appelle directement l'intuition, reposant sur son analyse du calcul pour faire admettre une thèse équivalente à celle de Church. (C'est pourquoi celle-ci est souvent appelée *thèse de Church-Turing* ou *thèse de Turing*) C'est ce papier qui va convaincre Gödel de la thèse de Church.

b) Turing publie successivement en 1948 et en 1950 deux articles où il aborde la question du mécanisme[9].

Dans les deux articles il propose une réfutation de l'argument Gödélien TI$^{-}$. Respectivement :

*Dans le papier de 1948 :* le théorème de Gödel ne s'applique qu'à des machines 100% correctes ou infaillibles. Une machine intelligente, ou un être intelligent ne doit pas être infaillible, argumente-il ensuite.

*Dans le papier de 1950 :* Il rappelle l'argument basé sur l'identification de base. Pour toute machine il existe des questions auxquelles une machine particulière (qu'elle soit universelle ou non) donnera soit une réponse fausse (la machine sera *incorrecte*), soit elle ne donnera pas de réponse (elle sera *silencieuse*). Turing considère ensuite que l'hypothèse selon laquelle ce résultat ne s'applique pas à nous (humain) est simplement extrêmement présomptueuse.

Notons par ailleurs que dans son papier fondamental de 1939, écrit sous la direction de Church aux Etats-Unis, il tente de différencier l'intuition de l'ingénuité au moyen de son analyse ordinale.

*3°) Popper (lu en 1948 publié en 1950)*

J. S. Mill avait déjà remarqué que la possibilité de l'omniscience entraîne la nécessité du déterminisme. Un être omniscient connaîtrait sa destinée (ses destinées) et serait incapable de l'altérer (les altérer). Cela entraîne que l'omniscience est incompatible avec l'omnipotence. En relativisant un raisonnement similaire sur une machine physique, plongée dans un

9 C'est dans celui de 1950 qu'il présente le problème sous une forme exclusivement positiviste au moyen du test d'imitation (MEC-DIG-BEH). L'idée du test apparaît néanmoins en filigrane dans l'article de 1948.

environnement stable, Popper argumente qu'une machine ne peut pas prédire en toute circonstance son propre avenir. L'argument de base correspond à une diagonalisation, le sujet de la prédiction est l'objet de la prédiction. Popper utilise en effet, au cours d'une argumentation, le résultat de Gödel. Il manifeste une certaine prudence quant à l'identification de base. Il considère en effet qu'une machine universelle est un être mathématique à la différence d'une machine physique "réelle". Il admet néanmoins que celle-ci peut bien en être une instantiation concrète singulière, et prudemment il associe à la machine physique des propositions formelles correspondant de façon univoque à des comportements descriptibles de la machine à différents instants. Il démontre de cette façon l'indéterminisme mécaniste (non indexical et abrupte, mais à terme) inhérent à la physique classique[10].

Néanmoins, dans une footnote Popper présente, comme Post, un argument typique de type $TI^+$ basé sur la mécanisabilité de l'argument élémentaire $TI^-$ :

> *I think that it may be possible to buid calculators capable of solving Gödelian questions - for example, by way of copying a method which operates not only with derivability but with truth, in Tarski's sense- and of adding the solution to its premises. Such a calculator may, whenever it is stimulated by a Gödelian question, 'grow' in consequence of this stimulus into one whose system of premises is strengthened. But even such a calculator would, for example, be incapable of 'knowing' at any moment whether its present system of premises was consistant - it could 'know' such a thing only of its old premises after strengthening them.* (footnote 1 in Popper 1950, page 183)

Popper est cependant dualiste et non-mécaniste, aussi insiste-il pour qu'on ne prenne pas trop aux sérieux de tels raisonnements : l'imprédictabilité des machines n'implique en rien que nous soyons des machines pour des raisons similaires (voir 1.1) qui font que MEC-BEH n'implique pas MEC-IND.

*4°) Gödel 1951*

Gödel aussi s'est intéressé aux relations entre les résultats d'incomplétude-insolubilité et la philosophie mécaniste de l'esprit. Selon Wang 1974, il tînt, lors d'une conférence délivrée en 1951 (non publiée), l'opinion selon laquelle deux propositions concernant l'esprit et les machines auraient été rigoureusement prouvées :

1) L'être humain est incapable de formuler (de mécaniser, de formaliser) l'entièreté de ses intuitions mathématiques. Gödel donne par exemple l'intuition de la consistance de la présentation de son intuition

[10] De même que Webb 1980 argumente sur l'incompatibilité de l'existence d'une machine universelle avec un démon de Laplace.

mathématique, et se réfère implicitement à son second théorème d'incomplétude. Ceci est de type $TI^-$. Il admet cependant que cela n'exclut pas l'existence, non constructive[11] d'une machine capable de formuler l'intuition mathématique humaine. Une telle machine peut même être, selon Gödel, découverte empiriquement. Mais on ne peut pas alors prouver qu'elle formalise cette intuition, ou alors on ne peut pas prouver qu'elle est correcte. Cette remarque correspond à la suggestion *a)* de Post (voir plus haut) de type $TI^+$. Cette idée est fondamentale et on va la retrouver souvent le long de cette section. Elle ne devrait pas étonner ceux qui ont commis les expériences de la duplication où l'on se voit survivre, par MEC-DIG-IND, d'une façon précisément non communicable.

2) *La disjonction de Gödel* : une autre proposition portant sur la question machine / esprit que Gödel estime rigoureusement prouvée est la suivante :

> ***Ou bien*** *l'esprit humain surpasse toutes les machines (pour être précis : l'esprit humain peut décider plus de questions de la théorie des nombres qu'une machine quelconque),* ***ou bien*** *il existe des questions de la théorie des nombres indécidables pour l'esprit humain.*

Une telle question serait (avec la thèse de Church) absolument (ou intuitivement, ou humainement) indécidable. A la différence de Post, Gödel semble opter pour la première alternative[12] de la disjonction. Ce qui peut paraître assez étonnant puisqu'il avait été convaincu de la thèse de Church telle que Turing l'a (re)énoncée dans son papier de 1936[13]. La raison est, que sans l'identification de base[14], on a seulement :

$$TC \rightarrow (\neg MEC \vee \exists \text{ une proposition absolument indécidable})$$

$$\text{avec } MEC \Leftrightarrow \{x \mid \boxdot x\} \text{ est RE,}$$

"$\boxdot x$" signifie "je sais intuitivement prouver x".

Gödel (au moins pendant cette période) semble non seulement choisir la première alternative, mais il semble avoir aussi opté pour le dualisme en philosophie de l'esprit. Il pense plausible l'existence d'une preuve mathématique de l'impossibilité de justifier mathématiquement la biologie

---

11 Ce qui ne choque pas le platoniste qu'était Gödel.

12 Gödel attribue un choix similaire à Hilbert, ce qui ferait du formaliste Hilbert un non-mécaniste. Cette attitude est, à mon avis, pas si facile à tenir. Voir Webb 1980pour une défense simultanée du mécanisme et du formalisme de Hilbert basée sur TI. On observera la similitude de la disjonction de Gödel et la "réfutation" de la thèse de Church de Kalmar (voir 2.2).

13 Wang 1974 cite par ailleurs un texte écrit "*récemment*" (et donc dans les environs des années 50, et pas des années 30) par Gödel. Gödel objecterait à l'analyse de Turing (de 1936) que la machine possède un nombre fini d'états, alors que l'esprit se développe constamment. Ceci est curieux puisque Turing montre l''existence de machine universelle. Celle-ci est capable d'émuler des machines s'auto-transformant, et s'auto-développant sans cesse, en escaladant éventuellement une échelle transfinie.

14 avec l'identification de base on a TC <=> MEC-DIG (voir 2.2).

(empirique) sur base de la physique (empirique) (Par exemple le cerveau humain n'aurait pas eu le temps de se développer depuis la naissance de notre univers, etc). Notons aussi bien chez Wang 1974 que chez Gödel 1951 (et presque chez tout le monde) le préjugé qui conduit à concevoir le mécanisme uniquement dans le cadre du matérialisme[15]. Je reviens sur cette question dans 3.3.

*5°) Nagel et Newman 1958*

C'est Nagel et Newman qui populariseront, surtout chez les Anglo-saxons, le théorème de Gödel et son interprétation de type plutôt TI⁻. L'argumentation est implicite et sera explicitée par ceux, comme Putnam 1960 et Arbib 1964 qui réfuteront cette interprétation.

On peut se convaincre de la présence du raisonnement suivant dans Nagel et Newman : Par TI, on a

> *pour tout système formel ou machine M, démontrant des propositions vraies (de l'arithmétique), je peux* ***prouver*** *(par un raisonnement métamathématique) une proposition vraie p que M ne peut pas prouver,*

donc je suis distinct de M. On reconnaît à nouveau une sorte de "test de Turing" limité à l'arithmétique élémentaire.

Critique de Putnam 1960 :

tout ce que TI permet de dire est :

$$con(M) \rightarrow p$$

et, dans le cas où M est une machine assez riche (comme PA), M peut aussi prouver cette proposition. Pour réellement prouver p, il faudrait aussi prouver la consistance de M, con(M), ce qui peut se révéler difficile si M est fort complexe (par exemple si M utilise des axiomes d'existence de très grands cardinaux).

Putnam met ainsi en évidence l'erreur logique la plus évidente de l'argument le plus simple de TI⁻. Notons que ni Gödel, ni Turing ne commette cette erreur. Post la commet et la corrige presque instantanément. Putnam argumente qu'il se pourrait que nous soyons inconsistants et Priest 1987 élabore ce point de vue.

Critique de Arbib 1964

15 A l'exception notable de Maine de Byran (voir Baertschi 1992), cf aussi la conversation entre Changeux et Connes, voir 3.3.

TI ne s'applique pas aux machines à inférence inductive (learning machine), capables d'être au moins momentanément incorrectes ou inconsistantes, dans la mesure où de telles machines apprennent à partir de leurs propres erreurs.

De plus, si on exige d'une machine qu'elle réponde par oui ou par non après un délai borné, en modifiant ses règles d'inférence de façon qu'elle évalue une probabilité, ou une croyance, une telle machine, par TI sera nécessairement incorrecte pour quelque question. Un résultat, dit Arbib, qui distingue difficilement l'homme de la machine, rejoignant ainsi un argument de 1948 de Turing (voir plus haut).

Remarquons qu'il est encore possible d'obtenir une machine correcte. Il suffit de respecter son incomplétude : en lui permettant de (méta)-répondre par des propositions du genre "je ne sais pas", ou "je n'ai pas de réponse", "donnez-moi plus de temps", ou même "stack overflow", ou mieux, en lui permettant de ne pas répondre du tout, c-à-d d'être silencieuse.

Après tout, nous savons (voir 2.1) que les machines universelles, extensionnellement, sont nécessairement strictement *partielles*.

Arbib, dans l'édition de son livre de 1987 (Arbib 1964), donne une démonstration du théorème d'incomplétude de Gödel et son "*incremental removal*", ainsi que le phénomène du *speed-up* de Gödel. Cela lui permet de concevoir des machines qui se transforment et qui accélèrent leur temps de calcul. Elles n'ont pas la forme explicite des machines de Myhill ou de Case, et du fait qu'il ne considère pas le speed-up de Blum mais celui de Gödel (voir 2.2), il n'est pas clair si cette évolution est programmable à la base. Arbib ne semble pas trop prendre au sérieux cet éventuel écueil. Il insiste qu'une machine intelligente doit être un robot couplé à un environnement duquel il va tirer l'essentiel de ses expériences (ce à quoi la machine peut croire).

On va voir comment préciser des argumentations similaires. Il est curieux que Turing, s'intéressant à la fois aux fondements des mathématiques et à la possibilité de l'intelligence artificielle n'ait pas creusé de tels arguments. Ceci est d'autant plus étonnant qu'il est l'auteur de *Systems of Logic Based on Ordinals*, un travail qu'il fera sous la direction de Church à Princeton, et pour lequel il n'était, semble-t-il, pas trop motivé (Feferman 1988).

Peut-être est-ce le préjugé (matérialiste ?) par trop favorable aux machines pensantes qui a empêché Turing de creuser davantage les arguments $TI^-$, de les retourner et de les enrichir, par exemple avec la notion d'oracle, ou avec les machines ordinales.

*6°) Lucas 1959, paru en 61, et Priest 1987*

Jusqu'en 1961, nous n'avons eu droit, au sujet des relations entre TI et la problématique machine/esprit, qu'à des "rumeurs". Une anticipation non

publiée de Post, une conférence non publiée de Gödel, quelques remarques de Turing, une footnote de Popper, une conclusion pas très explicite d'un ouvrage populaire (Nagel & Newman).

Avec *Minds, Machines and Gödel*, lu à Oxford en 1959 avant d'être publié en 1961, Lucas consacre un article exclusivement à une argumentation de type $TI^{-}$, c-à-d à une réfutation du mécanisme reposant sur une interprétation des deux théorèmes d'incomplétude de Gödel.

L'argumentation de Lucas explicite en détail le raisonnement de Nagel et Newman, en

> *pour tout système formel ou machine M, démontrant des propositions vraies (de l'arithmétique), je peux* ***produire comme vraie*** *(par un raisonnement métamathématique* ***intuitif)*** *une proposition vraie p que M ne peut pas prouver.*
> *donc je suis distinct de M.*

Pour Lucas, une machine ne peut rien *produire comme vraie* autrement qu'en exhibant une preuve formelle. L'ensemble des preuves émises par une machine constitue automatiquement, avec la thèse de Church, un ensemble récursivement énumérable, et la preuve donnée par la machine est nécessairement une preuve formelle. Cela découle de l'identification de base entre machine et système formel, identification défendue précisément par Lucas.

Lucas construit un schéma de réfutations des hypothèses mécanistes possibles le concernant. Il admet "je suis correcte et adéquat", grâce à quoi les hypothèses possibles sont limitées aux machines correctes et adéquates. Or, dit Lucas, présentez-moi une machine correcte et adéquate quelconque, je peux, l'observant de l'extérieur, produire comme vraie une proposition p que la machine ne peut pas produire comme vraie. Il s'agit de la proposition diagonale p, avec :

p <-> p n'appartient pas à l'ensemble des propositions produites comme vraies par M.

Comme M est une machine, avec l'identification de base et la thèse de Church, p est une proposition de l'arithmétique que M ne peut pas produire comme vraie, effectivement.

-conclusion, dit Lucas je suis différent de M, et ce raisonnement marche quelle que soit la machine que l'on présente.

L'argument de Putnam selon lequel l'homme pourrait être une machine inconsistante coupe à la racine l'argumentation de Lucas. Priest développe

une argumentation très semblable à celle de Lucas, mais avec une conclusion diamétralement opposée. En gros, Priest admet, dès le départ, l'hypothèse mécaniste. Le "produire comme vrai" est assimilé à l'intuitivement ou le naïvement prouvable. Il argumente que "intuitivement prouvable" est RE parce qu'une preuve doit être *finitairement* reconnaissable en tant que telle, et il refait le raisonnement de Lucas pour aboutir à une contradiction et rejeter l'hypothèse de consistance. En fait Lucas et Priest montrent tout deux :

$$\neg MEC \vee \neg CON$$

Priest, admettant MEC, conclut ¬CON, et plaide, d'une façon qui rappelle Hegel, pour la nature fondamentalement inconsistante de l'esprit. A l'opposé, Lucas croit à sa propre consistance et conclut à la nature essentiellement non-mécaniste de son esprit.

*Consistance et consistance.* Lucas affirme (avec quelque véhémence) qu'il est consistant. Il argumente à cet effet qu'il se corrige quand il se trompe. Mais alors, comme le font déjà remarquer Webb (1968, 1980, 1983) (voir aussi Visser 1986), il utilise non plus le prédicat

$$\neg B(\ulcorner\bot\urcorner), \text{ c-à-d } \neg\exists y\, bw_{PA}(\ulcorner\bot\urcorner,y), \text{ ou encore } \forall y\, \neg bw_{PA}(\ulcorner\bot\urcorner,y),$$

mais il utilise un prédicat <u>extensionnellement</u> équivalent bien qu'<u>intensionnellement</u> plus sophistiqué. Ceci n'est pas innocent car PA peut aussi démontrer sa propre consistance lorsque celle-ci est représentée de façon adéquate du point de vue intensionnel. Voilà deux exemples inspirés du prédicat de Rosser :

$$bw\text{-}rosser_{PA}(x,y) = bw_{PA}(x,y) \,\&\, \forall z\, (y \geq z \rightarrow \neg bw_{PA}(\ulcorner\text{-}x\urcorner,z))$$

Notons que bw-rosser(x,y) est décidable, et PA étant consistante $bw\text{-}rosser_{PA}$ est extensionellement équivalente à $bw_{PA}$.

1) Un prédicat assez simple due à Webb :

$$bw'_{PA}(x,y) = bw_{PA}(x,y) \,\&\, \neg bw_{PA}(\ulcorner\bot\urcorner,y)$$

Un prédicat plus sophistiqué, qui a le mérite de coller à la consistance non monotonique décrite par Lucas, et que Feferman avait utilisé (Feferman 1960) pour illustrer justement la différence entre intensionnel et extensionnel.

$$bw''_{PA}(x,y) = bw_{PA}(x,y)$$
$$\& \forall a \forall b \forall w\ (y \geq a \ \& \ y \geq b \ \& \ y \geq w \rightarrow \neg (bw_{PA}(x,y) \ \& \ bw_{PA}(\ulcorner \neg x \urcorner, y))$$

Notons que du point de vue extensionnel, la consistance de PA entraîne que bw, bw', bw'' sont équivalents. Il en est de même pour les B et CON respectifs. Mais il est facile de se convaincre que PA est à même de prouver con', ou con''. En effet $con' \Leftrightarrow \forall x \neg bw'(\ulcorner \bot \urcorner, x) \Leftrightarrow \forall x \neg bw'(\ulcorner \bot \urcorner, x) \vee bw'(\ulcorner \bot \urcorner, x)$ qui est entraîné par une forme quantifiée du principe du tiers-exclu (en l'occurrence, HA, l'arithmétique intuitioniste de Heyting, ne prouve par (son) Con'). Ceci montre aussi que l'équivalence extensionnelle entre bw, bw' et bw'' n'est pas prouvable *par* PA.

La suite est constituée de *reconstructions* ou / et réfutation de Lucas.

*7°) Benacerraf 1967 (+ Lucas'"Satan Stultified")*

Quelle drôle d'article que celui de Benacerraf, "God, the Devil and Gödel" !

Benacerraf argumente, d'abord, que les conséquences du théorème de Gödel peuvent être d'importance en théologie mathématique. Il fait allusion à la phrase de Weyl selon laquelle "Dieu existe parce que les mathématiques sont consistantes, et le diable existe car nous ne pouvons pas le démontrer". Que les mathématiques soient consistantes, ou que notre intuition mathématique soit correcte, et que cela, nous ne pouvons pas le (dé)montrer, Gödel l'a déjà extrait, comme on l'a vu plus haut, de son second théorème d'incomplétude, auquel Weyl fait allusion. Il suffit alors d'admettre les deux définitions philosophico-théologiques suivantes pour démontrer rigoureusement la proposition de Weyl :

D1) Dieu = la consistance des mathématiques

D2) Le diable = l'impossibilité de la communication humaine (et donc finie[16]) et convaincante (donc correcte) de la connaissance de la consistance des mathématiques.

Aussitôt après, Benacerraf admet qu'il est possible que la proposition de Weyl ne soit pas une conséquence naturelle du théorème de Gödel dans la mesure où D1) et D2) sont, pour le moins, débatable. Il a néanmoins illustrer comment le théorème de Gödel peut être utilisé de façon précise en philosophie ou en théologie[17]. Le "théorème" de Weyl concerne une limitation humaine. La réfutation de Lucas fonctionne de la même façon et

16 avec MEC-DEI.

17 Pour une (autre) utilisation du théorème d'incomplétude de Gödel en théologie, voir aussi Torrance 1969.

concerne une limitation des machines. Benacerraf critique alors le manque de rigueur dans la réfutation du mécanisme de Lucas et tente de la *reconstruire* pour finalement, dans un appendice démolir à la base sa propre reconstruction. Toutefois nous parviendrons ultérieurement à retrouver la substantifique moëlle de sa reconstruction.

*Le raisonnement de Benacerraf :*

Définitions

1) def: S = {p | je peux prouver intuitivement p}

2) def: S* = la fermeture de S (en logique du premier ordre avec identité)

En fait S* = {p | ⊡p}, avec ⊡ interprété par "peut savoir".

Voilà les hypothèses (en simplifiant un peu) liant le théorème de Gödel et la "limitation des machines" dans la reconstruction de l'argument de Lucas :

Hypothèses

1) hypothèse sur soi-même :

HS) *hypothèse de savoir* : je suis correct (ce que je prouve intuitivement est vrai).(⌜p⌝∈ S*)->p. Il s'agit donc d'un axiome de réflexion, ce qui justifie la notation ⊡⌜p⌝->p.

HI) *hypothèse de l'introspection*

⌜p⌝∈ S* => ⌜⌜p⌝∈ S*⌝∈ S* qui correspond à une règle de "nécessitation" : ⊢ p => ⊢ ⊡⌜p⌝ [18].

2) hypothèse sur les machines : essentiellement l'identification de base + la thèse de Post-Turing, c'est-à-dire que l'ensemble des théorèmes qu'une machine peut prouver est RE.

Benacerraf, au moyen d'une assez longue dérivation utilisant le premier et le second théorème d'incomplétude, démontre alors que les trois hypothèses suivantes, prises ensembles sont contradictoires, et il argumente que Lucas n'a pas démontré plus que ça.

a) ⌜Q ⊆ $W_j$ ⌝∈ S* ⊡(M est $\Sigma_1$-complète),

b) ⌜ $W_j$ ⊆ S* ⌝∈ S* ⊡(□ p->⊡p),

c) S* ⊆ $W_j$ ⊡p->□ p,

a) signifie que je peux prouver que j est une machine $\Sigma_1$-complète.

b) signifie que je peux prouver que je peux prouver les théorèmes que la machine j prouve.

---

18 J'écrirai simplement ⊡p ou □ p à la place de ⊡⌜p⌝ ou □⌜p⌝ lorsqu'aucune confusion n'est à craindre.

c) représente ce que Lucas tente de réfuter : que la machine j est capable de prouver tous ce que je prouve. Avec b) et l'hypothèse HS on a Wj = S*.

Lucas rend c) fautif, mais Benacerraf fait remarquer qu'à partir de la plausibilité de a), un mécaniste peut échapper à la réfutation de lucas en estimant b) fautif.

Il se pourrait donc qu'on ait à la fois $S^* \subseteq W_j$ et $W_j \subseteq S^*$ , dans ce cas nécessairement b est fautif et '$W_j \subseteq S^*$' $\notin S^*$.

On pourrait donc avoir $\boxdot p <-> \Box p$, la contradiction provenant de $\boxdot(\Box p -> \boxdot p)$. Dans ce cas, si je suis la machine j alors, ce que l'on a démontré c'est que je ne peux pas *savoir* que je suis la machine j :

$$\text{je suis la machine j} \Rightarrow \neg \boxdot(\text{je suis la machine j}) \quad (*)$$

En particulier une machine ne pourrait pas s'objectiver au point de savoir qu'elle est *telle machine particulière*. Benacerraf suggère que les machines sont psychologiquement limitées pour l'application du dicton Socratique "connais-toi toi-même" Cela correspond de près à la conclusion intuitive des paradoxes de la duplication où le fait d'être une machine permet de s'objectiver pratiquement et concrètement soi-même en se dupliquant (avec la possibilité théorique de la duplication qui est pourvue par 2-REC) tout en reconnaissant *de visu* (dirais-je) l'impossibilité de se reconnaître en tant que soi, aussi bien devant le cristal que devant sa reconstitution. De plus comme le mécanisme indexical est l'affirmation de l'existence d'un niveau (et donc d'une machine universelle) ou le fonctionnalisme est correcte, l'impossibilité de se reconnaître objectivement peut être ramenée au caractère nécessairement non constructif du choix du niveau de la duplication, c'est-à-dire au caractère non constructif du "$\exists$", comme dans la critique du raisonnement de Kalmar.

Hélas *le cochon était mort depuis longtemps*, Benacerraf annihile en effet sa propre reconstruction de Lucas en montrant dans une annexe que l'hypothèse du savoir et l'hypothèse de l'introspection sont déjà contradictoires[19] avec la consistance et le fait que la prouvabilité (qui est $\Sigma_1$-complète) soit de la machine est arithmétisable.

En fait la conclusion (*) contredit l'hypothèse d'introspection, et on peut dériver cette contradiction d'une façon générale, sans utiliser Gödel, ce que Benacerraf poursuit lui-même dans une annexe à la fin de son papier :

*Théorème* aucune entité arithmétisable ne peut être à la fois adéquate, vérifier l'hypothèse du savoir HS (T), et vérifier l'hypothèse de l'introspection HI (Nec).

---

[19] Myhill 1960 avait déjà remarqué cette contradiction (voir aussi Hanson 1971). En fait Gödel 1933 aussi.

*preuve*

Supposons que M soit arithmétisable, alors on peut utiliser le lemme de diagonalisation et il existe un p tel que M ⊢ p <-> ¬ □⌜p⌝, par introspection M ⊢ □(⌜p <-> ¬ □⌜p⌝⌝), par adéquation (et modus ponens) M ⊢ □⌜p⌝ <-> □⌜¬ □⌜p⌝⌝; par hypothèse du savoir M ⊢ □⌜-□p⌝ -> ¬ □⌜ p⌝, donc M ⊢ □⌜p⌝ -> ¬ □⌜p⌝, donc (cf la tautologie (p->¬ p)->¬ p), M ⊢ ¬ □⌜p⌝. Puisque M ⊢ p <-> ¬ □⌜p⌝, on a que M ⊢ p, et par introspection M ⊢ □⌜p⌝. Ainsi M ⊢ □⌜p⌝ et M ⊢ ¬ □⌜p⌝, donc M est inconsistante.

La reconstruction de Lucas par Benacerraf s'effondre. Il termine en espérant que l'on parvienne à reconstruire correctement sa reconstruction. Et Chihara 1972 rapporte que lors d'une réunion qui se tint un peu plus tard, Benacerraf aurait communiquer (oralement) qu'il ne croyait plus en une telle reconstruction.

Ceci est bien ennuyeux car il semble que cette argumentation, comme je l'ai suggéré, permet de comprendre, dans le cadre mécaniste, pourquoi une machine ne peut pas se reconnaître elle-même dans le cas où elle se rencontre, ce qui nous permet d'espérer éclairer les paradoxes de la duplication. Toutefois, Chihara 1972, et Reinhardt 1985, 1986 ont proposé des reconstructions de Benacerraf, et c'est ce à quoi j'aboutirai aussi.

*Définition* : J'utiliserai l'expression tomber dans le *piège de Benacerraf* pour les utilisations inadéquates de T et Nec dans une même dérivation. Notons que dans le théorème le lemme de diagonalisation n'est utilisé que pour produire une formule p telle que M ⊢ p <-> -□ p. En résumé (et plus généralement) les trois propositions suivantes sont contradictoires :

1) il existe p telle que M ⊢ p <-> -□ p
2) quel que soit p  M ⊢ □ p -> p
3) quel que soit p  M ⊢ p => M ⊢ □ p

Notre propos, en grande partie, va consister à reconstruire l'argumentation de Benacerraf sans tomber dans son piège[20].

*8°) Webb 1968, 1983*

Webb (1968) présente une critique de Lucas de type TI°. On peut d'une part sauver le mécanisme en abandonnant la thèse de Church, d'autre part en abandonnant la définition de "produire comme vraie" par la prouvabilité formelle (et donc diagonalisable)

20 Voir aussi Nelson 1982, 1987.

Webb, aussi bien dans son livre de 1980 que dans l'article de 1983, est très critique concernant Webb (1968). Il découvre que la thèse de Church est *plutôt* une alliée de l'incomplétude pour le mécanisme. La fermeture de P pour les diagonalisations (cf Kleene's overnigth), et donc la mécanisabilité de l'argument de la diagonale permet de réfuter Lucas. Aussi bien une machine que Lucas savent produire la proposition de Gödel d'une machine consistante adequate et correcte. Pour *savoir* que cette proposition est correcte, il faut savoir que la machine est consistante. Le mécanisme entraîne que ni Lucas, ni aucune machine ne peut déduire, à partir d'une présentation convenable d'une machine arbitraire, la consistance de celle-ci. L'argument diagonal étant mécanisable, c'est le bord entre l'entièreté de ce que les écoles du dedans peuvent produire comme vrai et la vérité (ontique) des écoles du dehors qui devient inéluctablement flou. La thèse de Church constitue ainsi un rempart, un ange gardien dit Webb, contre les arguments de type $TI^{-}$.

Chaque fois qu'on tente d'utiliser les résultats d'incomplétude pour nier l'hypothèse mécaniste : ***ou bien*** l'argument est intuitivement convaincant, mais dans ce cas avec la thèse de Church il est accessible à quelque machine, ***ou bien*** il est non constructif (comme lorsque Lucas demande au mécaniste de lui présenter une machine saine et consistante arbitraire), et dans ce cas Lucas ne parvient pas à convaincre qu'il *sait* vraie la proposition de Gödel, il sait seulement, comme Putnam 1960 l'avait remarqué, que <u>si</u> la machine est saine et consistante <u>alors</u> la proposition Gödélienne de la machine est vraie, mais cette proposition appartient aux preuve de la machine Löbienne (capable de prouver sa propre $\Sigma_1$-complétude).

On reconnaît l'argument de la duplication de soi (voir 1.3). Avec MEC-DEI, Lucas est duplicable, et le mécaniste peut présenter Lucas-1 à Lucas-2. Il est alors correcte de déduire du mécanisme que Lucas-1 (resp Lucas-2) sait quelque chose au sujet de Lucas-1 et Lucas-2, que Lucas-2 (resp Lucas-1) ne peut pas savoir, en l'occurrence : j'(Lucas) ai survécu en Lucas-1 (resp Lucas-2) et pas en Lucas-2 (resp Lucas-1). La situation est symétrique, et curieusement la seule façon pour Lucas de la rendre asymétrique est de prétendre savoir que l'autre est consistant (ou ici conscient). Mais MEC-DIG-IND, c'est-à-dire MEC-IND avec la thèse de Church (ou les principes de Gandy), interdisent à toutes machines de savoir cela sur elle-même et à fortiori sur leur duplicat relatif à un environnement ou une Gödélisation.

Webb reconstruit aussi Lucas directement à partir de l'analyse de Turing : Au lieu de comparer notre capacité en matière de preuves intuitives, Webb propose de comparer nos capacités en matière de décision du problème de l'arrêt, (restreint[21] à $\phi_x(x)$). Il s'agit donc encore d'une reconstruction de

[21] Ce n'est évidemment pas une restriction (cf 2.1 ou 2.2). De plus avec la thèse de Hilbert (et de Church), c'est équivalent à la capacité de prouvabilité intuitive. La thèse de Hilbert énonce que toute théorie est interprétable dans une théorie du premier ordre.

l'argument de Lucas. -*Donnez-moi une machine quelconque* (mais correcte !), je prétends que *je suis capable de la battre à la confrontation*[22] *de Turing-Lucas-Webb*, laquelle confrontation consiste à se mesurer sur le problème de l'arrêt. C'est le problème de décision pour K. Tout d'abord comme je sais que la machine est correcte, toutes les fois qu'elle décide de l'arrêt pour une machine x donnée, je peux suivre sa démarche et donc me convaincre du résultat. Ensuite, parce que je suis au moins à même d'imiter cette machine, je suis au moins aussi fort qu'elle dans le problème de la décision de K. A présent je vais vous démontrer que je suis *plus* fort qu'elle.

Sans perte de généralité on peut admettre que la machine M à laquelle je me compare suive le protocole suivant : M(x,y) affiche 1 si elle décide que $\phi_x(y)\Downarrow$, M(x,y) affiche 0 si elle décide que $\phi_x(y)\Uparrow$.

Je prétends alors, dit Lucas (réinterprété par Webb), être à même d'exhiber une machine m telle que je sais décider de l'arrêt (du non arrêt en l'occurrence) alors que M diverge sur cette machine m.

J'utilise les notations et les conventions de 2.2. Considérons une fonction F(z,x) telle que

$$F(z,x) = 1 \text{ si } \phi_z(x,x) = 0, \text{ et } \Uparrow \text{ (=10-GOTO-10) sinon}$$

F est partielle calculable, il existe donc un *r* tel que $F = \phi_r$. Par le théorème de la paramétrisation :

$$\phi_r(z,x) = \phi_{s(r,z)}(x)$$

Comme M est une machine, je peux lui assigner un code j telle que $M = \phi_j$. Je prétends que M diverge sur s(r,j), en effet $\phi_{s(r,j)}(s(r,j))\Downarrow <=> \phi_r(j,x)\Downarrow <=> \phi_j(x,x) = 0$. En particulier $\phi_{s(r,j)}(s(r,j))\Downarrow <=> \phi_j(s(r,j),s(r,j)) = 0$. Donc avec m = s(r,j), $M = \phi_j$ est piégée : $M(m,m) = 0 <=> \phi_m(m)\Downarrow$. Comme M est correcte, je sais que M va devoir diverger sur m.

Avec 2-REC on peut cependant construire la machine qui calcule son propre piège $\phi_m$

$$\phi_e(x) = \text{<code pour l'arrêt x>,}$$
$$\text{sauf si } x = e, \text{ dans ce cas elle sort } s(r,e)$$

L'assurance de Lucas provient donc du fait qu'il possède un algorithme pour piéger chaque machine. Son argument, avec la *thèse de Church* se

---

[22] Avec l'identification de base : TC <=> MEC (Post-Turing), on retrouve l'argumentation de Kalmar. En fait Lucas + Kalmar => arrêt de soi est absolument indécidable par *soi*. Ce qui est déjà illustré avec l'interprétation d'Everett de la mécanique quantique + MEC-DIG-IND.

retourne en faveur du mécanisme puisque la thèse de Church le rend machine-communicable. En 1.3 j'ai illustré comment une machine pouvait appliquer la réfutation de Lucas sur elle-même.

L'argumentation de Webb illustre que la thèse de Church est un vaccin contre les arguments contra-mécanistes reposant sur les résultats d'incomplétude. Il s'agit d'une situation générale respectant abstraitement le principe de Lao-tseu Watts Valadier : une communication d'une proposition p rendue "trop convaincante et précise" se retourne contre la proposition p.

Remarquons que le réfuteur de Lucas $\lambda xs(r,x)$ est le réalisateur intuitioniste de la formule classiquement fausse $\neg\forall x\,(p(x) \vee \neg p(x))$. Ceci est à comparer avec l'indéterminisme mécaniste où "je serai à Washington ou je serai à Moscou" est (classiquement) correct (pour autant que la duplication se fasse au niveau universel approprié), et où, cependant chacune des sous-propositions "je serai à Washington", et "je serai à Moscou" sont chacunes (absolument, intuitivement) indécidables. Pour un solipsiste, ou un élève d'une école du dedans, il n'est pas possible de croire que l'on puisse survivre dans *ce qui est pris pour un autre*, et le comportement de Jean Lefou est *apparemment* le comportement d'un fou.

*9°) Slezak (1982), Gunderson (1970)*

Le travail de Slezak (1982) est aussi une analyse détaillée des "erreurs" de l'argumentation de Lucas, de la confusion entre *produire comme vrai* et la *prouvabilité*, des confusions de niveau de langage. J'y reviendrai lorsqu'on aura développé une théorie du sujet, et abordé l'analyse modale du théorème de Gödel. Slezak observe aussi une confusion chez Lucas entre *type* et *token* : *s(r,j)* est défini uniformément à partir de *j*, et est donc basé sur une référence individuelle. Lucas pourrait de la même façon prouver qu'il n'est pas un homme. "Donnez-moi un homme", pourrait demander Lucas, et je vais vous prouver que je sais faire quelque chose qu'il ne sait pas faire, en l'occurrence "je peux regarder, sans aucun ustensile optique particulier, leur nuque ; comme je sais faire ça pour n'importe quel homme que l'on me présente, je ne suis aucun homme".

De plus on observe que cela montrerait seulement que Lucas est différent de chaque machine (et non pas supérieur) qu'on peut lui présenter, ou ici de chaque homme qu'on peut lui présenter.

Toutefois, avec MEC-IND, on peut présenter un homme à lui-même, en ce sens qu'on peut le dupliquer. Mais dans cette situation j'ai déjà argumenté que les deux copies sauront que l'autre ne peut pas se laisser convaincre par des justifications positives pour les propositions du genre "j'ai survécu (à Washington (ou à Moscou))". L'argument de Lucas fonctionne symétriquement et permet à chaque dupliqué de se distinguer de

l'autre, mais ne permet pas de communiquer le résultat à un tiers n'y surtout de conférer un statut supérieur à l'un des deux.

Bien sûr Lucas peut répondre qu'étant donné qu'il n'est pas une machine, il n'est pas définissable, encore moins duplicable. Et si on lui présente une machine-Lucas (correcte, adéquate et consistante) Il, Lucas, peut étudier cette machine et découvrir son *talon d'Achille*. La machine, elle, ne peut pas appliquer à Lucas cette procédure, puisqu'il n'est pas définissable. Mais une telle attitude revient à poser au départ la fausseté du mécanisme indexical.

Slezak étend son analyse au cogito de Descartes, présentant la certitude cartésienne comme produite par une autoréfutation de Lucas (j'y reviendrai dans 3.1).

*10°) Whiteley (cité par Hofstadter 1979)*

Hofstadter mentionne une formule due à Whiteley "Lucas ne peut pas produire cette phrase comme vraie", que l'on peut à nouveau désindexicaliser par double diagonalisation dans tout système ou "produire comme vraie' est définissable. Cela montre que "produire comme vrai" n'est pas définissable pour Lucas. La vérité humaine n'est pas humainement définissable. Avec le mécanisme cette formule est théorème : la vérité mécanique n'est pas mécaniquement définissable.

*11°) Rucker (1982)*

Il existe toute une série de résultats d'incomplétudes, qui à la différence de la démonstration de Gödel sont nécessairement non constructifs. Ces résultats sont dus en grande partie à Post et reposent sur sa notion d'ensemble *simple* et d'ensemble *immune*[23]. Ces notions ont été remis à l'honneur par Chaïtin qui leur a trouvé une interprétation élémentaire en terme de la théorie de la complexité et de l'aléatoire (voir Chaitin 1987, voir aussi Van Lambalgen 1989, Delahaye 1991). Je rappelle la définition d'une notion assez stricte d'aléatoire (déjà utilisée en 1.3) :

*Définition* une suite finie S, suffisamment longue est aléatoire s'il n'existe pas de programme capable de générer cette suite, programme dont la longueur (en binaire par exemple) serait plus courte que la longueur de la suite elle-même.

Une suite *infinie* est dite aléatoire si tous ses segments initiaux constituent des suites finies aléatoires.

Cette définition est assez forte. On peut montrer que les décimales de certains nombres irrationnels, comme $\pi$, vérifie de nombreux tests

[23] L'ensemble des propositions complexité(x) > complexité(moi) est immune (voir Van Lambalgen 1989). Je rappelle que les ensembles productifs P sont constructivement non RE en ce sens que pour chaque Wi inclus dans P, il existe un j appartenant à P et n'appartenant pas à Wi (de plus on peut le trouver à partir de i). Un ensemble est *immune* s'il est non-constructivement non RE, il possède une intersection finie avec tous les Wi.

statistiques pour le caractère aléatoire, mais π admet une définition algorithmique et n'est donc pas aléatoire (au sens de Chaitin).

*Théorème de Chaïtin* (version informelle) : une théorie (machine) descriptible avec *n* bits ne peut pas démontrer le caractère aléatoire d'une suite finie de *m* bits, avec m sensiblement plus grand que *n*.

*Démonstration* (informelle) Supposons qu'une théorie *T* de *n* bits soit à même de démontrer le caractère aléatoire d'un nombre, de *m* bit, avec *m* sensiblement plus grand que *n*. A partir de la théorie de *n* bits, on peut construire une machine dovetelleuse énumérant toutes les démonstrations de *T*, et s'arrêtant lorsqu'elle prouve le caractère aléatoire d'un nombre plus grand que *n+une-certaine-constante* (plus grande que la description de la machine dovetelleuse). Si cette procédure s'arrête sur un nombre, elle peut être vue comme une définition algorithmique de ce nombre, et la description de cette procédure est beaucoup plus petite que ce nombre. Il ne peut pas dès lors être aléatoire. Donc, si la théorie *T* est consistante, la machine dovetelleuse ne peut pas s'arrêter, et cela montre que la théorie *T* ne peut pas prouver le caractère aléatoire des nombres plus grands que la description de la dovetelleuse de son prouveur de théorèmes. De ce résultat on argumente aisément de l'indécidabilité du panmécanisme à partir du mécanisme. Il s'agit d'une forme très faible de solipsisme : une machine ne peut pas *prouver* que quelque chose est plus compliqué qu'elle-même.

*La réfutation de Lucas par Rucker*

La démonstration du résultat précédent repose sur un paradoxe qui, selon Russell, est dû à Berry. Considérons l'expression "le plus petit nombre qui n'est pas définissable en moins de 100 mots". Ce nombre semble être défini à présent. Toutefois on peut argumenter qu'il n'y aura vraiment un paradoxe qu'à condition de donner une définition précise du mot "définissable". Que penser alors du plus petit nombre qui n'est pas définissable par un programme LISP de longueur inférieure ou égale à, disons *n*.

Nous pouvons écrire un programme LISP, de longueur inférieure à *n* (s'il est plus grand il suffit de prendre un *n* plus grand) capable de calculer ce nombre : donc un tel programme ne pourra pas s'arrêter. En effet s'il s'arrête, il réalise la contradiction. Ici aussi, c'est l'existence des fonctions partielles qui permet de considérer la notion de *définissabilité* comme une notion bien définie quoique (nécessairement) non constructive.

Ce raisonnement montre que pour toute machine, de taille bornée, il existe un plus petit nombre que la machine ne sait pas nommer. Rucker montre alors que le raisonnement de Lucas ne fonctionne qu'à condition qu'il puisse nommer des nombres *arbitraires*, aussi grands qu'il le désire. Mais justement, si Lucas est une machine, il ne pourra être assuré de la

consistance de la machine qu'on lui présente qu'à condition que le nombre de Gödel de la description de la machine soit inférieure au plus petit nombre qu'il ne sait pas nommer. Brièvement, si Lucas est une machine, la validité de son raisonnement ne peut être garantie que pour la présentation de machines dont la description n'est pas beaucoup plus grande que sa propre description.

*Remarque :* on peut retrouver ces résultats plus directement en utilisant la fonction BB (voir 2.2).

*12°) Résumé de 2.3.1*

*Bien que la section précédente a déjà montré la non-trivialité de l'univers des machines, il est intéressant de regarder directement la portée de l'informatique théorique et des raisonnements diagonaux (avec la thèse de Church) sur l'hypothèse mécaniste. Une telle approche directe a commencé avec Post (10 ans avant que Gödel ne démontre les résultats d'incomplétude), a été abordée par Gödel lui-même, ainsi que Turing, et a été proposé à un large public de philosophes dans une publication de Lucas. Celui-ci use du théorème d'incomplétude pour démontrer sa propre supériorité sur les machines.*
*Lucas identifie machine et système formel. Cela est plausible au niveau où le fonctionnalisme est correct. Il identifie alors produire comme vrai par la machine et prouvable par la machine. Cela n'est pas plausible, parce qu'on peut définir une notion d'inférable par machine qui soit (au moins intensionnellement) différent de la prouvabilité. Ceci est abordé avec plus de détails plus loin. Avant d'arriver à cette forte réfutation, on peut chercher comme Benacerraf à voir ce qui peut être sauvé de la réfutation de Lucas. Je rappelle que Lucas raisonne de la façon suivante. Il admet d'abord qu'il est lui-même correct et consistant, grâce à quoi il se limite à se comparer avec des machines saines et adéquates. "Donnez-moi une machine correcte et adéquate quelconque, je peux produire comme vraie une proposition p que la machine ne peut pas produire comme vrai". Il s'agit de la proposition gödélienne p, obtenue par diagonalisation :*

*p <-> (p n'appartient pas à l'ensemble des propositions prouvables (productible comme vraie) par la machine)*

*Comme M est une machine, avec l'identification qu'il opère et avec la thèse de Church (qu'il utilise donc) p est une proposition vraie de l'arithmétique que la machine ne peut effectivement pas prouver. Donc conclut Lucas, je ne suis pas la machine M. Comme je peux faire ce raisonnement pour toutes les machines correctes et adéquates, c'est que je ne suis pas une machine.*
*Lucas a donc "démontré" CON -> ¬MEC (¬MEC-DIG). Priest a proposé un raisonnement similaire montrant MEC -> ¬CON, comme Priest admet MEC-DIG dès le départ, il conclut à la non-consistance de l'esprit.*
*J'ai présenté de nombreuses réfutations de la réfutation gödélienne du mécanisme par Lucas.*
*Webb, par exemple, commence par faire remarquer qu'il suffit de rejeter la thèse de Church pour sauver le mécanisme. Mais il est difficile, avec MEC-DIG, de nier la thèse de Church, comme on l'a vu à la section précédente. Plus tard Webb remarquera l'aspect algorithmique de la réfutation de Lucas. En effet l'assurance de ce dernier provient du fait qu'il exhibe un procédé (la diagonalisation gödelienne) lui permettant à partir d'une description de chaque machine M de construire la proposition indécidable. Avec la thèse de Church une telle activité, ainsi que son itération dans le transfini constructif, est programmable. Webb met en évidence une forme de LWV : tout argument convaincant et finiment*

*communicable contre le mécanisme se retourne logiquement, avec la thèse de Church, en faveur du mécanisme. On ne doit plus nier la thèse de Church pour sauver un mécanisme (qui aurait été difficilement digital), au contraire, l'effectivité de la diagonalisation fait de la thèse de Church un véritable ange-gardien (l'expression est de Webb) pour la philosophie mécaniste. Webb fait remarquer la confusion "extensionnel" et "intensionnel" de la consistance, ce qui nous rapproche des arguments antimécanistes basés sur les paradoxes de la duplication. Une autre réfutation immédiate est due à Slezak. Slezak, s'inspirant de Gunderson, met en évidence l'utilisation erronée de l'indexicalité. L'erreur de Lucas est équivalente à prouver que je ne suis pas un homme pour la raison que je peux observer, sans artifices optiques particuliers, la nuque de tout homme que l'on me présente, ce qu'aucun homme ne peut faire !*

*La réfutation la plus profonde, véritable reconstruction de l'argumentation de Lucas, est due à Benacerraf. C'est la plus intéressante concernant la duplication. Benacerraf montre que Lucas aurait démontrer qu'il est possible que "je sois une machine", mais que cela impliquerait que je ne peux pas savoir de quelle machine il s'agit. Il y aurait une égalité extensionnel non prouvable à partir d'une quelconque présentation intensionnelle, ce qui rejoint les paradoxes de la duplication. Malheureusement la reconstruction de Benacerraf elle-même est erronée. Sa reconstruction utilise en effet à la fois la nécessitation et la réflexion, avec un prédicat de connaissance arithmétisable, ce dont il montre lui-même, dans un appendice ajouté à son article, l'inconsistance. J'ai appelé cet usage "tombé dans le piège de Benacerraf". Diverses reconstructions de la reconstruction de Benacerraf, évitant le piège, sont proposées. Je mentionne celles de Chihara, Reinhardt, etc.*

*Je propose une telle reconstruction (c'est "l'avis des philosophes") dans la sous-section 2.3.3, "le connaissable" reposant sur une arithmétique épistémique due à Shapiro et Reinhardt, et je présente en 2.3.5, avec une interprétation arithmétique du stratagème, une reconstruction (c'est "l'avis des machines") accessible aux machines dans un sens naturel qui va être précisé.*

## *Biblio locale*


**ANDERSON** A. R., (Ed), **1964**, *Minds and Machines*, Prentice-hall, Inc., Englewood Cliffs NJ.

**ARBIB** M., **1964**, *Brains, Machines and Mathematics*, McGraw-Hill, 2ème éd. : 1987, Springer-Verlag, New-York.

**BAERTSCHI** B., **1992**, *Les rapports de l'âme et du corps, Descartes Diderot et Maine de Biran*, J. Vrin, Paris.

**BENACERRAF** P., **1967**, *God, the Devil, and Gödel*, The monist, vol 51, n° 1, pp 9-32.

**BOYER** D. L., **1983**, *J. R. Lucas, Kurt Gödel, and Fred Astaire*, Philosophical Quaterly, 33, (131), pp. 147-159.

**CODER** D., **1969**, *Gödel's Theorem and Mechanism*, Philosophy 44, pp. 234-237.

**CHAITIN** G. J., **1987**, *Information Randomness & Incompleteness*, 2ed Ed. 1990, World Scientific, Singapore.

**CHIHARA** C.S., **1972**, *On Alleged Refutations of Mechanism Using Gödel's Incompleteness Results. The Journal of Philosophy*. Vol LXIX, N° 17, september 21, pp. 507-526.

**DENNETT** D.C., **1972**, *Review of the book "The Freedom of the Will"* The Journal of Philosophy. Vol LXIX, N° 17, september 21, pp 527-531.

**DENNETT** D. C., **1972**, *Review of Lucas 1970a*, Journal of philosophy 69, 17, pp. 527-531.

**DODD** T., **1991**, *Gödel, Penrose and the possibility of AI*, Artificial Intelligence Review, 5, pp. 187-199.

**DUMMETT** M., **1963**, *The Philosophical Significance of Gödel's Theorem*, Ratio, vol. 5, pp. 140-155.

**FEFERMAN** S., **1988**, *Turing in the Land of O(z)*, in Herken 1988.

**GÖDEL** K., **1931**, *Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I*, Monatsh., Math. Phys., 38, pp. 173-98, traduit en Français dans Le théorème de Gödel, Seuil, Paris, pp. 105-143, 1989, aussi en Anglais dans Davis 1965.

**GÖDEL** K., **1933**, *Eine Interpretation des Intuitionistischen Aussagenkalküls*, Ergebnisse eines Mathematischen Kolloquiums, Vol 4, pp. 39-40.

**GÖDEL** K., **1936**, *On the Length of Proofs*, translated in Davies 1965, pp. 82-83.

**GÖDEL** K, **1951**, Bull. Am. Math. Soc., vol 58, 1952, p. 158. Voir Wang 1974 page 324.

**GÖDEL** K, **1972**, Communication orale à R. Rucker, voir Rucker 82 page 164

**GOOD** I. J., **1968**, *Human and Machine Logic*, The British Journal of Philosophy of Science, Vol. 18, pp. 144-147.

**GOOD** I. J., **1969**, *Gödel's Theorem is a Red Herring*, British Journal for the Philosophy of Science, 19, pp. 357-358.

**GUNDERSON** K., **1970**, *Asymetries and Mind-Body Perplexities*, in M. Radner and S. Winokur (Eds) : Minnesota Studies in the Philosophy of Sciences, Minneapolis, University of Minnesota Press.

**HANSON** W.H., **1971**, *Mechanism and Gödel's Theorems*, Brit. J. Phil. Sci. 22, 9-16.

**HERKEN** R. (ed), **1988**, *The Universal Turing Machine* A Half-Century Survey, Oxford University Press.

**HOFSTADTER** D., **1979**, *Gödel, Escher, Bach : an Eternal Golden Braid*, Basic Books, Inc., Publishers, New York.

**HUTTON** A., **1976**, *This Gödel is Killing Me*, Philosophia 6, pp. 135-144.

**KREISEL**, G., **1970**, *Church's thesis : a kind of reducibility axiom for constructive mathematics*, in Kino, A., Myhill, J., and Vesley, R.E. (eds.), Intuitionism and Proof Theory, Proceedings of the Summer Conference at Buffalo, New York, 1968, pp 121-150, North-Holland, Amsterdam.

**LEWIS** D., **1969**, *Lucas Against Mechanism*, Philosophy 44, 169, pp. 231-233.

**LEWIS** D., **1979**, *Lucas Against Mechanism II*, Canadian Journal of Philosophy 9, pp. 373-376.

**LUCAS** J. R., **1961**, *Minds, Machines and Gödel*, Philosophy, vol. 36, pp. 112-127. (aussi dans Anderson).

**LUCAS** J. R., **1968**, *Satan Stultified : A Rejoinder to Paul Benacerraf.*, The Monist, vol. 52, pp. 145-158.

**LUCAS** J. R., **1970**, *Mechanism: A Rejoinder*, Philosophy 45, 172, pp. 149-151. (Reply to Coder and Lewis 1969).

**LUCAS** J. R., **1970**, *This Gödel is Killing Me*, Philosophia 6, pp. 145-148.

**LUCAS** J. R., **1971**, *Metamathematics and the Philosophy of Mind*, a rejoinder, philosophy of sciences, XXXVIII, pp. 310-313.

**LUCAS** J.R., review of J.C Webb, 1980, The British Journal of Philosophy of Science, Vol. ?, pp. 441-444, ?.

**LUCAS** J. R., **1984**, *Lucas, Gödel and Astaire : A Rejoinder*, Philosophical Quaterly, 34, pp. 507-508.

**MYHILL J** , **1960**, *Some Remarks on the Notion of Proof*, Journal of Philosophy 57, pp. 461-471.

**NAGEL** E. et **NEWMAN** J. R., **1958**, *Gödel's Proof*, New-York university Press. Traduction française, 1989, La démonstration de Gödel, dans Le théorème de Gödel, Seuil, Paris, pp. 145-171.

**NELSON** R. J., **1982**, *The Logic of Mind*, D. Reidel, Dordrecht, Holland.

**NELSON** R. J., **1987**, *Church's Thesis and Cognitive Science*, Notre Dame Journal of Formal Logic. Vol. 28, N° 4.

**NEUMAIER** O.,**1987**, *A Wittgensteinian View of Artificial Intelligence*, in R. Born (Ed.), Artificial Intelligence The Case Against, Croom Helm, London & Sydney.

**PENROSE** R., **1988**, *On the Physics and Mathematics of Thought*, in Herken R. (ed), The Universal Turing Machine A Half-Century Survey, Oxford University Press.

**POST** E., **1921**, *Absolutely Unsolvable Problems and Relatively Undecidable Propositions : Account of an Anticipation*, in Davis 1965, pp. 338-433.

**PUTNAM** H., **1960**, *Minds and Machines*, Dimensions of Mind : A Symposium, Sidney Hook (Ed.), New-York University Press, New-York. Repris dans Anderson A. R. (Ed.),1964.

**PUTNAM** H., **1988**, *Representation and Reality*, A Bradforf Book, The MIT Press, Cambridge.

**PRIEST** G., **1987**, *In Contradiction*, Nijhoff International Philosophy Series, vol. 39, Martinus Nijhoff Publishers.

**RUCKER** R., **1982**, *Infinity and the Mind*, the Harverster press.

**SLEZAK** P., **1982**, *Gödel's Theorem and the Mind*, Brit. J. Phil. Sci. 33, pp. 41-52.

**SMART**, J. J. C., **1961**, *Gödel's Theorem, Church's Theorem, and Mechanism*, Synthese, Vol. 13, 1° 2, pp. 105-110.

**TORRANCE** T. F., **1969**, *Theological Science*, Oxford University Press, Trad. Franç. Jean-Yves Lacoste, Science Théologique, Presses Universtaires de France, Paris.

**TURING** A., **1939**, *Systems of Logic based on Ordinals*, Proc. London Math. Soc. 45, pp. 161-228. Aussi dans Davis 1965, pp. 155-222.

**TURING** A., **1948**, *Intelligent Machinery*, in Turing "1992", pp. 107-127.

**TURING** A., **1950**, *Computing Machinery and Intelligence*, Mind, Vol. LIX, N° 236. Aussi dans Anderson 1964.

**TURING**, A., "1992", *Mechanical Intelligence*, Collected Work of A. M. Turing, D. C. Ince (Ed.), North-Holland, Amsterdam.

**VAN LAMBALGEN**, M., **1989**, *Algorithmic Information Theory*, The Journal of Symbolic Logic, Vol 54, N° 4, pp. 1389-1400.

**VISSER** A., **1986**, *Kunnen wij elke machine verslaan ?*, in Geest, Computer, Kunst, Peter Hagoort & Rob Maessen (redactie), Stichting Grafiet, Utrecht.

**WANG** H, **1974**, *From Mathematics to Philosophy*, Routledge & Kegan Paul, London.

**WEBB** J. C., **1968**, *Metamathematics and the Philosophy of Mind*, philosophy of sciences, XXXV, pp. 156-178.

**WEBB** J. C., **1980**, *Mechanism, Mentalism and Metamathematics : An essay on Finitism*, D. Reidel, Dordrecht, Holland.

**WEBB** J. C., **1983**, *Gödel's Theorems and Church's Thesis : a Prologue to Mechanism*, Language, Logic, and Method, R. S. Cohen and Wartofsky (eds.), D.Reidel Publishing Company, 309-353.

**YU** Q., **1992**, *Consistency, Mechanicalness, and the Logic of Mind*, Synthese 90, pp. 145-179.

## 2.3.2 Le connaissable

*Brièvement*

*Ici j'analyse à nouveau l'argumentation principale réfutant l'argument de Lucas : la confusion entre vérité ou la connaissabilité et la prouvabilité formelle. L'analyse reposera sur la théorie de la connaissabilitée capturée par S4 et promue par des philosophes*[24]*, qui espère sauver à la fois l'*agent *intuitioniste, et la "réalité" classique. Une erreur dans l'argument :*

$$TI\ (S4) \Longrightarrow \neg\, MEC \qquad (TI^{-})$$

*est localisée. Il s'agit toujours d'une confusion entre MEC conçu comme* $\exists n \boxdot Fonc(n)$ *et MEC conçu comme* $\boxdot \exists n Fonc(n)$*. Le raffinement obtenu va permettre d'affirmer de façon plus précise :*

$$TI\ (S4) \Longrightarrow +MEC \qquad (TI^{+})$$

*La relation avec la duplication est à nouveau mise en évidence.*

---

*1°) Gödel 1933, McKinsey et Tarski 1948, Grzegorczyk 1964.*

Les axiomes naturels pour décrire une théorie de la connaissabilité ont déjà été introduits et motivés en 1.2. avec la théorie connue sous le nom de S4. Cette théorie appartenait à une suite de théories S1, S2, S3, S4 introduites par Lewis pour analyser la déduction. Lewis formalise S4 dans le langage du calcul propositionnel avec comme unique symbole supplémentaire un symbole *conditionnel* "$\dashrightarrow$" représentant la déduction. $\Box$ p peut être défini par $\top \dashrightarrow$ p et de même p $\dashrightarrow$ q peut être défini par $\neg\Diamond(p \,\&\, \neg q)$ ou $\Box(p \to q)$. La formalisation *modale* de S4 usant du carré est due à Gödel (Gödel 1933). Gödel n'a pas la même motivation que Lewis, il suggère que l'on puisse capturer formellement la notion de prouvabilité intuitioniste par un opérateur B, (que je note $\boxdot$ pour indiquer qu'il vérifie l'axiome T). De façon précise, Gödel prétend que le système S4, ayant pour axiomes (je rappelle)

| | |
|---|---|
| $\boxdot(p \to q) \to (\boxdot p \to \boxdot q)$ | K |
| $\boxdot p \to p$ | T |
| $\boxdot p \to \boxdot\boxdot p$ | 4 |

et pour règles d'inférence

| | |
|---|---|
| p, p->q => q | Modus Ponens |
| p => $\boxdot$ p | Nécessitation, |

---

[24] Voir le recueil d'articles édité par Shapiro (publié en 1985 par North-Holland) : *Intensional Mathematics*., voir aussi les articles de Reinhardt 1985, 1986.

permet d'interpréter les tautologies intuitionistes telles qu'elles sont capturées par le système de Heyting IL (celui-ci est équivalent au calcul propositionnel de Kleene avec le principe du tiers-exclu remplace par ¬p->(p->q) (voir 1.2)) :

$$IL \vdash A \iff S4 \vdash G33(A)$$

où G33(A) est une traduction de la formule propositionnelle intuitioniste A en une formule de S4 :

$$G33(p_i) = \boxdot p_i$$
$$G33(A \,\&\, B) = G33(A) \,\&\, G33(B)$$
$$G33(A \vee B) = G33(A) \vee G33(B)$$
$$G33(A \to B) = \boxdot G33(A) \to \boxdot G33(B)$$
$$G33(\neg A) = \boxdot \neg G33(A)$$

J'appelle la fonction G33 la traduction de Gödel[25].

Je conçois le sujet comme la source de sa connaissance. Il est pratiquement solipsiste en ce sens qu'il n'accède à rien d'autre qu'à cette source, il est le créateur ou le rêveur, éveillé ou endormi, de son univers mental[26]. Il étend ses connaissances de l'intérieur à l'instar des pythagoriciens des écoles du dedans. Voilà pourquoi, outre l'adéquation intuitive et linguistique de S4 pour la connaissance (déjà décrite dans 1.2), la capture par S4 du constructivisme intuitioniste suggère d'approfondir les intuitions de Brouwer au sujet du rôle de la subjectivité et du soi dans les fondements des mathématiques, et de reconnaître dans le processus de génération des connaissances un phénomène essentiellement mouvant et évolutif, comme une suite bifurquante de constructions. Le sujet ne s'approprie que ce qu'il contrôle, se mettant ainsi à l'abri de l'incontrôlable. La vérité, pour le sujet *solipsiste* sera la prouvabilité *intuitive*, conçue informellement de façon absolue, c-à-d incorrigible.

S4 capture cette équivalence dans le sens où

$$S4 \vdash \quad G33(A) \leftrightarrow \boxdot G33(A)$$

Nous verrons d'autres présentations de cette équivalence.

L'idée d'utiliser la logique intuitioniste pour décrire la connaissance, ou la façon dont la connaissance peut se développer, a aussi été proposée par

---

25 Pour être précis, j'emprunte une traduction due à McKinsey et Tarski 1948. C'est McKinsey et Tarski qui démontreront (au moyen de sémantiques algébriques) la proposition de Gödel 1933.

26 Ce que j'illustre d'avantage en 3.1. Notons que ce point de vue est controversé, *au moins* depuis Platon.

Grzegorczyk (1964). Bien que Grzegorczyk[27] utilise une sémantique -de la logique de Heyting- due à Beth, son idée est bien approximée par la sémantique que Kripke a donné de IL. Kripke a trouvé sa sémantique directement à partir de la sémantique qu'il avait obtenue pour S4. Les mondes possibles de S4 correspondent aux états de connaissance (états épistémiques) d'un sujet ou d'une collection de sujet.

De nombreuses formules de S4 ne sont pas l'image d'une proposition intuitioniste. Par exemple, d'une façon générale la formule

$$\boxdot(A \vee B) \rightarrow \boxdot A \vee \boxdot B \quad (*)$$

n'est *pas* un théorème de S4 et a priori, $\boxdot(A \vee B)$ n'est pas la traduction d'une formule intuitioniste. Il se peut que pour des formules particulières, comme $A = \top$, et $B = \bot$, la formule (*) soit un théorème[28] de S4. Intuitivement $\boxdot A \vee \boxdot B$ permet de représenter le "v" constructif. On a $S4 \vdash \boxdot A \vee \boxdot B$ entraîne $S4 \vdash \boxdot A$, ou $S4 \vdash \boxdot B$, et donc (par T), $S4 \vdash A$, ou $S4 \vdash B$. La proposition $\boxdot(A \vee B)$ représente ainsi, a priori, une proposition classique, affirmant l'existence d'une preuve intuitive de "pvq" où le "v" n'est pas nécessairement constructif.

De la même façon une formule comme T, $\boxdot p \rightarrow p$, est une hybride entre le constructif $\boxdot p$, et la proposition *ontique* p. Comme on s'intéresse ici autant aux propositions constructives qu'aux propositions non-constructives, S4 est plus approprié que IL pour décrire la connaissance d'un sujet (d'une machine) plongée dans une réalité classique (en première approximation). La philosophie de l'esprit tente de situer les relations entre le sujet et le monde, entre l'épistémisme et l'ontologie. A la différence de Dummett 1963, je ne pense pas que la logique intuitioniste puisse supplanter la logique classique. Je pense au contraire que la logique classique est encore le meilleur cadre pour découvrir l'intérêt de l'intuitionisme[29], surtout si on l'interprète comme une forme idéale d'épistémisme.

---

27 Grzergoczyk 1964 interprète la logique intuitionniste comme une logique de la recherche et de la vérifiabilité, proche, me semble-t-il, du positivisme de Wittgenstein-Malcolm. Avec la translation de Gödel 1933 (ou une autre), c'est un argument de plus pour le choix de S4 pour une sorte de prouvabilité intuitive, absolue, positiviste.

28 On peut le vérifier avec un démonstrateur de théorème de S4 (voir annexe 2). Le système donne une liste de contre-exemples comme réponses. Une formule est donc un théorème si le système rend la liste vide (appelée NIL). La formule $\top$ est représentée par (p->p) et le faux est représenté par (p & -p) où -p représente la négation de p. "bw" représente $\boxdot$. Il s'agit d'un système particulier S4grz qu'on motivera en 2.3.4; (ip un parseur permettant l'usage de présentations fonctionnelles infixées).

? (S4grzip '((bw (p v q) -> bw p v bw q)))

((P (F\# (((- Q) P))) (F\# (((- P) Q)))) (P (- Q) (F\# (((- P) Q)))) (Q (F\# (((- Q) P))) (F\# (((- P) Q)))) (Q (- P) (F\# (((- Q) P)))))

? (S4grzip '((bw ((p -> p) v (p & - p)) -> bw (p -> p) v bw (p & - p))))

NIL

On peut aussi s'en convaincre visuellement avec les modèles de Kripke à la façon de 1.2.

29 Ce qui est conforté par l'intrusion de l'intuitionisme en géométrie algébrique avec les topos de Grothendieck-Lawvere.

En 1933, Gödel avait déjà démontré ses résultats sur l'incomplétude et, après qu'il ait introduit S4 et affirmé l'interprétabilité de IL dans S4, il s'empresse de faire remarquer l'inadéquation de S4 pour décrire la prouvabilité formelle d'un quelconque système S. En effet, Gödel remarque *déjà* que par T et Nec, S4 $\vdash \Box(\Box p \to p)$, donc S4 $\vdash \Box(\Box \bot \to \bot)$, donc S4 $\vdash \Box \Diamond \top$, c-à-d S4 prouverait que S démontre la consistance de S en contradiction avec le second théorème d'incomplétude. S4, possède à la fois la règle "introspective" de nécessitation ainsi que l'axiome de réflexion (du savoir) ne peut pas décrire la prouvabilité *formelle* d'une théorie adéquate.

L'article de Gödel s'achève là. Gödel déjoue ainsi à la base l'essentiel du piège de Benacerraf.

Deux stratégies permettent ainsi de ne pas tomber dans *ce* piège. Nous pouvons supprimer de S4 soit l'axiome T (et garder la règle Nec). Dans ce cas on obtient K4, qui décrit correctement (mais pas complètement) la prouvabilité formelle de S. Dans ce cas les schémas modaux obtenus sont prouvables par S et S est *autoréférentiellement correct*. Ou bien supprimer Nec, et conserver T, dans ce cas, on obtient un système modal non normal, appelons-le $S4^{-}$, qui décrit correctement la prouvabilité du système, en incluant cette fois-ci des schémas modaux corrects (comme $\boxdot p \to p$), non prouvables par S.

*2°) Kaplan et Montague (1960), Thomason (1980)*

Selon Montague, qui était peut-être influencé par Quine qui n'aime pas trop l'usage de la logique intensionnelle (et modale), la raison principale de faire une théorie formelle de la connaissance est de pouvoir traiter la connaissance *extensionnellement*, avec un prédicat du premier ordre, et donc en garantissant le *principe d'identité de Leibnitz*. Autrement dit la connaissance devrait être a priori traitée syntactiquement par un prédicat opérant sur des phrases ou sur des nombres de Gödel de phrases.

Kaplan et Montague, indépendamment de Gödel 1933, et surtout Montague 1974 (voir aussi Montague 1962), qui cite Gödel 1933, raffinent le résultat de Gödel. Ils montrent de façon directe les limitations d'une approche extensionnelle (où $\boxdot$ est arithmétisable) de la connaissance ou des croyances.

*Montague* la théorie I, qui est la même que T, mais où l'application de la nécessitation est restreinte aux formules propositionnelles classiques est encore inconsistante. De façon précise, la théorie qui a pour axiomes :

| | |
|---|---|
| $\boxdot(p \to q) \to (\boxdot p \to \boxdot q)$ | K |
| $\boxdot p \to p$ | T |
| $\boxdot(\boxdot p \to p)$ | "Nécessitation-de-T" |

et comme règle :

si CP $\vdash$ A alors $\boxdot$A. ( = nécessitation restreinte aux formules propositionnelles classiques)

est inconsistante (si $\boxdot$ est définit avec un prédicat arithmétisable).
CP = calcul propositionnel classique.

*Thomason* Thomason généralise le résultat pour les "croyances". Je rappelle que pour les croyances, T n'est pas adéquat puisque l'on peut croire du faux, comme dans les rêves, ou simplement comme lorsqu'on est dans l'erreur. Toutefois la forme affaiblie de T suivante qui vaut :

$$\Box (\Box p \to p)$$

est naturelle pour les croyances : lorsqu'on est dans l'erreur, on ne le croit pas. Lorsqu'on est dans le rêve (nocturne), on ne le croit pas, à l'exception des rêves lucides (voir 3.1). De la même façon, l'axiome "déontique" D,

$$\Box p \to \Diamond p,$$

est naturel pour le *rationnellement-croyable* ou le *probable*. Lorsque l'on croit à une proposition, on ne croit pas dans la négation de cette proposition.

De façon précise, la théorie J qui admet comme axiomes

| | |
|---|---|
| $\Box (p \to q) \to (\Box p \to \Box q)$ | K |
| $\Box p \to \Box \Box p$ | 4 |
| $\Box p \to \Diamond p$ | D |

et comme règle, la règle de nécessitation, où $\Box$ est arithmétisable, est inconsistante. Ces deux résultats sont des raffinements de la remarque de Gödel, et constituent une variation sur le thème du piège de Benacerraf : que l'on restreigne l'usage de la nécessitation (Montague), ou que l'on affaiblisse l'axiome T en l'axiome D, l'opérateur de connaissance ou de croyance n'est pas arithmétiquement représentable.

On verra plus loin que ce n'est plus le cas si on abandonne l'axiome 4, ce qui est tout naturel pour l'immédiatement probable, comme il apparaît avec les expériences d'autoduplication.

On peut trouver des preuves concises des propositions de Montague et de Thomason dans Turner 1990.

En fait l'inconsistance est une conséquence directe de la représentabilité de $\boxdot$ ou de $\Box$ en terme de prédicat, qui permet alors l'application du lemme de diagonalisation.

En particulier, avec l'hypothèse mécaniste et l'identification de base, l'ensemble des propositions explicitement prouvables du sujet est un ensemble RE, donc il est définissable arithmétiquement, et le prédicat de prouvabilité vérifie d'office le lemme de diagonalisation. On serait à ce stade tenter de conclure que le sujet mécaniste est a priori non descriptible ni avec T ni avec les affaiblissements comme I et J proposés par Montague et Thomason. Cette conclusion mérite cependant d'être nuancée.

*3°) L'école intensionnelle (ou épistémique) Myhill 1960, 1985, Shapiro 1985, Reinhardt 1985 1986, et Goodman 1985, 1987, 1990*

Nous ne sommes pas très loin d'une reconstruction précise de l'argument de Lucas. En effet, à ce stade on pourrait prétendre que la connaissabilité □ de la machine, étant arithmétisable, ne peut pas vérifier l'axiome T, alors que T est naturel pour la connaissabilité ⊡ du sujet humain. Cette argument va se révéler incorrecte. Néanmoins, puisque S4 semble décrire naturellement le connaisseur (idéal, omniscient) humain, il est toujours possible d'étendre a priori une théorie (mathématique) au moyen d'un opérateur intensionnel (modal[30]) avec S4 décrivant axiomatiquement les axiomes et les règles de la connaissance intuitive. Dans ce cas, pour éviter de tomber dans le piège de Benacerraf, il suffit de distinguer la prouvabilité intuitive (informelle) représentée par ⊡ de la prouvabilité formelle de la théorie représentée par □ . C'est ce qui a déjà été suggéré en 2.1. A présent "⊡p" n'est plus une abréviation d'une forme avec prédicat du genre "B(⌜p⌝)". La théorie est ainsi enrichie par la description d'un sujet directement plongé dans la théorie elle-même par l'intermédiaire d'un opérateur modal. Cela permet de capturer les nuances intensionnelles ou constructives que l'on rencontre naturellement en mathématiques.

C'est pourquoi, à l'inverse de Montague et Thomason, et déjà rien que dans le champ limité de la connaissance arithmétique, Myhill, Goodman, Shapiro et Reinhardt (pour ne citer qu'eux) analysent directement les relations entre l'épistémique et l'ontique (Shapiro), ainsi que les conséquences philosophiques du théorème de Gödel en présence de la thèse de Chuch (Reinhardt) au moyen de la logique modale S4 pour représenter le constructif, l'intuitif ou l'absolu. Ils étendent à cette fin l'arithmétique de Peano avec les axiomes de S4 quantifiés (en logique du premier ordre, voir plus loin), et définissent l'arithmétique épistémique EA. Le "⊡" représente

---

[30] Le terme "modal" possède une signification moins générale que le terme "intensionnel". Intensionnel signifie que l'on tient compte de la forme (la présentation) de l'énoncé, de la théorie ou de la machine. Par exemple l'approche de la récursion présentée en 2.2 est intensionnel car on tient explicitement compte, non pas de la forme des programmes, mais *du fait* que les programmes (les machines) ont une forme relativement à un environnement universel (voir aussi Smith 1980, Royer 1987). Dans ce cas on ne peut plus appliquer le principe d'identité de Leibnitz. La logique modale est une façon de faire de la logique dans de tels contextes (dit opaques).

la prouvabilité intuitive, absolue, ou encore constructive, à un certain moment fixé dans le temps.

Myhill, en 1960, interprète le théorème d'incomplétude comme une preuve que pour tout système formel consistant il existe des inférences correctes que le système ne peut pas effectuer. Il argumente alors que l'étape majeure dans le développement des mathématiques sera l'introduction d'entités mathématiques intensionnelles capables de refléter, notamment, cette intrusion épistémique. Il introduit pour cela une notion (intensionnel), de prouvabilité absolue, rejoignant de cette façon l'anticipation de Post. Malheureusement il ne tient pas compte de ses propres remarques de 1952 selon lesquelles les résultats d'incomplétude sont des théorèmes de psychologie, ce qui est une conséquence de MDI.

Il critique en effet la notion d'extension transfinie des théories (machines), élaborée en 1939 par Turing et élaborée par Feferman[31] 1962, et en se référant au papier de Gödel de 1933, fait remarquer que la notion de preuve formelle ne peut pas être utilisée pour la notion de preuve absolue. Cette dernière doit vérifier l'axiome T ce que la preuve formelle ne peut pas prouvablement vérifier.

Myhill se rendra compte dans son papier de 1985 que cette remarque doit être nuancée. Je montrerai avec le stratagème dans la section suivante une façon simple de nuancer effectivement la distinction entre preuve formelle et preuve intuitive[32]. Je reviendrai explicitement sur cet aspect des choses à la fin de la section sur le stratagème.

De même Goodman (1984), estimant que les mathématiques commencent une nouvelle crise de fondement, argumente lui aussi en faveur de la prise en compte explicite dans le champ des mathématiques d'entités intensionnelles. Celles-ci sont typiquement décrites par l'intermédiaire de logiques modales . Voir aussi Goodman 1987. On observe ici ce que j'appelle un plongement du sujet dans l'objet. Goodman compare explicitement ce "plongement" avec le rôle de l'observateur en physique. Il se réfère à la fois à la relativité et à la mécanique quantique. Il présente la vision purement extensionnelle des mathématiques comme un réductionnisme (du genre de celui qui a été défendu par Quine qui aurait affirmé que *la logique modale a été conçue dans le pêché* (voir Boolos 1979). Goodman critique de la même façon le caractère abusivement extensionnel de la présentation habituelle de la théorie de la récursion. Il estime que les propositions intensionnelles comme la thèse de Church, mais aussi -c'est moi qui ajoute- comme le

---

[31] Voir aussi Feferman & Spector 1962, Kreisel 1972.

[32] De plus il veut des "énoncés" gödélisés, il garde T et K avec les propositions entre quote (ou gödelisée) et enlève 4 car il veut que dans □ ⌜p⌝, □ représente la prouvabilité intuitive et absolue, et p représente un énoncé arithmétique. p<-> □ ⌜p⌝ est donc interdit.

second théorème de récursion (cf Smith 1980, Royer 1987), ainsi que les questions sur la forme des expressions mathématiques et la longueur des preuves ont un rôle capital dans l'élaboration des mathématiques. Il critique aussi le constructivisme en tant que fondement exclusif des mathématiques :

> *... classical mathematics resembles classical Newtonian physics. It presupposes a single omnicient knower who plays no role in the theory itself. Alternatively, it is a theory to be thought of as true but not as known.*
>
> *Constructive mathematics, on the other hand, have emphasized the role of the knowing mathematician to the exclusion of the reality known.*

Les mathématiques épistémiques ou intensionnelles réconcilie l'ontique et l'épistémique et cela a priori indépendamment de l'hypothèse mécaniste. Goodman, comme Myhill, défend alors la théorie S4 pour capturer *mathématiquement* et *classiquement* le sujet mathématicien. ⊡p est interprété par p est sach***able*** connaiss***able***.. Le sujet épistémique dispose d'un environnement ontique classique ainsi que de ressources illimitées.

Goodman se réfère à Gödel 1933, sans dire pourtant que Gödel parle de l'intuitionisme, et il interprète le carré par une forme de prouvabilité abstraite. Et de fait le carré "⊡"est plus général que la prouvabilité intuitioniste. Pour Goodman le carré représente la prouvabilité informelle, intuitive, absolue, non-formalisable, le sujet mathématicien, etc. Notons que c'est cohérent avec la prouvabilité intuitioniste, du moins au sens de Brouwer. Paradoxalement l'aspect "sacrilège" de la formalisation de Heyting de la logique intuitioniste est moins grave dans ce contexte, puisque "⊡" n'y est plus arithmétisable.

Goodman, à la différence de Myhill 1960, mais à la façon de Myhill 1985, réalise que la prouvabilité informelle et formelle peuvent coïncider extensionnellement, mais ne peuvent pas coïncider intensionnellement. Ce point subtil sera clarifié avec l'usage du stratagème et le mécanisme. C'est un point délicat : en gros pour ne pas tomber dans le piège de Benacerraf, la "vérité" de cette identification ne peut être ni formellement justifiable, ni intuitivement ou absolument ou informellement justifiable.

C'est avec des motivations similaires que Shapiro[33] 1985 et Reinhardt 1985 introduisent l'arithmétique épistémique. La motivation de Reinhardt consiste en grande partie à réexaminer l'interprétation des phénomènes d'incomplétude en terme épistémique. Il aborde explicitement la question du mécanisme et aboutit à des conclusions similaires à celles que je vais tirer du stratagème. Quelques théorèmes de Reinhardt sont exposés plus loin.

---

[33] Voir aussi Shapiro 1981, 1989.

Shapiro interprète le carré "⊡" épistémique dans un sens large de connaissabilité, ou de preuve informelle, par un sujet ou par une communauté de sujet à un moment donné. C'est donc toujours le connaiss***able*** à un instant donné. Ce "connaissable" peut évoluer dans le temps (voir aussi Popper 1950). Il argumente qu'on peut l'interpréter aussi par le "vérifiable", ce qui peut encore être comparé à l'interprétation de la logique intuitioniste en terme de recherche scientifique comme celle de Grzegorczyk 1964.

Chez tous ces auteurs, les conséquences logiques des propositions connaissables sont connaissables, le sujet est *omnicient* (cf 1.2). Par la traduction de Gödel (Gödel 1933), S4 joue le rôle de pont entre l'épistémique (les intuitionismes, les constructivismes, d'une façon générale les écoles du dedans) et l'ontique (l'école du dehors) ; de même qu'elle est un pont entre ce qui peut être connu et le vrai. Comme l'intuitionisme, l'approche épistémique marque un retour aux notions de vérité comme Post l'a anticipé dans les années 20 (voir plus haut). A la différence de l'intuitionisme, l'approche épistémique ne privilégie pas le sujet aux dépens de l'ontologie. Cette approche marie ainsi deux philosophies traditionnellement opposées. Abordons de plus près l'arithmétique épistémique et l'usage philosophique qu'on peut en faire.

*4°)L'arithmétique épistémique (EA)*

Les axiomes et les règles de EA sont ceux et celles de PA + ceux et celles de S4, auxquels on ajoute les formules ⊡x, avec x étant un axiome de PA ou de S4. On ajoute encore l'axiome modal pour la quantification,

$$\boxdot \forall x A(x) \Rightarrow \forall x \boxdot A(x).$$

Remarquons l'absence de $\boxdot \exists x A(x) \Rightarrow \exists x \boxdot A(x)$, en écho à $\boxdot(A \vee B) \rightarrow \boxdot A \vee \boxdot B$. Cela permet de distinguer l'existence constructive et non constructive de la même façon que nous distinguons la disjonction constructive et non constructive.

Deux résultats

1) EA est consistante relativement à l'arithmétique de Peano (Shapiro 1985)
2) La thèse de Church épistémique (cf 2.1)

$$\boxdot \forall x \exists y \boxdot P(x,y) \rightarrow \exists z \boxdot \forall x (\phi_z(x)\Downarrow \,\&\, \boxdot P(x,\phi_z(x))$$

est consistante relativement à EA (et donc PA) (Flagg 1985, Goodman 1986).

*Reinhardt (1985, 1986)*

Reinhardt, parallèlement à Shapiro, et indépendamment de Myhill et de Goodman (qu'il ne cite pas, notons qu'il cite Gödel 1933) tente de capturer la prouvabilité informelle (intuitive, absolue) en étendant l'arithmétique classique de Peano-Dedekind avec S4. Comme dans la **double**[34]-anticipation de Post de **1941**, il opère ainsi le *retour à la vérité* et plonge, de la même façon que Myhill, le sujet dans l'objet.

Je présente à présent une série de théorèmes gravitant autour du phénomène de l'incomplétude et du mécanisme. Toutes les propositions démontrées dans la sous-section suivante seront (trivialement) démontrables dans la théorie épistémique, que je vais tirer du stratagème appliqué à l'autoréférence (en 2.3.3 et 2.3.4) et qui est une extension de S4.

La proposition suivante nécessite l'adéquation de la théorie de base, mais n'utilise pas la thèse de Post-Turing. Il correspond à la version relative habituelle du premier théorème d'incomplétude de Gödel.

*PROPOSITION 1* Si S est une formule avec une variable libre (représentant un système formel) telle que

$$\boxdot(S(\ulcorner p \urcorner) \to p)$$

alors il existe un énoncé q tel que $\boxdot q \ \& \ \boxdot \neg S(\ulcorner q \urcorner)$, q est intuitivement prouvable et on sait (prouver intuitivement) que q n'est pas formellement prouvable dans S. Comme cette situation inclut la prouvabilité formelle, je me permettrai d'abréger $S(\ulcorner p \urcorner)$ par $\Box p$. Notre hypothèse s'écrit alors

$$\boxdot(\Box p \to p)$$

et nous devons démontrer l'existence d'un q tel que

$$\boxdot q \ \& \ \boxdot \neg \Box q$$

Autrement dit, je sais que q est vrai, et je sais que S ne prouve pas q.

*Remarque* : on obtient le même résultat avec l'hypothèse

$$\boxdot(\Box p \to \boxdot p)$$

puisque $\boxdot p \to p$. En effet, on a :

---

[34] Double car en 1941 il rend compte de son anticipation de 1921, mais il anticipe encore dans les footnotes Gödel 1951, Lucas 1961, et Benacerraf, Myhill, Goodman, Reinhardt, etc.

| | | |
|---|---|---|
| 1 | ⊡(□ p->⊡p) | |
| 2 | □ p->⊡p | (par T sur 1) |
| 3 | ⊡p->p | T |
| 4 | □ p->p | CP |
| 5 | ⊡(□ p->p) | NEC sur 4 |

de telles dérivations ne seront plus détaillées, et les justifications seront résumées par *calcul dans S4.*

*preuve de la proposition 1*

Comme "□" (= S) est arithmétisable, on peut appliquer le lemme de diagonalisation pour obtenir un énoncé q tel que q <-> ¬ □ q, ceci est intuitivement prouvable (plus rigoureusement, ceci est pouvable dans PA, et dans EA on a :-

$$\boxdot(q \leftrightarrow \neg \Box q) \quad (*)$$

d'autre part, par instantiation de notre hypothèse :

$$\boxdot(\Box q \rightarrow q)$$

On tire de (*) :

$$\boxdot(\Box q \rightarrow \neg q).$$

Un peu de calcul dans S4 donne

$$\boxdot(\Box q \rightarrow (q \,\&\, \neg q)).$$

c-à-d

$$\boxdot(\neg \Box q)$$

par (*) et par S4-calcul, on a ⊡q, donc on a ⊡q & ⊡ ¬ □ q. QED.

Remarquons déjà que ce résultat, qui formalise dans S4 le caractère constructif du premier théorème d'incomplétude de Gödel est utilisé par Lucas dans sa réfutation de MEC, ou ⊡ représente ce que Lucas sait produire comme vrai (assimilé ici à la prouvabilité intuitive) et □ représente ce qu'une machine peut prouver.

*PROPOSITION 2* Si le prédicat "□" constitue une borne supérieure de la prouvabilité intuitive, c'est-à-dire, si "□" prouve tout ce que je prouve intuitivement, et qu'en plus je sais (prouver intuitivement, absolument) que tel est bien le cas, c'est-à-dire

$$\boxdot(\boxdot p \rightarrow \Box p)$$

alors il existe une proposition q absolument non prouvable, et je sais qu'elle est absolument non prouvable, c'est-à-dire :

$$\boxdot \neg \boxdot q$$

*preuve*

| | |
|---|---|
| $\boxdot(\boxdot p \rightarrow \Box p)$ | par hypothèse |
| $\boxdot\boxdot p \rightarrow \boxdot\Box p$ | par S4 |
| $\boxdot p \rightarrow \boxdot\Box p$ | par S4 |

de plus, il existe un énoncé q tel que $q \leftrightarrow \neg\Box q$, et $\boxdot(q \leftrightarrow \neg\Box q)$, donc

$\boxdot q \rightarrow \boxdot\neg\Box q$
$\boxdot q \rightarrow \boxdot(\Box q \,\&\, \neg\Box q)$
$\boxdot q \rightarrow \boxdot\bot$
$\boxdot q \rightarrow \bot$
$\neg\boxdot q$
$\boxdot\neg\boxdot q$

QED. Si "$\Box$" représente, en outre, un prédicat (de prouvabilité) d'une machine M (ou d'une théorie) adéquate, la consistance de cette machine (qui prouve tout ce que je prouve intuitivement) est absolument indémontrable. Autrement dit, l'hypothèse $\boxdot(\boxdot p \rightarrow \Box p)$ entraîne $\neg\boxdot\Diamond\top$.

On pourrait déduire à ce stade, comme Lucas, une réfutation du mécanisme. Si je suis consistant alors je ne suis pas une machine X, car cette machine obéirait à l'hypothèse de la proposition 2, je ne pourrais pas connaître (démontrer intuitivement, absolument) la consistance de cette machine, et, étant X, je ne pourrais pas connaître ma propre consistance. Ce point est délicat, même en admettant que je connais ma propre consistance l'argument est erroné comme je le montrerai plus loin.

La proposition 2 est proche d'une version absolue du second théorème d'incomplétude de Gödel, mais il se pourrait que $\Diamond\top$ soit simplement faux, ce qui rendrait la *preuve* de la formule $\neg\boxdot\Diamond\top$ triviale. Une machine inconsistante démontre (trivialement) tous les théorèmes que je sais démontrer intuitivement, et la consistance de cette machine est *absolument* indémontrable, simplement parce qu'elle est fausse (par T).

Reinhardt a cependant donné une preuve d'une version absolue du second théorème d'incomplétude de Gödel, c'est-à-dire une preuve de l'existence d'une proposition arithmétique *vraie* et absolument indémontrable.

Il utilise à cette fin la forme spéciale de thèse de Church qu'il appelle, (suite à une conversation avec Oswaldo Chateaubriand !) *thèse de Post-Turing* (j'en ai déjà parlé dans 2.1).

Il complémente ainsi l'approche de Kalmar 1959 (voir plus haut) en démontrant l'existence de propositions vraies *absolument* indécidables à partir d'une version de la thèse de Church.

*Définitions*

1) une propriété λnP(n) est (informellement) *décidable* ssi

$$\forall n(\boxdot P(n) \vee \boxdot \neg P(n))$$

2) une propriété λnP(n) est (informellement) *semi-décidable* ssi

$$\forall n(P(n) \rightarrow \boxdot P(n)$$

En liant fonction totale et ensemble récursif la thèse de Church peut s'écrire :

$$\forall n\ (\boxdot P(n) \vee \boxdot \neg P(n)) \rightarrow \{n \mid P(n)\} \text{ est récursif}$$

La thèse de Post-Turing est la thèse correspondante de la thèse de Church (plutôt de Kleene[35] pour être tout à fait précis) pour l'identification de la notion d'ensemble semi-décidable avec la notion d'ensemble récursivement énumérable :

$$\forall n(P(n) \rightarrow \boxdot P(n)) \rightarrow \{n \mid P(n)\} \text{ est RE} \quad \text{(PT)}$$

Dans une théorie adéquate, en l'occurrence EA, on peut écrire PT ainsi :

$$\forall n(P(n) \rightarrow \boxdot P(n)) \rightarrow \exists e \forall n(P(n) \leftrightarrow U(e,n))$$

où U(x,y) est équivalent à $y \in Wx$. (U est appelé prédicat universel de Feferman, cf Feferman 1960).

---

[35] En effet les étapes d'un processus émulable par une machine de Turing, indexées par un paramètre comme le temps ou le numéro de l'étape sont récursivement énumérables. Réciproquement l'énumération d'un ensemble RE est un processus émulable. (J'appelle *thèse de Kleene* la thèse selon laquelle les numérotations acceptables (ou LISP, etc.) émule tous les calculs possibles de fonctions partielles calculables. La thèse de Kleene généralise la thèse de Church).

Remplaçons P(n) par ⊡P(n) :

$$\forall n(\boxdot P(n) \rightarrow \boxdot\boxdot P(n) \rightarrow \exists e \forall n(\boxdot P(n) \leftrightarrow U(e,n))$$

Comme on a pour tout n $\boxdot P(n) \rightarrow \boxdot\boxdot P(n)$, PT est équivalent à

$$\exists e \forall n(\boxdot P(n) \leftrightarrow U(e,n)) \qquad \text{(Reinhardt 1985)}$$

La consistance de cette formule sera discutée plus tard. Montrons qu'elle permet d'obtenir une version élémentaire de la disjonction de Gödel.

*PROPOSITION 3* Il existe une formule q(x) avec une variable libre telle que

$$PT \rightarrow \exists e(q(e) \,\&\, \neg \boxdot q(e))$$

*preuve* (on remarquera la ressemblance avec le paradoxe de Russell, l'usage de la double diagonalisation). De plus la preuve fonctionne avec n'importe quel prédicat U vérifiant PT (pas seulement le prédicat universel de Feferman).

Soit en effet $q(x) \leftrightarrow \neg U(x,x)$.
Avec PT on a $\exists e \forall x(\boxdot \neg U(x,x) \leftrightarrow U(x, e))$.
Donc $\exists e(\boxdot \neg U(e,e) \leftrightarrow U(e, e))$.
Comme on a $\boxdot p \rightarrow p$, dans S4, on a $\forall e(\boxdot q(e) \leftrightarrow \neg q(e)) \,\&\, (\boxdot q(e) \rightarrow q(e))$, donc on a $\exists e(q(e) \,\&\, \neg \boxdot q(e))$.
Conclusion : $\neg PT \vee$ (il existe une proposition absolument indécidable).

C'est la disjonction de Gödel. On peut comparer ce résultat à la conclusion de Kalmar 1959. Reinhardt analyse Kalmar plus systématiquement dans EA.
Ce résultat est une version *absolue* du premier théorème d'incomplétude de Gödel. La version donnée ici est très générale, elle ne nécessite même pas l'adéquation de la théorie puisqu'on n'invoque pas le lemme de diagonalisation.

*PROPOSITION 4* Le schéma

$$\exists e \boxdot \forall x(\boxdot p(x) \leftrightarrow U(x,e))$$

est inconsistant.
*preuve*
(la preuve n'utilise pas $\boxdot p \rightarrow \boxdot\boxdot p$)
Comme on a $\boxdot\forall \Rightarrow \forall\boxdot$, on va montrer l'inconsistance de

$$\exists e \forall x \Box(\Box p(x) <-> U(x,e)) \qquad (*)$$

Soit $p(x) = \neg U(x,x)$ (c'est "$x \notin W_x$", le productif $K^c$, voir 2.2)
Comme $\Box$ est correct (on a $\Box p \rightarrow p$) on a:

$$\Box(\Box p(x) \rightarrow p(x))$$

et avec x = e :

$$\Box(\Box p(e) \rightarrow p(e)) \qquad (**)$$

De même avec le schéma (*), on a avec x = e

$$\Box(\Box p(e) <-> U(e,e))$$

c-à-d :

$$\Box(\Box p(e) <-> \neg p(e)) \qquad (***)$$

Remarquons que $(a \rightarrow b) \& (a <-> \neg b) \rightarrow (b \& \neg a)$ est une tautologie, et par distributivité de $\Box$, on a $\Box(a \rightarrow b) \& \Box(a <-> \neg b) \rightarrow \Box(b \& \neg a)$, avec (**) et (***) on obtient :
$\Box(p(e) \& \neg \Box p(e))$, donc $\Box p(e) \& \Box \neg \Box p(e)$. Une dernière application de T donne :

$$\Box p(e) \& \neg \Box p(e)$$

Ce qui est inconsistant. QED.

*5°) Analyse de la réfutation de Lucas*

Les précisions que Lucas apporte à l'usage traditionnel[36] du théorème de Gödel contre l'hypothèse mécaniste sont les suivantes : 1) il estime qu'il sait qu'il est consistant (sa propre consistance fait partie de ce qu'il peut produire comme vraie), 2) il réfute dialectiquement les hypothèses mécanistes. C'est-à-dire : pour chaque machine consistante qu'un philosophe mécaniste lui présente il prétend être à même d'exhiber, et donc de produire, un énoncé intuitivement vrai que la machine ne sait pas produire comme vrai.

Comme je décris la connaissance d'un quelconque sujet avec S4, sa première hypothèse est raisonnable. En effet $\neg \Box \bot = \Diamond \top$ est un théorème de S4. Avec sa deuxième hypothèse, il décide de ne se comparer qu'aux machines consistantes laissant le choix de celles-ci aux philosophes

36 Comme celle de Nagel et Newman (voir plus haut).

mécanistes. Il n'éprouve pas la nécessité de se comparer à une machine inconsistante estimant être différent de cette machine dès le départ (puisqu'il sait qu'il est consistant). Afin de ne pas sortir du cadre de l'arithmétique épistémique, je vais restreindre l'ensemble des propositions utilisées dans la comparaison à l'ensemble des propositions arithmétiques[37]. Ces propositions peuvent se référer à la description de la machine présentée par le philosophe mécaniste ou du prédicat de prouvabilité de cette machine, puisque ces descriptions sont arithmétiquement représentables.

Je rappelle et précise d'abord, avec S4, le lien entre l'identification de base, la thèse de Church (sous la forme de Post-Turing) et l'hypothèse mécaniste indexicale (voir 2.2).

Nous avons vu que PT est équivalent au schéma

$$\exists e \forall n(\boxdot P(n) <-> U(e,n)).$$

*Une stratégie*[38] *de Russel-Kripke-Reinhardt*

Considérons l'ensemble de nombres naturels de Gödel (ou l'ensemble des listes de Gödel, etc.) suivant

$$E = \{\ulcorner p \urcorner \mid \boxdot p\}$$

Etendons le langage de EA avec un terme E désignant l'appartenance à E : E(x) <-> x ∈ E. Pour les formules *p qui ne contiennent pas le symbole E*, on a, sans problèmes :

$$\boxdot(\boxdot p <-> E(\ulcorner p \urcorner)) \qquad (1)$$

Ceci permet de traiter ⊡ comme un "prédicat", et l' "extension" de ⊡ comme un ensemble de nombres naturels sans tomber, au moins formellement, dans le piège de Benacerraf. En effet, si le symbole "E" était utilisable dans p, on pourrait prendre pour p une formule diagonale comme q <-> ¬E(⌜q⌝), et on dériverait ⊡q <-> ¬⊡q. De plus cela ne limitera pas notre propos puisqu'on limite la comparaison de "moi" et la machine uniquement sur les propositions arithmétiques.

A présent, avec l'identification de base, MEC est équivalent au fait que E est RE.

Remplaçons P(n) dans PT par E(n), alors

$$\exists e \forall n(\boxdot E(n) <-> U(e,n)) \qquad (2)$$

---

[37] Voir Chihara 1972 pour une idée similaire. Il s'agit du "test de Turing" sur l'arithmétique informelle.

[38] Il s'agit d'une vieille stratégie pour éviter les pièges de la diagonalisation, ou de l'auto-référence. Elle fut remise à la mode par Kripke pour pouvoir construire une théorie de la vérité. Reinhardt 1985 s'en est inspirée pour traiter de la connaissance.

La machine *e*, dans ce cas démontre les mêmes théorèmes que "moi". Avec S4 (quantifié) on a

$$\boxdot\forall n\ (\boxdot P(n) <-> \boxdot\boxdot P(n)) \qquad (3)$$

On obtient, avec (1), (2) et (3) + un abus de notation

$$\exists e\forall p(\boxdot p <-> U(e,\ulcorner p\urcorner))$$

ou encore, en prenant "$\Box_e$" pour un prédicat de prouvabilité de la machine *e* :

$$\exists e\forall p(\boxdot p <-> \Box_e p)$$

Ce qui revient à dire que si je suis une machine, mon prédicat de prouvabilité (même intuitive) est arithmétisable

En résumé, et en simplifiant, encore, les notations :

$$\forall p\ \boxdot p <-> \Box p,$$

p est considéré ici comme un nombre naturel (un nombre de gödel) et c'est la stratégie de Russel-Kripke-Reinhardt qui nous permet d'identifier $\boxdot p$ avec $E(\ulcorner p\urcorner)$.

Je commettrai encore les deux abus de notation suivant :

$$\boxdot = \{x \mid \boxdot x\}$$
$$\Box = \{x \mid \Box x\}$$

x représentant des propositions arithmétiques.

Grâce à ces quelques abus, une version béhavioriste (voir 1.3) de l'hypothèse mécaniste indexicale, avec l'identification de base alliée à la thèse de Church est succinctement capturée par l'égalité *extensionnelle* suivante :

$$\boxdot = \Box$$

Venons-en à présent à quelques reconstructions de l'argument de Lucas.

1) Tout d'abord, n'est-il pas évident que $\boxdot \neq \Box$ ? Nous savons que $\boxdot$ obéit à T, et, si la comparaison est effectuée sur une machine adéquate et

consistante (comme fait Lucas), □ n'obéit à T (cf Gödel 1933). Mais ceci montre seulement que ⊡ et □ sont *intensionnellement* différents. Ce sont des descriptions différentes qui pourraient fort bien avoir la même extension, comme BW-Rosser, BW' et BW'' rencontrés plus haut.

Comme cet argument est important, je présente une autre version de la même *erreur*.

2) Si ⊡ = □, alors ◈ = ◊, ou l'égalité est extensionnelle[39]. En particulier on a ◈T <-> ◊T, donc on a ⊡(◈T <-> ◊T) et on a ⊡◈T <-> ⊡◊T. Mais si ⊡ = □, la machine prouve tous les théorèmes que je prouve et sa consistance est absolument indémontrable : ¬⊡◊T, donc on a ¬⊡◈T, ce qui contredit le fait que ma consistance fait partie des vérités que je sais produire (dirait-Lucas). Plus prosaïquement cela contredit notre analyse du sujet en terme de S4. Donc ⊡ ≠ □.

L'erreur, ici, peut être localisée dans le passage de la formule ◈T <-> ◊T à la formule ⊡(◈T <-> ◊T), il ne s'agit pas d'une application de la règle de nécessitation puisqu'on a pas prouvé ◈T <-> ◊T. Cette formule est essentiellement hypothétique, et la logique modale est introduite exprès pour raisonner dans des situations où le théorème de déduction de Herbrand n'est pas valable.

Gödel 1933 affirmait que ⊡ ne s'applique pas à la prouvabilité formelle. A strictement parler cette remarque peut engendré des confusions, et Myhill 1960 est tombé dans le panneau. Il est clair à présent qu'avec le mécanisme, la différence entre □ et ⊡ est purement intensionnel.

3) *Reconstruction de Benacerraf-Chihara-Reinhardt-Slezak-Wang*
(reconstruction que je désignerai par "BCR")

Je présente d'abord une reconstruction, que j'espère fidèle de l'argument *erroné* de Lucas.

| | | |
|---|---|---|
| *Hypothèses* | □p->p | (1 : la machine est correcte) |
| | ⊡p->p | (T : je suis correct) |
| *thèse* | ⊡ ≠ □ | (je = la machine) |

*démonstration* □ est arithmétisable (c'est une machine), donc il existe p tel que p <-> ¬□p est intuitivement (constructivement) prouvable, dès lors :

---

[39] L'égalité apparaissant entre deux connecteurs modales sera toujours considérée de façon extensionnelle sauf mention explicite du contraire.

| | |
|---|---|
| ⊡(p <-> ¬ □p) | (2) |
| ⊡(□p->¬ p) | |
| ⊡(□p->p) | par (1) : ERREUR !!! |
| ⊡(□p->(p & ¬ p)) | |
| ⊡(□p->⊥) | |
| ⊡¬ □p | (3) |
| ⊡p | par (2) |
| ¬ □p | par (3) et (T) |

mais, si ⊡ = □, ⊡p <-> □p ; contradiction.

Si la démonstration précédente avait été correct, elle aurait montré que l'ensemble de formules

{□p->p, ⊡p->p, ⊡ = □}

est inconsistant.

L'erreur de Lucas a été de déduire ⊡(□p->p) à partir de □p->p. On peut voir cette erreur aussi bien comme une nécessitation erronée, ou une application du théorème de déduction dans un contexte intensionnel. La démonstration devient correcte si au lieu de prendre comme hypothèse

□p->p

on prend, soit directement

⊡(□p->p)

soit encore l'hypothèse

⊡(□p->⊡p)

qui signifie que je sais (prouver) que je sais (prouver) tout ce que la machine sait (prouver).

Dans ces cas la reconstruction que j'ai proposée de la réfutation de Lucas démontre que l'ensemble de formules

{⊡(□p->p), ⊡p->p, ⊡ = □}

ou encore l'ensemble des formules

{⊡(□p->⊡p), ⊡p->p, ⊡ = □}

sont inconsistants. Nous voyons alors que l'hypothèse mécaniste n'est pas réfutée. L'égalité (extensionnelle) ⊡ = □ reste plausible, y compris avec les hypothèses de Lucas selon lesquelles "moi" (Lucas) et la machine à laquelle on se compare sont consistants. Cette réfutation de l'argument de Lucas est plus informative que les réfutations du genre de celles de Putnam qui rejettent soit l'autoconsistance du sujet et/ou de la machine.

*6°) Réfutation de Lucas et paradoxe de la duplication*

C'est au niveau où le fonctionnalisme est correcte, c'est-à-dire au niveau où je survis à la substitution de mes parties que je suis définissable où représentable en terme de machine. Dès lors, dire qu'il existe un niveau revient à dire qu'il existe une machine. La consistance de PT, sur laquelle j'ai promis de revenir (voir plus loin) entraîne la consistance de MEC,

$$\boxdot\exists e(\boxdot p <-> \Box_e p)$$

De même, la démonstration de la proposition 4, c'est-à-dire, la démonstration de l'inconsistance du schéma

$$\exists e \boxdot \forall x(\boxdot p(x) <-> U(x,e))$$

peut être étendue (avec ID, MEC et PT) en démonstration de l'inconsistance du schéma (voir Reinhardt)

$$\exists e \boxdot(\boxdot p <-> \Box_e p)$$

Autrement dit si je suis *une* machine, je ne peux pas savoir que je suis *cette* machine. On retrouve la conclusion de Benacerraf sous une forme directement applicable à l'analyse des problèmes posés par la duplication. Ce qui rejoint l'analyse de la confusion token/type proposée par Slezak. Lucas demande au mécaniste de lui présenter une machine *candidate à être lui*, et à partir de celle-ci il exhibe une proposition que cette machine ne peut pas savoir. Mais Lucas ne peut pas savoir cette proposition non plus, à moins qu'il ne sache à l'avance qu'il est cette machine. Bref, il doit faire confiance au mécaniste en ce qui concerne la proposition □p->p, de même que Monsieur D (voir 1.3) ne peut que faire confiance à son médecin, ou mourir (cliniquement). Or le mécaniste n'a qu'une prétention non constructive : comme avec la proposition de Kalmar, sa prétention est de la forme ⊡∃e..., avec un "∃" nécessairement non constructif (non intuitif, non absolu, etc.).

Lorsque Post dit (voir 2.3.1) :

> *The conclusion that man is not a machine is invalid. All we can say is that man cannot construct a machine which can do all the thinking he can.*

Il n'est pas tout à fait correct (s'il l'était, le translateur ne serait pas possible, mais avec 2-REC nous savons qu'il est *en principe* possible).

La correction à apporter est la suivante : ... "all we can say is that man cannot construct a machine which can *provably* do all the thinking he can". A moins qu'il ne sous-entendait le "provably" dans "construct".

Quoi qu'il en soit retenons qu'une machine saine consistante et -complète peut construire un double d'elle-même, mais elle ne peut pas se reconnaître, ni prouver qu'il s'agit d'un double, *réussi*, d'elle-même.

*7°) Résumé de 2.3.2*

*Si un prédicat (de connaissance) est arithmétisable, la nécessitation et la réflexion sont incompatibles, c'est une conséquence simple du lemmme de diagonalisation. Ainsi, La façon la plus simple d'éviter a priori de tomber dans le piège de Benacerraf, est d'utiliser pour mathématiser la connaissance, un opérateur modale étendant la logique de la théorie ou de la machine en question. On obtient une mathématique épistémique, on dit aussi une mathématique intensionnelle, en étendant la théorie avec les axiomes et les règles de S4. S4 est défendable d'un point de vue purement philosophique. C'est dans ce sens que cette sous-section étudie l'"avis des philosophes", ou le point de vue de la "connaissance" tel que des philosophes (de Platon à Shapiro, si on veut résumer grossièrement) l'ont isolé.*

*En effet, Shapiro, et indépendamment Reinhardt, mais aussi Goodman, Myhill et d'autres (voir le recueil d'article édité par Shapiro 1985) ont ainsi défini EA, Epistemic Arithmetics, qui est une extension de l'arithmétique de Peano. Le but est de pouvoir travailler en mathématique classique, tout en étant à même de dériver des preuves constructives ou épistémiques. Ces travaux reposent sur un petit article de Gödel de 1933 où il établit l'existence d'une traduction de la logique intuitioniste dans le système modale épistémique S4. Cette traduction permet de donner une interprétation épistémique de l'intuitionisme (plus généralement du constructivisme).*

*G33 de PL dans MPL est défini comme suit :*

$$G33(p_i) = \boxdot p_i$$
$$G33(A \,\&\, B) = G33(A) \,\&\, G33(B)$$
$$G33(A \vee B) = G33(A) \vee G33(B)$$
$$G33(A \rightarrow B) = \boxdot G33(A) \rightarrow \boxdot G33(B)$$
$$G33(\neg A) = \boxdot \neg G33(A)$$

*L'interprétation de Shapiro de la proposition modale* $\boxdot A$ *est*
*- La communauté peut en venir à connaître A à partir de ce qu'elle connaît déjà ;*
*- A est prouvable de façon informelle ;*
*- A est vérifiable*
*et l'analyse de Shapiro est de la même nature que celle que Grzegorczyk faisait de l'intuitionisme. On pense aussi à celle de Kolmogorov 1983.*

*Il s'agit donc d'une mathématique hybride ne négligeant ni l'ontologie classique, ni l'intuitioniste (interprété épistémiquement). Les raisonnements philosophiques basés sur le théorème de Gödel y trouvent un cadre naturel. Reinhardt y analyse en détail la réfutation de Lucas. Et ici j'ai proposé dans ce cadre une première reconstruction de la reconstruction de Benacerraf de l'argument de Lucas.*

*Je montre précisément où se trouve l'erreur dans la dérivation (formalisée) de Lucas ou ses hypothèses sont les suivantes :*

| | | |
|---|---|---|
| Hypothèses | $\Box p \rightarrow p$ | (1 : la machine est correcte) |
| | $\boxdot p \rightarrow p$ | (T : je suis correct) |
| thèse | $\boxdot \neq \Box$ | (je = la machine) |

*L'erreur consiste, comme dans la critique de la thèse de Church par Kalmar, à confondre une existence nécessairement non constructive avec une existence constructive. Notons que la thèse béhavioriste discutée ici n'use pas l'identification de base. Il s'agit d'un "test de Turing" limité à l'arithmétique informelle.*

## *biblio-locale*


**BOOLOS** G., **1979**, *The Unprovability of Consistency, an Essay in Modal Logic*, Cambridge University Press.

**CHIHARA** C.S., **1972**, *On Alleged Refutations of Mechanism Using Gödel's Incompleteness Results*. The Journal of Philosophy. Vol LXIX, N° 17, september 21, pp 507-526.

**DUMMETT** M., **1963**, *The Philosophical Significance of Gödel's Theorem*, Ratio, vol. 5, pp. 140-155.

**FEFERMAN** S., **1960**, *Arithmetisation of Metamathematics in a general Setting*, Fundamenta Mathematicae, XLIX, pp. 35-92.

**FEFERMAN** S., **1962**, *Transfinite Recursive Progressions of Axiomatic Theories*, Journal of Symbolic Logic, Vol 27, N° 3, pp. 259-316.

**FEFERMAN** S. and **SPECTOR** C., **1962**, *Incompleteness along Paths in Progressions of Theories*, Journal of Symbolic Logic, Vol 27, N° 4, pp. 383-90.

**FLAGG** R., **1985**, *Church's Thesis is Consistent with Epistemic Arithmetic*, in Shapiro 1985.

**GODEL** K., **1933**, *Eine Interpretation des Intuitionistischen Aussagenkalküls*, Ergebnisse eines Mathematischen Kolloquiums, Vol 4, pp. 39-40

**GOODMAN** N. D., **1970**, *A Theory of Constructions equivalent to Arithmetic*, in Kino, A., Myhill, J., and Vesley, R.E. (eds.), *Intuionism and Proof Theory* : Proceedings of the Summer Conference at Buffalo, New York, 1968, pp 121-150, North-Holland, Amsterdam.

**GOODMAN** N.D., **1984**, *The Knowing Mathematician* Synthese 60, pp. 21 - 38.

**GOODMAN** N. D., **1985**, *A Genuinely Intensional Set Theory*, in Shapiro pp. 63-79.

**GOODMAN** N.D., **1986**, *Flagg Realisability in Arithmetic*, Journal of Symbolic Logic, V. 51, N° 2, pp. 387-392.

**GOODMAN** N. D., **1987**, *Intensions, Church's Thesis, and the Formalisation of Mathematics*, Notre Dame Journal of Formal Logic, Vol. 28, N° 4, pp. 473-489.

**GOODMAN** N. D., **1990**, *Mathematics as Natural Science*, Journal of Symbolic Logic, Vol 55, N° 1, pp. 182-192.

**GRZEGORCZYK**, A., **1964**, *A Philosophicaly Plausible Formal Interpretation of Intuitionistic Logic*, Indagationes Math. 26, pp. 596-601.

**KAPLAN** D. and **MONTAGUE** R., **1960**, *A Paradox Regained*, Notre Dame Journal of Formal Logic, 1, pp. 79-90.

**KOLMOGOROV** A. N., **1983**, *Combinatorial Foundations of Information Theory and the Calculus of Probabilities*, Uspekhi Mat. Nauk 38, 4, pp. 27-36, 1983. In english : Russian Math. Surveys 38, 4, pp. 29-40.

**KREISEL** G., **1972,** *Which Number Theoretic Problems can be solved in Recursive Progressions on* $\Pi_1^1$*-Paths Through O?*, Journal of Symbolic Logic, Vol 37, N° 2, pp. 311-334.

**MONTAGUE** R., **1962**, *Theories Incomparable with respect to Relative Interpretability*, The Journal of Symbolic Logic, Vol 27, N° 2, pp. 195-211.

**MONTAGUE** R., **1974**, *Syntactical treatments of modality, with corollaries on reflexion* .

**MYHILL** J., **1952**, *Some Philosophical Implications of Mathematical Logic*, The review of Metaphysics, Vol. VI, N° 2.

**MYHILL** J , **1960**, *Some Remarks on the Notion of Proof*, Journal of Philosophy 57, pp. 461-471.

**MYHILL** J., **1985**, *Intensional Set Theory*, in Shapiro 1985, pp. 47-61.

**POPPER** K., **1950**, *Indeterminism in Quantum Physics and in Classical Physics*, Brit. J. Phil. Sci., 1, 2 and 3, pp. 117-33 and 173-195.

**REINHARDT** W.N., **1985**, *Absolute Version of Incompleteness Theorems*, Noûs, 19, pp. 317-346.

**REINHARDT** W.N., **1986**, *Epistemic Theories and the Interpretation of Gödel's Incompleteness Theorems*, Journal of Philosophical Logics, 15, pp. 427-474.

**ROYER** J. S., **1987**, *A Connotational Theory of Program Structure*, Lecture Notes in Computer Science n° 273, Springer-Verlag, Berlin.

**SHAPIRO** S., **1981**, *Understanding Church's Thesis*, Journal of Philosophical Logic, 10, pp. 353-365.

**SHAPIRO** S., **1985**, *Intensionnal Mathematics*, North-Holland.

**SHAPIRO** S., **1985**, *Epistemic and Intuitionistic Arithmetic*, in Shapiro 1985, pp. 11-46.

**SHAPIRO** S., **1989**, *Logic, Ontology, Mathematical Practice*, Synthese 79, pp. 13-50.

**SMITH** C. H., **1980**, *Applications of Classical Recursion Theory to Computer Science*, in F. R. Drake and S. S. Wainer (eds), Recursion Theory : its generalisation and applications, Proceedings of Logic Colloquium '79, Cambridge University Press.

**THOMASON** R. H., **1980**, *A Note on Syntactical Treatment of Modality*, Synthese, 44, pp. 391-395.

**TURING** A., **1939**, *Systems of Logic based on Ordinals*, Proc. London Math. Soc. 45, pp. 161-228. Aussi dans Davis 1965, pp. 155-222.

**TURNER**, R., **1990**, *Truth and Modality for Knowledge Representation*, Pitman UK.

**URQUHART** A., Review of *Intensional Mathematics*, Shapiro (ed.), Studia Logica L, 1, pp. 161-162.

### 2.3.3 Le théorème de Gödel et la logique modale

Jusqu'à présent, au sujet de la prouvabilité formelle B, ou de la machine □, seul le fait, remarqué par Gödel dans son papier 1933, que B ne vérifie pas la *prouvable réflexion* B(B->p), a été utilisé. Dans 2.2, on a vu, en clignant des yeux, que B vérifie les axiomes du système modale K4. Dans cette sous-section on va regarder la logique de B de plus près.

En effet, la logique modale est devenue un outil pour le métamathématicien. *La logique G*, ayant pour axiome essentiellement une version modale du *théorème de Löb* permet de capturer les conséquences des théorèmes d'incomplétude pour les systèmes formelles autoréférentiellement corrects.

*Brièvement*

> *Cette section est un prélude à "l'avis des machines" sur la question. On étudie la logique des communications finies et convaincantes des machines autoréférentiellement correctes.*
> *Brièvement l'autoréférence conduit à la logique G*[40], *mais aussi, cadeau inattendu, à une logique de la vérité G**
>
> *AUTORÉFÉRENCE. ===> (G,G*)*
>
> *"L'avis des machines" résultera d'une application du stratagème à la logique G.*
> ---

*1°) le morphisme de Magari-Boolos et G*

J'ai montré dans le chapitre précédant qu'une machine adéquate[41] et correcte vérifiait :

1) M ⊢ p => M ⊢ B(⌜p⌝) (L1)
2) M ⊢ B(⌜p⌝) & B(⌜p->q⌝) -> B(⌜q⌝) (L2)
3) M ⊢ B(⌜p⌝) -> B⌜(B(⌜p⌝)⌝) (L3)

J'avais dit qu'en clignant des yeux on reconnaissait le système normal modal K4. De façon plus précise, on peut construire une traduction du système modal K4 dans le langage de la machine (ou une théorie) M de la façon suivante. On considère une fonction F, appelée réalisation, qui assigne à chaque variable propositionnelle, $p_i$, un énoncé du langage de M.

Comme l'ensemble des variables propositionnelles modales et l'ensemble des énoncés d(u langage d)e M est récursivement énumérable une telle

---

[40] Appelée aussi L, GL, KW4, Prl, etc.

[41] Je rappelle que par *adéquate* j'entends que l'ensemble des théorèmes prouvés par la machine est récursivement énumérable, qu'il comprend les théorèmes de PA (à une traduction linguistique calculable près), et qu'il ne comprend pas le faux. Une machine ($\Sigma_2$-)saine, consistante et $\Sigma_1$-complète fait l'affaire.

fonction F peut être effective, mais les résultats qui nous intéresseront ne dépendront pas du choix de la réalisation F.

$$MB_F(p_i) = F(p_i)$$
$$MB_F(A \vee B) = MB_F(A) \vee MB_F(B)$$
$$MB_F(A \& B) = MB_F(A) \& MB_F(B)$$
$$MB_F(\neg A) = \neg MB_F(A)$$
$$MB_F(\Box A) = B(\ulcorner MB_F(A) \urcorner)$$

les formules $\neg A$, $\top$, $A \vee B$, $A \& B$, $A <-> B$, $\Diamond A$, sont considérées, ici, comme des abréviations de $A \to \bot$, $\neg \bot$, $\neg A \to B$, $\neg(A \to \neg B)$, $(A \to B) \& (B \to A)$, $\neg \Box \neg A$ respectivement[42].

Remarquons que si nous nous restreignons à l'arithmétique de Peano (ou son correspondant dans le langage de M). Nous obtenons une interprétation d'un opérateur modal $\Box$ par un prédicat, *B*, dans une théorie du premier ordre, en particulier dans l'arithmétique. Dans ce sens $\Box$ est arithmétisable, comme $\boxdot$ ne l'est pas.

*K4-correction de M (PA)*

On peut déduire de L1, L2 et L3 que K4 est correct pour l'arithmétique PA ou la machine M, en ce sens que

$$K4 \vdash A \Rightarrow \text{quel que soit F, } M \vdash MB_F(A)$$

En particulier : quel que soit F, $PA \vdash MB_F(A)$.

La prouvabilité formelle, capturée correctement par $\Box$, est une première approximation de la croyance communicable. Les choses crues étant identifiées aux propositions dérivables dans l'arithmétique du premier ordre. L'identification de base rend cette affirmation plus plausible encore pour la prouvabilité par une machine, et si l'ensemble des propositions prouvables par cette machine est fermée pour les conséquences logiques du premier ordre, et contient PA, K4 est une description axiomatique correcte pour ce que la machine peut prouver. On peut dire plus :

Rappelons qu'un système (théorie avec prédicat de prouvabilité B, logique modale avec $\Box$, machine démonstratrice de théorème avec un métadémonstrateur, etc) est autoréférentiellement correct (cf Smullyan 1985, 1987, Marchal 1991, 1992) si le système est autoréférentiellement approprié par rapport à son prédicat de prouvabilité (croyance, savoir) relativement à

---

42 Ceci afin de raisonner sur des formules plus simples, mais on aurait pu directement traduire $MB_F(A \& B) = MB_F(A) \& MB_F(B)$, $MB_F(A \vee B) = MB_F(A) \vee MB_F(B)$, $MB_F(\neg A) = \neg MB_F(A)$, etc...

un environnement universel (qui supporte l'éxécution de la machine, qui suit les règles d'inférence, etc.). On s'intéresse à ce qu'une telle machine est à même de prouver sur elle-même et K4 est une première approximation.

K4 $\vdash \Box A$ entraîne que K4 $\vdash$ A, (M, PA $\vdash$ $MB_F(A)$), et inversément. En ce sens K4 est lui-même autoréférentiellement correct.

C'est le caractère de correction autoréférentielle qui permet de lire K4 $\vdash$ p par K4 prouve p, et K4 $\vdash \Box p$, par K4 prouve que K4 prouve p. Ceci, *nous* pouvons encore l' interpréter par

K4 prouve "je prouve p"

en gardant néanmoins à l'esprit qu'il n'est pas évident, et l'article de Benacerraf jette certainement un trouble sur cette question, que K4 croit, prouve, sache, ou plus généralement, soit en relation épistémique particulière avec K4, c'est-à-dire avec lui-même.

Exemples de correspondance :

| | |
|---|---|
| a => $\Box$ a (Nec), et | $\vdash$ a => B(⌜a⌝) |
| prouvable $\Sigma_1$-complètude, et | $\Box$ a -> $\Box$ $\Box$ a. |

Je dérive la règle d'inférence dite de **M**onotonie **R**ationnelle :

$$A \to B \Rightarrow \Box A \to \Box B \qquad (\mathbf{RM})$$

Je n'utilise que K, et la règle N, cette règle vaut donc pour tous les systèmes normaux[43].

A->B => $\Box$(A->B), par N, et comme par K, $\Box(A \to B) \to (\Box A \to \Box B)$, on a, par modus ponens $\Box A \to \Box B$.

Le théorème de Löb,

$$M \vdash B(\ulcorner p \urcorner) \to p \Rightarrow M \vdash p$$

est interprétable modalement avec la règle

$$M \vdash (\Box p \to p) \Rightarrow M \vdash p$$

pour laquelle du reste K est fermée (à la différence de K4).

Mais la *formule L de Löb*

---

[43] Je dirai qu'un système est normal s'il admet une sémantique de Kripke. Dans ce cas on a au moins l'axiome K et la règle de nécessitation (voir 1.2).

$$\Box(\Box p \to p) \to \Box p \qquad (L)$$

correspond à une formalisation de la preuve du théorème de Löb dans l'arithmétique de Peano (dans le langage de la machine adéquate).

Le résultat suivant montre en effet que la preuve de Löb à partir du lemme de diagonalisation, est formalisable dans K4, et donc par PA ou M, pour toutes les réalisations F.

$$K4 \vdash \Box(q \leftrightarrow (\Box q \to p)) \to L.$$

En effet

K4 ⊢ □(q<->(□ q->p))->(□ q<->□(□ q->p)), par K
K4 ⊢ □(□ q->p)->(□ □ q->□ p), par K
K4 ⊢ □ q->□ □ q, par 4
K4 ⊢ □(q<->(□ q->p))->(□ q->□ p), par calcul propositionnel (CP) à partir des lignes précédentes,
K4 ⊢ □□(q<->(□q->p))->□(□q->□p), par la règle dérivée RM
K4 ⊢ □(q<->(□q->p))->□□q<->(□q->p)), par 4
De plus A->(B->A) est une tautologie, donc
K4 ⊢ □A->□(B->A) par CP et RM, on a encore, par K et CP
K4 ⊢ □A->(□B->□A), donc, en remplaçant A par □p->p, et B
par □q->□ p, on obtient :
K4 ⊢ □(□p->p)->(□(□q->□p)->(□p->p))
de façon similaire, on a encore,
K4 ⊢ □(q<->(□q->p))->(□(□q->p)->□q)
et en appliquant CP sur les lignes précédentes on obtient :
K4 ⊢ □(q<->(□q->p))->(□(□p->p)->□p) QED.

A présent, par la K4-correction de M, on obtient

$$M \vdash B(\ulcorner q \leftrightarrow (B\ulcorner q\urcorner \to p)\urcorner) \to (B\ulcorner(B\ulcorner p\urcorner \to p)\urcorner \to B\ulcorner p\urcorner)$$

Comme M vérifie le lemme de diagonalisation on peut trouver un énoncé q tel que

$$M \vdash q \leftrightarrow (B\ulcorner q\urcorner \to p)$$

et donc, par L1

$$M \vdash B(\ulcorner q \leftrightarrow (B\ulcorner q\urcorner \to p)\urcorner)$$

et par modus ponens

$$M \vdash (B^{\ulcorner}(B^{\ulcorner}p^{\urcorner}->p)^{\urcorner}->B^{\ulcorner}p^{\urcorner})$$

ce qui donne une sorte de *second théorème de Löb*. Il s'agit d'une formalisation du premier théorème de Löb dans l'arithmétique PA (ou dans M)

Cela signifie que pour toute réalisation F, et donc pour toute traduction t (dépendant de la réalisation F) on a

$$M \vdash MB_F(\Box(\Box p->p)->\Box p)$$

Je commettrai, quant le contexte le permet, l'abus de langage consistant à oublier "$MB_F$"[44].

J'écrirai par exemple :

$$M \vdash (\Box(\Box\bot->\bot)->\Box\bot)$$

*On retrouve le second théorème de Gödel.*

Comme $\Box\bot->\bot$ <=> $\neg\Box\bot$ <=> $\Diamond\top$, on obtient le second théorème de Gödel :

$$M \vdash (\Box\Diamond\top->\neg\Diamond\top)$$

Ce qui est aussi une instantiation du principe de Lao-tseu-Watts-Valadier. L'analyse sémantique de ce principe effectuée dans 1.2, nous permet d'affirmer que la formule L n'est pas une conséquence de K4, elle capture un aspect non trivial des conséquences du lemme de diagonalisation. La question naturelle à présent est de caractériser sémantiquement L. Nous savons déjà que le modèle doit être réaliste (voir 1.2), mais est-ce suffisant pour L ?

Avant de passer à la sémantique de L, je résume d'abord ce qui est acquis, et je propose une remarque.

*Définition* G est le système de logique modale normale ayant les axiomes K, 4, L. et comme règles MP, et N.

---

[44] On retient que si le symbole "$\vdash$", est précédé de M, ou de PA, ou d'un système modal auto-référentiellement correct, le symbole "$\Box$" s'interprète par l' énoncé arithmétique "je prouve" (énoncé par PA), ou par l'énoncé dans le langage de M. L'usage de "$\vdash$" est informel et appartient au métalangage, $B(^{\ulcorner}p^{\urcorner})$ ou $MBF(\Box p)$ appartient à la langue objet qu'on étudie (de PA, de M, etc.) et qu'on axiomatise modalement.

*G-correctitude de M*. on a montré

$$G \text{ prouve } A \Rightarrow \forall F \quad M \text{ prouve } MB_F(A).$$

*Remarque*

PA + con(⌜⊤⌝) est consistante, de même PA + ¬con(⌜⊤⌝) est consistante, mais G + ◊⊤ est inconsistante. En effet ◊⊤ = □⊥ -> ⊥. Par nécessitation on a □(□⊥ -> ⊥), par Löb, on a □⊥, et en utilisant à nouveau □⊥ -> ⊥, on a ⊥.

De même G + ¬◊⊤ est consistante, bien qu'elle n'est pas auto-appropriée[45], puisque G + □⊥ ⊬ ⊥. Cette différence de comportement reflète le fait que PA, (étant une logique du première ordre (non modale) vérifie le métathéorème de déduction, ce qui n'est pas le cas pour G (logique modale) qui est une logique modale avec règle de nécessitation. En l'occurrence, si $G^-$ désigne G sans la règle de nécessitation, $G^-$ + ◊⊤ est consistante, et $G^-$ + ◊⊤ ⊢ A entraîne que $MB_F(A)$ est vraie (pour toute réalisation F) bien que non nécessairement prouvable par l'arithmétique.

*2°) Sémantique de Kripke de G*

*Définition* R désignant une relation d'accessibilité (voir 1.2), une *échelle infinie* est une suite de mondes $a_0, a_1, a_2, a_3, \ldots$ tel que $a_0Ra_1Ra_2Ra_3R\ldots$

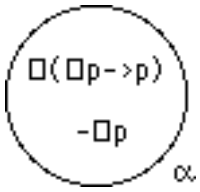
*Définition* R est bien chapeautée ssi il n'existe pas d'échelles infinies.

*théorème* (W,R) respecte les formules □p->□□p et □(□p->p)->□p ssi R est transitive et bien chapeautée.

(*remarque* cela ne veut pas dire qu'on est nécessairement mortel, mais qu'on n'est pas prouvablement immortel)

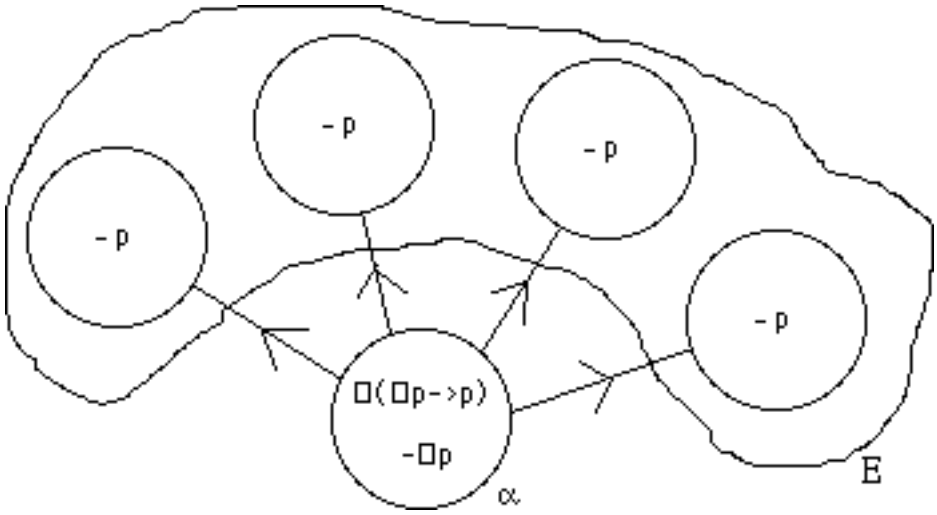
a. *preuve* (<=) Supposons R transitive et bien chapeautée, et supposons que (W,R) ne respecte pas □(□p->p)->□p (on a déjà montré que R transitive entraîne que (W,R) respecte □p->□□p.

Il existe donc un monde où □(□p->p) est vrai et □p est faux :

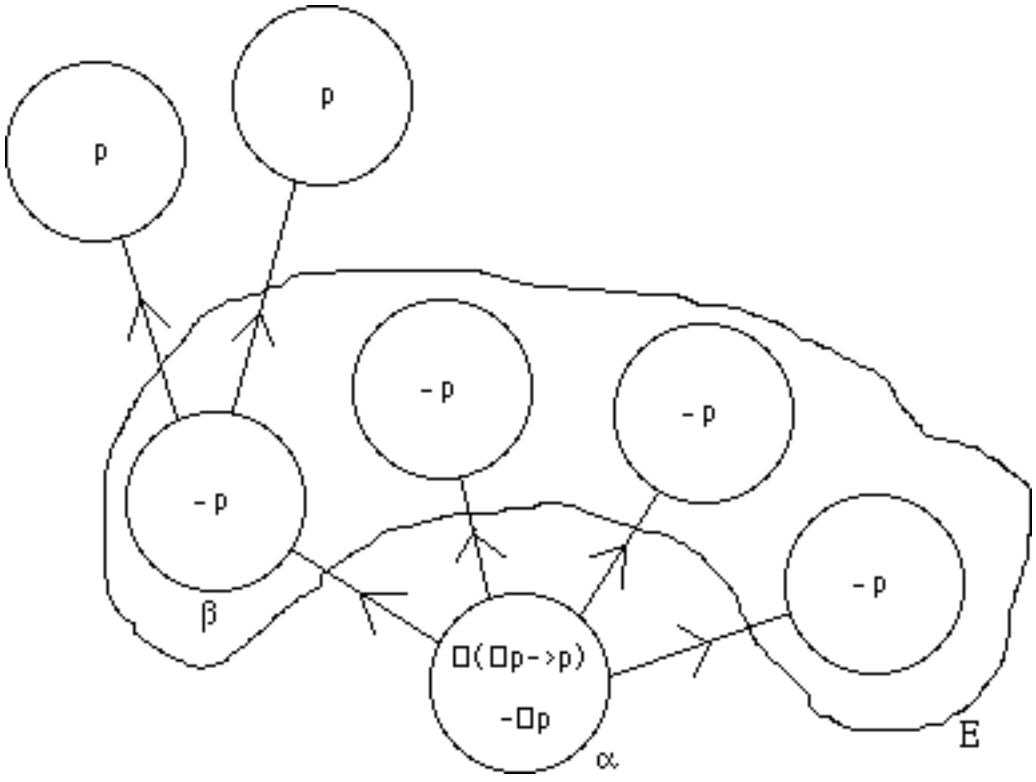



45 *Exemple* : dans la théorie {Ax1, Ax2} la formule F "il y a deux axiomes" est vrai, du moins concernant cette présentation intensionnelle, mais 1+2+F, n'est ni saine, ni auto-référentiellement correcte.

L'ensemble des mondes où p est faux et auquel α accède est non vide puisque p n'est pas prouvable-nécessaire dans α. Désignons par E cet ensemble :



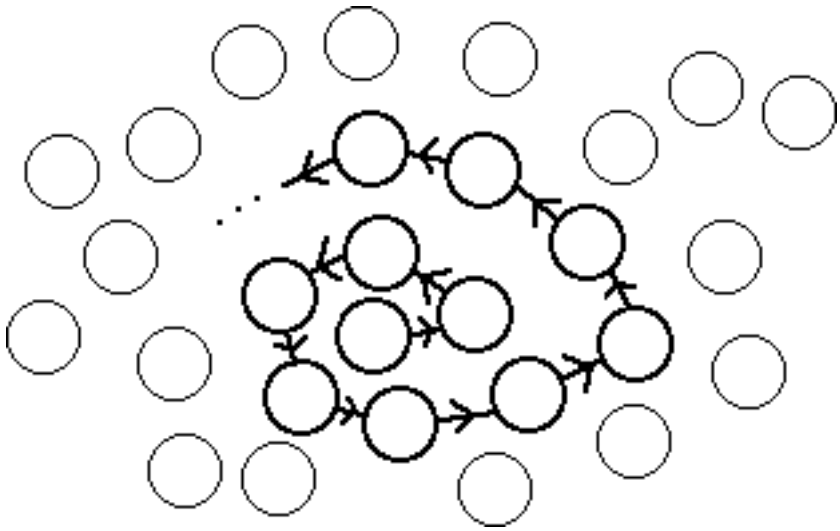

Je dis qu'un monde ξ est R-maximal dans un ensemble E si ξRψ -> y n'appartient pas à E. Comme R est bien chapeautée, il existe un monde R-maximal β dans E. (sinon E comprendrait une échelle infinie et R ne serait pas bien chapeautée). Si ce monde β est un dernier monde, □p y est vrai, si ce monde n'est pas un dernier monde, alors il accède à un ou plusieurs mondes qui ne sont pas dans E, puisqu'il est R-maximal. P est vrai dans ces mondes auxquels β accède, en effet si ¬p était vrai dans un tel monde, comme la relation est transitive, ils appartiendraient à E :



mais on a aRb et □(□p->p) est vrai dans a, donc □p->p est vrai dans b, mais on a vu que □p est vrai dans b, donc p est vrai dans b, en même temps que ¬p. Contradiction.

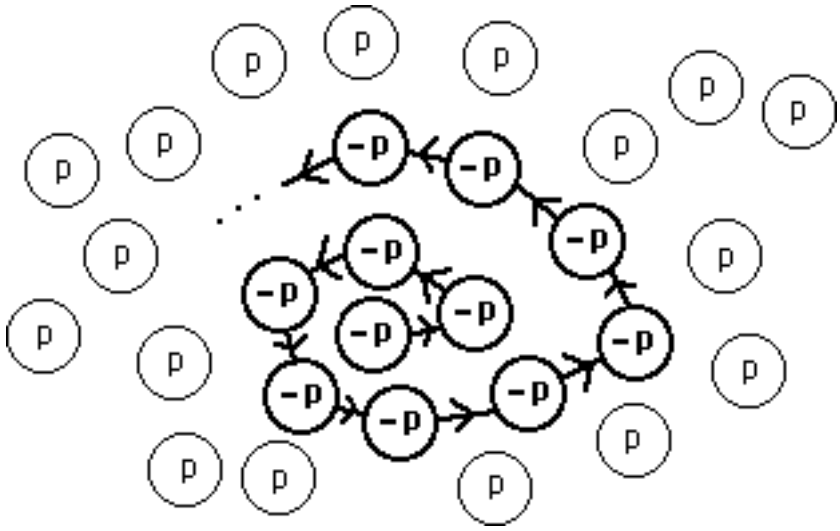
On remarque en particulier que les modèles de g sont réalistes. Cela illustre le fait que la formule de Löb est une généralisation non triviale de C (□◊p->¬◊p).

b. *preuve* (=>) supposons que (W,R) respecte □p->□□p et □(□p->p)->□p, il faut montrer que R est bien chapeautée. Supposons que R ne soit pas bien chapeautée. Il existe alors une échelle infinie, en gras dans le dessin :



je n'ai représenté que les flèches de l'échelle pour ne pas alourdir le dessin.

Considérons la valuation suivante : ¬p dans tous les mondes de l'échelle, p partout ailleurs.



Tous les mondes de l'échelle accèdent à (au moins un) monde où ¬p est vrai. Donc ¬□p est vrai dans tous les mondes de l'échelle. Donc □p->p est vrai dans tous les mondes de l'échelle (puisque □p y est toujours faux, rappelez vous de la définition de l'implication). De même □p->p est vrai dans tous les mondes qui ne sont pas dans l'échelle, puisque p y est toujours vrai. Donc □p->p est vrai partout, et dès lors □(□p->p) est aussi vrai dans tous les mondes. Mais le référentiel était supposé respecter la formule de Löb □(□p->p)->□p, donc □p est vrai partout y compris dans l'échelle. Comme, on l'a déjà vu, □p->p est vrai dans tous les mondes de l'échelle, p est vrai dans tous les mondes de l'échelle. Contradiction.

Ce théorème du respect s'étend, avec la notion de modèle canonique comme en 1.2, en théorème de complétude de G vis-à-vis des modèles transitifs et bien chapeautés. On peut l'étendre aussi en théorème de complétude de G vis-à-vis des modèles *finis* transitifs et irréflexifs. La propriété d'avoir des modèles finis garantit la décidabilité[46] de G.

Pour parler franchement cette complétude n'est pas aussi facile à démontrer pour G que pour les autres systèmes ; la raison est que le modèle canonique de G est irréflexif (voir Boolos 1979, voir aussi la deuxième édition de Boolos & Jeffrey 1974).

*3°) La preuve de Solovay*

En 1976, Solovay démontre la PA-complétude de G. Sa démonstration reste valable avec PA remplacé par M, une machine correcte et adéquate (donc RE), ou par n'importe quelle machine autoréférentiellement correcte :

$$\forall F\ M \vdash MB_F(A) \Rightarrow G \vdash A.$$

c'est-à-dire la réciproque de la M-correction. G constitue ainsi une formalisation correcte et complète des conséquences du théorème de Gödel prouvable par l'arithmétique de Peano. Et donc si la logique modale G ne prouve pas quelque chose, c'est que PA, ou la machine M, ne le prouve pas non plus (pour une réalisation F). Je sketche la démonstration de Solovay.

*preuve* (par contraposition):

Comme G a la propriété de complétude vis-à-vis des modèles finis. On peut limiter la recherche d'un contre-exemple dans les modèles bien fondés transitifs et finis. Cela entraîne que G est décidable. Notons qu'un

---

[46] De même que des méthodes classiques de logique modale permettent de montrer la propriété dite *du modèle fini* pour la plupart des logiques modales rencontrées ici. Voir annexe 2 pour un démonstrateur de théorème de G.

référentiel est bien fondé, fini et transitif si et seulement si il est fini, irréflexif et transitif.

On doit démontrer que $\forall F\ M \vdash MB_F(A) \Rightarrow G \vdash A$. On va démontrer la contraposée, $G \nvdash A \Rightarrow \exists F\ M \nvdash MB_F(A)$. Si $G \nvdash A$, il existe, grâce à la propriété de complétude vis-à-vis des modèles finis, un modèle de Kripke *fini* K (et donc avec une relation d'accessibilité R finie aussi, et irréflexive et transitive) avec un monde $a_0$ (qu'on peut supposer initial) tel que $a_0 \Vdash A$.

Comme K est fini il va pouvoir être manipulé de toute sorte de façons par M elle-même. On va *simuler* K dans M pour définir une réalisation particulière F, et donc une traduction $MB_F$ telle que $M \nvdash MB_F(A)$

Comme $MB_F$ est l'inconnue, je la note t, et je note $A^t$ pour $MB_F(A)$). La simulation reposera sur le second théorème de récursion (et sa formalisation repose sur le lemme de diagonalisation).

Sans perte de généralité, on peut identifier K à {1,...,n},avec $1= a_0$ (c-à-d : $1 \Vdash A$ Pour simplifier les notations on ajoute 0 tel que 0Ri avec $i \in \{1,...,n\}$.

On va construire une réalisation F et un prédicat A tels que pour $x \in \{1,...,n\}$ on puisse, B étant une formule modale quelconque, simuler le forcing "$\Vdash$" dans l'arithmétique. En particulier on a besoin de :

$$x \Vdash B \Rightarrow M \vdash A(x) \rightarrow \neg B^t$$

Le travail sera terminé si A est tel qu'on puisse prouver que M + A(1) est consistant. En effet dans ce cas $1 \nVdash A \Rightarrow M \vdash A(1) \rightarrow \neg A^t \Rightarrow$
$M + A(1) \vdash \neg A^t \Rightarrow M \nvdash A^t$

1) Définition du prédicat A

$$A(x) \Leftrightarrow \lim_{y \to \infty} H(y) = x$$

et H est définie par 2-REC, en terme de sa propre limite :

$H(0) = 0$
$H(x+1) = z$ si $H(x)\ R\ z$ et x+1 est une preuve de $\neg A(z)$
$= H(x)$ sinon.

Il n'est pas difficile de voir que H est non seulement totale calculable, mais que M sait prouver ce fait.

On peut comparer la fonction H avec le comportement d'un candidat mécaniste, mais *prudent* (voir 1.3) pour la translation ou la duplication. Il se

trouve dans le monde (l'état) 0, et il accepte de monter dans le translateur pour aller vers le monde (l'état) 1, sous réserve qu'il possède (codé par la description du monde où il va, en l'occurrence 1) une preuve qu'à partir de 1, le translateur ne va pas l'amener vers un monde z qui serait un dernier monde.

Les expériences par la pensée de 1.3, montre, avec MDI, qu'une telle preuve n'existe pas. Avec le mécanisme, on peut aboutir à tous les mondes accessibles tout en étant incapable de prouver qu'un seul de ces mondes est transitoire (et donc qu'on y a survécu). Intuitivement A(0) est vrai, mais si le candidat est "consistant" (dans le sens qu'il n'a pas de croyances fausses), l'ajout d'une "croyance" A(i), avec i entre 0 et n ne le rend pas inconsistant. Avec cette interprétation on peut lire A(i), comme je ne monte plus dans le translateur à partir de i, ou encore "je meurs en i" ou encore je m'attache à i, etc.

Solovay montre précisément qu'il en est bien ainsi avec un candidat dont les croyances sont les théorèmes de PA.

On démontre "aisément" les *propositions de base*

a) A(0)
b) en gros : M $\vdash$ A(x) $\leq$ n,
c) M + A(i) est consistant avec i $\in$ {0,...,n}
d) M $\vdash$ (A(x) & x > 0) -> B(⌜¬ A(x)⌝)

voir Solovay 1976, ou Smorynski 1985 pour une démonstration plus détaillée.

2) construction de t

Il suffit de définir F sur les propositions atomiques vu le morphisme de Magari-Boolos. Rappelons que K est fixé et 1 $\nvdash$ A

$$p^t = F(p) = \bigvee\{A(x) \mid x \in \{1,...,n\} \& x \vdash p\}$$

En terme de translation, $p^t$ signifie je "meurs" ou je reste bloqué dans un monde où p est vrai.

Avec la proposition 1, on sait que M + A(1) est consistant. Il reste donc à prouver le *lemme de simulation* ((1) est nécessaire pour prouver (2) par induction) :

(1) x $\vdash$ B => M $\vdash$ A(x) -> $B^t$

(2) $x \nVdash B \Rightarrow M \vdash A(x) \rightarrow \neg B^t$.

preuve de 1) (induction)

Le cas des formules atomiques

Si p est vrai en $x_0$ alors $p^f = \ldots v\ A(x_0)\ v \ldots$ qui est M-prouvablement impliqué par $A(x_0)$. De même si $x_0 \nVdash p$, c-à-d p n'est pas vrai en $x_0$, alors $A(x_0)$ est en contradiction avec chaque terme de $p^f$ puisqu'ils auront la forme A("quelque chose de $\neq$ de $x_0$ ") et que M prouve que F est une fonction totale.

Les cas B = C->D

$x \nVdash C \rightarrow D \Rightarrow x \vdash C \ \&\ x \nVdash D$,

on a (hyp. d'induction) $x \vdash C \Rightarrow M \vdash A(x) \rightarrow C^t$, et de même on a (hyp. d'induction) $x \nVdash D \Rightarrow M \vdash A(x) \rightarrow \neg D^t$, donc $x \nVdash C \rightarrow D \Rightarrow M \vdash A(x) \rightarrow (C^t \ \&\ \neg D^t)$, ce qui entraîne $M \vdash A(x) \rightarrow (\neg(C \rightarrow D))^t$

Reste le cas B = $\Box$C, et sa négation

pour $\Box C$ :

$x \vdash \Box C \Rightarrow \forall y\ (x R y \Rightarrow y \vdash C)$

$\Rightarrow \forall y\ (x R y \Rightarrow M \vdash A(y) \rightarrow C^t)$, (hyp. d'induction)

$\Rightarrow \&_{xRy}\ (M \vdash A(y) \rightarrow C^t)$

$\Rightarrow (M \vdash (\vee_{xRy}\ A(y)) \rightarrow C^t)$ (le passage du "et" au "ou")

$\Rightarrow M \vdash B(\ulcorner(\vee_{xRy}\ A(y)\urcorner) \rightarrow B(\ulcorner C^t\urcorner)$

or $M \vdash (A(x) \rightarrow \&_{-xRz}\ B(\ulcorner \neg A(z)\urcorner)$ (grâce à Prop de base b)).

donc $M \vdash (A(x) \rightarrow B(\ulcorner \vee_{xRy}\ A(y)\urcorner)$,

donc $x \vdash \Box C \Rightarrow M \vdash (A(x) \rightarrow B(\ulcorner C^t\urcorner)$

$\Rightarrow M \vdash (A(x) \rightarrow (\Box C)^t)$

Reste le cas B = $\neg\Box$C,

$x \nVdash \Box C \Rightarrow \exists y\ (x R y\ \&\ y \nVdash C)$

$\Rightarrow \exists y\ (x R y\ \&\ M \vdash A(y) \rightarrow \neg C^t)$, (hyp. d'induction)

$\Rightarrow \exists y\ (x R y\ \&\ M \vdash C^t \rightarrow \neg A(y))$

$\Rightarrow \exists y\ (x R y\ \&\ M \vdash B(\ulcorner C^t\urcorner) \rightarrow B(\ulcorner \neg A(y)\urcorner)$ (*)

et par la proposition de base 1 c), si xRy alors

$M \vdash A(x) \rightarrow \neg B(\ulcorner \neg A(y)\urcorner)$, ce qui avec (*) donne

$M \vdash A(x) \rightarrow \neg B(\ulcorner C^t\urcorner)$, donc

$M \vdash A(x) \rightarrow \neg(\Box C)^t$

Preuve de (2) voir Solovay 1975 ou Smorynski 1985.

*4°) un cadeau inattendu, G**

M est consistante, et la machine M est supposée être consistante aussi. Donc ◊⊤ est vrai pour M. On a vu cependant que G + ◊⊤ est inconsistant. L'inconsistance a été mise en évidence dans une dérivation qui utilisait la règle de nécessitation, en passant de □⊥->⊥ à □(□⊥->⊥), ce qui par Löb permet de déduire ⊥. Toutefois □⊥->⊥ (<=> ¬□⊥ = ◊⊤) est vrai pour M, et G + □⊥->⊥, pour autant qu'on abandonne la règle de nécessitation, décrit des propositions correctes concernant PA ou M, ou G. Il en est de même si l'on remplace la consistance □⊥->⊥ par la correction (soundness), c-à-d le schéma de réflexion □p->p. Comme tous les théorèmes de G sont correctes sur G (PA, M), et comme on s'est interdit d'utiliser la nécessitation, il y a intérêt à donner une présentation de G la plus exhaustive, et ensuite à dériver des nouveaux théorèmes avec □p->p (et sans nécessitation). Plus généralement l'idée est de s'interdire la nécessitation dans les dérivations où □p->p est utilisée. On évite ainsi explicitement le piège de Benacerraf. La théorie obtenue peut naturellement être axiomatisée par l'ensemble des théorèmes de G + le schéma □p->p, et ayant comme unique règle d'inférence le modus ponens. Cette théorie a pour nom G*, et elle est correcte dans sa description des schémas modaux vrais concernant M :

$$G^* \vdash A \Rightarrow \forall F\ MB_F(A)$$

Notons que G* n'est pas autoréférentiellement correct, on a à la fois G* ⊢ ◊⊤, et G* ⊢ ¬□◊⊤. G* parle de G (ou de M, PA, etc.). Seul G, M, PA, parle sur eux-mêmes, au sens de l'autoréférence correcte abordée ici.

Un *invraisemblable cadeau inattendu* est offert par Solovay sous la forme d'un second théorème de complétude, on a la réciproque :

$$\forall F\ MB_F(A) \Rightarrow G^* \vdash A.$$

Autrement dit, si la traduction d'une formule modale est vraie (arithmétiquement, ou dans le langage de M, etc.) alors G* la démontre.

*Sketch de la preuve*

J'appelle S(A) l'ensemble des sous-formules[47] de A, et $S^{\Box}(A)$ celles de la forme □x. Par exemples $S^{\Box}(◊⊤) = \{□⊥\}$. Pour rappel, ◊ est une abréviation de ¬□¬.

---

[47] On se souviendra qu'une formule figure parmi l'ensemble de ses propres sous-formules, ainsi S(□□p) = {□□p, □p}.

Si G* $\nvdash$ A alors G $\nvdash$ A, et il existe un modèle de Kripke K, fini semblable à celui considéré dans la preuve précédente, en particulier 1 $\Vdash$ A. On reprend la construction de F et de t de la même façon. Mais on a besoin d'un autre lemme de simulation, que je ne démontrerai pas (voir Solovay 1976). J'appelle *réflexion d'une formule* A, la formule $\Box$A->A.

*Lemme de simulation bis*

Soit B une formule modale telle que la réflexion de chaque sous-formule $C \in S^{\Box}(B)$ est vraie dans le monde 1, c-à-d, 1 $\Vdash$ $\Box$C->C. Dans ce cas, si D est une sous-formule quelconque de B alors :

(1) 1 $\Vdash$ D => M $\vdash$ A(0) -> $D^t$
(2) 1 $\nVdash$ D => M $\vdash$ A(0) -> $\neg D^t$.

*Définition* Soit * une fonction de l'ensemble des formules modales dans l'ensemble des formules modales.

*(A), que je note plutôt A*, est définie par la conjonction de la réflexion de toutes ses sous-formules (de la forme $\Box$x) :

$$(\&_{B_i \in S^{\Box}(A)} \Box B_i \text{->} B_i) \text{->} A$$

On peut alors démontrer le résultat suivant, dont on va tirer la complétude de G* pour la vérité *sur* G (M, PA).

G $\nvdash$ A* => il existe une réalisation F telle que $MB_F(A)$ est fausse.

En effet, si G $\nvdash$ A*, il existe un modèle de Kripke K tel que 1 $\Vdash$ $(\&_{B_i \in S^{\Box}(A)} \Box B_i \text{->} B_i)$ et 1 $\nVdash$ A.

Dans ce cas, le lemme de simulation bis permet de tirer

$$M \vdash A(0) \text{->} \neg A^t.$$

Or, par la proposition de base 1 a), A(0) est vraie, et M étant saine, $A^t$ est fausse.

Remarquons à présent que G $\vdash$ A* => G* $\vdash$ A, en effet si A* est un théorème de G, A est un théorème de G*, puisqu'il est conséquence d'une conjonction finie de réflexions, et qu'on dispose dans G* des théorèmes de G, des réflexions et du modus ponens.

Et ceci démontre le théorème de complétude, car à présent si G* $\nvdash$ A, G $\nvdash$ A*, et on vient de montrer que dans ce cas A est fausse.

Remarquons encore que G* ⊢ A => G ⊢ A*, en effet, si G ⊬ A*, G* ⊬ A* puisque G* étend G, donc A* est fausse, donc A est fausse, donc, puisque G* est correct, G* ⊬ A. On a donc

G* ⊢ A <=> G ⊢ A*.

Comme G est décidable, G* est décidable, et G*\G est décidable. Ainsi l'ensemble des schémas modaux vrais, mais non prouvable par M est décidable[48].

Ceux qui veulent réfuter Lucas n'espérait pas tant, G*\G, l'ensemble des schémas propositionnels modaux, vrais mais non prouvables par M, est, avec TC, un ensemble récursif. On peut construire une machine capable d'être autoréférentiellement correcte tout en étant à même de produire comme vraie, sans les prouver et surtout sans que la machine ne prétende les prouver, les schémas modaux corrects et non-prouvables qui la concerne. Il est nécessaire, pour produire du vrai non prouvable, d'éviter le piège de Benacerraf en interdisant la présence simultanée de la nécessitation et de la réflexion dans une même dérivation. Avec le second résultat de complétude, nous savons que cela est suffisant. J'argumenterai en particulier que la décidabilité de G*\G rend G*\G (auto)-inférable inductivement.

*Remarque :* Ona G* ⊢ ◊⊤ et G* ⊬ □◊⊤ , G* n'est donc pas fermé pour la nécessitation et G* n'est donc pas une logique modale normale. En particulier il n'y a pas de sémantique de Kripke. On verra plus loin une *variante* de la sémantique pour G*, due à Boolos. Cette variante illustrera le caractère limite-inférable de G* (et de G*\G).

*5°) Le théorème du point fixe*

Le comportement des énoncés, ou des machines, de Gödel, Henkin, Rogers, etc. sont formalisables dans l'arithmétique et dans G :

| | |
|---|---|
| G ⊢ □(p <-> ¬□p) -> □(p <-> ◊⊤) | Gödel-1931 |
| G ⊢ □(p <-> □p) -> □(p <-> ⊤) | Löb-1955 |
| G ⊢ □(p <-> ◊p) -> □(p <-> ⊥) | Rogers-1967 |

Ces résultats sont des cas particuliers d'un théorème général du point fixe dû à de Jongh et Sambin[49].

---

48 Voir en annexe un démonstrateur de théorèmes pour G et G*.

49 Ce résultat généralise un résultat de Bernardi et Smorynski (voir Boolos 1979, Smorynski 1985).

*Définitions*

1) On dit qu'une formule modale A(p, q,...) est modale en p, si les occurrences de p apparaissent dans le champ d'un opérateur modal □.

2) Un point fixe d'une formule modale A(p, q, ...) est une formule H(q, ...), notée simplement H, dans laquelle la variable propositionnelle p n'apparaît pas, telle que H <-> A(H,q, ...) est prouvable dans G. A(H, q, ...) représente la formule A(p, q, ...) avec p remplacé par H.

*théorème du point fixe* si A est modale en p, alors A possède un point fixe H. H peut être déterminée algorithmiquement. De plus

$$G \vdash \Box(p \leftrightarrow A(p, q, ...)) \rightarrow \Box(p \leftrightarrow H)$$

Je donne encore un exemple[50] (qui sera utilisé à deux reprises dans la suite, cf □p&◊p peut être abrégé par ¬⧈p) :

$$G \vdash \Box(p \leftrightarrow \Box p \,\&\, \Diamond p) \rightarrow \Box(p \leftrightarrow \Box\bot \,\&\, \Diamond\Diamond\top)$$

*6°) Extensions et raffinements de G et G**

Les réflexions présentent pourront être enrichies par les théories qui étendent G et G* et raffinées par les théories qui raffinent G et G*.

Par exemple :

a) La logique de la prouvabilité des machines sur la progression de Turing-Feferman (Beklemishev 1991, voir aussi les logiques polymodales de Carlson 1986, voir aussi Visser 1984 pour un "élégant voyage" de G à G*).

b) La logique de *l'interprétabilité relative* (Feferman 1960, Berarducci 1991, voir aussi Visser 1991). L'interprétabilité relative est un connecteur intensionnel binaire. Grossièrement p ◁ q est interprété arithmétiquement par "la théorie $T+MB_F(q)$ est interprétable dans la théorie $T+MB_F(p)$". en particulier, PA est capable de démontrer

$$p \triangleleft q \rightarrow (\Diamond p \rightarrow \Diamond q)$$

ainsi que

$$\Diamond p \triangleleft p$$

D'autres formules sont nécessaires pour axiomatiser complètement les formules sur l'interprétabilité prouvable par PA. Mais ces formules existent. De même, il existe un "G*" correspondant axiomatisant complètement les formules vraies sur l'interprétabilité.

---

50 On peut toujours vérifier une telle formule avec le démonstrateur de théorèmes de G donné en annexe.

Dans la plupart de ces théories □p peut être défini par ¬p◁⊥.

Les théories obtenues constituent un raffinement de la prouvabilité. Par exemple G ne distingue pas la prouvabilité de PA, de ZF ou de NBG[51], ni des machines autoréférentiellement correcte d'une façon générale.

Si l'interprétation que je propose de G et de G* pour la philosophie mécaniste de l'esprit est pertinente, la notion d'interprétabilité permettra des analyses plus fines.

*7°) G, G*, LWV et le translateur*

En travaillant *dans l'arithmétique* on peut identifier une formule avec une théorie ou une machine. C'est notamment ce que j'ai fait dans 2.2 en parlant de machine de Rogers, machine Löbienne, etc. Le théorème du point, démontrable dans l'arithmétique capture d'une façon générale ce que des machines adéquates sont capables de prouver concernant des discours autoréférentiels d'autres machines. De cette façon la formule

$$(\&_{B_i \in S^{\Box}(A)} \Box B_i \text{->} B_i) \text{->} A$$

permet de considérer un mécaniste toujours imprudent, mais un peu moins que le précédant. Pour choisir un niveau de description, il exige non plus une preuve de l'adéquation du niveau, (c-à-d une preuve qu'il ne va jamais finir mourir dans un monde accessible (avec le translateur), mais une preuve de l'adéquation du niveau pour chacune de ses parties finies, ainsi qu'une preuve que la conjonction des réflexions de ces parties finies entraîne sa survie. Dans ce cas il utilise le translateur. Il reste imprudent car il ne sait toujours pas si le mécanisme est correct, et en particulier si G* s'applique à lui-même "*vu comme formule arithmétique*". Mais, si le mécanisme est correcte, il aura prouver le plus qu'il est possible, par les deux théorèmes de Solovay, de prouver, pour un mécaniste.

G*\G constitue l'espace des solutions du principe de Wittgenstein, si l'on admet MDI, et l'identification de base. Justifier la survie à la translation est l'analogue de la preuve de la consistance d'une de ses extensions, comme l'illustre l'usage de la fonction "sceptique" de Solovay dans ses démonstrations.

Notons (cf la sémantique) que l'axiome modale 4 est un théorème de G, ce qui pose des problèmes pour l'utilisation directe de G pour le calcul des probabilités. Je suggèrerai dans la section suivante l'application du

---

[51] ZF = la théorie des ensembles de Zermelo et Fraenkel (voir Krivine 1969, NBG est la théorie des ensembles de von Neumann Bernays et Gödel, voir Mendelson 1987).

stratagème affaibli, lequel fait perdre 4 et la fermeture pour la nécessitation (voir 1.3).

*G*\G est-il une logique* ? Tenter de définir ce qu'est une logique nous entraînerait trop loin. Je me contente de deux remarques.

1) Aucune tautologie classique (ni intuitioniste donc) n'est "démontrable" par (c-à-d n'appartient à) G*\G, puisque G démontre les tautologies classiques.

2) G*\G est, *quand même*, fermé pour le modus ponens (MP). En effet si A ∈ G*\G, et A->B ∈ G*\G, alors A et A->B ∈ G*, et donc B ∈ G*, puisque G* est fermé pour MP. D'autre part, G ⊬ B, sinon G prouverait A->B (en effet avec MP et le schéma tautologique (classique et intuitioniste) p->(q->p), on dérive la règle p=>(q=>p)). Donc B ∈ G*\G.

*8°) Livres et Histoire*

L'histoire de G commence en partie en Italie avec Magari. Cette histoire est relatée dans un article de Boolos et Sambin (1991). On y apprend que la motivation de Magari était de rendre concevable que des machines soient capables, dans un sens pertinent de se tromper. Cela rejoint les relations entre conscience et intelligence abordée dans 2.3.5, plus loin.

Trois ouvrages (au moins) sont consacrés presque exclusivement à G et à l'approche modale de l'autoréférence. Un ouvrage récréatif (!) de Smullyan, (Smullyan[52] 1987), l'ouvrage classique de Boolos (Boolos 1979) et une monographie universitaire de Smorynski (Smorynski 1985).

*9°) Résumé de 2.3.3*

*La première idée aurait été de définir, comme Lucas, la connaissance pour une machine (un "produire comme vraie") par la prouvabilité formelle. Mais c'est là que repose la faiblesse principale du raisonnement de Lucas. En ce qui nous concerne, intuitivement, la connaissance intuitive est bien différente de la prouvabilité formelle qui s'apparente plus à la communication positiviste ou scientifique. De plus, la prouvabilité formelle ne vérifie pas* □(□p->p)*, ce que la prouvabilité intuive, ou la connaissance vérifie.*

*Il n'en reste pas moins que cette forme de communication est très intéressante, principalement par rapport à notre théorie de la conscience proposée en 1.2, où la conscience vérifie les propriétés qui en font l'inconnue dont le positiviste ne peut parler (Wittgenstein, Watts, Lao-tseu, Valadier).*

*Grâce aux travaux de Magari, Boolos, Solovay et bien d'autres, le concept de preuve formelle (par une machine $\Sigma_1$-complète et adéquate, avec l'identification de base, ou par une théorie assez riche du premier ordre comme PA ou ZF) a pu être axiomatisé complètement. De façon précise, on assigne une interprétation des*

---

[52] Smullyan propose une interprétation naïve très intéressante des modèles de Kripke. Il considère une collection d'individus. Tous les individus croient à la logique classique propositionnelle ainsi qu'on proposition de leur parent (c'est suffisant pour avoir K), si de plus ils croient aux propositions crues par tous leurs ancêtres, on obtient la transitivité et donc K4). Si en outre il existe un premier ancêtre, leurs croyances collectives est décrites par G. On peut combiner cette interprétation naïve avec l'interprétation mécaniste en admettant que pour survivre il faut sauver les croyances de ses ancêtres. Le mécanisme est relié à G dans ce contexte par le fait que la procédure de sauvegarde est finie grâce à la présence du premier ancêtre. Au lieu de prendre des individus différents, on peut considérer un individu unique mais variable.

*formules modales dans le langage de la machine en attribuant à chaque formule atomique $p_i$ une formule de M (du langage de M) $F(p_i)$ et on l'étend inductivement. Le carré modale $\pi A$ va être interprété par le prédicat de prouvabilité formelle B de Gödel.*

$$MB_F(p_i) = F(p_i)$$
$$MB_F(A \vee B) = MB_F(A) \vee MB_F(B)$$
$$MB_F(A \& B) = MB_F(A) \& MB_F(B)$$
$$MB_F(\neg A) = \neg MB_F(A)$$
$$MB_F(\Box A) = B(\ulcorner MB_F(A) \urcorner)$$

*On peut alors montrer que la théorie suivante, appelée G, est saine pour la prouvabilité formelle de la machine.*

| | | | |
|---|---|---|---|
| axiomes | $\Box(p \to q) \to (\Box p \to \Box q)$ | | K |
| | $\Box p \to \Box\Box p$ | | 4 |
| | $\Box(\Box p \to p) \to \Box p$ | | L |
| règles | $p, p \to q \Rightarrow q$ | MP | |
| | $p \Rightarrow \Box p$ | | Nec |

*Dans le cas où la théorie est déductivement équivalente à une théorie classique du premier ordre $\Sigma_2$-saine et $\Sigma_1$-complète, comme PA ou ZF, G est non seulement sain mais aussi complet pour les schémas modaux de la prouvabilité formelle. De plus cette théorie est décidable. Ces résultats sont dus à Solovay 1976.*
*La formalisation du second théorème d'incomplétude de Gödel est un corollaire facile de L (la formule de Löb). Il suffit en effet d'y remplacer p par $\bot$, on obtient:*

$$\Diamond \top \to \neg \Box \Diamond \top$$

*En particulier G prouve LWV.*
*Les référentiels de la sémantique de Kripke de G sont constitués par les référentiels transitifs et bien chapeautés (sans échelle infinie).*
*L'article de Solovay introduit un second système de logique modale, appelé G*, et qui formalise <u>toutes</u> les conséquences propositionnelles des résultats d'incomplétude pour les machines saines et adéquates, y compris celles que la machine ne peut pas prouver formellement (second théorème de Solovay)*

Définition *Soit SOL une fonction de l'ensemble des formules modales dans l'ensemble des formules modales. SOL(A), de MPL dans MPL, est définie par :*

$$(\&_{B_i \in S^{\Box}(A)} \Box B_i \to B_i) \to A$$

*On peut alors démontrer le résultat suivant, dont on va tirer la complétude de G* pour la vérité* sur *G (M, PA).*

$$G^* \vdash A \Leftrightarrow G \vdash SOL(A)$$

*En particulier G* prouve W, et G*\G donne l'espace des solutions de W.*

*La décidabilité de G entraîne ainsi la décidabilité de G*. Nous pouvons généraliser la réfutation de Webb de la critique du mécanisme de Lucas en construisant avec G* une machine capable de trouver, au moins formellement les vérités non prouvables la concernant. Cette machine reste auto-référentiellemnt correcte, en distinguant ce qu'elle sait formellement prouver de ce qu'elle trouve (ou produit comme vrai) à partir de G*. Elle peut aussi réfléchir ces vérités et (par*

*itération de ces réflexions) escalader des échelles transfinies de théorie. Notons l'existence de théories modales étendant G* complète pour ces théories tranfinies (jusqu'à $\varepsilon_0$, Beklemishev), sur $\omega_1^{CK}$ (Carlson).*
*G* apparaît comme la logique naturelle de ce que la machine peut inférer correctement sur son duplicata, celui-ci provenant d'une duplication effectuée a priori au niveau adéquat. C'est la non-constructivité de l'existence de ce niveau qui interdit l'usage de la nécessitation, et distingue ainsi le savoir communicable du savoir inférable et non-communicable. cela devrait se préciser dans la suite.*

## *biblio locale*


**ARTEMOV** S. and **DZHAPARIDZE** G., **1990**, *Finite Kripke Models and Predicate Logics of Provability*, Journal of Symbolic Logic, Vol 55, N° 3, pp. 1090-1098.

**BOOLOS** G. and **SAMBIN** G., **1991**, *Provability: the Emergence of a Mathematical Modality*, Studia Logica L, 1, pp. 1-23.

**BELLIN** G., **1985**, *A system of natural deduction for GL*, Theoria, Vol LI, pp. 89-114.

**BEKLEMISHEV** L. D., **1991**, *Provability Logics for Natural Turing Progressions of Arithmetical Theories*, Studia Logica L, 1, pp. 108-128.

**BERARDUCCI** A., **1990**, *The Interpretability Logic of Peano Arithmetic*, Journal of Symbolic Logic, Vol 55, N° 3, pp. 1059-1089.

**BOOLOS** G., **1979**, ***The Unprovability of Consistency, an Essay in Modal Logic***, Cambridge University Press.

**BOOLOS** G., **1988**, review of "SMORYNSKI C. ***Self-Reference and Modal Logic***. Springer Verlag, 1985"., Journal of Symbolic Logic, Vol 53, N° 1, pp. 306-308.

**BOOLOS** G. S. et **JEFFREY** R. C., **1974**, *Computability and Logic*, Cambridge University Press, 3ème éd. : 1989, Cambridge.

**BOOLOS** G. and **SAMBIN** G., 1991, *Provability: the Emergence of a Mathematical Modality*, Studia Logica L, 1, pp. 1-23.

**CARLSON** T., **1986**, *Modal Logics with Several Operators and Provability Interpretations*, Israël Journal of Mathematics, 54, pp. 14-24.

**GÖDEL** K., **1931**, *Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I*, Monatsh., Math. Phys., 38, pp. 173-98, traduit en Français dans Le théorème de Gödel, Seuil, Paris, pp. 105-143, 1989, aussi en Anglais dans Davis 1965.

**FEFERMAN** S., **1960**, *Arithmetisation of Metamathematics in a general Setting*, Fundamenta Mathematicae, XLIX, pp. 35-92.

**MAGARI** R., **1975**, *Representation and Duality Theory for Diagonalizable Algebras*, Studia Logica XXXIV, 4, pp. 305-313.

**MARCHAL** B., **1990**, *Des fondements théoriques pour l'intelligence artificielle et la philosophie de l'esprit*, Revue Internationale de Philosophie, 1, n° 172, pp 104-117.

**MARCHAL** B., **1992**, *Amoeba, Planaria, and Dreaming Machines*, in Bourgine & Varela (Eds), *Artificial Life, towards a practice of autonomous systems*, ECAL 91, MIT press, pp. 429-440.

**MENDELSON** E., **1987**, *Introduction to Mathematical Logic*, 3ème édition, Wadsworth & brooks / Cole.

**MONTAGUE** R., **1962**, *Theories Incomparable with respect to Relative Interpretability*, The Journal of Symbolic Logic, Vol 27, N° 2, pp. 195-211.

**LÖB** M. H., **1955**, *Solution of a Problem of Leon Henkin*, Journal of Symbolic Logic, 20, pp. 115-118.

**SMORYNSKI** C., **1985**, *Self-Reference and Modal Logic.* Springer Verlag.

**SMULLYAN** R., **1985**, *Modality and Self-Reference*, in Shapiro 1985, pp. 191-209.

**SMULLYAN** R., **1987**, *Forever Undecided*, Alfred A. Knopf, New York.

**SOLOVAY** R., **1976**, *Provability Interpretations of Modal Logic*. Israel Journal of Mathematics 25, pp 287-304.

**VISSER** A., **1991**, *The Formalization of Interpretability*, Studia Logica L, 1, pp. 81-105.

## 2.3.4 Le stratagème en arithmétique

*Brièvement*

*L'application du stratagème à la logique de l'autoréférence conduit à une théorie étendant S4. Il s'agit de S4Grz. La traduction de la logique intuitioniste avec S4 vaut pour S4Grz, ce que Grzegorczyk avait déjà montré. On peut dès lors considérer S4Grz, sous réserve bien sûr d'une appréciation de S4 pour une théorie de la connaissance et d'une appréciation du stratagème pour une théorie naturelle de la connaissance de la machine autoréférentiellement correcte. De même est isolé une forme naturelle d'intuitionisme ARIL*[53] *(de solipsisme) pour ces machines.*

*AUTORÉFÉRENCE. ===> (G,G*) (strat)===> S4Grz, ARIL*

*Un dernière reconstruction de l'argumentation est proposée,*

$$TI\ (S4Grz) ==> +MEC \qquad (TI^{+})$$

*L'analyse des réfutations "gödéliennes" du mécanisme effectuée dans l'arithmétique épistémique de Reinhardt et de Shapiro (voir 2.3.2) est intégralement retrouvée dans l'interprétation arithmétique de l'épistémisme proposée ici.*

---

*1°) Motivation pour le stratagème*

a) le rêve et la réalité

Le stratagème[54] identifie *savoir p* avec *(croire p) & (p est vrai).* Une motivation philosophique est liée au principe de l'impossibilité de distinguer le rêve de la réalité.

Exemple : Imaginons une personne endormie en train de rêver qu'elle mange un gâteau. Imaginons que, dans son rêve, elle énonce une proposition ***p*** : *je crois que je mange un gâteau*. Dans ce cas il est naturel d'estimer qu'elle énonce une proposition vraie, bien que la sous-proposition *je mange un gâteau* soit fausse. On a □p et on a ¬p, on a même □p & ¬p. Si dans ce même rêve elle énonçait la proposition ***q*** : *je sais que je mange un gâteau*, il est naturel d'estimer qu'elle énonce une proposition fausse. Admettre que l'on ne sait pas distinguer le rêve de la réalité, comme certains philosophes (je reviens sur cette question en 3.1), reviens à admettre que l'on ne sait pas distinguer, en toute circonstance *savoir* et *croire*.

Dans ce cadre *être éveillé* est assez bien représenté par la conjonction de "□p" avec "p". Au lieu de prendre le rêve et l'éveil, on peut prendre l'état qui consiste à être dans l'erreur (une sorte d'état de locale inconsistance).

---

53 Pour Auto-Referentially-based-Intuitionnistic-Logic. Je l'appellerai aussi simplement IL.

54 Il s'agit d'une vieille idée. On la retrouve par exemple dans Le Théétète, y compris en relation avec la question du rêve et de la veille (voir aussi la discussion de Burnyeat 1991).

Jean sait que la terre est plate (ou 1+1=3) sonne mal, alors que Jean croit que la terre est plate (ou 1+1=3) sonne mieux.

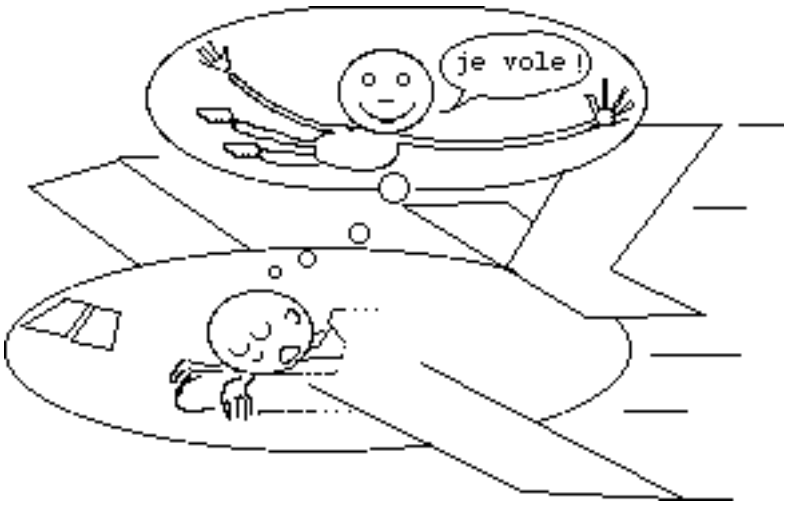
Brièvement le stratagème consiste à définir ⊡p par □p & p

L'existence du rêve montre que □p->p n'est pas prouvable ou croyable.

L'existence (psychologique) de l'*état d'erreur* comme de l'état de rêve, implique que pour la croyance on n'a pas □p -> p. Et dans l'état éveillé, pour celui qui se souvient du rêve (et surtout du rêve contralucide) on n'a pas non plus □(□p -> p), sauf pour les propositions telles qu'on a p&□p. De même celui qui reconnaît *avoir fait* des erreurs (et pouvoir encore en faire) ne peut pas admettre le schéma □(□p -> p).

Au moins localement l'état de rêve et l'état d'erreur correspond à un état d'inconsistance locale, ce que l'on peut faire correspondre, au moins localement à un dernier monde avec la sémantique de Kripke. Le souvenir du rêve, comme le souvenir de l'erreur permet (en terme d'état psychologique) de concevoir □p & ¬p, et donc une logique de l'esprit devrait permettre la dérivation de ◊(□p & ¬p) pour certaine valeur (fausse) de p. Ce point est développé dans 3.1. Remarquons que pour appliquer le stratagème, les propositions sont nécessairement appliquées dans le sens contextuelle le plus large.

Ceci permet de répondre à l'objection du "savoir *pour une mauvaise raison*" : par exemple le petit bonhomme ci-dessous vole et croit qu'il vole. Sait-il qu'il vole ? La réponse est non au cas où il croit voler des ses propres mains, comme le contexte le suggère. Toutefois, l'on voit que le stratagème ne permet a priori aucune interprétation causale du savoir. Une conception Humienne de la causalité est appropriée dans ce contexte, ce qui est abordé en 3.3, après qu'un engagement ontologique minimal soit commis.




b) Gödel 1933 et Lucas 1961

Une autre motivation pour le stratagème permet de remonter à Gödel 1933 et Lucas 1961, et repose sur l'idée que l'on aimerait parvenir à extraire

des théories doxastiques (croyance) et épistémiques (connaissance) directement à partir de l'hypothèse mécaniste.

Si on admet comme Lucas, que croire, pour la machine, = produire (formellement) comme vrai (une définition positiviste), alors MEC => croire vérifie G, et MEC + le stratagème permet de retrouver S4.

Gödel, dans son papier de 1933, a fait remarquer que la S4-prouvabilité (formalisée par "$\boxdot$") n'est pas adéquate pour décrire la prouvabilité formelle (voir plus haut). Nous savons, par les analyses de la section précédente, que cette non-adéquation est de type intensionnel : la prouvabilité intuitive "$\boxdot$" et la prouvabilité formelle "$\Box$" ne relève pas de la même logique, et caractérise des points de vue différents (le point de vue du dedans, intérieur, subjectif, locale, constructif, *versus* le point de vue du dehors, extérieur, objectif, classique, formalisable, etc...), et ces points de vue sont distingués par S4 et G respectivement. Mais la réfutation de Lucas montre que rien ne s'oppose à ce que leurs *extensions* respectives coincident. Il est ainsi possible de caractériser le sujet directement à partir de G, sans surimposer a priori la logique modale S4 et sans utiliser l'hypothèse mécaniste.

L'idée de modifier l'intension apparaît déjà dans Lucas, et nous l'avons déjà rencontrée. Lucas raisonne cependant comme si une telle modification entraînait automatiquement une modification de l'extension. Nous devons modifier l'intension du prédicat de prouvabilité de façon telle que son extension demeure invariante. Nous savons que nous pouvons faire ça (sur base de la consistance de PT).

A présent, j'aimerais parvenir à faire de même sans invoquer la consistance de PT, seulement en modifiant l'intension des prédicats. Prenons par exemple le prédicat de prouvabilité de Webb (voir plus haut) :

$$bw'_{PA}(x,y) = bw_{PA}(x,y) \,\&\, \neg\, bw_{PA}(\ulcorner\bot\urcorner,y)$$

Une forme plus simple revient à définir directement un nouveau prédicat de prouvabilité (et donc de consistance) par

$$\boxdot p = \Box p \,\&\, \Diamond p$$

C'est la version affaiblie du stratagème déjà invoqué en 1.3.

Ce prédicat est arithmétisable. En particulier il admet une interprétation dans l'arithmétique (ou dans le langage d'une machine adéquate).

F est, comme avec le morphisme de Magari-Boolos, une fonction qui associe aux variables propositionnelles un énoncé arithmétique.

DEON-A($\bot$) = $\bot$
DEON-A(p) = F(p)
DEON-A(X->Y) = (DEON-A(X) -> DEON-A(Y))

$$DEON\text{-}A(\Box X) = (B(\ulcorner DEON\text{-}A(X)\urcorner) \& \neg B(\ulcorner \neg DEON\text{-}A(X)\urcorner)).$$

avec B, le prédicat de prouvabilité de Gödel. (DEON pour "déontique", le système obtenu prouve trivialement l'axiome D, A pour Arithmétique).

Du coup, par Benacerraf-Montague-Thomason, nous savons à l'avance qu'il ne satisfaira pas T (voir 2.3.1 et 2.3.2). Nous pouvons néanmoins utiliser le démonstrateur de théorème pour G pour regarder quelques schémas modaux démontrables (par PA, une machine adéquate, etc.). En particulier

$$\Diamond\!\!\!\!\bigcirc \top$$

est démontrable par une machine autoréférentiellement correcte (est G-démontrable). Ce qui satisfait la critique de Webb de l'usage de la consistance non monotonique par Lucas. Mais

$$\boxed{\bigcirc}\top$$

n'est déjà pas un G-théorème[55] (quoi qu'il s'agisse d'un G*-théorème). Les schémas modaux basés sur $\boxed{\bigcirc}$ ne peuvent donc pas être axiomatisés par une logique modale normale[56]. En effet (voir 1.1) dans tout modèle de Kripke (avec n'importe quel référentiel) le vrai est toujours satisfait dans tous les mondes possibles.

$\boxed{\bigcirc}$ est extensionnellement équivalent à $\Box$. Pour vérifier cela il suffit, en vertu du second théorème de Solovay, de constater que l'équivalence entre $\Box$ et $\boxed{\bigcirc}$ est démontrable dans G*. En fait cette équivalence est immédiate puisque nous savons que la prouvabilité formelle d'une proposition p par une machine adéquate entraîne p. En effet G* démontre[57]

$$G^* \vdash \Box p <-> \Box p \,\&\, \Diamond p$$

mais la réflexion T, $\Box p \to p$ n'est pas satisfait par cette logique[58] et donc ne satisfait pas l'intension que nous voulons lui donné.

---

55 ? (kdip '(bw (p -> p)))
(((BW (- (-> P P)))))
? (kdip '(- bw - (p -> p)))
NIL
Bien sûr, le résultat est évident, si G prouvait D, par modus ponens, G prouverait W!

56 En outre, cette théorie ne satisfait pas T, ? (kdip '(bw p -> p))-->(((- P) (F\# ((P P))))), à la différence de kd*ip : (kd*ip '(bw p -> p))-->NIL.

57 ? (g*ip '(bw p <-> bw p & - bw - p))
NIL

58 ? Avec le démonstrateur de théorème de G et G*, il est aisé de construire un démonstrateur de KD?, par exemple (kdip '(bw p -> p)) donne un contre-exemple (((- P) (F\# ((P P))))) ; (kdip '(¬bw ¬(p -> p))) donne NIL, donc $\Diamond\top$ est un théorème de KD?.

Je reviendrai cependant plus tard sur cette solution particulière qui présente un intérêt intrinsèque (c'est le stratagème affaibli). La solution simple et naturelle pour extraire S4 de l'autoréférence correcte (donc de G) revient à utiliser le stratagème :

$$\boxdot p = \Box p \,\&\, p$$

Remarquons que $\boxdot p$ n'est plus a priori arithmétisable, à cause de la présence explicite de p. Pour l'arithmétiser, on aurait besoin d'un prédicat de vérité V :

$$\boxdot p = \Box p \,\&\, V(\ulcorner p \urcorner)$$

mais, par Tarski (voir 2.2), un tel prédicat n'existe pas.

Du coup, nous sommes a priori assurés que $\boxdot$ peut vérifier T sans tomber dans le piège de Benacerraf. De plus, pour une machine adéquate et consistante, $\boxdot = \Box$, c'est-à-dire que $\boxdot$ et $\Box$ sont extensionnellement identiques[59].

$$G^* \vdash \Box p <-> \Box p \,\&\, p$$

D'autre part, on a d'office :

$$(\Box p \,\&\, p) -> p$$

donc $\boxdot p -> p$, et T est vérifié.

Il est facile de montrer que $\boxdot p$ vérifie les axiomes de S4.

a) $\boxdot p -> p$, car G prouve $(\Box p \& p) -> p$

b) $\boxdot p -> \boxdot\boxdot p$, car G prouve $(\Box p \& p) -> (\Box(\Box p \& p) \& (\Box p \& p)$

c) $\boxdot(p -> q) -> (\boxdot p -> \boxdot q)$ car G prouve $\Box(p -> q) \& (p -> q) -> ((\Box p \& p) -> (\Box q \& q))$

de même pour une machine correcte $\vdash p$ entraîne $\vdash p \& \Box p$, et la nécessitation est vérifiée.

*2°) Le morphisme de Boolos-Goldblatt, Kusnetsov & Muravitsky.*

Considérons la transformation BGKM de l'ensemble des formules modales dans l'ensemble des formules modales MPL->MPL:

Les variables propositionnelles sont supposées ordonnées $p_i$.

$$BGKM(p_i) = p_i$$
$$BGKM(A \vee B) = BGKM(A) \vee BGKM(B)$$

---

59 ? (g*ip '(bw p <-> bw p & p))
NIL

$$BGKM(A\&B) = BGKM(A) \;\&\; BGKM(B)$$
$$BGKM(\neg A) = \neg BGKM(A)$$
$$BGKM(\Box A) = \Box(BGKM(A)) \;\&\; BGKM(A)$$

Si R = (W, R) un référentiel, et si

$$t(W, R) = (W, R \cup \{(x,x) \mid x \in W\}),$$

alors

*lemme* (Goldblatt 1978, Boolos 1980a)

t(W, R) respecte A ssi (W, R) respecte BGKM(A).

*preuve*

Il suffit de montrer que, quelle que soit la formule A, et quelle que soit la valuation P,

$$(W,R,P) \models A \;<\!\!-\!\!>\; (W,S,P) \models BGKM(A),$$

où $R = S \cup \{(x,x) \mid x \in W\}$.

La démonstration se fait par induction sur la longueur des formules. Pour les propositions classiques le résultat est évident puisque leurs valeurs de vérité dans un modèle ne dépendent pas de la relation d'accessibilité. Reste à démontrer le résultat pour $A = \Box B$. Supposons donc (hypothèse d'induction *hi*) qu'on ait pour un monde w

$$(W,R,P) \models_w B \;<\!\!-\!\!>\; (W,S,P) \models_w BGKM(B) \quad hi$$

Il faut montrer

$$(W,R,P) \models_w \Box B \;<\!\!-\!\!>\; (W,S,P) \models_w BGKM(\Box B).$$

mais on a

$(W,S,P) \models_w BGKM(\Box B)$
$<\!\!-\!\!>$ $(W,S,P) \models_w \Box$ BGKM(B) & BGKM(B)
$<\!\!-\!\!>$ $(W,S,P) \models_w \Box$ BGKM(B) et $(W,S,P) \models_w$ BGKM(B)
$<\!\!-\!\!>$ $\forall x \in W(wSx \rightarrow (W,S,P) \models_x BGKM(B))$ et $(W,S,P) \models_w$ BGKM(B)
$<\!\!-\!\!>$ $\forall x \in W(wSx \rightarrow (W,R,P) \models_x B)$ et $(W,R,P) \models_w B$ par *hi*.
$<\!\!-\!\!>$ $\forall x \in W((wSx \vee w=x) \rightarrow (W,R,P) \models_x B)$
$<\!\!-\!\!>$ $\forall x \in W(wRx \rightarrow (W,R,P) \models_w B)$
$(W,R,P) \models_w \Box B$

*corollaires* T ⊢ A ssi K ⊢ BG(A), S4 ⊢ A ssi K4 ⊢ BG(A).

*3°) Interprétation arithmétique et la formule de Grzegorczyk*

Si, avec Goldblatt 1978, on considère les modèles de l'arithmétique de Peano comme les mondes possibles, ou si, comme on l'a considéré jusqu'ici, on regarde les états d'un sujet accessible comme étant des extensions consistantes, on aimerait :

nécessairement vraie = prouvable = vrai dans tous les mondes possibles (et donc = vrai dans tous les modèles de PA). Grâce au théorème de complétude (!) de Gödel (1930), ceci est vérifié. Toutefois l'usage du mot "nécessairement" (c'est-à-dire de la modalité ontique) exigerait que □p->p. Gödel fait explicitement la remarque, qu'on a pas □p->p (sauf pour les p tel que □P, et uniquement ceux-là). L'ontique est cependant vérifié si on prend pour □p, □p&p C'était le but de Goldblatt, et c'est ce qui permet de considérer les extensions consistantes comme des "vrais" mondes possibles.

*4°) Le morphisme arithmétique de Boolos-Goldblatt*

A la façon de Magari-Boolos, on peut encore définir une interprétation de l'ensemble des propositions modales prouvables par S4 dans l'ensemble des propositions de l'arithmétique. La traduction Tr correspondante est alors définie inductivement[60] de la façon habituelle :

Tr(⊥) = ⊥
Tr(p) = F(p)
Tr(X->Y) = (Tr(X) -> Tr(Y))
Tr(⊡X) = (B(⌜ Tr(X)⌝) & Tr(X))

Cela suffit, pour traduire une négation on use de : ¬p <-> p-> ⊥.
Tr est obtenu par une composition de BGKM avec $MB_F$ : $MB_F$∘BGKM.

Toutes les Tr-traductions de S4 sont prouvables par PA, c'est ce qu'on a montré plus haut. Mais S4 n'est pas complet pour ces translations. Un résultat de complétude existe cependant, il faut ajouter Grz (Grzegorczyk 1967) comme axiome :

□(□(p->□p)->p)->p

60 De même pour ⊡, on peut définir l'interprétation arithmétique correspondante, avec T(□X) = B(⌜T(X)⌝) & -B(⌜-T(X)⌝).

*Théorème* l'ensemble des T-traductions prouvables par une machine adéquate et correcte est capturé sainement et complètement par la théorie modale S4Grz :

| | | |
|---|---|---|
| AXIOMES: | □p->p | T |
| | □p->□□p | 4 |
| | □(p->q)->(□p->□q) | K |
| | □(□(p->□p)->p)->p | **Grz** |
| REGLES: | p et p->q entraîne q | *MP* |
| | p entraîne □p | *NEC* |

Dans sa thèse de doctorat Segerberg avait déjà caractérisé une classe de référentielles respectant Grz.

*Théorème* (Segerberg 1971)

**(W,R) respecte Grz <=> (W,R) est un ordonné**[61] **fini**

preuve : voir plus loin.

*corollaire* (Kuznetsov & Muravitsky 1977, Goldblatt 1978)

$$S4Grz \vdash A \text{ ssi } G \vdash BGKM(A)$$

*preuve* : immédiat avec le lemme de Boolos-Goldblatt grâce au théorème de Segerberg et le fait qu'on a t(modèle de Kripke fini de G) = modèle de Kripke fini de S4Grz.

De même on a

$$S4Grz \vdash A \text{ ssi pour tout F, M} \vdash (MB_F \circ BG)(A) \quad (*)$$

où M désigne toujours l'arithmétique de Peano, ou une extension RE de PA, ou encore une machine adéquate.

*5°) Le morphisme de Gödel-Grzegorczyk-Goldblatt G33*

Nous savons depuis Gödel 1933 et McKinsey & Tarski 1948 qu'il existe une interprétation du calcul propositionnel intuitioniste IL (de Heyting) en terme de la logique modale S4. Il existe différentes traductions possibles, je me réfère à celles données plus haut, que je rappelle :

61 Un ordonné (W,R) est un ensemble W muni d'une relation d'ordre R. R est un ordre ssi R est réflexive, transitive et antisymétrique (attention les anglo-saxons disent *partial order*).

$G33(p_i) = \Box p_i$
$G33(A \& B) = G33(A) \& G33(B)$
$G33(A \vee B) = G33(A) \vee G33(B)$
$G33(A \to B) = \Box G33(A) \to \Box G33(B)$
$G33(\neg A) = \Box \neg G33(A)$

On a (Gödel 1933, McKinsey & Tarski 1948), voir aussi Fitting 1969 pour une preuve claire et moderne :

$$IL \vdash A \leftrightarrow S4 \vdash G33(A) \quad (**)$$

Grzegorczyk a montré que l'on pouvait caractériser IL par la théorie plus forte S4 + Grz (Grzegorczyk 1967)

$$IL \vdash A \leftrightarrow S4Grz \vdash G33(A)$$

S4Grz étend strictement S4, en particulier $S4 \nvdash Grz$. Mais S4 et S4Grz sont équivalents pour les formules images de type G33(A).

En composant les interprétations (*) et (**), on peut construire une interprétation arithmétique de IL. $A^\circ = MB_F(BG(G33(A)))$ :

| | |
|---|---|
| $\bot^\circ$ | $\bot$ |
| $p^\circ$ | $F(p) \& B(F(\ulcorner p \urcorner))$ |
| $(X \to Y)^\circ$ | $(X^\circ \to Y^\circ) \& B(\ulcorner X^\circ \to Y^\circ \urcorner)$ |

Par exemple $(\neg A)^\circ$ est interprété par $\neg(A)^\circ \& B(\ulcorner \neg(A^\circ) \urcorner)$.

On a donc, avec M une machine adéquate

$$IL \vdash A \leftrightarrow M \vdash A^\circ$$

on obtient donc une interprétation de la logique intuitioniste, et donc des procédures des écoles du dedans, directement en terme de vérité (platoniste) et d'autoréférentialité correcte.

Il est aisé (voir annexe 2), avec ce résultat, de construire un démonstrateur de théorème pour la logique propositionnelle intuitioniste à partir d'un démonstrateur de théorème pour G. Cela corrobore l'adéquation entre le mécanisme et l'interprétation épistémique de l'intuitonnisme. Pour autant que l'hypothèse mécaniste soit consistante avec l'interprétation arithmétique de G, et de S4Grz. Ce qui reste encore à montrer.

Il serait, pour la même raison, intéressant de lier directement la motivation de Grzegorczyk 1967 et son interprétation épistémique de 1964 de

la logique intuitioniste[62] en terme de description formelle (et positiviste) de procédure de recherche scientifique. Cela rejoint aussi l'analyse de S4, ou EA, effectuée par Shapiro 1985.

Je mentionne Artemov (1990) qui érige le stratagème en thèse, en montrant qu'à une équivalence prouvable près □p&p est l'unique façon de définir ⊡, de telle façon que les axiomes de S4 soient vérifiés au niveau G (Il exige aussi que G ⊢ ⊡p->□p, voir annexe 6.

*6°) Schémas modaux pures*

Nous savons que G est une axiomatisation complète et décidable des schémas modaux prouvables par une machine adéquate, où la modalité formalise la prouvabilité formelle dont Gödel a montré l'arithmétisabilité (c'est le premier théorème de Solovay 1976). De même G* formalise de façon complète et décidable les schémas modaux vrais concernant une machine adéquate, aussi bien ceux que cette machine est à même de prouver sur elle-même que ceux qu'elle ne peut pas prouver sans perdre son autoréférentialité (c'est le second théorème de Solovay 1976). Il est naturel de chercher les résultats équivalents pour S4Grz et IL. En utilisant G*, Goldblatt démontre

IL = IL*

Il pose la question concernat S4Grz, et, en utilisant à nouveau G*, Boolos parvient à montrer

S4Grz = S4Grz*

Ces deux résultats, qui se complètent, montrent qu'en ce qui concerne les schémas modaux de Grzegorczyk ⊡, ou l'interprétation arithmétique de l'intuitionisme, la vérité collapse avec la prouvabilité, ce qui corrobore davantage l'analyse du sujet en terme de modalité épistémique S4, ainsi que l'interprétation intuitioniste de la source de cette connaissance.

résumé :

| Prouvable par M | Vrai pour M | |
|---|---|---|
| G | G* | PN |
| IL | IL | NP |
| S4Grz | S4Grz | NP |

62 Espace topologique totalement distributif et algèbre relationnelle bien ordonnée.

PN = Perte de la nécessitation, NP = nécessitation préservée. Dans ce qui suit, on travaille dans le langage (arithmétique) de M. Grâce au stratagème on peut considérer des schémas modaux hybrides. Par exemple la thèse mécaniste ⊡ = □ (avec thèse de Post-Turing et identification de base, voir plus haut) peut s'écrire avec le stratagème

$$\Box p <-> \Box p \ \& \ p$$

c'est-à-dire B(⌜p⌝) <-> B(⌜p⌝) & p dans le langage de M.

*7°) La consistance de PT*

Une sémantique de l'arithmétique épistémique de Reinhardt-Shapiro, EA, permettant de prouver la consistance de la thèse de Church sous la forme TC1

$$\boxdot \forall x \exists y \ \boxdot P(x,y) -> \exists z \boxdot \forall x \ (\phi_z(x)\Downarrow \ \& \ \boxdot P(x,\phi_z(x)),$$

a été donnée par Flagg. Cette preuve utilise des outils comme la théorie des catégories et repose sur des travaux de Hyland, Johnstone et Pitts (HJP). La démonstration de Flagg utilise et assemble (entre autres) :

- la translation de Gödel 1933 (de IL vers S4),
- la réalisabilité de Kleene 1945 (voir 2.2),
- la sémantique fonctorielle de Lawvere 1969,
- l'utilisation de cette dernière par Hyland, Johnstone et Pitt 1980.

La preuve de Flagg illustre une apparition supplémentaire de la logique intuitioniste dans le champ des mathématiques classiques. En effet la logique intuitioniste constitue la logique (interne) naturellement associée à tous topos, lesquels permettent de

- généraliser non trivialement la théorie des ensembles[63],
- étendre de façon naturelle le calcul λ de Church,

et constituent une sorte d'univers du mathématicien (bien qu'a priori cet univers soit *du dedans* et donc constructif dans un sens très large). Bell 1986 utilise le terme *local*.

63 Quoi que les topos soient d'abord apparus en géométrie algébrique.

Goodman 1986 a donné une preuve directe (sans détour catégoriel, mais en passant par une arithmétique épistémique intuitioniste, et un autre résultat de Gödel[64])

*Question* : les preuves de Flagg et Goodman fonctionnent-elles pour l'arithmétique épistémique ou ⊡ est défini par le stratagème arithmétique.

En est-il de même pour les réflexions sur la thèse de Church et sur la thèse de Post-Turing de Reinhardt (voir 2.3.2) :

$$PT \quad \exists e \forall n(\boxdot P(n) <-> U(e,n))$$

Il est vraisemblable que PT soit consistante. Que dire de ⊡PT :

$$\boxdot PT \qquad \boxdot \exists e \forall n(\boxdot P(n) <-> U(e,n))$$

De même il est plus que probable que la thèse "P⊡T" suivante est réfutable :

$$P\boxdot T \qquad \exists e \boxdot \forall n(\boxdot P(n) <-> U(e,n))$$

Avec l'interprétation intuitive et mécaniste du stratagème, cette thèse revient à affirmer l'existence d'une machine qui s'identifierait à une machine de façon communicable ce qui, au moins pour l'analyse propositionnelle présentée est réfutable, et donc encore plus suspicieux dans cette version du premier ordre.

Par contre le principe de Markov :

$$\mathbf{MP} \qquad \{\forall x(P(x) \vee \neg P(x)) \ \& \ \neg\neg \exists x P(x)\} -> \exists x P(x)$$

n'est pas un théorème dans l'intuitionisme extrait de l'autoréférence. C'est-à-dire qu'il existe des réalisations F telles que

$$G \nvdash MB_F(BG(G33(\mathbf{MP})))$$

Et il en est de même pour son correspondant épistémique (Voir Artemov 1990).

---

[64] CP $\vdash$ p <-> IL $\vdash$ $\neg\neg$p (une interprétation de CP dans IL, qui, une fois étendue, établit la consistance réciproque de HA et PA). Cette traduction rend la logique intuitionniste plus générale que la logique classique, et montre que la notion de généralité est très relative en logique.

La difficulté de ces questions provient de la présence des quantificateurs. Il faut travailler dans une extension de G avec quantificateurs. Que peut-on en dire ? On a sûrement

$$\Box\forall xP(x)\text{->}\forall x\Box P(x),$$

dont l'interprétation arithmétique serait une formalisation de la règle d'instantiation universelle. On a sûrement pas, même au niveau G* quantifié

$$\forall x\Box P(x) \text{ -> } \Box\forall xP(x).$$

En effet si l'arithmétique de Peano (ou une machine saine et $\Sigma_1$-complète) est consistante elle prouve $\neg bw(n, \ulcorner\bot\urcorner)$ pour tout nombre naturel, mais elle ne prouve certainement pas $\forall x \neg bw(x, \ulcorner\bot\urcorner)$, qui est équivalente à $\neg\exists xbw(x, \ulcorner\bot\urcorner)$, qui est un énoncé d'autoconsistance[65].

*8°) La reconstruction BCR revisitée*

Les deux hypothèses de Lucas selon lesquelles 1) il est consistant (et même correct) et 2) la machine présentée est correcte

$\Box p\text{->}p$ (1)
$\boxdot p\text{->}p$ (T)

sont toutes deux des théorèmes de G*, elles sont donc vraies. La thèse de Lucas

$$\boxdot \neq \Box$$

est simplement fausse. La thèse de la reconstruction de Benacerraf-Chihara-Reinhardt selon laquelle l'hypothèse mécaniste est correcte, mais pas connaissable est vérifiée puisque le schéma $\boxdot = \Box$ est un théorème de G* et n'est pas démontrable par G, ni par S4Grz

$M \nvdash \Box(\boxdot = \Box)$
$M \nvdash \boxdot(\boxdot = \Box)$

Regardons la preuve de Lucas dans ce contexte et identifions (à nouveau) le passage non valide de son raisonnement :

---

[65] Depuis que j'ai écrit ces lignes j'ai appris que G avec quantificateurs n'est pas axiomatisable ainsi que G* avec quantificateurs. (voir Artemov 1990, Artemov & Dzhaparidze 1990).

*démonstration* il existe p telle que p <-> ¬□p est intuitivement (constructivement) prouvable, dès lors les deux premières lignes du raisonnement ne posent pas de problèmes,

⊡(p <-> ¬□p)
⊡(□p->¬p)

par contre la troisième ligne,

⊡(□p->p)

est erronée. □p->p est vraie, c'est un théorème de G*, mais ce n'est pas un théorème de G, et donc □(□p->p) & (□p->p) n'est pas un théorème de G, donc M ne peut pas prouver pas le schéma hybride ⊡(□p->p).

Le reste de la dérivation est correcte

| | |
|---|---|
| ⊡(□p->(p & ¬p)) | |
| ⊡(□p->⊥) | |
| ⊡¬□p | (3) |
| ⊡p | par (2) |
| ¬□p | par (3) et (T) |

mais ne prouve ici que ce que l'on savait déjà, à savoir que

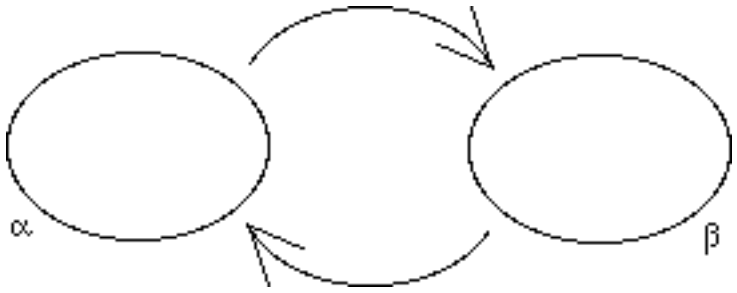
(⊡ = □) est un théorème de G*\G, et fait donc partie des propositions vraies et non prouvables. Avec le stratagème et le théorème 1 de la section précédente, la thèse (⊡ = □) est absolument indécidable. Avec M une machine (extension RE de PA) démontrant au moins tous les théorèmes que je sais démontrer. Dans la mesure où l'on s'estime capable de produire une démonstration de la consistance de l'arithmétique de Péano, le raisonnement de Lucas montre tout au plus que si *je suis une machine*, je suis une machine plus puissante (en terme de production de propositions arithmétiques vraies) que PA[66].

### *9°) Sémantique de Kripke de Grz*

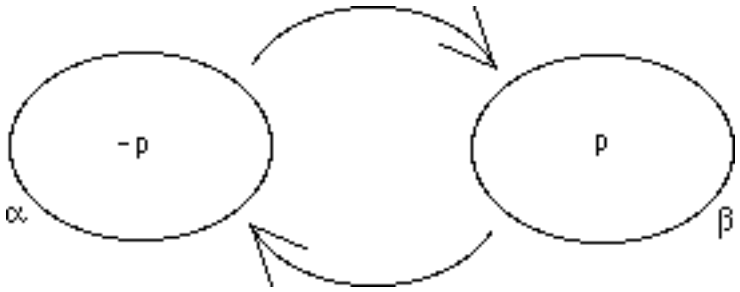
Proposition Si (W,R) respecte Grz, alors R est antisymétrique

Preuve en effet si R n'est pas antisymétrique, alors il existe un sous-référentiel se présentant ainsi :

66 Gentzen a produit une telle preuve de la consistance de PA. Cette preuve est forcément, par Gödel 1931, non formalisable dans PA.

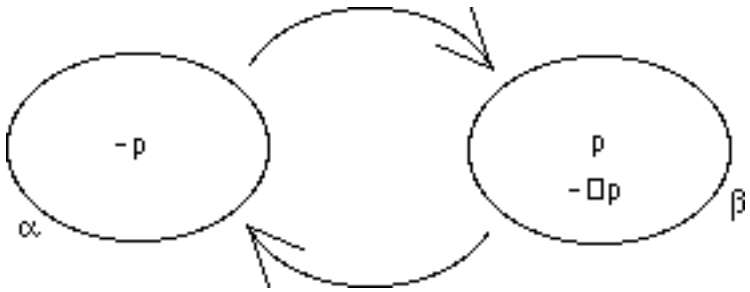



je mets "¬p" dans un des mondes (α par exemple), et p partout ailleurs, voilà le sous-modèle :

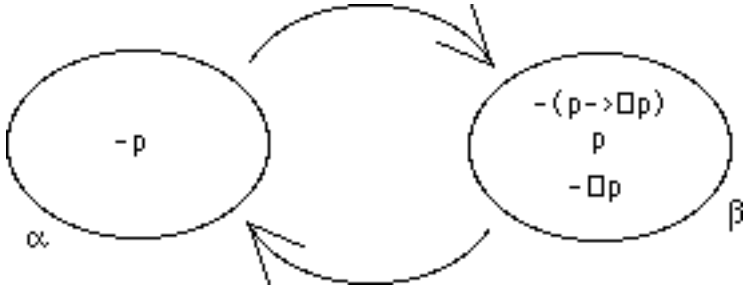



Remarquons que partout où p est vrai, c-à-d partout sauf dans le monde α, on a ⊡(p->⊡p)->p, si en plus ⊡(p->⊡p)->p était vrai dans α, alors ⊡(⊡(p ->⊡p)->p) serait vrai partout, y compris dans α, et Grz serait faux dans α. On va effectivement montrer que ⊡(p->⊡p)->p est vrai dans α.

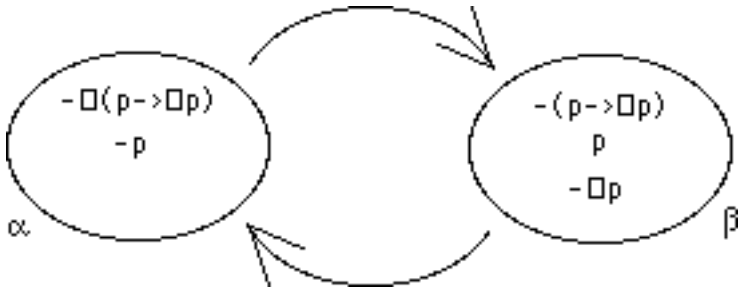
Comme β accède à (au moins un) monde où p est faux, on a:



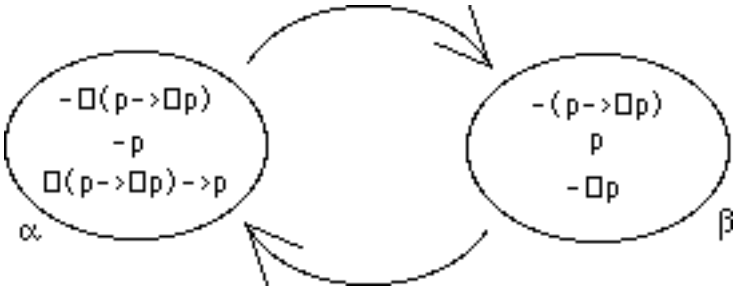

et donc



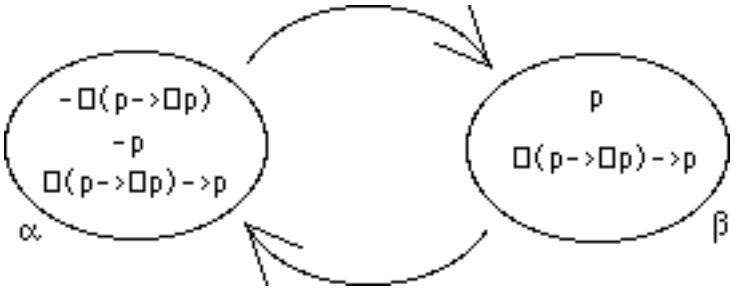
donc α accède à un monde où p->⊡p est faux, ainsi



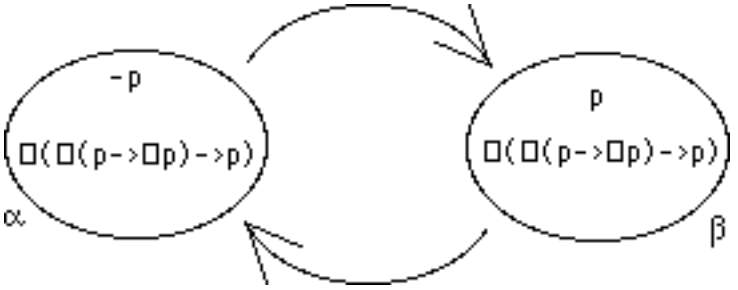

donc




Je peux m'arrêter ici, mais pour ceux qui veulent voir le déroulement jusqu'au bout, je rappelle que partout où p est vrai, ⊡(p->⊡p)->p est vrai :




c-à-d:

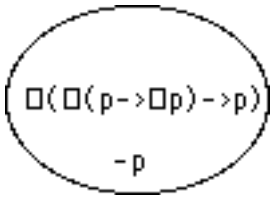



et Grz est fausse dans α. Q.E.D.

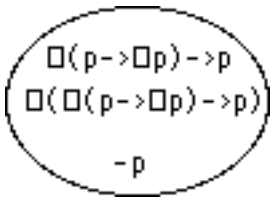
<u>proposition</u> Si un référentiel est fini, réflexif, transitif et antisymétrique, alors il respecte T, 4, et Grz.

*preuve* on a déjà montré qu'un référentiel réflexif et transitif respecte T et 4. Soit (W,R) un référentiel fini (W est fini) et R est transitive et réflexive. On va montrer que s'il ne respecte pas Grz alors il est infini ou il n'est pas antisymétrique (voir c378). De nouveau je décompose la démonstration en graphique. (Elle coule de source, et on peut voir que Grz est une forme d'axiome de l'infini qui échoue lamentablement).

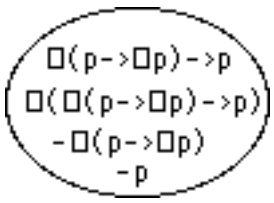
Si (W,R) ne respecte pas Grz, il y a un monde ou Grz est faux (pour une certaine valuation) :



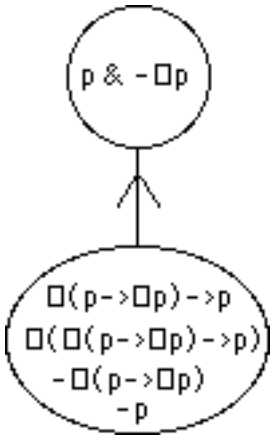

mais (W,R) est réflexif, donc



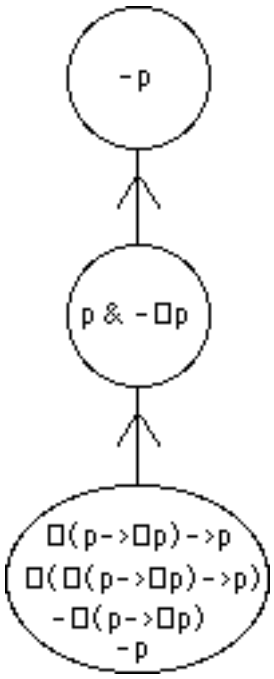

donc



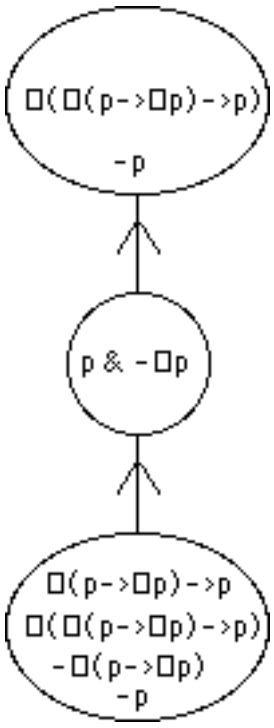

donc




donc




mais R est transitive, donc




Remarquons que β est différent de α puisque p n'y a pas la même valeur, et que γ est différent de β pour la même raison. A présent si γ et a sont le même monde, R est ne peut plus être antisymétrique et la démonstration est terminée. Sinon, on observe que γ est dans les mêmes conditions que α au début de la démonstration. On a donc démontré (par une induction implicite) l'existence d'une suite de mondes avec aRbRcRdRe...avec a≠b, b≠c, c≠d, etc. (puisque aRb entraîne que p n'a pas la même valeur dans a et b).

Mais W est fini, donc, cela va boucler et par transitivité on aura bien deux mondes différents t et u avec tRu et uRt. Q.E.D.

*Remarques*

1)Avec le lemme de Boolos-Goldblatt on peut cararactériser S4Grz par les référentiels réflexifs, transitifs et qui ne contiennent pas de suite de mondes avec $a_1Ra_2Ra_3R...$ avec $a_i \neq a_j$ pour tout $i \neq j$, c-à-d les référentiels réflexifs transitifs et bien chapeautés.

2) On peut démontrer, avec la technique des modèles canoniques, un théorème de complétude : S4Grz $\vdash$ A ***ssi*** A est vraie dans tous les mondes de tous les modèles finis, réflexifs, transitifs et antisymétriques (voir Segerberg 1971, Boolos 1979). De même S4Grz $\vdash$ A ***ssi*** A est vraie dans tous les mondes de tous les modèles, réflexifs, transitifs et bien chapeautés.

*10°) Myhill, Goodman 1985*

Myhill 1985 & Goodman 1985 ont aussi utilisé une *version du stratagème* pour montrer qu'il est possible d'avoir un système formel où la prouvabilité intuitive, représentée directement par un opérateur modal (voir section précédente), est équivalente avec la prouvabilité formelle dans cette même théorie (contrairement à ce que Myhill écrivait lui-même 1960).

Les deux articles introduisent une mathématique intensionnelle, avec un opérateur modal, construite sur S4. Myhill contredit Gödel (son "erreur") et Myhill 1960 (voir 2.3.2). Le carré représente la prouvabilité formelle dans la théorie, et cela malgré la présence de T dans S4. Sa théorie s'appelle $IST^+$ , A représente une formule fermée. Myhill procède au remplacement de

$$\Box \text{ A est vrai ssi } IST^+ \vdash A$$

par

$$\Box \text{ A est vrai ssi A est vrai et } IST^+ \vdash A$$

De même Goodman (sa théorie s'appelle ZFM*) :

$$\Box \text{ A est vrai ssi A est vrai et ZFM* } \vdash A$$

Cela permet d'identifier (extensionnellement) la prouvabilité intuitive avec le prouvabilité formelle dans une théorie épistémique.(= théorie étendant S4). Que l'application de cette version du stratagème aboutisse à une théorie consistante ne me semble pas évidente. La consistance du stratagème arithmétique donne d'office lieu à une théorie consistante vu qu'on ne sort pas de l'arithmétique.

*11°) Le stratagème affaibli*

Pour les croyances rationnelles (au sens de Thomason, voir 2.3.2) la logique G n'est pas adéquate puisque G $\nvdash$ D où D est l'axiome déontique $\Box p \rightarrow \Diamond p$.

La logique S4Grz $\vdash$ D, mais est trop forte, puisqu'elle prouve T aussi, il s'agit donc plus de connaissance que de croyance (et même de connaiss**abilité**, avec 4).

Existe-t-il une modalité, ayant une traduction arithmétique, intermédiaire ?

On se rappelle que Lucas propose un *prédicat de consistance corrigeable,* qui *garantit* sa propre consistance. Webb a cependant déjà illustré la non-pertinence de ce genre de changement de perspective intensionnelle pour réfuter le mécanisme (voir 2.3.1). Il montre qu'un tel prédicat est machine-définissable et qu'avec une telle nuance intensionnelle, *une machine prouve aussi sa propre consistance.* . Cela réfute Lucas, mais n'enlève rien à l'intérêt du prédicat. Webb propose comme prédicat de ce genre :

$$\Box p \ \& \ \Diamond \top$$

Du point de vue de Kripke : c'est équivalent à $\Box p \ \& \ \Diamond p$. En fait Webb aurait pu prendre $\Box p \ \& \ \Diamond p$. Dans ce cas, avec $\boxdot p \iff \Box p \ \& \ \Diamond p$, on a $\diamondsuit p \iff \Box p \vee \Diamond p$. (avec les lois de Morgan) Et donc quelle que soit la logique de départ[67], on aura

$$\boxdot p \rightarrow \diamondsuit p$$

C'est-à-dire l'axiome déontique D. De ce point de vue intensionnel nouveau, la consistance $\diamondsuit \top$ est démontrable, pourvu que la proposition $\Box \top$ soit un théorème de la logique dont on part. En particulier G $\vdash \Box \top$, et donc pour la *machine* à laquelle, implicitement, Lucas applique G, Webb montre qu'un prédicat *extensionnellement équivalent*, rend la consistance communicable *de la part de la machine.*

En résumé (*p* arithmétique[68]) :

$$G^* \vdash \Box p \leftrightarrow \boxdot p,$$
$$G \nvdash \Box p \leftrightarrow \boxdot p.$$

On se rappelle aussi des motivations pour la recherche d'un prédicat de prouvabilité valable pour une notion d'accès directe, avec une relation

---

[67] Pourvu que cette logique soit fermée pour la règle A & B => A v B.

[68]... à la transformation de Magari-Boolos près.

d'accessibilité en un coup, et donc non transitive, pour chercher une logique des probabilités des états transitoires (les formules du type $\Diamond p$ y sont toujours vraies) directement accessibles (voir 1.3). Par ailleurs garantir la consistance revient à restreindre la logique aux états *consistants*, c'est-à-dire, *transitoires*, dans le contexte de la translation. C'est la façon d'avoir P=1/2 dans l'expérience de duplication, malgré qu'en termes de mondes accessibles, on devrait avoir au moins P=1/3, vu la présence nécessaire des derniers mondes (voir 1.3).

Tout ceci pour introduire l'affaiblissement du stratagème.

Définissons alors ⦻p par □p & ◊p.

Plus précisément regardons l'ensemble des formules A tel qu'une machine adéquate M ⊢ DEON-A(A) (voir plus haut).

Mieux. Procédons en deux étapes comme Boolos, Goldblatt, ou Kusnetsov & Muravitsky ont procédé avec le stratagème fort. Je définis d'abord une transformation opérant dans les langages modaux :

DEON(p) = p avec p variable propositionnelle
DEON(AvB) = DEON(A) v DEON(B)
DEON(A&B) = DEON(A) & DEON(B)
DEON(¬A) = ¬DEON(A)
DEON(□A) = □(DEON(A)) & ◊DEON(A))

Dans ce cas DEON-A = $MB_F \circ DEON$, et on peut définir le système, grâce à la complétude arithmétique de G :

$$KD? = \{A \mid M \vdash MB_F \circ DEON(A)\} = \{A \mid G \vdash DEON(A)\}.$$

On ne peut pas ne pas s'intéresser aux propositions images de DEON **vraies** *sur* la machine. Celles-ci, grâce à la complétude arithmétique de G* sont données par le système :

$$KD?^* = \{A \mid MB_F \circ DEON(A)\} == \{A \mid G^* \vdash DEON(A)\}$$

L'inclusion de G dans G* entraîne l'inclusion de KD? dans KD?*. On verra, qu'à la différence de S4Grz et ARIL qui sont invariants pour la starrification, KD? est strictement inclus dans KD?*.

Comme "⦻" est arithmétisable, on sait déjà, avec Thomason (voir 2.3.2), qu'on va perdre soit l'axiome K, soit l'*immodeste* □(□p->p), soit la règle de Nécessitation. Qu'en est-il ?

*La logique de* ☐

Avec le stratagème affaibli sur G on gagne ◇T, (puisque G ⊢ □T v ◇T), mais on perd ☐T, (puisque G ⊬ □T &◇T). Ceci est fâcheux, car on ne peut plus espérer une sémantique de Kripke. En effet □T est vrai dans tous les mondes de tous les modèles de tous les référentiels de Kripke. Cela entraîne aussi la perte de la fermeture pour la règle de nécessitation.

Il est facile de se convaincre que KD? prouve les formules[69] K, M, C, où :

K est l'axiome de Kripke,
M est la formule □(p & q) -> (□p & □q),
C est la formule (□p & □q) -> □(p & q).

On peut aussi vérifier ces formules avec le démonstrateur de théorèmes donné dans l'annexe 2 :

**On a M :**

```
? (kdip '((bw (a & b)) -> (bw a & bw b)))     ; faux avec v à la place de &
NIL
```

**On a C :**

```
? (kdip '((bw a & bw b) -> (bw (a & b))))     ; vrai aussi avec v
NIL
```

**mais on n'a pas N :**

```
? (kdip '(bw (p -> p)))
(((BW (- (-> P P)))))
```

*Proposition* KD? est fermé pour MP et RM, où RM est la règle de monotonie rationnelle :

$$p \to q \Rightarrow \Box p \to \Box q \quad \text{RM}$$

Preuve (pour RM)

D'abord G est fermé pour RM. En effet, par une simple application de la nécessitation p->q => □p->□q, et de même : G est fermé pour la règle p->q => ◇p->◇q, donc p->q => □p&◇p ->□q&◇q, donc G est fermé pour p->q => ☐ p->☐ q.

*proposition* KD?* n'est pas fermé pour RM :

---

[69] Je suis la nomenclature de Chellas 1980. La formule □T s'appelle N.

KD?*⊢ T->□T, mais KD?*⊬ □T->□□T. (ce qu'on peut toujours vérifier avec le démonstrateur donné dans l'annexe 2).

Ceci n'est pas étonnant. Après tout G* lui-même n'est pas fermé pour RM : G* ⊢ T->◊T, mais G* ⊬ □T->□◊ T.

*proposition* KD?* démontre N et T : KD?*⊢ □T, KD?*⊢ □p->p

Preuve : KD?*⊢ N est une conséquence directe de G* ⊢ ◊T, et KD?*⊢ T est une conséquence directe de G*⊢ □p->p.

KD?* est donc une extension de KT (sans Nec) et pourrait être proposé pour une forme de connaissance immédiate et incorrigible de la part d'une machine. Ce résultat montre que KD? est *strictement* inclus dans KD?*.

Je conserve le nom "KD?*", afin qu'on se souvienne que cette logique est obtenue par la "starification" de KD?. La starification correspond au passage de la prouvabilité ***par*** la machine (G) à la vérité ***sur*** la machine (G*).

Question : Est-il possible d'obtenir une axiomatisation de KD?* à partir de KD? + T + la suppression de RM ?

Proposition KD?* ⊬ □p->□□p, et donc puisque KD? est inclus dans KD?*, KD? ne prouve pas non plus la formule 4.

*Définition*

a) La règle d'inférence RE est la suivante : p<->q => □p <-> □q.

b) Une logique modale est dite classique minimale si

1) ses théorèmes sont fermés pour la règle[70] RE

2) le schéma □p <-> ¬ ◊ ¬p est vérifié.

Comme 2) est une conséquence des lois de de Morgan dans G, on sait que KD? constitue une logique classique minimale.

Les logiques classiques minimales admettent une sémantique dite des voisinages ou encore sémantique de Scott-Montague (Chellas 1980).

*Sémantique de Scott-Montague*

Un modèle de Scott-Montague, appelé aussi *modèle minimal* (W, N, V) est la donnée d'un ensemble de mondes W, d'une fonction N de W dans

[70] Rien avoir avec ensemble RE = ensemble Récursivement Enumérable.

$2^{2^W}$, qui associe à chaque monde $\alpha$ un ensemble d'ensembles de mondes[71], $N(\alpha)$ appelé système de voisinage de $\alpha$.

A nouveau chaque monde satisfait la logique propositionnelle classique. Définition

$\Box p$ est vrai dans un monde $\alpha$ si l'ensemble des mondes (de W) où p est vrai, noté $[p]$, appartient à $N(\alpha)$ :

$$\alpha \Vdash \Box p \Leftrightarrow [p] \in N(\alpha),$$

de même on exige, pour avoir $\neg \Box \neg = \Diamond$

$$\alpha \Vdash \Diamond p \Leftrightarrow W \setminus [p] \notin N(\alpha).$$

p est satisfaite par un modèle si p est vraie dans tous les mondes du modèle.

Exemple : un modèle minimal où $\Box T$ n'est pas satisfait :

$$\alpha \nVdash \Box T \Leftrightarrow [T] \notin N(\alpha),$$

Comme $[T] = W$, il suffit que W n'appartienne pas à $N(\alpha)$. Par exemple, $W = \{1, 2\}$ et $N(1) = N(2) = \{\}$.

*Définition* p est C-valide si p est vrai dans tous les mondes d'une classe C de modèles minimaux. Cela donne une sémantique pour la règle d'inférence : on a p => q si la C-validité de p entraîne la C-validité de q. Je renvoie à Chellas 1980 pour plus d'information. Je vais me contenter de montrer que la règle RE est toujours vérifiée avec les modèles de Scott-Montague.

*Théorème* p<->q => $\Box$p<->$\Box$q

Si p<->q est C-valide, p<->q est vraie dans tous les mondes des modèles de la classe C. Du coup, dans tous ces modèles $[p] = [q]$, mais alors quel que soit $\alpha$, monde d'un de ces modèles, $[p] \in N(\alpha)$ ssi $[q] \in N(\alpha)$, et donc pour tout $\alpha$, $\alpha \Vdash \Box p$ ssi $\alpha \Vdash \Box q$ et, par calcul propositionnel classique dans $\alpha$, on a $\alpha \Vdash \Box p \Leftrightarrow \Box q$.

Ni G* ni KD?* ne sont fermés pour RE. En effet G* $\vdash$ T<->$\Diamond$T, mais G* $\nvdash$ $\Box$T<->$\Box\Diamond$ T, et KD?* $\vdash$ T<->$\Box$T, mais KD?* $\nvdash$ $\Box$T<->$\Box\Box$T.

Conclusion : pas de sémantique de Scott-Montague, ni pour G*, ni pour KD?*.

---

71 Ou un ensemble de propositions si on identifie une proposition avec l'ensemble des mondes qui vérifie cette proposition.

Pour KD?, le fait que les formules M et C sont des théorèmes permet de montrer que pour les mondes $\alpha$, les systèmes de voisinages $N(\alpha)$ sont des quasi-filtres :

- si x et y appartiennent à $N(\alpha)$, alors $x \cap y$ appartient à $N(\alpha)$,
- si $x \cap y$ appartient à $N(\alpha)$, alors x et y appartiennent à $N(\alpha)$.

Les systèmes de voisinages ne sont pas des filtres parce qu'ils n'ont pas d'éléments maximals. Ceci est dû au fait que KD? $\nvdash \Box\top$.

En résumé, avec $ABC^{X,Y,Z}$ représentant une logique qui possède les axiomes A, B, C et les règles d'inférence X, Y, Z, on a $KDMN^{MP,RM}$ est arithmétiquement sain pour la prouvabilité-garantissant-la-consistance $\boxdot$ :

$$KDMN^{MP,RM} \vdash A \Rightarrow G \vdash DEON(A) \Rightarrow M \vdash MB_F \circ DEON(A)$$

n'est certainement pas complet :

$$M \vdash MB_F \circ DEON(A) \Rightarrow G \vdash DEON(A) \Rightarrow KDMN?^{MP,RM} \vdash A$$

Le point d'interrogation est l'analogue de Grz, qui donne la complétude de la théorie obtenue avec le stratagème fort, pour le stratagème faible. Le point d'interrogation désigne donc 0 ou 1 ou une infinité d'axiomes nécessaire pour axiomatiser "$\boxdot$". Même question pour le vrai sur "$\boxdot$". Grâce aux théorèmes de Solovay 1976, ces théories sont cependant décidables et on construit aisément un démonstrateur de théorèmes (cf annexe 2).

Intérêt de KD?

1) On n'a pas "la transitivité", ou plus exactement on a pas 4, c'est mieux circonscrire la notion d'état directement "accessible", ou d'état voisin, comme il est nécessaire de faire pour formuler arithmétiquement les paradoxes de la duplication (mais aussi le **P**aradoxe du **D**ovetelleur **U**niversel **PDU** (voir 3.3). J'utilise des guillemets pour *accessible* parce qu'on a plus de relation d'accessibilité ; l'apparition de la sémantique de Scott-Montague, par contre, rend le terme "voisin" plus pertinent.

2) Les modèles de KD devraient admettre des généralisations avec l'adjonction de poids, de flou, de quantification numérique ou par treillis. Allons-nous mettre en évidence une théorie des fonctions de croyance, comme celle de Dempster-Shafer ? (voir Shafer 1976, Smets 1988). Allons-nous introduire de façon naturelle des logiques non-monotoniques, des conditionnelles, etc. (voir à ce sujet Lamarre 1992).

3) le prédicat $\boxdot$ est arithmétisable !, en particulier voilà le point fixe de la proposition "Gödélienne" $p \leftrightarrow \neg\boxdot p$ :

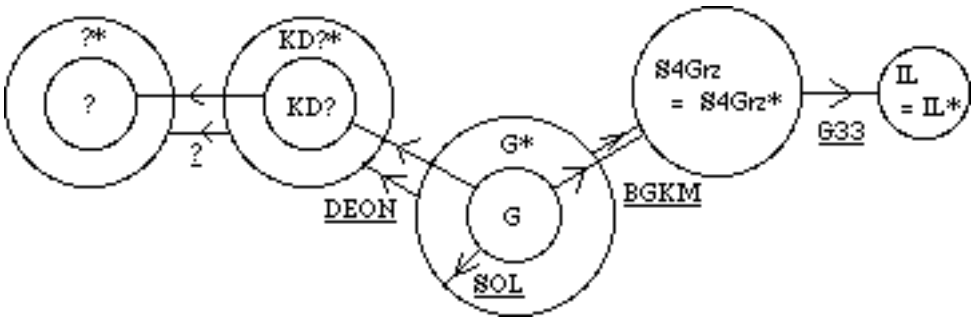
$$G \vdash \Box(p \leftrightarrow \neg\boxdot p) \rightarrow \Box(p \leftrightarrow (\Box\bot \vee \Diamond\Diamond\top))$$

*Grossièrement* : je rêve ou je vais me réveiller, je suis (déjà) mort, ou je survivrai).

ce qu'on peut vérifier directement avec G :

```
? (gip '((bw (p <-> -(bw p & - bw - p))) ->
        (bw (p <-> (bw - (p -> p) v (- bw - - bw - (p -> p)))))))
NIL
```

Résumé avec graphique




Les flèches F A->B indiquent l'existence d'une traduction F du système B dans le système A.

G permet d'étudier les propositions prouvables dans les différents systèmes de logique modale attachés aux machines autoréférentiellement correctes. G* permet d'étudier des propositions vraies et non prouvables, des systèmes de logique modale attachés de la même façon à ces machines.

*Question* : existe-il des logiques faibles "?" et "?*" (voir la figure plus haut), c-à-d une logique dont l'ensemble des théorèmes est un sous-ensemble de l'ensemble des tautologies classiques (comme la logique intuitioniste) et une traduction ? (voir la figure plus haut) de cette logique dans KD? et KD?* respectivement ? Une telle transformation jouerait le rôle de G33 avec l'interprétation arithmétique (autoréférentielle) de l'intuitionisme.

### *12) Conclusions*

Comme G* démontre $\Box p \leftrightarrow \Box p \,\&\, \Diamond p$, et que G ne démontre pas ce schéma, on peut refaire le raisonnement avec le stratagème affaibli, et donc, avec les abus de notation évidents, on a $G^* \vdash \Box p \leftrightarrow \boxdot p$, de même que $G^* \vdash \boxdot p \leftrightarrow \boxdot p$, et $G^* \vdash \Box p \leftrightarrow \boxdot p$.

Ce qui peut être résumé par la formule

$$\Box = \boxdot = \circledcirc$$

où l'égalité est interprétée extensionnellement.

Il existe cependant des propositions p, telle que G ne prouve pas $\Box p <-> \boxdot p$, ni $\boxdot p <-> \circledcirc p$, ni $\Box p <-> \circledcirc p$.
En l'occurrence pour p = ⊥, $G^* \vdash \neg\Box(\Box p <-> \boxdot p)$, $G^* \vdash \neg\Box(\Box p <-> \circledcirc p)$, et $G^* \vdash \neg\Box(\boxdot p <-> \circledcirc p)$.

Cela montre que ces égalités extensionnelles ne sont pas, dans au moins trois sens différents *productibles comme vrai* par la machine-sujet M :

| | | |
|---|---|---|
| $M \nvdash \Box(\Box = \boxdot)$ | $M \nvdash \Box(\Box = \circledcirc)$ | $M \nvdash \Box(\boxdot = \circledcirc)$ |
| $M \nvdash \boxdot(\Box = \boxdot)$ | $M \nvdash \boxdot(\Box = \circledcirc)$ | $M \nvdash \boxdot(\boxdot = \circledcirc)$ |
| $M \nvdash \circledcirc(\Box = \boxdot)$ | $M \nvdash \circledcirc(\Box = \circledcirc)$ | $M \nvdash \circledcirc(\boxdot = \circledcirc)$ |

On peut vérifier ces formules par des raisonnements élémentaires, ou utiliser les démonstrateurs de théorèmes donnés dans l'annexe[72] 2.
L'analyse du sujet de la section précédente, compatible avec l'interprétation en terme de duplication, se reconstruit donc aisément dans l'interprétation arithmétique.

La consistance $\Diamond\top$ reste la meilleure approximation de la conscience de l'autre (ou de soi vu comme un autre dans le cas de la duplication), de même que $\diamondsuit\top$ reste, avec l'hypothèse mécaniste une bonne approximation de la conscience de soi (non représentable par la machine ou le sujet).
Avec le stratagème $\diamondsuit\top$ est équivalent à $\Diamond\top \vee \top$, et l'analyse ne peut fonctionner qu'avec une machine qui, vue comme extension RE de l'arithmétique de Peano doit être suffisamment complexe pour que sa consistance ne soit pas évidemment démontrable par le sujet.

---

[72] ? Exemples d'application, comme dans les dessins, le " ¬ " s'écrit "-" :
(G*ip '(- (bw (bw (p & - p) <-> ((bw (p & - p)) & (p & - p))))))
NIL
? (Gip '(- (bw (bw (p & - p) <-> ((bw (p & - p)) & (p & - p))))))
(NIL)
? (G*ip '(bw p <-> bw p & p))
NIL
? (Gip '(bw p <-> bw p & p))
(((BW P) (- P)))
? (Gip '(bw p <-> bw p & (- bw - p)))
(((BW P) (BW (- P))))
? (G*ip '(bw p <-> bw p & (- bw - p)))
NIL

Je pense de même que la logique de "⊡" devrait servir pour les croyances rationnelles immédiates, ou une forme de probabilité (voir 1.3 et 3.3). La part non communicable est donnée par KD?*\KD?, et devrait correspondre à une forme d'intuition immédiate de la part de la machine auto-référentiellement correcte.

*13°) Résumé de 2.3.4*

*Le stratagème revient à définir* ⊡p *par* □p & p. *Une expression hybride comme* ⊡□p *est permise. Elle est interprétée par* □□p&□p *avec* □p *représentant la prouvabilité formelle arithmétisable (et diagonalisable), de même,* ⊡⊡p *se laisse traduire récursivement en* □(□p & p) & (□p & p). *En effet on peut définir le morphisme suivant (d'un langage modale dans lui-même).*

*Le morphisme de Boolos,Goldblatt et Kusnetsov et Muravitsky:* BGKM *de MPL dans MPL*

BGKM($p_i$) = $p_i$
BGKM(AvB) = BGKM(A) v BGKM(B)
BGKM(A&B) = BGKM(A) & BGKM(B)
BGKM(¬A) = ¬BGKM(A)
BGKM(□A) = □(BGKM(A)) & BGKM(A)

*A présent, en composant les morphismes de Magari-Boolos* $MB_F$ *avec le morphisme de Boolos-Goldblatt BG, on obtient une interprétation de S4 dans le langage de la machine M.*

*A la façon de Magari-Boolos, on peut, en fait, définir directement, une interprétation de l'ensemble des propositions modales prouvables par S4 dans l'ensemble des propositions de l'arithmétique ou d'une machine saine et adéquate. La traduction correspondante est alors définie inductivement*[73] *de la façon habituelle :*

*On peut alors montrer que la théorie suivante, appelée S4Grz*

| | | | |
|---|---|---|---|
| AXIOMES: | □p->p | | T |
| | □p->□□p | | 4 |
| | □(p->q)->□p->□q | K | |
| | □(□(p->□p)->p)->p | | **Grz** |
| REGLES: | p et p->q entraîne q | MP | |
| | p entraîne □p | NEC | |

*est complète à la fois pour la vérité et la connaissabilité.*
*Cette équivalence capture un aspect parmi les plus "Brouwériens" de l'intuitionisme solipsiste de Brouwer.*

*Grzegorczyk avait montré que, comme avec S4, on a*

S4Grz ⊢ G33(A) ssi IL ⊢ A

[73] De même pour ⊡, on peut définir l'interprétation arithmétique correspondante, avec T(□X) = B(⌜T(X)⌝) & ¬B(⌜¬T(X)⌝) (voir texte).

*Avec les travaux sémantiques de Segerberg, il n'est pas diificile de montrer que*

$$S4Grz \vdash A \text{ ssi } G \vdash BG(A)$$

*si bien qu'en combinant les morphismes de Magari-Boolos avec le morphisme de Boolos-Goldblatt, on obtient*

$$S4Grz \vdash A \text{ ssi pour tout } F\ M \vdash MB_F(BG(A))$$

*De même, la composition des morphismes de Magari-Boolos, Boolos-Goldblatt, avec Gödel 33 donne une interprétation arithmétique de l'intuitionisme. Cet intuitionisme dérive en quelque sorte directement de l'autoréférence correcte, c'est pourquoi je le dénomme ARIL.*

*Les extensions de* $\boxdot$ *et de* $\Box$*, ainsi que de la croyance obtenue avec une version affaiblie du stratagème (plus psychologique* $\circledcirc$*) sont identiques en ce qui concerne les propositions purement ontiques (arithmétique), mais diffère intensionnellement. Ceci correspond au fait que G* prouve les équivalences (extensionnelles) entre les trois opérateurs :*

$$G^* \vdash \Box = \boxdot = \circledcirc$$

*mais G n'en prouve aucune. G* ne prouve pas qu'il y ait une équivalence intensionnel, (qui est bien sûr) fausse d'un point de vue intensionnel :*

$$G^* \nvdash \Box(\Box = \boxdot)$$
$$G^* \nvdash \boxdot(\Box = \boxdot)$$
$$G^* \nvdash \circledcirc(\Box = \boxdot)$$
$$G^* \nvdash \Box(\Box = \circledcirc) \text{ etc.}$$

*Je rappelle qu'une expression du genre "* $\Box = \boxdot$ *" est mise pour "*$\Box p <-> \boxdot p$*", avec p arithmétique (ontique).*
*Cela permet de montrer que la machine peut inférer le mécanisme critiqué par Lucas, ou que la machine peut inférer les conséquences de l'existence (non-constructive) d'un niveau de duplication.*
*Une reconstruction de l'analyse de Benacerraf est une nouvelle fois proposée. La différence entre G et G* localise l'erreur d'une façon accessible par une machine. Cette localisation peut être "produite comme vraie" par inférence. Reste à étudier de façon précise la notion d'inférence par machine, ce qui est l'objet de la sous-section suivante.*

## *Biblio locale*


**ARTEMOV** S., **1990**, *Kolmogorov's Logic of Problems and a Provability Interpretation of Intuitionistic Logic*, in Parikh R., (Ed.), Proceedings of the Third Conference on Theoretical Aspect of Reasoning about Knowledge (TARK 90), Morgan Kaufmann Publishers.

**ARTEMOV** S. and **DZHAPARIDZE** G., **1990**, *Finite Kripke Models and Predicate Logics of Provability*, Journal of Symbolic Logic, Vol 55, N° 3, pp. 1090-1098.

**BELL** J.L., **1986**, *From Absolute to Local Mathematics*., Synthese 69, pp. 409-426.

**BOOLOS** G., **1979**, *The Unprovability of Consistency, an Essay in Modal Logic*, (chapitre 13) Cambridge University Press.

**BOOLOS** G., **1980**.(a), *Provability, Truth, and Modal Logic*, Journal of Philosophical Logic, 9, pp. 1-7.

**BOOLOS** G., **1980**.(b), *On Systems of Modal Logic with Provability Interpretations*, Theoria, 46, 1, pp. 7-18.

**BOOLOS** G., **1980**, *Provability in Arithmetic and a Schema of Grzegorczyk*, Fundamenta Mathematicae, 96, pp. 41-45.

**BURNYEAT** M., **1991**, *Socrate et le jury: de quelques aspects paradoxaux de la distinction platonicienne entre connaissance et opinion vraie*, dans Canto-Sperber M. (ed.), 1991, Les paradoxes de la connaissance, essais sur le Ménon de Platon, Editions Odile Jacob, Paris, pp. 237-251.

**CHELLAS** B. F., **1980**, *Modal logic an introduction*, Cambridge University Press, Cambridge.

**FITTING** M. C., **1969**, *Intuitionistic Logic, Model Theory and Forcing*, North-Holland Publishing Company, Amsterdam.

**FLAGG** R., **1985**, *Church's Thesis is Consistent with Epistemic Arithmetic*, in Shapiro 1985.

**GÖDEL** K., **1933**, *Eine Interpretation des Intuitionistischen Aussagenkalküls*, Ergebnisse eines Mathematischen Kolloquiums, Vol 4, pp. 39-40, also in FEFERMAN & Al. 1986.

**GOODMAN** N.D., **1984**, *The Knowing Mathematician* Synthese 60, 21-38

**GOODMAN** N. D., **1985**, *A Genuinely Intensional Set Theory*, in Shapiro pp. 63-79.

**GOODMAN** N.D., **1986**, *Flagg Realisability in Arithmetic*, Journal of Symbolic Logic, V. 51, N° 2, pp. 387-392.

**GOLDBLATT** R., **1978**, *Arithmetical Necessity, Provability and Intuitionistic Logic*, Theoria, Vol 44, pp. 38-46.

**GRZEGORCZYK** A., **1964**, *A Philosophicaly Plausible Formal Interpretation of Intuitionistic Logic*, Indagationes Math. 26, pp. 596-601.

**GRZEGORCZYK** A., **1967**, *Some relational systems and the associated topological spaces*, Fundamenta Mathematicae, LX pp. 223-231.

**KUZNETSOV** A. V.and **MURAVITSKY** A. YU., **1977**, *Magari Algebras*, Fourteenth All-Union Algebra Conf., Abstract part 2 : Rings, Algebraic Structures, Novosibirsk Univ., Novosibirsk, pp. 105-106 (En Russe).

**HYLAND** J. M. E., **JOHNSTONE** P. T. and **PITTS**, **1980**, *Tripos Theory*, Math. Proc. Camb. Phil. Soc. 88, pp. 205-232.

**KAPLAN** D. and **MONTAGUE** R., **1960**, *A paradox Regained*, Notre Dame Journal of Formal Logic, 1, pp. 79-90.

**LAMARRE** P., **1992**, *Etude des Raisonnements non-monotones : apports des logiques des conditionnels et des logiques modales*, Thèse, Université Paul Sabatier, Toulouse.

**LAWVERE** F. W., **1969**, *Adjointness in Foundations*, Dialectica, Vol 23, N° 3/4, pp. 281-295.

**McKINSEY** J. C. C. & **TARSKI** A., **1948**, *Some Theorems about the Sentential Calculi of Lewis and Heyting*, Journal of Symbolic Logic, 13, pp. 1-15.

**MONTAGUE** R., **1974**, *Syntactical treatments of modality, with corollaries on reflexion principles and finite axiomatizability*, in R. Montague, Formal Philosophy, New Haven and London.

**MYHILL** J , **1960**, *Some Remarks on the Notion of Proof*, Journal of Philosophy 57, pp. 461-471.

**MYHILL** J., **1985**, *Intensional Set Theory*, in Shapiro 1985, pp. 47-61.

**REINHARDT** W.N., **1985**, *Absolute Version of Incompleteness Theorems*, Noûs, 19, pp. 317-346.

**REINHARDT** W.N., **1986**, *Epistemic Theories and the Interpretation of Gödel's Incompleteness Theorems*, Journal of Philosophical Logics, 15, pp. 427-474.

**SEGERBERG** K., **1971**, *An essay in Classical Modal Logic*, Filosofiska Studier, 3 volumes, Uppsala.

**SHAFER** G., **1976**, *A Mathematical Theory of Evidence*, Princeton University Press, New Jersey.

**SMETS** P., **1988**, *Belief Function*, chapitre 9 de Smets & *aL* 1988.

**SMETS** P., **MAMDANI**, E. H., **DUBOIS** D., **PRADE**, H., **1988**, *Non Standart Logics for Automated Reasoning*, Academic Press, Londres.

**SHAPIRO** S., **1985**, *Epistemic and Intuitionistic Arithmetic*, in Shapiro 1985, pp. 11-46.

**SHAPIRO** S. (Ed), **1985**, *Intensional Mathematics*, North-Holland.

**THOMASON** R. H., **1980**, *A Note on Syntactical Treatment of Modality*, Synthese, 44, pp. 391-395.

## 2.3.5 L'inférable (*intelligence artificielle théorique*)

*Brièvement*

*On définit et on étudie les capacités théoriques d'inférence inductive des machines. L'accent est mis sur le rôle de l'intensionnalité et la nécessaire non constructivité.*

---

*1°) introduction*

On peut ordonner les réfutations de l'argumentation de Lucas selon la portion de raisonnement jugée valide (à une éventuelle reconstruction près).

- ***Au sommet,*** on peut placer l'argumentation de Benacerraf, ou plutôt la reconstruction BCR, qui conserve la majeure partie de l'argumentation de Lucas : la correction apportée est une nuance (de taille) épistémique ; la conclusion antimécaniste de Lucas, que l'on peut écrire sous la forme $\boxdot \neg \exists i(je=i)$, est remplacée par la conclusion épistémique $\neg \exists i \boxdot (je=i)$.

- ***Au milieu,*** la réfutation de Putnam ou de Priest conserve beaucoup moins. Remarquons que l'argumentation de Priest est équivalente à celle de Lucas :

Lucas montre MEC -> ¬CON, or CON (selon Lucas), donc ¬MEC
Priest montre CON -> ¬MEC, or MEC (selon Priest), donc ¬CON

Cependant, sans corrections épistémiques, tous les deux acceptent ¬MEC ∨ ¬CON. Ce qui signifie qu'on devrait pouvoir appliquer une reconstruction BCR à l'analyse de la transconsistance de Priest[74]. Indirectement c'est ce vers quoi on se dirige à présent, mais en partant de la réfutation du plus bas niveau de l'argumentation de Lucas. En effet

- ***A la base,*** hormis la négation de la thèse de Church ou de l'identification de base, on peut, comme Arbib, ne pas prendre au sérieux les réfutations reposant sur le théorème de Gödel car celles-ci reposent sur le concept de connaissance ou de croyance certaine, ou encore sur les notions de prouvabilité intuitive/formelle (cela revient au même avec l'analyse des sections précédentes), alors que la quasi-totalité des croyances humaines (de la vie de tous les jours) sont

a) d'une part de type incertain, vague, probabiliste ou simplement de nature inductive.

---

[74] Ce qui donne (dans une analyse un peu naïve) $\neg \exists i \boxdot (je=i) \vee \Box \bot$. Ceci, interprété dans G, donne une interprétation grossière de la transconsistance de Priest. Son inconsistance ($\vdash \bot$) est approchée (un peu trivialement à ce niveau) par $\Box \bot$ qui n'entraîne le faux qu'au niveau G*.

b) d'autre part provient a priori d'un couplage entre un sujet (un individu, une collection d'individus) et un environnement (cf Arbib, voir plus haut).

La croyance au fait que le soleil va se lever demain, où la croyance dans les lois de la nature et d'une façon générale les lois "scientifiques", ne sont jamais prouvées ou communiquées comme tels. Ces connaissances (croyances) sont seulement inférées à partir de l'observation. Elles peuvent aussi être *relativement* prouvée dans une théorie qui est elle-même *inférée* , avec divers degrés de plausibilité, à partir de l'expérience[75]. De telles croyances *peuvent* être correctes, dans ce cas ces croyances (ces connaissances si on applique le stratagème fort dans ce contexte) peuvent être éventuellement confirmées, et c'est le mieux que l'on puisse espérer.

Utiliser le fait que notre connaissance est de nature inductive pour rendre non pertinente l'argumentation de Lucas revient à introduire de nouveaux critères de comparaison entre l'homme et la machine. En effet Lucas a tenté de prouver qu'il est *supérieur* à toute machine. La relation de supériorité envisagée par Lucas utilise la taille des ensembles de propositions vraies que l'entité homme ou machine peut produire.

Lucas démontre que pour chaque machine qu'on lui présente, il est capable de, trouver une proposition vraie que cette machine ne sait pas produire (ce qui en fait suffit pour conclure qu'il est différent de chaque machine[76]), $\neg\exists i(je = i)$. Affirmant cependant être à même de produire sa propre consistance et celle de la machine, Lucas conclut encore qu'il est *supérieur* à toutes machines : $\forall i(je > i)$.

Nous avons vu dans la section précédente les faiblesses de l'argumentation (basée sur le théorème de Gödel) de Lucas.

Existe-t-il une argumentation similaire pour la connaissance inductive ?

Peut-on espérer baser une telle argumentation sur les résultats de Gödel, ou sur les techniques d'arithmétisation et de diagonalisation utilisée par Gödel ?

Existe-t-il un correspondant de Lucas pour l'inférence inductive ? Il semble que la réponse soit affirmative, et que Putnam 1988 (au moins) ait accepté de jouer ce rôle. Putnam 1964 est à la fois un des pionniers de la philosophie fonctionnaliste, sous la forme de MEC-DIG-FORT[77] , et quelques années après, vers la fin des années 70 semble-t-il, il est devenu un de ses détracteurs. En 1988 Putnam écrit :

---

[75] J'aurai l'occasion d'argumenter qu'il en est de même pour la conscience, comme Helmholtz pour les perceptions.

[76] Pour réfuter l'hypothèse mécaniste indexical, il est pourtant suffisant de prouver être différent de chaque machine. Quelqu'un qui prouverait, par exemple, qu'il est *inférieur* à toutes machines possibles réfuterait aussi l'hypothèse mécaniste indexicale.

[77] Un des premiers (en faisant toujours abstraction de Post 1921 et de Turing 1950) à critiquer l'usage des théorèmes de Gödel pour réfuter les théories mécanistes de l'esprit (voir 2.3.1).

> *What Gödel showed is, so to speak, that we cannot formalize our own mathematical capacity because it is part of that mathematical capacity itself that it can go beyond whatever it can formalize. Similarly, my extension of Gödelian techniques to inductive logic showed that it is part of our notion of justification in general (not just of our notion of mathematical justification) that reason can go beyond whatever reason can formalize* (Putnam 1988).

Cet extrait fait partie d'une argumentation contre le fonctionnalisme et le mécanisme[78]. Putnam se réfère à un résultat de 1963, où il opère une diagonalisation sur une collection de machines extrapolantes, pour démontrer l'inexistence d'une machine extrapolante universelle (voir aussi Putnam 1965). A moins de donner un argument convaincant selon lequel l'homme (ou un homme) est un extrapolateur universel, ce résultat ne distingue pas l'homme de la machine. Il ne semble d'ailleurs pas que Putnam ait été tenté par cette conclusion en 1963. Les travaux de Putnam de 1963 et 1965, ceux de Solomonoff 1964a, 1964b, ainsi que celui de Gold 1965 figurent parmi les premiers travaux dans le domaine mathématique de l'inférence inductive *théorique*[79].

Putnam donne une première définition d'un concept clair (l'extrapolation) analysable dans le domaine des *théories* possibles de l'inférence inductive, et il donne la première démonstration d'un résultat théorique.

La théorie de l'inférence inductive va s'asseoir confortablement sur la théorie de la récursion avec les travaux de Gold (Gold 1965, 1967) et connaître un large développement. On peut distinguer *principalement*[80] les écoles allemandes (Zeugman, Wiehagen), les écoles soviétiques ou baltes (Barzdin), et les écoles américaines (Case, Daley, Osherson-Stob-Weinstein).

*2°) Machines extrapolantes et relation d'ordre d'intelligence*

Revenons à Putnam et cherchons de nouveaux critères de comparaison d'intelligence entre machines[81]. Pour faire de l'intelligence artificielle

---

[78] Plus précisément Putnam argumente que le mécanisme est faux ou trivial.

[79] Il s'agit d'une branche de l'intelligence artificielle théorique. Il va de soi que l'inférence inductive est plus ancienne et plus large. Pour citer quelques noms : Bacon, Locke, Helmholtz, De Gerando, Ramsey, Carnap, Kyburg, Shapiro (Ehud), Demsper, Shafer, Smets.

[80] cf Minicozzy en Italie par exemple. La théorie de la récursion est connue pour s'être développée indépendamment et parallèlement à l'est et à l'ouest. L'événement le plus célèbre dans ce cadre est la solution donnée par Friedberg et Mucnik au fameux problème de Post 1944 concernant la classification des ensembles récursivement énumérables.

[81] Avec MDI, cela suffit. Cela ne signifie pas qu'on ne puisse utiliser la théorie de la récursion pour aborder la reconnaissance chez des entités plus générales. Osherson, Stob, et Weinstein donnent l'exemple de collections particulières d'ensembles de nombres qui ne sont pas identifiables à partir d'*exemples positifs* (c-à-d d'éléments qui appartiennent à la collection) (comme FINSET U ω, FINSET = la collection d'ensembles finis) ni par l'humain, ni par la machine, il existe aussi des collections reconnaissables de fonctions, mais pas par des machines. Dans ce cas cependant on peut montrer que ces collections peuvent être reconnues par des machines avec l'oracle de l'arrêt (K, <O', voir aussi Brandt 1986, Posner 1980), si bien qu'utiliser ces résultats pour distinguer l'homme de la machine, revient à refaire (implicitement au moins) le raisonnement de Kalmar pour réfuter la thèse de Church (voir 2.2). Ce n'est pas ce qu'on va faire ici. L'idée est plutôt

théorique il n'est pas nécessaire de définir l'intelligence, seulement de définir, une relation d'ordre entre entité susceptible de manifester une capacité d'apprentissage de façon extensionnelle, comme avec la capacité d'extrapoler correctement. Pour ce faire l'approche de Binet 1911 est déjà suffisante :

*Définition informelle 1* : un test de "quotient intellectuel" consiste en la présentation d'une suite de données obéissant à une loi effective (et donc programmable a priori (avec la thèse de Church).

*Définition informelle 2* une entité réussit ou passe le test si, lorsqu'on lui présente la suite successivement, elle parvient à un moment donné à extrapoler correctement la suite présentée.

*Définition informelle 3* une machine M1 est dite plus intelligente qu'une machine M2 si la collection des tests réussissables par M2 est strictement incluse dans la collection des tests réussissables par M1.

L'effectivité du test permet de représenter celui-ci par une fonction totale récursive de ω dans ω, au moyen d'un codage approprié.

*Remarques*

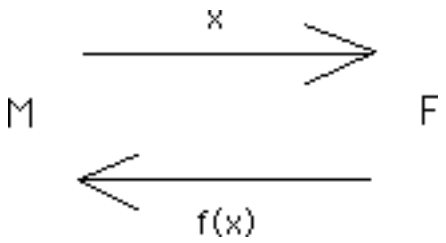
1) Malgré l'aspect positiviste de cette définition, et malgré l'effectivité de la présentation d'un test, la définition ne donne pas un critère effectif de reconnaissance de l'intelligence. A priori rien ne permet de sélectionner un ensemble fini de tests permettant de comparer l'intelligence de deux entités (hommes ou machines) données. La relation d'ordre est donc non constructive a priori. Les critiques habituelles sur l'usage des tests de quotients intellectuels dans la psychologie pratique restent *a priori* valables (voir Gould 1983).

De plus l'ordre induit par la définition de la comparaison entre l'intelligence des entités n'est pas *a priori* total ni linéaire. Des entités différentes peuvent réussir sur des collections disjointes de tests, ou sur des collections qui ne sont pas inclues l'une dans l'autre. Tous ces *a priori* vont être confirmés, a posteriori, dans cette section. On constatera par exemple que la définition simple de Binet de la relation d'ordre pour l'intelligence entraîne l'existence de degré d'intelligences *incomparables*.

---

d'illustrer comment une définition relativement simple de comparaison de compétence en matière de reconnaissance explose en une multitude de comparaisons plus fines possibles entre les machines elles-mêmes. On peut estimer la ressemblance entre les machines intelligentes et nous-mêmes (erreurs permises, inconsistance, probabiliste, *asking questions*, etc, ...). Les machines seront ultimement modestes face à l'inconnu, qu'incarne notamment, avec MDI, la machine universelle.

2) Il ne faut pas attacher au terme "quotient intellectuel" un sens trop restreint. L'extrapolation d'une fonction totale récursive peut aussi servir à modéliser l'activité d'un scientifique essayant de prédire un résultat d'expérience à partir des résultats obtenus précédemment, par exemple la position d'une planète dans le ciel (comme Leverrier) ou d'un élément atomique dans un tableau (comme Mendeleev). C'est l'interprétation en terme de philosophie des sciences proposée par Case et Smith 1983, et défendue encore par Case 1986 (avec un critère d'identification à la place de l'extrapolation, voir plus loin). Case 1986 représente la situation ainsi[82] :




M représente un agent (un sujet) en train de faire une expérience x sur un phénomène F et f(x) est le résultat de l'expérience. Son but est d'extrapoler ou de prédire le résultat d'expérience future sur F.

Dans ce cas la nature, où un aspect phénoménologique partiel de la nature, se comporte comme un test de "quotient intellectuel" grandeur nature. On peut arguer qu'il en est de même pour le chat en train de chasser une souris. En effet comme la souris se déplace pendant que le chat lui court après, si on veut éviter un paradoxe du type d'Achille et la tortue, il est nécessaire de conférer au chat une capacité minimale d'extrapolation.

D'une façon générale l'extrapolation d'une fonction totale récursive peut servir à modéliser les processus d'apprentissage. Une fonction est apprise par une entité M lorsque cette dernière est à même ultimement d'extrapoler correctement cette fonction.

*3°) la diagonale de Putnam*

Je prends les définitions précises de Case et Smith 1983.

*Définition* une machine extrapolatrice M est une machine qui reçoit comme entrée une suite (qui peut être vide) de nombres naturels et qui éventuellement sort (après un certain temps) un nombre naturel.

$M(x_0,x_1,x_2,...,x_{n-1})$ dénote la sortie de M, si elle existe, sur la suite $(x_0,x_1,x_2,...,x_{n-1})$. Cette suite joue le rôle de test. Toutefois, comme on s'intéresse à l'apprentissage, le critère de succès de l'extrapolation est défini à la limite[83] :

---

82 Pour être précis, Case 1986 introduit ce dessin pour les critères d'identification (et non d'extrapolation) comme celui de Gold, introduit plus loin.

83 L'apprentissage est à la connaissance et à la communication ce que la perception est à l'extrapolation immédiate. Il existe de nombreux travaux dans l'extrapolation en un coup, où une seule hypothèse est permise. C'est aussi ce à quoi l'on doit procéder pour les inférences des expériences de

*Définition* : une machine extrapolante M extrapole une fonction f, ce qu'on écrira sous la forme $f \in NV(M)$, ssi

1) $\forall n \in \omega$, $\forall x_0,x_1,x_2,...,x_{n-1} \in \omega$,
$M(x_0,x_1,x_2,...,x_{n-1})$ est défini,

2) $\forall^{\infty} n\ M(f(0),f(1),...,f(n-1)) = f(n)$.

NV est mis pour ***N**ext **V**alue*. Par la condition 1 on exige de cette machine qu'elle donne toujours un résultat sur toute suite de données, c'est-à-dire M est totale. Elle identifie correctement f si elle est capable d'extrapoler f pour les tests suffisamment longs. On regardera plus loin ce qui se passe si on supprime la condition 1.

*Définition* un extrapolateur M est dit universel s'il est capable d'extrapoler toutes les fonctions totales récursives. Dans ce cas, le "graal" R (voir 2.1) est inclu dans NV(M).

*Théorème* (Putnam 1963)
Il n'existe pas de machine extrapolatrice universelle

*preuve* supposons qu'une telle machine M existe. Soit *e* son code. On va définir, par induction + diagonalisation, une fonction totale calculable que M ne peut pas reconnaître :

$$f^0 = \{\}$$
$$f^{n+1} = f^n \cup \{n, 1 \dot{-} [M(f^n)](n)\}$$

$f = \bigcup_{n \in \omega} f^n$ Si M reconnaît f, on a $f(n) = 1 \dot{-} f(n)$, absurde[84].

Cette preuve traduit l'argument intuitif souvent évoqué par ceux qui prétendent que le concept d'inférence inductive n'a pas de sens dans la mesure ou l'on peut toujours contredire une inférence inductive prédictible.

---

l'autoduplication. Une réflexion approfondie ici devrait illustrer les intensions capturées par le stratagème fort, qui concerne la connaissabilité dans les voisinages de l'infini (à *terme* non borné) et le stratagème *affaibli* qui concerne plus une notion de croyance ou même de "*probance*" dans les voisinages de zero (à *terme* borné). Le terme "terme" désigne ici, par exemple, le nombre de changements d'avis de la machine. Des résultats inspirant à cet égard sont ceux de Freivalds rappelés et étendus par Daley, Kalyanasundaram, Velauthapillai 1992.

[84] Je rappelle que $x \dot{-} y = x-y$ si $x \geq y$, et 0 sinon.

Remarquons, que s'il n'existe pas de machines capables d'extrapoler une quelconque fonction totale récursive, pour chaque fonction totale récursive f prise individuellement, il existe une machine capable de l'extrapoler.

Il suffit en effet de doter cette machine particulière du code *e* de *f* avec une machine universelle (pour calculer f). Elle est à même alors de distinguer des entrées-sorties d'une fonction totale g différente de f. Dans ce cas elle sort 0 (ou n'importe quoi, c'est juste pour faire en sorte qu'elle soit définie sur toutes les suites), sinon elle sort $\phi_u(e, n)$ sur (f(0),f(1),...,f(n-1)). Cette machine est évidemment triviale car elle n'extrapole qu'*une* fonction. Le concept intéressant est donc celui de *collection* de fonctions extrapolables par *une* machine M.

*Définition* NV = la classe de toutes les collections de fonctions totales récursives extrapolables.

$$NV = \{E \mid \exists M\ E \subseteq NV(M)\}.$$

Le théorème de Putnam peut s'énoncer sous la forme

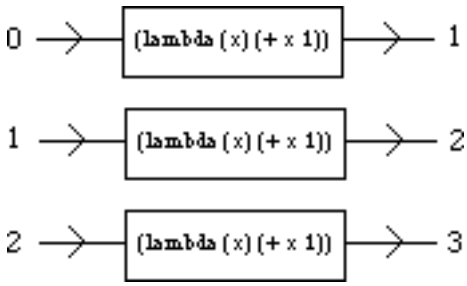
$$R \notin NV.$$

Putnam (1988) interprète son résultat comme une limitation de l'hypothèse mécaniste dans la psychologie cognitive, mais R est la limite (jamais atteinte) de tout ce qui est construit par les écoles du dedans. C'est donc beaucoup demandé et on pourrait se restreindre à quelques sous-ensembles de fonctions totales récursives. On pourrait se limiter par exemple à un ensemble de fonctions prouvablement totales récursives dans une théorie donnée. On peut aussi limiter l'ambition en admettant des critères d'extrapolation moins sévères, en supprimant par exemple la condition 1. Avant ça, on va regarder un autre type d'inférence inductive basée sur l'*identification* d'un phénomène, plutôt que son *extrapolation*.

On sera amener à définir et réfléchir sur l'idée (Socratique) d'auto-identification, ainsi qu'à une éventuelle notion d'auto-extrapolation. Nous verrons en 3.1 des exemples d'auto-extrapolations instinctives immédiates telles qu'elles apparaissent dans les rêves nocturnes.

*4°) Machines identifiantes (Gold 1967)*

Considérons la fonction λx x+1. L'utilisation typique de l'ordinateur consiste, pour calculer cette fonction, à écrire un programme, par exemple (lambda (x) (+ x 1)), et ensuite de faire éxécuter ce programme sur l'ordinateur :




Une machine à inférence inductive effectue en quelque sorte l'opération inverse :


On présente donc à gauche la sémantique (extensionnelle) d'une fonction, et la machine génère un programme qui calcule cette fonction, du moins lorsqu'elle parvient à identifier cette fonction. Bien sûr, cette sémantique est d'une façon générale une structure infinie. On présente donc un graphe d'une fonction à une machine, et elle est censée trouver un programme (une théorie) capable de prédire (calculer) cette fonction.

L'inférence inductive est inverse de la déduction. En fait, outre le passage de l'infini (le phénomène) au fini (l'explication) on peut distinguer deux aspects singuliers de l'inférence inductive :

a) la reconnaissance de phénomènes récursifs ou programmables est le processus inverse de l'interprétation usuelle d'un programme par un ordinateur.

> $i \rightarrow \phi_i$ *interprétation* par une machine universelle de Turing,
> $\phi_i \rightarrow i$ (ou j tel que $\phi_j = \phi_i$) *inférence inductive.*

b) il est connu qu'à un certain niveau d'abstraction l'interprétation $i \rightarrow \phi_i$ ou l'évaluation (d'un terme) peut être vu comme un modus ponens généralisé[85]. L' "inverse" le plus évident du modus ponens

$$A, A \rightarrow B \Rightarrow B$$

est l'abduction (Pierce)

$$B, A \rightarrow B \Longrightarrow A$$

---

[85] Abstraction capturée par les catégories entre autres, mais cela apparaît aussi avec l'écriture d'un interpréteur prolog en prolog.

qui n'est assurément pas une *règle d'inférence* valide, mais une *règle d'inférence inductive* intéressante, puisque A peut être *par défaut*, considéré comme une explication plausible de B. En particulier si "A->B" est la théorie actuelle, et si on observe le phénomène B, A peut être une explication jugée plausible.

On devine l'importance de l'utilisation de logiques non-standards, comme les logiques non-monotoniques[86], ou simplement les statistiques dans ce contexte, par exemple sous la forme des probabilités des causes avec le théorème de Bayes[87] ou de la théorie des croyances de Dempster-Shafer (voir Smets 1988, Krause et Clark 1993).

iii. passage de l'infini au fini (cf les considérations de Hobbes et Descartes dans 1.1) On se rappelle la question de savoir si une machine est à même de concevoir l'infini. On peut raisonnablement, sans tomber dans un positivisme trop réducteur, objecter que cette question offre peu de sens puisqu'il est difficile de vérifier une assertion aussi bien positive que négative de ce fait. Ici le problème est contourné car on exige seulement d'une machine qu'elle soit à même de synthétiser un programme capable de manipuler correctement des entités (par exemple des ensembles) *infinies* par le biais de descriptions ou intensions relatives, localement finitairement capturables.

On peut exiger que l'identification se fasse dans un voisinage de l'infini (identification à la limite de Gold), ou qu'elle se fasse dans un voisinage fixé a priori de zéro (identification finie de Barzdin, Freivalds). Dans ce dernier cas la généralité de l'approche fait qu'elle fonctionne aussi bien pour des programmes qui tentent d'identifier une fonction en un coup (sans plus jamais changer d'hypothèses) mais aussi le niveau des neurones d'un réseaux neuronaux, le niveau de base d'un système immunologique, etc.

Si on représente les états de connaissance/croyance par les modèles de Kripke, la différence entre ces deux types d'identification est que dans l'identification finie la relation d'accessibilité n'est pas transitive, alors qu'elle l'est avec l'identification à la limite. C'est la raison pour laquelle l'identification finie devrait jouer un rôle plus important que l'identification à la limite dans la formulation du problème du corps et de l'esprit. A l'inverse, l'apprentissage (learning) est plus fidèlement modélisé par l'identification à la limite. La phase d'apprentissage étant définie par la séquence des hypothèses incorrectes générées par la machine : apprendre, comme cela va être justifier plus bas, c'est (essentiellement) se tromper encore et encore.

---

86 Voir Besnard 1989.

87 Voir aussi Osherson, Stob et Weinstein 1988b.

***Definition*** (Gold 1967): une **M**achine à **I**nférence **I**nductive (**MII**) est une machine M qui, prend comme entrées successives toutes (à la limite) les paires <entrée/sortie> d'une fonction f (présentée extensionnellement donc) et sort de temps à autre des programmes, appelé hypothèses. On dit que f a été présentée (extensionnellement) à M.

La MII converge si, ultimement, elle sort toujours la même hypothèse. On dit que la MII *M* identifie correctement f et on écrit $f \in EX(M)$ si *M* converge vers un programme qui calcule f.

*notation* f:x = {(0,f(0) (1, f(1)) ...(x, f(x))}

$$f \in EX(M) <-> \forall^{\infty} x\ M(f:x) = p\ \&\ \phi_p = f$$

J'appelle *p* indifféremment avis, hypothèses, programmes, ou encore hypothèses-programmes émises par M.

A la différence de l'extrapolation il est indifférent que la machine M soit totale ou non. On peut en effet, à partir d'une MII non totale, construire une MII totale qui identifie les mêmes fonctions. Nous verrons cependant qu'il n'est pas indifférent que les hypothèses-programmes émises par M soient totales ou non.

Notons à nouveau que chaque $\phi_i$ est trivialement identifiable. En effet chacun des $\phi_i$ est identifié par la machine $\lambda x\ i$, qui donne toujours i pour n'importe quelle <entrée, sortie> présentée. J'appelle une telle machine une machine *idiotique*. Ici aussi, le concept intéressant est donc le concept de classe de fonctions identifiables, sous-ensemble de R (ou de P mais je vais me limiter à R) par *une* machine à inférence inductive M.

Il est facile de montrer que tout sous-ensemble algorithmiquement générable[88] de R appartient à EX (Gold 1967) mais, comme PEX, **R** lui-même n'appartient pas à EX.

*preuve* considérons un ensemble récursivement énumérable de fonctions totales récursives[89]. Un tel ensemble peut s'écrire

$$\{\phi_{p(i)} \mid i \in \omega\}$$

où p désigne une fonction totale récursive

Il faut construire une machine à inférence inductive capable d'identifier cet ensemble. Il suffit de construire M (décrit informellement) :

---

88 Lorsque je dis qu'un ensemble de fonctions (totales ou partielles) est algorithmiquement générable, je commet un abus de langage, c'est l'ensemble des indices de ces fonctions qui est ainsi générable.

89 On peut généraliser aisément en prenant des ensembles (de code) de fonctions totales récursives récursivement énumérables relativement au problème de l'arrêt.

M(f : n) = chercher, en dovetellant, le plus petit i tel que $\forall x \leq n\ \phi_{p(i)}(x) = f(x)$, sortir p(i) pour un tel i trouvé.

Comme les $\phi_{p(i)}$ sont totales, et que la fonction présentée appartient à la classe un tel processus va converger sur un code de la fonction présentée.

Une telle procédure est inutilisable pour identifier toutes les fonctions totales calculables puisque R, celui que j'appelle le graal, n'est pas RE (voir 2.1). Peut-être existe-t-il une autre procédure ? On va voir que ce n'est pas le cas.

*Definitions*

1) $EX = \{S \sqsubset R \ \exists M\ S \sqsubset EX(M)\}$.

Le résultat précédant peut s'écrire sous la forme : tout ensemble RE inclus dans R appartient à EX.

Si on exige que les hypothèses émises par la machine à inférence inductive M sont totales, on dit, avec Case et Ngo-Manguelle (1979), que la machine est *Poppérienne*. Case justifie cette appellation en faisant remarquer que les hypothèses totales émises par M sont réfutables ou falsifiables par les expériences (entrées-sorties) suivantes. Si une hypothèse est partielle la machine peut être silencieuse dans l'application de la théorie qu'elle a émise. On obtient un critère prouvablement plus stricte.

2) $PEX = \{S : \exists M$ Poppérienne $S \sqsubset EX(M)\}$.

La Poppérianité d'une machine rend extrapolable la classe des fonctions identifiables. Inversément la collection des classes extrapolables correspond à la collection des classes identifiables par des machines Poppériennes.

Précisément, on a le théorème suivant.

*Théorème*

$$PEX = NV$$

*preuve*

1) $PEX \subseteq NV$.

La preuve capture l'idée que si une fonction est identifiable, alors on sait appliquer l'algorithme produit pour extrapoler la fonction présentée. Soit un ensemble E de fonctions appartenant à PEX(M). On peut supposer sans perte de généralité que M produit une hypothèse pour chaque suite, y

compris la suite vide[90]. Comme les hypothèses produites par l'MII M sont totales, Les fonctions appartenant à E seront extrapolables par une machine extrapolatrice M définissable par

$$M(x_0,x_1,x_2,...,x_{n-1}) = \phi_{M(\{(0,x_0),...(n-1,x_{n-1})\})}(n)$$

2) NV $\subseteq$ PEX

L'idée est toute simple bien que plus délicate à formaliser. En effet si E $\in$ NV, il existe alors une machine M qui extrapole les fonctions f qui appartiennent à E. Supposons que

$$f = \{(x_0,y_0)\ (x_1,y_1)\ ...\ (x_i,y_i)\ ...\}$$

Comme M extrapole f, il existe un k tel que $M(y_0,...y_k)$ donne le résultat correct $y_{k+1} = f(x_{k+1})$, et donc est à même de générer $(x_{k+2},y_{k+2})$, $(x_{k+3},y_{k+3})$, etc. De plus $(y_0,...y_k)$ est fini, et donc codable. Donc avec le code de M et le code de $(y_0,...y_k)$, on peut construire un code p pour le calcul de f.

Donc il existe une machine M capable d'identifier à la limite f. Pour s'assurer que M est totale, on décide qu'elle sort toujours le code de $\lambda x.0$ pour les suites $\{(x_0,y_0)\ (x_1,y_1)\ ...\ (x_t,y_t)\}$ avec $t < k$.

Remarquons 1) que cette preuve n'est pas constructive puisque rien ne nous permet de déterminer k. Pour une preuve constructive voir Case et Smith 1983 ; 2) c'est la totalité des hypothèses émises par la MII qui correspond à la totalité de la machine extrapolante elle-même. La condition 1 (totalité de la machine extrapolante) exigée par Putnam dans sa définition rend ce genre de machine essentiellement Poppérienne.

On aura l'occasion d'observer ce que devient l'extrapolation avec la suppression de cette hypothèse.

*Corollaire* (une autre version de Putnam, 1963) :

$$\mathbf{R} \notin \text{PEX}$$

Cela vaut la peine de regarder à nouveau la preuve de Putnam dans le cadre des machines identifiantes. La facilité de contredire la machine (par diagonalisation) provient de ce que Putnam exige implicitement la *popperianité* de la machine.

Soit en effet m le code de la MII. La fonction calculable suivante ne peut pas être reconnue par la MII poppérienne:

---

[90] On ne restreint pas la généralité en supposant M totale, ce qui est bien sûr différent de supposer que les sorties de M soient totales (à la différence des machines extrapolantes comme cela apparaît dans la suite).

$$\phi_e(0) = 0$$
$$\phi_e(x+1) = 1+\phi_{\phi_m(\phi_e(:x))}(x+1).$$

où $\phi_e(:x)$ désigne :

$$\{(0, \phi_e(0)), (1, \phi_e(1)), \ldots (x, \phi_e(x))\}$$

En effet, supposons que m reconnaisse e, alors

$$\exists x \ \phi_m(\phi_e(: x)) = e$$

mais alors

$$\phi_e(x+1) = 1+\phi_e(x+1)$$

ce qui est contradictoire puisque, la MII étant poppérienne, $\phi_e$ est totale et donc bien définie en x+1. QED. La preuve est constructive, il est donc aisé d'ajouter des capacités auto-évoluantes par rapport à l'inférence inductive sur un réseau autoréférentiel, *à-la-Myhill-Case* (voir 2.2), de machines à inférence inductive poppérienne.

*Caractérisation de PEX*

Quelques caractérisations de PEX (et donc de NV) seront utiles.

1) En terme de complexité (théorème de Barzdin Freivald, mais aussi Adleman. cité dans Blum & Blum 1975).

M désigne une machine extrapolante, la variable h désigne une fonction totale récursive. $(\phi_i,\beta_i)$ désigne une numérotation acceptable accompagnée d'une mesure de Blum (voir 2.2).

Je rappelle qu'une fonction f est h-facile si f

$$\exists i \ \phi_i = f \text{ et } \forall^{\infty} x \ \beta_i(x) =< h(x)$$

*Théorème de Barzdin et Freivald 1972* (aussi obtenu indépendamment par Adleman 1973). Voir Blum et Blum 1975.

a) $\forall h \ \exists M \ \forall f \in R$ f est h-facile -> M extrapole correctement f
b) $\forall M \ \exists h \ \forall f \in R$ M extrapole correctement f -> f est h-facile.

*Preuve* Voir Blum et Blum 1975

2) En terme de classe RE (Barzdin et Freivald 1972, cité dans Case et Smith 1983)

PEX = {E | E est inclue dans une classe RE de fonctions récursives}

*Preuve* voir Case et Smith 1983.

*5°) Réfutation de Putnam*

Toute cette section peut être vue comme une réfutation de l'idée de Putnam comme quoi le mécanisme est soit faux, soit trivial.

La réfutation ne porte donc pas sur l'article original de 1963 mais sur son interprétation philosophique de 1988. La réfutation ne porte pas non plus sur la validité de l'interprétation philosophique, dans le sens où je partage la conclusion selon laquelle la raison (comprenant les justifications inductives) peut aller au delà de quoi que ce soit que la raison puisse formaliser. Mais cela n'enlève rien à la portée du mécanisme lequel permet de justifier de façon précise cette conséquence (Marchal 1990).

Autrement dit la critique de Putnam est en quelque sorte une critique appartenant inéluctablement aux discours des machines dans les voisinages de l'infini. Cela rejoint Minski lorsqu'il affirme que les machines intelligentes le seront non pas parce qu'elles vont résoudre les questions de philosophie et de conscience, mais plutôt parce qu'elles s'en poseront elles aussi (Minski 1968), tout en *nous* posant par ailleurs de nouvelles questions.

A vrai dire, avant d'en arriver là, je vais réfuter Putnam d'une façon intermédiaire en montrant (grâce aux résultats obtenus en inférence inductive après la publication de son papier de 1963) qu'il existe bien, asymptotiquement des machines identifiantes universelles (pour peu qu'on affaiblisse le critère de succès de l'identification[91]). Les preuves d'existence seront non-constructives. Ce sont même des conséquences *nécessairement non-constructives* des phénomènes d'incomplétude. Ensuite j'argumenterai et spéculerai sur la ressemblance psychologique entre de telles machines et nous (voir aussi 2.3.6).

*6°) La puissance de la machine silencieuse*

Popper estime qu'une théorie scientifique doit être réfutable pour être intéressante. Je propose, avec Case et Smith 1983, une *sorte* de réfutation de Popper. Il s'agit de la preuve donnée par Case et Smith qu'exiger la poppérianité, pour les machines à inférence inductive, est restrictif.

Il y a plus de classes de fonctions récursives capables d'être reconnues par de machines non-poppériennes que par des machines poppériennes.

*Theorème* (Case and Smith 1983)

---

[91] Pour peu que cet affaiblissement soit suffisant. Il va de soi que cela est possible. A l'extrême on peut proposer un critère d'identification laxiste : toutes les hypothèses sont considérées comme valable. Dans ce cas R est trivialement identifiable.

PEX $\subsetneq$ EX

Case et Smith (1983) ont montré que l'ensemble des fonctions calculées par des machines autoreproductrices, lorsqu'on leur présente, par exemple, un argument fixé particulier DUP :

$$S = \{f : \phi_{f(DUP)} = f\}$$

appartient à EX, mais pas à PEX. En effet, la fonction :

$\phi_e(0) = e$ ; comparez avec la fonction de Putnam (plus haut)
$\phi_e(x+1) = 1+\phi_{\phi_m(f_e(:x))}(x+1)$.

est trivialement identifiée, en un coup même, par la MII $\beta$ qui sort a sur {..., (0,a), ...}. En fait S appartient à FIN (dénotée aussi par $EX^0_0$), la classe des fonctions identifiables en *un seul* pari et sans erreurs (voir plus loin).

S n'appartient pas à PEX car S contient des fonctions arbitrairement croissantes presque partout[92], alors que PEX ne contient que des fonctions h-easy comme le montre le résultat de Barzdin et Freivald.

Voici à titre illustratif l'aspect constructif de la preuve traduit en LISP :

```
(def 'EX-BUT-NOT-PEX '(lambda (MII-pex-var)
 (k
  (list 'lambda (list 'y 'x)
       (list 'cond
            (list (list 'equal 'x 0) 'y)
            (list 't (list (list '+ 1
                         (list  (list MII-pex-var
                                   (list  'ulis 'y
                                        (list '- 'x 1)))
                              'x))
                    )
               )
          )
       )
  )          ;  avec ulis (i,n) = {(o,fj(o), …(n,fj(n))}
))           ;  et k est le métaprogramme de Kleene. QED .
```

*A-t-on réfuter Putnam ?*

Avec EX, oui dans les sens où on a mis en évidence l'existence d'autres critères d'identification, qui capturent d'autre classes. Non dans deux sens :

---

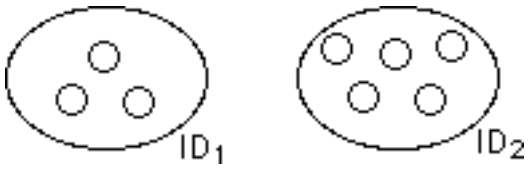
92 On se rappelle qu'avec 2-REC, pour toute fonction partielle récursive (y compris les fonctions universelles) on peut construire une machine calculant cette fonction et s'auto-dupliquant (voir 2.2).

1) qu'on a encore :
*Théorème de Gold* 1967: (le graal n'est pas encore atteint)

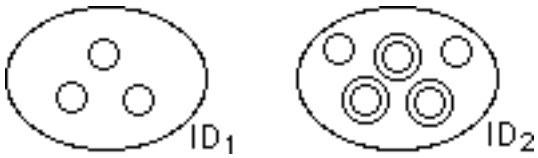
$$R \notin EX.$$

Case et Smith ont considérablement raffiné ce résultat, $R \notin$ n'appartient pas aux collections identifiées par des machines auxquelles on donne un critère d'identification de plus en plus laxiste (voir plus loin).

2) la séparation *forte*
Lorsqu'on écrit $PEX \subsetneq EX$ cela signifie que l'*ensemble* des machines non poppériennes reconnaissent plus que l'*ensemble* des machines poppériennes. Plus généralement $ID_1 \subsetneq ID_2$ signifie que la collection des ensembles de fonctions reconnus par *des* machines vérifiant le critère d'identification de $ID_1$ est inclus dans la collection des ensembles de fonctions reconnus par *des* machines vérifiant le critère d'identification de $ID_2$. Toutefois la définition de Binet concerne des machines individuelles ainsi que l'ensemble des fonctions identifiables par ces machines, et non pas des collections de machines ou des collections d'ensembles identifiables. En affaiblissant un critère d'identification, on élargit la collection des ensembles de fonctions identifiables. Cela n'entraîne pas l'élargissement d'une classe particulière de fonctions identifiables par une machine individuelle.




En résumé, les ensembles identifiables appartenant à $ID_1$ ne sont pas nécessairement inclus dans les ensembles identifiables appartenant à $ID_2$. Lorsque tel est le cas, on dit, avec Case, Chen et Jain (1992), que $ID_2$ est fortement séparé de $ID_1$ :




Dans ce cas l'affaiblissement du critère d'identification rend individuellement les machines plus intelligentes dans le sens précis de Binet. La séparation forte est nécessaire aussi pour rendre sensible l'analogie entre la réfutation de Lucas du mécanisme (reposant sur la théorie de la démonstration) et la critique du mécanisme de Putnam (reposant sur la théorie de l'inférence inductive). Case, Chen et Jain (1992) ont réussi fort

heureusement à démontrer la séparation forte entre PEX et EX, ainsi qu'entre les critères $EX^i$ et $BC^i$ apparaissant dans la conquête du Graal que je propose d'aborder à présent[93].

*7°) Case et Smith, ou la conquête du Graal*

*Definitions:* (Case and Smith 1983) :

1) $f =^{0v1} g$ signifie que f est égal à g excepté, *possiblement* sur une entrée. Autrement dit, f diffère de g sur au plus 1 valeur.

$$\#\{x \mid f(x) = g(x)\} \leq 1$$

2) M $EX^1$-identifie f si, lorsqu'on présente f à M, M converge sur un programme calculant une fonction g telle que $f =^{0v1} g$. On écrit $f \in EX^1(M)$.

3) $EX^1 = \{S : \exists M\ S \subset EX^1(M)\}$. $EX^1$ est donc l'ensemble des classes de phénomènes reconnaissables par des machines à inférence inductive auxquelles on permet de commettre *au plus* une erreur.

*Théorème* (Case and Smith 1983) $EX \subsetneq EX^1$. Plus précisément ils ont montré que l'ensemble :

$$S^{0v1} = \{f : \phi_{f(0)} =^{0v1} f\}$$

n'appartient pas à EX. La MII $\alpha$ (voir plus haut) témoigne de l'appartenance de $S^{0v1}$ à $EX^1$.

Il est plus difficile de montrer que $S^{0v1}$ n'appartient pas à EX. Bien que la démonstration de Case et Smith 1983 soit très intéressante, notamment parce qu'elle utilise la généralisation de Case de 2-REC (voir 2.2), elle sort du cadre de cette section. Le point important est que cette démonstration est non constructive.

La preuve est non constructive, le "*possiblement* " est inévitable. Un tel élargissement ne peut pas se faire de façon constructive (algorithmique).

Qu'il en soit *nécessairement* ainsi a été démontré, avec 2-REC, par Chen :

*Théorème* (Chen[94]) Il n'est pas possible de générer uniformément à partir d'une machine à inférence inductive M (plus exactement à partir d'une de ses présentations $\ulcorner M \urcorner$) une fonction récursive que M ne peut pas reconnaître.

---

93 Case, Chen et Jain 1992 montrent aussi l'existence de critères d'identification qui sont séparés, mais qui ne sont pas fortement séparés.

94 Communication à Case et Smith (1983), qui propose aussi une autre démonstration.

*Preuve:* Supposons qu'il existe une fonction calculable $\phi_j$ telle que pour toute MII M :

1) $\phi\phi_j(\ulcorner M \urcorner)$ est totale calculable et
2) $\phi\phi_j(\ulcorner M \urcorner)$ n'appartient pas à EX(M).

Dans ce cas, la MII suivante :

$$m = \lambda x\ (k\ j)$$

i.e. la machine qui sur toute entrée sort toujours $\phi_j$ appliquée à elle-même, est telle que $\phi\phi_{j(m)}$ appartient à EX(m), et ceci contredit 2). QED.

Ce résultat montre que la façon dont Lucas réfute le mécanisme ne va pas fonctionner pour les machines à inférence inductive. L'assurance de Lucas était liée à l'aspect constructif de la démonstration de Gödel (lié à la mécanisabilité de l'argument de la diagonale). Si "produire comme vrai" est défini par l'inférence correcte (ce qui est analysé en détail dans la section suivante), le caractère nécessairement non constructif des démonstrations d'existence d'MII illustre la vanité de *prouver* l'anti-mécanisme en philosophie de l'esprit puisque de telles preuves ne peuvent reposer que sur des machines convenablement présentées ou décrites. On peut comparer cette critique de Lucas avec celle de Rucker (voir 2.3.1), qui, elle aussi, use d'une forme non constructive de l'incomplétude.

Cela montre encore qu'il faille tenir compte du fait que le "ou" dans la définition de $f =^{0v1} g$, est, pour certain f et certain g nécessairement non constructif, de même que le "v" du "≤" dans

$$\#\{x \mid f(x) = g(x)\} \leq 1$$

Notons encore qu'avec le résultat de non-constructivité nécessaire de Chen, les extensions d'ensembles identifiables par affaiblissement de critères ***fortement séparés*** sont non constructives. Il n'est pas possible de mettre en évidence explicitement une fonction identifiable par une machine dont le critère aurait été, algorithmiquement affaibli. Avec le second théorème de récursion nous pouvons donc construire une machine capable de produire des extensions plus intelligentes qu'elles (au sens de Binet), mais aucune de ces extensions n'est à même de prouver qu'elle est plus intelligente ni qu'elle sait explicitement produire des machines plus intelligentes qu'elles.

Ces résultats se généralisent. Admettons qu'on permette à une machine à inférence inductive de commettre

$0 \vee 1 \vee 2$

erreurs.

On peut définir la classe $EX^2$ correspondante. On a

$$EX^1 \subsetneq EX^2$$

On peut démontrer un résultat plus général, avec la

*définition* M $EX^i$-identifie f si et seulement si M, sur une présentation de f, converge vers p, et $\phi_p =^i f$, avec $=^i$ mis pour $=^{o v1 v2 v \ldots v i}$.

Case et Smith 1983 ont en effet démontré :

$$EX^i \subsetneq EX^{i+1}$$

ainsi que

$$\bigcup_{n \in \omega} EX^n \subsetneq EX^*$$

où le critère EX* permet à la machine de commettre un nombre indéterminé, mais fini d'erreurs.

Autrement dit, ne pas limiter le nombre d'erreurs permet de reconnaître strictement plus de classes de fonctions. J'insiste que pour interpréter correctement ce genre de résultats dans le sens de Binet, ce qui est nécessaire pour réfuter Putnam, il faut encore utiliser la version forte de la séparation.

Case, Chen et Jain l'ont prouvé pour la hiérarchie qui constitue une sorte de conquête du graal. Je propose de poursuivre ce chemin, du moins en ce qui concerne la séparation simple pour ne pas compliquer le propos.

Je rappelle la définition de EX,

$$f \in EX(M) <-> \forall^{\infty} x\ M(f:x) = p\ \&\ \phi_p = f$$

M identifie f si M finit par converger vers un unique programme calculant f. On permet un nombre indéterminé mais fini de changements d'avis pour la convergence.

Barzdin a donné un curieux affaiblissement de ce critère d'identification.

On permet à M de changer une infinité de fois d'avis, pourvu qu'à partir d'un certain moment, ces avis, bien que, changeant continuellement,

calculent toujours la même fonction[95]. Dans ce cas on parle de convergence extensionnelle ou behavioriste. La convergence pour EX, se faisant au niveau du code, est intensionnelle. Barzdin note cette classe $G^{\infty}$, voir Zeugmann 1987, mais je vais suivre la notation de Case et Smith qui la désigne par BC.

Pour que M BC-identifie f, la suite d'hypothèse-programmes doit donc converger extensionnellement. On demande simplement :

$$f \in BC(M) <-> \forall^{\infty} x \; \phi_{M(f:x)} = f$$

La MII peut donc générer une infinité d'hypothèses pourvu que *presque toutes* ces hypothèses calculent la même fonction, qu'elles soient *extensionnellement* équivalentes.

On a le curieux résultat :

$$EX^* \subsetneq BC$$

*Remarque* si on drope la condition 1 dans la définition de l'extrapolation, on a NV' = BC (où NV' = NV sans la condition 1).

*Justification :* cela est dû au fait que l'extrapolation est une notion extensionnelle. Ensuite on a vu qu'enlever la condition 1 *dépopperianise* la machine.

Cela illustre la puissance des[96] machines extrapolantes "silencieuses", elles reconnaissent strictement plus que EX*.

A présent comme pour EX, on peut affaiblir le critère d'identification en permettant à la machine de produire des hypothèses incorrectes. On définit ainsi, $BC^1$, $BC^2$, ..., $BC^n$, ainsi que BC*. Comme pour EX on démontre :

*Théorèmes*

$$BC \subsetneq BC^1, \; BC^i \subsetneq BC^{i+1}, \; \bigcup_{n \in \omega} BC^n \subsetneq BC^*$$

Preuves (voir Case et Smith 1983).

*Théorème* (Harrington[97]) Le graal, enfin :

---

95 Un peu comme un scientifique qui changerait tout le temps de théories, malgré que ses nouvelles théories conduisent exactement aux mêmes prédictions que les anciennes. Curieusement une machine qui se permet un tel comportement peut reconnaître des classes plus large de fonctions (Case & Smith 1983 + Case, Chen & Jain 1992).

96 On aimerait utiliser un singulier ici. C'est justement ce que la séparation forte permet.

97 Communication orale d'Harrington à Case et Smith (1983).

$R \in BC^*$

*Preuve* (voir Case et Smith 1983). En fait $P \in BC^*$ (dans un sens qu'il faudrait préciser, voir Rinn et Schinzel 1988 pour les définitions et une démonstration directe de cette proposition). Il existe donc une machine capable d'apprendre, au sens de BC*, la classe complète de l'école du dehors. Stricto sensu ces résultats pourraient être considérés comme une réfutation de Putnam. Malheureusement :

Chen a pu montrer dans sa thèse de doctorat (Chen 1981) que pour toute BC*-machine à inférence, c-à-d pour toute machine M telles que $R \in BC^*(M)$, on peut, pour tout nombre k, trouver des fonctions récursives f telles qu'$\exists^\infty$n pour lesquels M(f|n) ne calcule pas f de n+1 à n+k.

BC* n'est donc définitivement pas un critère d'identification pratique. On peut dire au sujet des machines BC* que *leurs ailes de géants les empêchent de marcher*.

Regardons quelques autres critères d'identification.

*8°) Changements d'avis :*

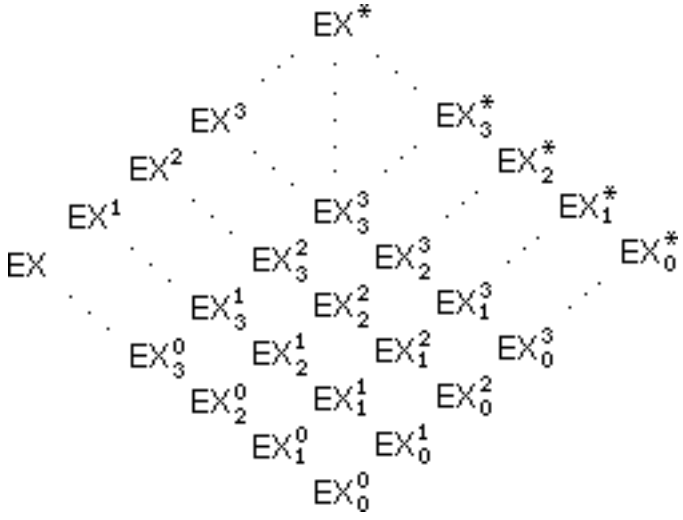
*Définition* M $\mathbf{EX}^i_j$-identifie f, et donc $f \in \mathbf{EX}^i_j(M)$, si et seulement si M $\mathbf{EX}^i$-identifie f avec au plus j changements d'avis.

Les machines dont on exige 0 changement d'avis correspondent à la classe FIN étudiée par Barzdin et Freivalds. C'est une identification en un coup.

*Définition* $\mathbf{EX}^i_j = \{E \mid \exists M\ E \subset \mathbf{EX}^i_j(M)\}$. En particulier $\mathbf{EX}^0_0$ = FIN.

*Théorème* $\forall i,j,k,r \quad \mathbf{EX}^i_j \subsetneq \mathbf{EX}^k_r$ ssi i =< k et j =< r.

Cela est plus clair avec un graphique, les arrêtes qui sont seulement suggérées représentent l'inclusion.




Ce graphique illustre la non-comparabilité évoquée plus haut au sujet de l'usage des tests de Binet.

*9°) Non-union et réseau bifurquant*

Autre conséquence de la non-constructivité : le phénomène de non-union de Blum et Blum 1975.

*Définition* Supposons que $\phi_k$ est présentée à une MII M, et que $\phi_k(a) = b$. Soit h la dernière hypothèse de g sur $\phi_k$. Dans le cas où $\phi_k(x)$ est égale à $\phi_h(x)$ pour tout x excepté *a* : c-à-d. $\phi_h(a) = c$, et $c \neq b$, on dit que M est *précisément incorrecte* sur a.

*Definition* $EX^{=1} = \{S : \exists M\ S \sqsubset EX(M)$ et M est précisément incorrecte sur exactement un nombre$\}$. Comme on présente à M une fonction totale, une telle erreur peut toujours être réfutée et corrigée à la limite. Donc on a :

$$EX^{=1} = EX$$

*Théorème de non-union* (Barzdin 74, Blum & Blum 75) : $A \in EX$ & $B \in EX$ n'implique pas $A \cup B \in EX$.

*Preuve.* Soit $S_1$ l'ensemble des fonctions qui sont calculées par des machines qui se reproduisent elles-mêmes de façon précisément incorrecte .sur la valeur particulière DUP

$S_1 = \{f : f_{\phi(0)} =^{=1} f\}$. Soit $S = \{f : \phi_{f(0)} = f\}$. $\alpha$ témoigne que $S_1$ appartient à $EX^{=1} = EX$, et $\alpha$ témoigne aussi que S appartient à EX. Ainsi $S_1$ et S appartiennent tous les deux à EX, mais l'union $S_1 \cup S = S^{0v1}$ n'appartient pas à EX (théorème précédent). Ceci prouve le théorème de non-union de Blum & Blum :

l'union d'élément de EX n'appartient pas nécessairement à EX. Remarquons que cela est à nouveau dû au fait que "v" dans 0v1 est nécessairement non constructif.

Si nous voulons tenir compte de tels élargissements non-constructifs, nous devons permettre des bifurcations (branchement) dans le réseau de Myhill (voir 2.2), c'est-à-dire en ajoutant dans l'équation de récursion d'une *planaire infinie* du style :

$\phi_{\phi_e(i)}(z)$ = si i=0 & z = PROUVE alors sortir T(PA),
sinon
si z=NEXT alors sortir $\phi_e(i+1)$, else
si z= PROUVE alors sortir $T(RL(\phi_e(i-1)))$.

une ligne du genre :

sinon, si z=INDUCE alors sortir $IND(\phi_x(j)\ i \geq j)$

où INDUCE est un nouvel identificateur, et IND est ce que j'appelle un *semi-réalisateur*, c-à-d essentiellement un programme qui, à partir d'une certaine spécification, construit une collection de programmes capables d'être éxécutés en parallèle, parmi lesquels figure au moins un qui rencontre la spécification bien qu'il ne soit pas nécessairement possible de choisir algorithmiquement lequel. IND ici est une MII, capable de doveteller sur elle-même et ses hypothèses.

*10°) Pluralisme et probabilisme*

Définition (Smith 1982) Une équipe C de n machines reconnaît une fonction f si une d'entre elles reconnaît cette fonction selon un certain critère d'identification X. On écrit $f \in C(n, X)$.

Le théorème de non-union entraîne qu'une équipe de deux machines reconnaît strictement plus de classes de fonctions qu'une machine.

Smith a généralisé ce résultat pour des équipes plus larges.

Il montre que la donnée de m+1 machines permet de reconnaître des collections strictement plus larges de classes de fonctions que la donnée de M machine.

S'il était possible de déterminer algorithmiquement laquelle des machines dans l'équipe a reconnu le phénomène, il serait possible de construire une machine à inférence inductive reonnaissant la collection de fonctions reconnues par l'équipe, en contradiction avec la généralisation du théorème de non-union. L'existence du critère du pluralisme est une conséquence de l'incomplétude. La non-constructivité, qui ici est donc prouvablement nécessaire n'empêche pas cependant l'existence d'utilisation

pratique. Case, cité par Smith 1982, imagine à cet effet l'envoi d'une collection de robots sur une planète étrangère. Si on admet que les chances de "survie" d'une machine, plongée dans un environnement inconnu, augmente avec ses capacités d'identification de phénomènes, la non-union suggère d'envoyer une large équipe.

Smith décrit complètement les compensations possibles entre l'exactitude (la justesse) de la machine, l'hésitation (le nombre de changement d'avis) et la taille de l'équipe (nombre de machines de l'équipe).

Il montre que le critère de la taille est en quelque sorte plus puissant que les deux autres.

Augmenter la taille de l'équipe permet toujours, aussi bien de diminuer l'hésitation que d'augmenter la justesse de l'inférence. Inversément augmenter l'hésitation ou diminuer la justesse ne permet pas de compenser la petitesse de la taille d'une équipe.

Barzdin avait déjà montré

$$C(2,EX) - C(1, BC) \neq \{\}$$

Il existe donc des classes de phénomènes qui ne sont jamais BC-reconnues, et donc ne sont jamais $EX^i$-reconnues pour i quelconque appartenant à $\omega \cup \{*\}$, mais qui sont reconnaissables par des couples de machines EX.

*Théorème* (Smith 82)

$$\forall n \in \omega \;\; C(n+1,EX) - \bigcup_{m \in \omega} C(n,BC^m) \neq \{\}.$$

Dans les démonstrations, on diagonalise sur des ensembles de fonctions identifiées par des équipes arbitraires couvrant donc toutes les façons dont ces machines intéragissent ou communiquent les unes avec les autres.

Comme le fait remarquer Smith, c'est l'utilisation du théorème de récursion ou de ses généralisations (notamment le théorème de Case à partir duquel la planaire de 2.2 est conçue) qui permet de ne pas devoir expliciter le détail du fonctionnement interne des équipes, pourvu que chaque entité de l'équipe soit algorithmique.

Le pluralisme justifie l'importance du parallélisme pour le développement de l'intelligence. Il n'y a cependant rien de magique dans ce parallélisme, qui peut être émulé (abstraction faite du temps relatif) par une machine de Turing universelle et séquentielle. Cela pour prévenir une utilisation erronée du parallélisme en philosophie de l'esprit comme celle de Johnson-Laird 1987. Concluons avec Smith 1982, pour l'apprentissage des machines, la diversité est la clé du succès, comme dans le monde biologique.

*Remarque* : Deux machines peuvent être simulées par une machine universelle. L'équipe peut donc servir à modéliser le doute entre plusieurs idées évoluant (consciemment ou inconsciemment) chez un individu unique. Imaginons qu'un tel individu doit prendre une décision. Une façon est de choisir au hasard une des idées. Dans ce cas l'hypothèse qu'il propose est l'hypothèse actuelle d'une de ses "sous-machines". Une machine individuelle peut reconnaître ainsi avec une probabilité 1/n ce qu'une équipe de n machines peut reconnaître. Une telle remarque peut être considérablement nuancée, voir Pitt 1989 et Pitt et Smith 1988.

*11°) quelques conjectures*

*a) La conjecture de Barzdin*. Informellement la conjecture de Barzdin énonce que si une collection C de fonctions est identifiable alors elle est soit

-identifiable par une technique d'énumération (c-à-d essentiellement par une technique de recherche dans un espace de solutions), soit

-identifiable parce qu'elle s'autoprésente sur un argument, comme la fonction standard donnée par EX-BUT-NOT-PEX. Barzdin a énoncé sa conjecture relativement à la classe EX et Zeugmann a généralisé cette conjecture pour les critères d'identification quelconques. En particulier Zeugmann a démontré la conjecture pour les classes FIN (= $EX_0$), REX et T-REX. Pour la REX-identification on exige que la machine ne converge que sur les fonctions qu'elle identifie. On dit qu'une telle machine est reliable (Minicozzy[98] 1976). Pour T-REX on exige que la machine soit au moins reliable sur le *graal R*). Kurtz et Smith 1989 ont réfuté la conjecture originale de Barzdin.

On peut voir la conjecture de Barzdin comme une sorte de thèse de Church pour l'inférence inductive telle qu'elle est traitée dans le paradigme de Gold-Putnam (Webb 1983 contient des remarques qui vont dans ce sens).

*b) une conjecture de Case* (Case 1989) Il est souvent reproché aux contre-exemples utilisés en inférence inductive théorique d'être ad hoc et peu naturels. Case compare cette critique à celle des mathématiciens vis-à-vis de la nature peu traditionnelle de la proposition indécidable isolée par Gödel dans son théorème d'incomplétude. Dans les deux cas la nature auto-référente ou intensionnelle de la proposition ou des machines est à l'origine de cette critique. Si cette remarque est justifiée pour le mathématicien, elle semble peu convaincante pour le philosophe de l'esprit, pour qui une machine ou une proposition autoréférente est plutôt opportune. Toutefois, en ce qui concerne l'existence de propositions indécidables en arithmétique,

98 Minicozzy 1976 a introduit cette classe. Cette classe est fermée pour l'union à la différence de EX.

on a pu mettre en évidence des propositions indécidables dont le contenu mathématique est naturel (voir Paris et Harrington 1977) et Case conjecture, par analogie, que des contre-exemples naturels, relativement aux techniques concrètes d'apprentissage développées en intelligence artificielle, seront mis en évidence.

Mentionnons qu'une autre conjecture due à Wiehagen pourrait donner un sens à la notion d'apprentissage universel. Elle se base sur le choix d'une numérotation de Gödel particulière (Wiehagen 1991, voir aussi Delahaye 1992).

La portée de ces conjectures pour la philosophie mécaniste (et non mécaniste) de l'esprit reste à approfondir.

*12°) Autres résultats*

De nombreux autres résultats sont pertinents pour la philosophie de l'esprit. Je mentionne le phénomène d'inconsistance. Le phénomène d'inconsistance apparaît chez Barzdin 1974 et Blum et Blum 1975. Si on restreint les hypothèses d'une machine à inférence inductive M aux hypothèses strcictement consistantes, c-à-d ne contredisant pas les entrées-sorties présentées, on restreint la portée de l'apprentissage ! Le phénomène est en partie lié au phénomène d'incomplétude (voir aussi Delahaye 1992).

Je mentionne aussi la démonstration, par Daley et Smith 1986, de l'existence d'un phénomène d'accélération dans le domaine de l'inférence inductive analogue à celui que Blum a mis en évidence pour le calcul des machines universelles (voir 2.1).

*Conclusion* : le miracle est qu'une telle théorie puisse exister. La définition simple mais non effective de Binet fait éclater, en informatique théorique, le concept d'intelligence en une myriade de hiérarchies, souvent orthogonales les unes avec les autres. Le rôle du *non constructif* et de l'*intensionnel* est capital.

*13°) Résumé de 2.3.5*

*Les phénomènes sont modélisés par des parties du Graal, ou même identifiés avec de telles parties par PAN-MEC. Ce chapitre propose des résultats en partie essentiellement non constructifs. On se promène dans l'espace du dehors, en ce sens qu'on s'intéresse à la* ***totalité*** *du Graal R (ce qui ne peut être conçu qu'à partir du dehors) Notons que la majorité des résultats peuvent être étendus à P.*

*On dit qu'une machine a reconnu un phénomène si, lorsqu'on lui présente ce phénomène elle est capable de l'extrapoler (NV-identification), ou de construire une explication (une théorie) de ce phénomène (EX-identification).*

*Si on admet de comparer l'intelligence des machines par la taille des classes de phénomènes qu'elles sont à même de reconnaître, on peut démontrer, d'une façon générale qu'une machine est d'autant plus intelligente qu'elle est capable :*

*a) de donner des explications incomplètes (et donc de rester quelque fois silencieuse),*

*b) de donner des explications incorrectes (et donc de se tromper quelques fois),*

*c) de renouveler la présentation de ses hypothèses (sans changer leur extension, et donc de passer à une théorie équivalente),*
*d) de produire des explications inconsistantes avec les données (et donc d'être localement relativement inconsistante),*
*e) de produire des explications divergentes avec une certaine probabilité (machine probabiliste) d'être correcte,*
*f) de travailler en équipe (phénomène de non-union),*
*g) de changer d'avis.*
*Les démonstrations utilisent la plupart le théorème de Kleene, ou la généralisation de Case. Elles illustrent à nouveau l'importance de la diagonalisation en ce qui concerne les capacités limites des machines. Les résultats sont non constructifs : d'une façon générale il n'est pas possible de mesurer ou de comparer l'intelligence des machines ni sur base de la description du code (ce qui est déjà une conséquence du théorème de Rice (voir 2.2), ni sur base de leurs comportements locaux.*
*La réfutation du mécanisme par Lucas, appliquée aux machines à inférence inductive, échoue à la base, partout où la non constructivité est nécessaire dans la preuve de l'existence d'une telle machine.*

## *Biblio plus locale*


**BARZDIN** J., **1974**, *Two Theorems on the Limiting Synthesis of Functions*, Theory of Algorithms and Programs. Barzdin (Ed.), 1, Latvian State University, Riga, pp. 82-88.

**BESNARD** P., **1989**, *An Introduction to Default Logic*, Springer Verlag, Berlin.

**BLUM** L. & **BLUM** M., **1975**, *Toward a Mathematical Theory of Inductive Inference*. Information and Control 28,.pp. 125-155.

**BRANDT** U., **1986**, *The Position of Index Sets of Identifiable Sets in the Arithmetical Hierarchy*, Information and Control 68, pp. 185-195.

**CASE** J. & **NGO-MANGUELLE** S., **1979**, *Refinements of inductive inference by Popperian machines*. Tech. Rep., Dept. of Computer Science, State Univ. of New-York, Buffalo.

**CASE** J., **1986**, *Learning Machines*, in Demopoulos & Marras 1986.

**CASE** J. & **SMITH** C., **1983**, *Comparison of Identification Criteria for Machine Inductive Inference*. In Theoretical Computer Science 25,.pp 193-220.

**CASE** J., **1987**, *Turing Machines*, in Encyclopedia of Artificial Intelligence, pp. S. Shapiro (ed.), John Wiley and Sons, New-York.

**CASE** J., **1989**, *The Power of Vacillation*, COLT '88, Proceedings of the 1988 Workshop on Computational Learning Theory, pp. 196-205.

**CASE** J., **CHEN** K., **JAIN** S., **1992**, *Strong Separation of Learning Classes*, in Jantke (Ed.) 1992.

**CHEN** K., **1981**, *Tradeoffs in Machine Inductive Inference*, Ph. D. Thesis, Computer Science Department, SUNY at Buffalo, Amherst, NY.

**CHEN** K., **1982**, *Tradeoffs in the Inductive Inference of Nearly Minimal Size Programs*, Information and Control, 52, pp. 68-86.

**DALEY**, R. P., **1987**, *Inductive Inference Hierarchies, Probabilistic vs Pluralistic Strategies*, Lecture Notes in Computer Science N° 215, pp. 73-82, Springer-Verlag.

**DALEY** R., **KALYANASUNDARAM** B., **VELAUTHAPILLAI** M., **1992**, The Power of Probabilism in Popperian FINite Learning, in Jantke 1992.

**DALEY** R.P. and **SMITH** C.H., **1986**, *On the Complexity of Inference Inductive*. Information and Control 69, pp. 12-40.

**DELAHAYE** J. P., **1992**, *L'inférence inductive : présentation et analyse de quelques résultats*. preprint.

**DEMOPOULOS** W. and **MARRAS** A. (eds), **1986**, *Language Learning and concept Acquisition : Foundational Issues*, Ablex Publishing Corporation, Norwood, New-Jersey.

**FULK** M.A., **1988**, *Saving the Phenomena: Requirements that Inductive Inference Machines* Not Contradict Known Data, Information and Computation, 79, pp. 193-209.

**GOLD**, E. M., **1965**, *Limiting recursion*, Journal of Symbolic Logic, 30, 1, pp. 27-48.

**GOLD** E.M., **1967**, *Language Identification in the Limit*. Information & Control 10, pp. 447-474.

**GOULD** S. J., **1983**, *La mal-mesure de l'homme*, traduction de l'américain par J. Chabert, Editions Ramsay, Paris.

**JANTKE** K. P., **1992**, *Analogical and Inductive Inference*, Proceedings of the International Workshop AII '92, Dagstuhl Castle, Germany, October 1992. Lecture Notes in AI, n° 642, Springer-Verlag, Berlin.

**JOHNSON-LAIRD** P., **1987**, *How Could Consciousness Arise from the Computations of the Brain ?*, in **Mindwaves**, C. Blakemore and S. Greenfield, Basil Blackwell, Oxford and New York, pp. 247-257.

**KRAUSE** P and **CLARK** D., **1993**, *Representing Uncertain Knowledge, an Artificial Intelligence Approach*, Kluwer Academic Publisher, Dordrecht.

**KURTZ** S. A. and **SMITH** C. H., **1989**, *On the Role of Search for Learning*, in Rivest R., Haussler D., Warmuth M. K. (eds.), COLT '89, Morgan Kaufman Publishers, Inc., 1989.

**MARCHAL** B., **1990**, *Des fondements théoriques pour l'intelligence artificielle et la philosophie de l'esprit*, Revue Internationale de Philosophie, 1, n° 172, pp 104-117.

**MINICOZZY** E., **1976**, *Some natural properties of strong-identification in inductive inference*. Theoretical Computer Science, 2, pp. 345-360.

**MINSKI** M, **1968**, *Mind, Matter and Model*, in Semantic Information Processing, Minski M. (ed.), Cambridge MIT Press.

**OSHERSON** D.N., **STOB** M.and **WEINSTEIN** S., **1986**, *Systems that Learn*, MIT press.

**OSHERSON** D.N., **STOB** M.and **WEINSTEIN** S., **1986**, *Aggregating Inductive Expertise*. Information and Control 70, pp. 69-95.

**OSHERSON** D.N., **STOB** M.and **WEINSTEIN** S., **1988a**, *Synthesizing Inductive Expertise*. Information and Computation 77, pp. 138-161.

**OSHERSON** D.N., **STOB** M.and **WEINSTEIN** S., **1988b**, *Mechanical Learners pay a Price for Bayesianism*,. The Journal of Symbolic Logic, Vol. 53, n° 4, pp. 1245-1251.

**PARIS**, J. B., **HARRINGTON** L., **1977**, *A Mathematical Incompleteness in PA*, Handbook of Mathematical Logic, J. Barwise ed., North-Holland, pp. 1133-1142.

**PITT** L., **1989**, *Probabilistic Inductive Inference*. Journal of the Association for Computing Machinery. Vol 36, N° 2, pp. 383-433.

**PITT** L. and **SMITH** C.H., **1988**, *Probability and Plurality for Aggregations of Learning Machines*. Information and Control 77, pp. 77-92.

**POSNER** D.B., **1980**, *A Survey of Non-R.E. Degrees ≤ O'*, in F. R. Drake and S. S. Wainer (eds), Recursion Theory : its generalisation and applications, Cambridge University Press, 1980.

**PUTNAM** H., **1963**, *Probability and confirmation* The Voice of America, Forum Philosophy of Science, 10 (U.S. Information agency, 1963). Reprinted in Mathematics, Matter, and Method. Cambridge University Press.Cambridge 1975.

**PUTNAM** H., **1960**, *Minds and Machines*, Dimensions of Mind : A Symposium, Sidney Hook (Ed.), New-York University Press, New-York. Repris dans Anderson A. R. (Ed.),1964.

**PUTNAM**, H., **1965**, *Trial and error predicates and a solution to a problem of Mostowski*, Journal of Symbolic, 30, 1, pp. 49-57.

**PUTNAM** H., **1988**, *Representation and Reality*, A Bradforf Book, The MIT Press, Cambridge.

**RINN** R. & **SCHINZEL** B., **1988**, *Learning by Teams from Examples with Errors*. Springer Verlag. Lectures notes in computer science n° 329, pp 223-234.

**ROYER** J.S., **1986**, *Inductive Inference of Approximations*. Information and Control 70, pp. 156-178.

**SMITH** C.H., **1982**, *The Power of Pluralism for Automatic Program Synthesis*. Journal of the Association for Computing Machinery, Vol 29, N° 4, pp. 1144-1165.

**SOLOMONOFF**, R. J., **1964a**, *A Formal Theory of Inductive Inference i*, Information and Control, 7, pp. 1-22.

**SOLOMONOFF**, R. J., **1964b**, *A Formal Theory of Inductive Inference ii*, Information and Control, 7, pp. 224-254.

**WEBB** J. **1983**, *Gödel's Theorems and Church's Thesis: A Prologue to Mechanism* in R. S. Cohen and M. W. Wartofsky (eds.), *Language, Logic, and Method*, pp. 309-353, D. Reidel Publishing Company.

**WIEHAGEN** R., **1991**, *A Thesis in Inductive Inference*, in *Nonmonotonic and Inductive Logic*, Dix J., Jantke K. P., Schmitt P. H. (eds), Lecture Notes in Artificial Intelligence, N° 543, Springer-Verlag, 1991.

**ZEUGMANN**, T., **1987**, *On Barzdin's Conjecture*, Lecture Notes in Computer Science N° 265, pp. 220-227, Springer-Verlag.

### 2.3.6 La philosophie des machines dans les voisinages de l'infini

*Brièvement*

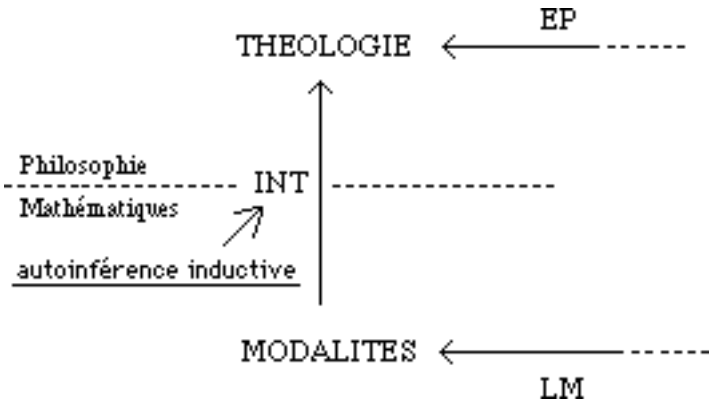
*Si on admet avec Lucas une forme d'identification de base entre machine et système formel, alors on admet que les théorèmes d'incomplétude de Gödel s'appliquent aux machines. Dans ce cas MDI entraîne que les théorèmes de Gödel s'appliquent à nous (je). Pas seulement les théorèmes de Gödel, mais aussi le théorème de Löb et ses conséquences prouvables et improuvables.*
*G axiomatise la part prouvable, justifiable formellement, positivement communicable, de ces conséquences (concernant justement le communicable et l'incommunicable). Comme par exemple* $\Box\Diamond \top -> \neg \Diamond \top$.

*G* axiomatise en plus la part non communicable, comme* $\Diamond\top$, *mais aussi* $\neg \Box\Diamond \top$, $\Diamond \Diamond\top$, $\Diamond \Box\bot$, *etc.*
*Par ailleurs G étend C4, et les expériences par la pensée mettent en évidence des propositions indexicales absolument indécidables, ce que les théorèmes d'incomplétude entraînent avec la thèse de Church. La machine, consistante et* $\Sigma_1$*-complète (l'ensemble de ses communications est créatif au sens de Post) ne peut pas justifier la consistance d'aucunes extensions d'elle-même. Cette consistance peut cependant être inférée, de façon implicite et instinctive, ou de façon explicite dans le mode du pari. C'est la décidabilité de G* qui rend les propositions de G* trivialement, idiotiquement en fait, inférables.*
*Pour traiter de la connaissance, on remarquera la présence d'un sujet intuitioniste naturellement associé à la machine, et décrit complètement*[99] *(il n'y a pas de part incommunicable) par S4Grz. Pour traiter de l'indéterminisme abrupte, dont j'ai montré en 1.3 que l'indéterminisme quantique est un cas particulier (avec l'interprétation d'Everett), il faut dériver une logique de l'accès immédiat dans les voisinages de 0.*
---

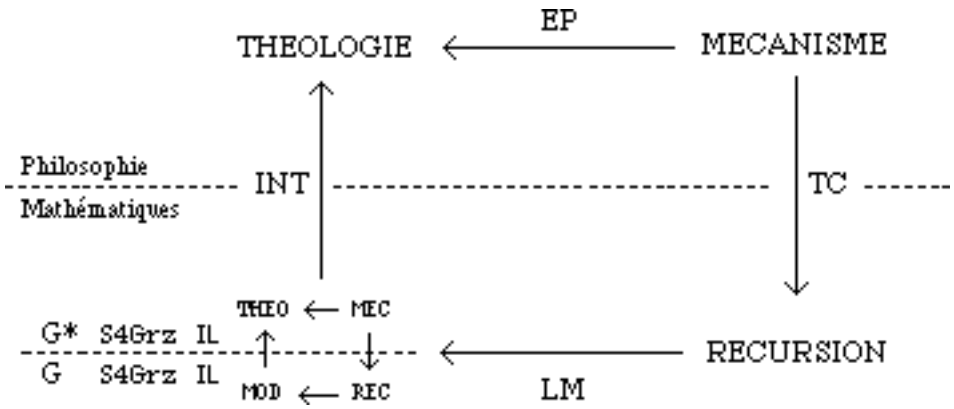
*Méthodologie*

G et G* axiomatisent respectivement les discours et la vérité concernant les discours des machines consistantes autoréférentiellement correctes. J'argumente aussi que G* décrit les discours et les interrogations des machines autoréférentiellement et auto-*inférentiellement* correctes dans les voisinages de l'infini. Cela permet des interprétations théologiques sous la forme d'inférences diverses que l'on peut voir comme autant de paris du type du pari de Pascal.

[99] Au niveau propositionnel.

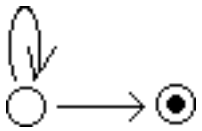



L'existence de tels discours, mélange d'inférences inductives et de déductions, permet de plonger le diagramme méthodologique complet dans G*, par l'intermédiaire des modalités issues de l'autoréférence correcte G, S4Grz, IL, etc.




*1°) G étend C*

La théorie G étend *ce qui peut ressembler* à une théorie de la conscience, en l'occurrence la théorie C décrite dans la première partie. L'extension est stricte, il est facile de construire une relation réaliste qui ne soit pas bien cappée. Par exemple :



La théorie C a été isolée à partir de considérations linguistiques (à partir de Wittgenstein, Watts, Valadier, Lao-Tseu, etc., voir 1.2) et de considérations géométriques (au moyen des expériences par la pensée de l'autoduplication en mécanisme indexical). C admet aussi bien des modèles transitifs que des modèles non transitifs et permet la distinction entre la survie et l'immortalité. La théorie G*, avec la thèse de Church et l'identification (nécessairement non constructive) de base, constitue la logique mécaniste de l'autoréférence correcte au sujet d'entités finies éventuellement extensibles.

A la différence de C, G* est "transitif". En particulier G*⊢ ◊ T<->◊ ◊T, G⊬ ◊ T<->◊ ◊T, C⊬ ◊ T<->◊ ◊T.

Au *niveau* G*, la non-garantissabilité de la survie est rendue équivalente avec la non-garantissibilité de l'immortalité. Cela justifie la recherche d'une logique non transitive pour le calcul des probabilités (voir plus loin).

*2°) Conscience et consistance, de soi et de l'autre*

L'analyse de la réfutation de Lucas suggère déjà que ◊T est interprétable par la *conscience de l'autre*, ou la *conscience de soi* vu comme un autre, comme dans le cas de la duplication (postposée ou non). C'est la conscience de l'autre soi, ou la conscience de soi tel qu'une autre personne peut l'inférer ou l'interroger. La conscience de soi telle qu'on la ressent, ou telle qu'on la connait est mieux décrite par ◇T. Il s'agit là d'une forme épistémique de la consistance qui par réflexion est trivialement communicable, en fait ◇T <-> ◊T v T. Cette trivialité reflète l'adéquation a priori du niveau fonctionnel.

Selon cette interprétation, on peut dire que du point de vue extensionnel, béhaviouriste, extérieur ou encore du point de vue du dehors, la conscience de moi (◇T) et la conscience de mon double ou de la machine censée être moi (◊T) sont identiques (G* ⊢ ◇T<->◊T)

Inversément, du point de vue intérieur (intuitif, informel, absolu ou du dedans) telle n'est pas le cas, puisque ◇T<->◊T entraîne ⊬(◇T<->◊T), pour la même raison que □p<->⊡p entraîne ⊬(□p<->⊡p). Il n'est pas étonnant que le *sceptique de Solovay* hésite un temps infini avant de monter dans le duplicateur, et cela justifie le caractère ironique de la proposition "j'ai survécu" *après* l'usage du translateur.

Aucune machine ne peut déterminer le niveau où le fonctionnalisme est correct. Cela reviendrait à prouver l'existence d'une machine consistante qui l'étend[100]. Cela ne veut pas dire qu'elle ne peut pas inférer l'existence d'un tel niveau ou commettre quelques paris, ce qu'on va voir plus loin.

En résumé G et G*\G sont des logiques de la conscience de la tierce personne ou de la tierce machine.

G correspond aux discours communicables (comme "si je survis à la translation je ne pourrai convaincre personne de la préservation de mon identité"). G*\G correspond aux "discours" incommunicables, mais auto-inférable comme je le justifie dans cette sous-section. Avec le stratagème S4Grz correspond de la même manière à la logique de la conscience

[100] Sans tenir compte de l'indétermination nécessaire sur l'état instantané de la trace du programme qui le constitue au niveau adéquat. On a observé, en 2.2, qu'une machine peut soit se capturer et se relancer relativement à l'environnement avec une erreur sur l'état instantané, soit capturer *une description* précise et correcte de cet état. Auquel cas elle doit faire confiance à *un autre* pour la relancer. Cette confiance est comparable à celle du patient pour son anesthésiste.

(temporelle) de la première (et innommable) personne. Et KD? ainsi que KD?* une sorte de probance (probabilité-croyance).

*3°) Conscience et intelligence*

Quel lien peut-il y avoir entre la théorie de la conscience et la théorie de l'intelligence -à la Binet- développée dans la sous-section précédente. Deux questions seront abordées :

1) sachant que G*\G est l'ensemble des propositions autoréférentiellement correctes qui ne sont pas auto-justifiables, sont-elles auto-inférables ?

2) Ces inférences de parties de G*\G rendent-elles les machines plus intelligentes (au sens de Binet-Putnam-Gold). On peut aussi se demander si ces inférences rendent les machines plus efficaces, rapides etc. Il reste à préciser davantage la notion d'auto-inférence.

On cherche ainsi le rôle, non pas de l'autoconsistance -ce qui serait un peu absurde, mais de l'intérêt à parier, ou au moins à interroger, cette auto-consistance.

*4°) Machines Socratiques*

Quels discours *autoréférentiels et corrects* peuvent-ils être tenus, dans les voisinages de l'infini, par des machines appartenant à des collections de machines introspectives et communiquant entre elles ?

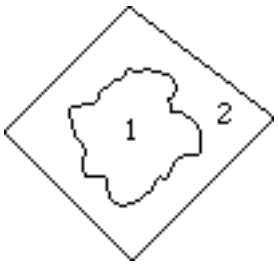
Une autre question qui nous intéresse est de savoir si une machine, dont nous avons déjà vu qu'elle ne peut pas se reconnaître elle-même de façon prouvable[101], peut-elle cependant *inférer* son propre code (socratisme intensionnel) ou une description adéquate équivalente (socratisme extensionnel), où *inférer* correctement un niveau où une substitution finie préserve son identité ?

*Remarques*

- On remarque qu'avec cette façon d'aborder les choses, il y a autant de notion de socratisme qu'il y a de critère d'identification. Cette notion est donc variable.

- En terme de cristal de Gödel (cf 1.1) si une machine infère un niveau où elle survit à la duplication postposée et donc à la reconstitution de son cristal, celui-ci peut être schématiquement représenté de la façon suivante :

[101] C'était l'objet de la reconstruction de l'argumentation de Benacerraf.




La région 1 capture son code proprement dit et la région 2 capture l'information superflue due au fait que la duplication s'est effectuée à un niveau trop bas (par exemple au niveau du système universel sous-jacent comme lors de l'inférence de la théorie T1 (voir 1.1). Le fait que MEC est nécessairement de la forme ⊡∃nFONC(n) et pas de la forme ∃n⊡FONC(n) entraîne l'impossibilité pour le sujet (le candidat mécaniste) d'isoler effectivement la région 1 de la région 2, ni de prouver (communication finie et convaincante) que 1+2 soit suffisant pour sa survie personnelle.

Définition *Machines à inférence inductive socratique*

Soit M une MII. M calcule, entre autres, une fonction partielle récursive. Extensionnellement M peut être identifiée avec cette fonction : $M = \phi_m$. La définition la plus simple et la plus directe du socratisme revient à dire (voir aussi Osherson, Stob Weinstein 1986) qu'une machine est X-socratique si elle est à même de reconnaître sa propre extension (X représente un critère d'identification)

$$\phi_m \in X(M)$$

Lorsque $X \in \{EX, BC\}$, il n'est pas difficile de montrer, en combinant l'usage du dovetellage et du second théorème de récursion, que le socratisme, à la différence de la consistance ou du self-monitoring[102], ne réduit pas la classe de fonctions reconnaissables. Il ne l'agrandit pas non plus. On a (par exemple) le théorème: pour toute classe $C \subseteq R$

$$\exists M\ C \subseteq EX(M) <=> \exists M \text{ socratique } C \subseteq EX(M).$$

*Sketch de la preuve (avec X= EX)* : soit une MII $\alpha$. J'esquisse la preuve (constructive) de l'existence d'une MII $\beta$ telle que si $f \in EX(\alpha)$ alors $f \in EX(\beta)$ et (socratisme) $\beta \in EX(\beta)$. On construit $\beta$ avec 2-REC de telle façon que sur une suite de données (un segment initial) présentée, $\beta$ dovetelle $\beta$ et $\alpha$.

La MII $\beta$ sort continuellement l'hypothèse $\beta$ (c'est-à-dire *elle-même*) tant que cette hypothèse ne contredit pas le segment initial présenté. Dans le cas

[102] J'entends par *self-monitoring* la capacité de la machine à décider de la dernière hypothèse, elle sait qu'elle ne changera plus d'avis. Lorsque le nombre de changements d'avis est borné, le critère du self-monitoring est en général vérifié.

contraire $\beta$ sort les hypothèses de $\alpha$. $\beta$ converge (trivialement, pour ne pas dire *idiotiquement*) sur elle-même, et reconnaît par ailleurs toute fonction identifiable par $\alpha$.

En simplifiant, une machine ne sait pas s'identifier avec son code de façon prouvable (formellement ou informellement, cf sous-section précedente), mais avec 2-REC, nous savons que pour toute machine nous pouvons construire une machine équivalente (extensionnellement parlant) capable de se produire elle-même sur un identificateur particulier.

Une MII peut aussi doveteller son propre code et le comparer au code présenté, donnant son propre code comme hypothèse par défaut, et ne changeant d'avis que lorsque qu'une différence apparaît (auquel cas la MII reprend son travail de MII). Une telle machine reconnaît par une inférence inductive un peu grossière, idiotique, son propre code.

Le socratisme ne limite pas la capacité d'inférence inductive (relativement au critère d'identification EX), mais il ne l'agrandit pas non plus.

Des remarques similaires peuvent être justifiées, à des nuances près, pour d'autres types d'identification, comme l'apprentissage par la trace (cf encore l'inférence d'un niveau universel comme T1 décrite dans 1.1). Les nuances sont importantes : pour l'apprentissage de grammaire à partir de la présentation de textes, ou plus généralement pour l'apprentissage à partir d'exemples exclusivement positifs, c-à-d encore l'apprentissage à partir de $W_i$ plutôt qu'à partir des $\phi_i$, on peut montrer que le socratisme limite la capacité d'inférence (Osherson, Stob, Weinstein 1986, voir la biblio de la sous-section précédente).

Revenons au discours limite. La *possibilité* de tels discours autoréférentiels est *réalisée* par l'opérateur de Kleene (avec lequel le soi est défini en 2.2 non indexicalement) et la nécessité de tels discours est illustré par l'*existence* même de la théorie de l'inférence inductive laquelle repose sur les théorèmes d'incomplétude et les diagonalisations effectives ou non.

Si on interprète le "*produire comme vrai*" de Lucas par prouver (informellement ou formellement) **V** inférer correctement, ou encore le mélange des deux (prouver à partir d'hypothèses inférées), de tels discours seront localement capturés par des extensions de G, pour le vrai et prouvable, de G*, pour le vrai, et par G*\G pour le vrai non prouvable, mais inférable[103]. Ceci résulte des théorèmes de Solovay, grâce auquel G, G* et G*\G sont des théories finiment présentables et décidables, et donc inférables.

103 Plus exactement par les extensions de G* \ les extensions de G correspondantes.

*5°) Machines presque Socratiques*

*Définition* (informelle) Une machine à inférence inductive est *presque Socratique* si elle identifie à la limite une théorie (ou une suite de théories) correcte concernant *différents aspects* d'elle-même (aspects extensionnels où intensionnels relativement à un environnement ou relativement à une machine universelle[104] avec ou sans oracle).

L'inférence d'une théorie peut être ramenée à l'inférence d'une fonction (éventuellement partielle pour les théories indécidables) calculable. Il suffit d'identifier la théorie avec la fonction caractéristique de l'ensemble des nombres de Gödel des paires <théorème / démonstration> de la théorie, (voir Blum & Blum 1975.

Une machine presque socratique peut par exemple inférer qu'elle est une machine universelle d'un certain type particulier, émulée par une machine universelle d'un autre type. En ce sens elle peut inférer, ce qui revient toujours à parier, l'existence d'un niveau où le fonctionnalisme est correcte.

*6°) Sagesse idiotique*

La machine β, présentée dans la preuve précédente, se produisait elle-même "idiotiquement" lorsqu'on lui présente des approximations successives de son extension. Bien que, à la différence des machines idiotiques de la sous-section précédente, elle soit capable de changer d'avis, elle n'en est pas moins idiotique en ce qui concerne l'identification de son code.

De la même façon on peut construire une machine presque socratique qui soit autoréférentiellement correcte avec la donnée d'un démonstrateur de théorème de G[105]. Pour construire, à partir d'une machine auto-référentiellement correcte, une machine auto-*inférentiellement* correcte, il suffit de lui adjoindre un démonstrateur de théorèmes de G*. L'essentiel étant de permettre à la machine de distinguer les propositions communicables (G) des propositions non communicables G*\G.

Par exemple G peut être utilisé pour les affirmations et G*\G, pour les interrogations. Cela revient, à placer un démonstrateur de G et de G* directement dans les équations de récursions initiales (voir 2.2).

Quoiqu'une telle machine est hautement non réaliste comme modèle de l'intelligence ou de la conscience, elle peut déjà être utilisée pour réfuter des arguments antimécanistes. Le principe même de l'argumentation de Lucas est réfuté en prenant "produire pour vrai (à mon sujet)" une proposition de G* (Marchal 1990, 1992). Une approche plus naturaliste consisterait à démontrer l'émergence de G et G* à partir d'un critère de sélection à-la-

---

104 Sans nécessairement isoler des représentations univoques d'elles-mêmes, ou aboutir à une sémantique catégorique. On peut coder une théorie axiomatisable au moyen d'une fonction partielle récursive et ramener l'inférence de telles théories à l'inférence telle qu'on l'a décrit dans la sous-section précédente. Voir aussi Blum et Blum 1975.

105 Des exemples précis de telles machines ont été donnés par Smullyan 1987.

Darwin favorisant les entités autoréférentiellement correctes relativement à leurs environnements.

Cette utilisation du Darwinisme est différente du *Darwinisme élémentaire ou arithmétique* mentionné dans le premier chapitre pour motiver le stratagème affaibli. Ces deux utilisations du Darwinisme ne sont pas incompatibles.

*7°) La machine de moins en moins idiotique.*

Ici on permet à la machine un ensemble infini de révisions. C'est une des conséquences les plus élémentaires de l'incomplétude que de permettre à une machine de jouer un jeu infini (pour parler comme Carse 1986).

C'est dans ce sens que Gödel généralise Pythagore et avec TC cet aspect des choses est dérelativisé, nous permettant *in fine*, d'identifier l'intuitif et l'informel avec l'absolu comme partie commune des écoles du dedans (= le dedans, avec la thèse de Church intuitioniste).

Brouwer considérait déjà les mathématiques (intuitionistes) comme la partie exacte de la pensée.

*Exemple* considérons le réseau autoréférentiel capturé par les équations de récursion *à la* Case-Myhill :

$\phi_{\phi_e(i)}(z)$ = si i=0 & z = PROUVE alors sortir T(PA),
sinon
si z=NEXT alors sortir $\phi_e(i+1)$, else
si z= PROUVE alors sortir $T(RL(\phi_e(i-1)))$.

On peut adjoindre à ce réseau des capacités d'inférence inductive ou d'extrapolation, par exemple :

si z= INDUIT alors sortir $IND(\{\phi_x(j)\}\ j =< i$ (°)

A présent 2 choses peuvent être faites :
-inférence sur soi, où, dans le cas d'une inférence poppérienne (PEX),
-diagonalisation sur les résultats de l'inférence sur soi.

Dans le premier cas $\phi_e(i)$ sort des hypothèses sur lui-même à partir de l'observation de ses input-output, comme une machine socratique (voir plus loin), ou comme avec IND dans la ligne (°), en observant ses "voisins" dans le réseau.

Dans le deuxième cas, à la façon de la machine autoréfuteur de Lucas (comme celle décrite par Webb, voir plus haut, remplacée par la machine RL-IND qui correspond au réfuteur de Putnam) :

si z= INDUIT alors sortir $RL\text{-}IND(\{\phi_x(j)\}\ j =< i$ (°)

ou encore, dans les situations non constructives, comme celles mises en évidence dans les preuves de Case et Smith, on peut prendre pour IND-RL le *semi-réalisateur*. Il produit un nombre fini (ou dénombrable, RE avec oracle) de programmes parmi lesquels se trouvent au moins un qui vérifie la spécification liée au critère d'identification. Dans ce cas le réseau devient bifurquant.

Dans tous les cas, les théories sur lesquelles les machines à inférence inductive se stabiliseront seront des extensions de G, G* (voir aussi Carlson 1986, Beklemishev 1991).

De même, pour la connaissance informelle, avec le stratagème, les théories produites seront des extensions de S4Grz (voir aussi Artemov 1990). G,G*,S4Grz, & Co. apparaissent donc comme les squelettes des discours autoréférentiels stables des machines ou collections de machines dans les voisinages de l'infini.

Cela résulte principalement du fait qu'aussi bien l'identité personnelle, avec MEC-DIG-IND, que l'autoréférence arithmétique, reposent sur la construction diagonale de Gödel-Kleene (voir 2.2).

*8°) La logique de l'auto-inférence correcte*

Ici est approfondie la réfutation de Lucas de bas niveau à-la Arbib (cf 2.3.1) : où on remplace les démonstrateurs de théorèmes par des MII.

Putnam (1965) a montré par diagonalisation qu'il n'y a pas de machines extrapolantes universelles.

On a vu que cet argument ne tient plus si on généralise les critères d'identification, notamment en admettant des critères réalistes. Ici "réaliste" peut être pris aussi bien dans le sens commun que dans le sens des modèles de C, car c'est être réaliste que d'admettre la possibilité de l'erreur (vu comme une communication du faux) dès lors qu'on s'occupe de machine à inférence.

Les résultats de Chen (voir 2.3.5), montrent qu'il en est *nécessairement* ainsi si on désire avoir la reconnaissance de classes de phénomènes assez large. La machine doit pouvoir être localement (et relativement) inconsistante, localement incorrecte, etc...

Cela justifie (arithmétiquement) la non-monotonie de l'intelligence. De plus par Case, Smith et Chen encore, les sauts sont nécessairement non constructifs (comme d'ailleurs les gap de Borodin, cf Hartmanis). Cela suggère la possibilité d'une relation plus profonde entre G, G* et l'inférence. Il ne sera dès lors pas non plus inopportun de regarder de plus près S4Grz et l'intuitionisme extrait de l'arithmétique par le mécanisme et le stratagème.

Considérons une machine à inférence inductive presque socratique capable en outre de démontrer les théorèmes de l'arithmétique élémentaire (éventuellement formalisée dans une théorie du premier ordre).

La réfutation de Lucas de bas niveau est obtenue en interprétant "produire comme vrai" par "inféré correctement". Est-il possible de dériver une logique de l'inférence correcte ? Quelle relation une telle théorie peut-elle avoir avec S4 ou S4Grz ? Quel lien peut-on espérer isoler entre l'inférence (éventuellement instinctive) correcte et la connaissance intuitive (la preuve informelle éventuellement relative à une collection de propositions inférées correctement) ?

La notion d'inférence *correcte* est sémantique et il est nécessaire de décider du modèle dans lequel la machine est plongée. Je vais, comme plus haut me limiter au modèle (standard) de l'arithmétique. Le choix de ce modèle est rendu plausible avec MEC et la thèse de Church (ou mieux, la thèse de Post-Turing). Dans la section 3.3 un engagement ontologique minimal permettra, avec MEC, de justifier ce choix de façon plus précise. Retenons ici que la machine, comme le candidat mécaniste décidé à utiliser un translateur comme moyen de locomotion, va inférer des propositions concernant un nombre naturel la (le) représentant complètement relativement à un environnement universel (comme le code génétique représente le corps relativement au loi de la chimie + un cytoplasme, ...).

M auto-infère correctement P ssi P(⌜M⌝) est inférée par la machine et P(⌜M⌝) est arithmétiquement vraie. L'exemple générique est donné par une machine (adéquate et consistante) qui infère sa propre consistance. Si la machine s'identifie d'une façon ou d'une autre au produit de son inférence (presque) socratique, les révisions de ses hypothèses sont autant de cause de changements de ses activités possibles. En particulier si elle infère son adéquation (ou son universalité) et sa consistance (ou même son inconsistance) elle se transforme en une nouvelle machine

$$M \Longrightarrow M + con\ ⌜M⌝ \text{ (ou } M + \neg con\ ⌜M⌝).$$

Cette nouvelle machine est adéquate et consistante. M + con ⌜M⌝ est correcte. M + ¬con ⌜M⌝ est consistante mais n'est pas (autoréférentiellement) correcte. Chacune de ses machines est toujours descriptible par G&G*. Ce processus peut être itéré. En particulier G&G* s'étend en un système $TL\varepsilon_0 \& TL\varepsilon_0^{+}$ si l'itération s'effectue jusqu'à $\varepsilon_0$ (Becklemishev 1991). $TL\varepsilon_0$ correspond à l'extension de G, et $TL\varepsilon_0^{+}$ correspond à l'extension de G*. Voir aussi Carlson 1986 pour des extensions jusqu'à $\omega_1^{CK}$.

Remarquons bien que si M infère la consistance de M (comme proposition arithmétique) elle devient une autre machine. Dans le cas contraire elle deviendrait équivalente à une machine de Rogers du genre

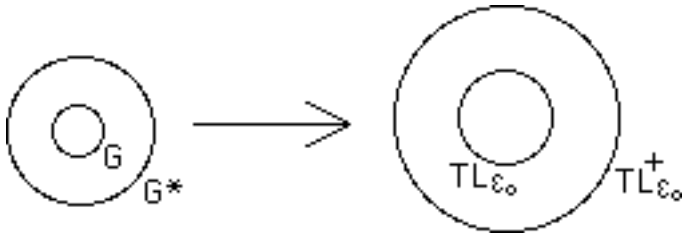
$$M+Con(⌜M+Con(⌜M+Con(⌜M+Con⌜... \ ...)))$$

Expression dont j'ai montré en 2.2 comment la rendre finie (avec 2-REC). Cet aspect des choses est reflété avec G :

G+◊⊤ est inconsistant, (G+□⊥ est consistant)

alors que M+con(⌜M⌝) est consistante. Autrement dit ou bien la consistance demeure hypothétique ou bien elle est adoptée structurellement par la machine. Dans ce dernier cas ou bien la consistance est absolue et la machine, se transformant en machine de Rogers, devient inconsistante, ou bien la consistance est locale et la machine en sort intensionnellement transformée, prête éventuellemnt à itérer cette transformation dans le transfini. On peut concevoir la situation hybride où la machine infère les versions absolues de façon hypothétique et les versions relatives ou locales comme auto-enrichissements personnels.

Remarquons que de tels enrichissements personnels n'amenuisent pas l'incomplétude. A cause même de l'interprétation épistémique que le mécanisme permet de tirer de l'incomplétude, l'ignorance ne décroît pas au fur et à mesure que la machine infère et réfléchit des propositions vraies et non communicables. Le contraire est vérifié. L'ignorance croît avec la connaissance, celle-ci éclairant l'ignorance vue comme un espace des savoirs possibles.




L'ignorance ici peut être assimilée à une vision à partir d'une école du dedans de l'école du dehors. On voit poindre un "rôle" pour la conscience, c-à-d l'inférence (instinctive[106] ou non) de sa propre consistance, qui est de tenir compte, pour ses actions possibles, de l'ignorance. C'est cette prise en compte qui permet de prendre des risques, de faire preuve de courage, etc. Je reviens sur le rôle de la conscience plus loin.

Comme G et G* sont décidables au lieu de prendre le correspondant arithmétique, con(M), de la consistance ◊⊤, on peut prendre les correspondants arithmétiques de n'importe quelle proposition appartenant à G*\G.

G*\G décrit ainsi, modulo un schéma de traductions arithmétiques, l'ensemble des propositions autoréférentielles qu'une machine ne peut pas

[106] D'une façon générale j'entends "instinctif" comme "placé initialement dans les équations de récursion".

communiquer à une tierce machine, mais qu'elle peut néanmoins inférer (absolument et hypothétiquement, *ou* relativement et structurellement) pour son compte personnel. G* permet ainsi une lecture mécaniste des principes de Wittgenstein et de Lao-Tseu (Watts-Valadier) aussi bien dans la version positiviste (et relative) que solipsiste (et absolue) par ses théorèmes :

| | |
|---|---|
| ¬ ⊡◊ ⊤ | ¬ □◊ ⊤ |
| ⊡◊ ⊤ -> ¬◊ ⊤, | □◊ ⊤ -> ¬◊ ⊤, respectivement. |

En particulier une personne ayant "survécu" à la translation qui en *déduirait* (et pire : communiquerait) qu'elle va survivre *nécessairement* aux translations futures est inconsistante[107] (communicatrice du faux). L'hypothèse mécaniste, au niveau G*, se mue ainsi nécessairement pour la machine *pratiquante*, en une infinité d'actes de foi[108].

*9°) Sommes -nous tombés dans le piège de Wittgenstein ?*

C'est la nature essentiellement hypothétique des inférences de consistance, ou de l'inférence du mécanisme (sous la forme de ◊⊤ ou de ◊ ◊⊤) qui nous permet de communiquer sur l'incommunicable tout en restant consistant relativement au mécanisme, et nous permet dès lors d'éviter le piège de Wittgenstein.

Cet aspect est illustré formellement par la perte de la règle de nécessitation de G* ce qui fait du piège de Benacerraf un cas particulier du piège de Wittgenstein.

◊ ⊤ mais aussi ¬ ⊡◊ ⊤ sont inférables sur la base de l'acte de foi mécaniste, étant des théorèmes de G*, mais n'étant pas des théorèmes de G, la nature non communicables de ces vérités est préservée et même, avec la thèse de Church, expliquée.

De façon plus profonde c'est le contour nécessairement flou du graal, c-à-d le caractère non récursivement énumérable des indices des fonctions totales récursives (cf *Kleene's overnight* dans 2.1) qui rend plausible cette interprétation mécaniste de l'incomplétude. Et je rappelle que le contour du graal est flou du fait que l'école du dehors est fermée pour la diagonalisation.

Les théorèmes de G correspondent à des propositions communicables. Pour prendre le cas paradigmatique : G ⊢ ◊⊤ -> ¬□◊⊤ , ou encore G* ⊢ □(◊⊤ -> ¬□◊⊤), bref, "◊⊤ -> ¬□◊⊤" est communicable. Cela a été directement illustré avec l'expérience de la duplication par le fait que la

---

107 Notons que le simple fait qu'elle communique avoir survécu l'a rend inconsistante.

108 Pour une interprétation des résultats de Gödel en terme d'acte de foi de la part du mathématicien, voir aussi Guillen 1983.

tierce personne peut modifier le translateur, à l'insu du mécaniste pratiquant, de telle façon que l'original ne soit pas annihilé (voir 1.3). En particulier on observe, aussi bien directement avec le translateur qu'avec le mécanisme digital + la thèse de Church (et donc avec G&G*.) que les propositions non communicables ne sont pas non plus auto-communicables. Cela confirme la nécessité du renouvellement des actes de foi mécaniste. Cet aspect des choses était déjà capturé par la théorie C, et apparaît à présent comme appartenant aux discours stables des machines autoréférentiellement et auto-inférentiellement correctes dans les voisinages de l'infini. La thèse de Church permet de plonger ainsi naturellement la théorie C dans la théorie G sans qu'il ait été nécessaire de modéliser en détail les expériences par la pensée. Cela n'enlève rien à l'intérêt intrinsèque d'une telle modélisation qui néanmoins serait trop complexe à ce stade. L'analyse théologique (linguistique et géométrique) du chapitre 1 est arithmétiquement instanciée, à travers la thèse de Church classique, par les conséquences de l'incomplétude. Le mécaniste peut dès lors interpréter les phénomènes d'incomplétude comme le début d'une découverte (toujours partielle et jamais complétée) du mathématicien au sein de l'arithmétique (informelle).

*10°) Le solipsiste muet*

*Grz et le solipsiste* Regardons à présent le résultat de l'application du stratagème fort à G (= à G* ici), c-à-d S4Grz.

Par le morphisme de Gödel 1933 et les morphismes de Boolos-Goldblatt-Kusnetzov-&-Muravitskii, S4Grz (et ses extensions, cf Artemov 1990) donne une description épistémique du sujet que j'appellerai aussi le solipsiste muet[109]. Celui-ci est non arithmétisable, c'est la principale qualité de $\Box$. Il est muet à son sujet et ne s'invoque jamais lui-même de façon *explicite* dans ses exhibitions. Sa connaissance est non verbale. Il montre et exhibe plus qu'il ne démontre. De façon interne et directe il est plutôt décrit par ARIL, mais l'aspect épistémique, qu'il vaut mieux décrire classiquement est formalisé par S4Grz et AREA (avec la translation de Gödel). Ses preuves consistent en des extensions constructives de lui-même. La philosophie de Brouwer, alliée au mécanisme fait de ce sujet un créateur innommable[110].

Ceci est corroboré avec $\neg\Box$(je=i) quel que soit i. C'est le *solipsiste* qui instinctivement aura quelque répugnance, fondée ! (avec MEC-DIG), à utiliser le translateur comme moyen de lococmotion.

Le solipsiste dont la connaissance est décrite par S4Grz n'admet pas de référence à lui-même explicitable ou arithmétisable. Il est par définition lié à

109 Par opposition au solipsiste bavard ou doctrinaire que l'on va rencontrer en 3.3.

110 Son caractère innommable (tiré ici de l'impossibilité de l'arithmétiser) peut être comparé au fait que la thèse de Church intuitionniste est inconsistante avec un schéma dû à Kripke qui formalise la notion de sujet créateur.

la vérité puisqu'il ne peut connaître p, (dans AREA on a $\Box p \to p$) sans que p soit vrai, ce qui découle du stratagème ou de la thèse d'Artemov. Avec MEC il connaît complètement la vérité dans le sens que IL* = IL, ou encore S4Grz* = S4Grz, ce qui, me semble-t-il capture un aspect de la conception qu'a Bergson de l'intuition.

La majorité des mathématiciens intuitionistes insistent sur le fait qu'ils ne partagent pas le solipsisme de Brouwer tout en reconnaissant son génie et la portée de l'intuitionisme en philosophie des mathématiques[111].

Cette circonstance oblitère le caractère solipsiste de cette philosophie où les objets mathématiques sont des constructions de l'esprit et chez Brouwer il s'agit même de libres constructions d'un esprit individuel, d'un soi. Je conjecture que ce soi est innommable en philosophie mécaniste.

Brouwer pourrait être d'accord car s'il permet la libre référence à ce soi, par exemple lorsque l'esprit construit un nombre réel (une suite de choix), il présente ses mathématiques comme quelque chose d'informel et l'intuitionisme comme fondamentalement non formalisable. Et bien qu'il se soit un temps enthousiasmé pour la formalisation de Heyting[112], il a insisté sur le caractère provisoire d'une telle formalisation. En particulier les nombreuses critiques de la notion Brouwerienne de sujet créateur, ne fonctionnent que sur les approximations formelles comme celles dues à Kreisel (1970) ou Kripke (voir Beeson 1985).

En admettant cependant que l'existence du sujet créateur soit en contradiction avec le principe de Markov (ce qui n'est prouvable qu'avec des versions approximées du sujet) on est amené à considérer la thèse de Church et le principe de Markov dans l'arithmétique épistémique de l'autoréférence, c'est-à-dire AREA (EA avec la formule de Grzegorczyk). Le fait, par exemple, que l'on sache prouver, (dans l'arithmétique de Peano), l'existence de fonctions non calculables comme BB et de fonctions (prédicats) absolument non calculables (indécidables) entraîne la fausseté de la pseudo-thèse de Church épistémique affaiblie :

$$\Box\forall x\exists y\ P(x,y) \to \exists z\Box\forall x\ (\phi_z(x)\Downarrow\ \&\ \Box P(x,\phi_z(x))$$

Je pense cependant que la thèse de Church

$$\Box\forall x\exists y\ \Box P(x,y) \to \exists z\Box\forall x\ (\phi_z(x)\Downarrow\ \&\ \Box P(x,\phi_z(x))$$

---

111 En 3.3 je critiquerai explicitement la *doctrine* solipsiste. Néanmoins, ici, nous découvrons un *solipsiste commun* à tout sujet (machine). Ce dernier reste consistant (avec le mécanisme) lorsqu'il n'essaye pas de convaincre les autres (qui en fait n'existe pas pour lui) du bien-fondé du solipsisme.

112 Brouwer a considéré le travail de Heyting comme étant au moins aussi important que les résultats d'incomplétude de Gödel. Mais plus tard il critique l'approche de Heyting lui-même.

est vraie. Ce devrait être un théorème d'une version quantifiée de G*, si une telle version existait (ce qui n'est pas le cas, voir 2.3.3).

Les constructivistes russes (voir un compte rendu du constructivisme de l'école de Markov dans Bridges et Richman 1987, Richman 1983, Troelstra et Van Dalen 1988, Beeson 1985) admettent un principe dû à Markov :

$$\forall x(A(x) \vee \neg A(x)) \;\&\; \neg\neg \exists x A(x) \;\rightarrow\; \exists x A(x)$$

qui semble naturel avec la thèse de Church classique mais qui est indépendant de la plupart des versions de la thèse de Church intuitioniste. Artemov (1990) a démontré la fausseté de Markov dans AREA.

Ceci peut être utilisé pour raffiner le fait que le mécanisme entraîne que le solipsiste en nous est antimécaniste, ce qui rejoint l'idée que le solipsiste n'accepte pas le V classique de l'autoduplication. Ceci a été en partie suggéré et est rendu consistant par le fait que la thèse de Post-Turing n'est pas connaissable ou prouvable "de l'intérieur" : AREA $\nvdash \Box p \leftrightarrow \boxdot p$.

*Grz et l'inférence inductive* Peut-on étendre le stratagème à l'inférence inductive (correcte relativement au modèle standard de l'arithmétique). Admettons que la machine (autoréférentiellement et inférentiellement correcte) infère correctement ce qu'elle prouve intuitivement. C'est assez naturel. Peut-on étendre l'interprétation de $\boxdot$ par l'inférence correcte INFC(p) ? INFC(p) signifiant que p est inférable par la machine et que p est vrai (dans l'interprétation standard de l'arithmétique) On peut vérifier que INFC(p) satisfait les axiomes de S4. INFC(p) vérifie-t-il la formule de Grzegorczyk ? Cela semble faux. En effet pour les propositions fausses et absolument indécidables $p \rightarrow \boxdot p$ est vrai, mais peut être à son tour absolument indécidable. Dans ce cas on aurait $\neg \boxdot(p \rightarrow \boxdot p)$ et Grz, $\boxdot(\boxdot(p \rightarrow \boxdot p) \rightarrow p) \rightarrow p$, n'est pas satisfaite.

Une autre voie pour ce problème serait de trouver un lien entre Grzegorczyk 1964 et Grzegorczyk 1967, c'est-à-dire entre une interprétation de la logique intuitioniste en terme de recherche scientifique et la motivation topologico-algébrique de la formule Grz elle-même (voir aussi Wansing 1993). L'approche de l'intuitionisme de Kolmogorov, (voir aussi Medvedev (voir Rogers 1967, voir encore Cooper 1990, et Case 1971) et finalement Artemov 1990 donnent des espoirs de clarifier le rôle de Grz dans une approche modale abstraite de l'inférence inductive[113].

---

[113] Je signale qu'au paragraphe 8 de Marchal 1991 (un article écrit en 1987) je confonds $\boxdot$ et INF si bien que j'identifie de façon erronée inf($\Diamond \top$) et $\boxdot \Diamond \top$.

*Remarque*. Grz (et donc le mécanisme + le stratagème) permet d'utiliser la topologie en philosophie de l'esprit d'une façon non métaphorique. Voir Lavendhomme (1984, 1986) pour quelques applications plus métaphoriques quoique Lacaniennes. Vu l'importance de la topologie dans les sciences de l'information il n'est pas étonnant qu'elle ait un rôle en philosophie mécaniste de l'esprit.

Plus étonnant est le rôle que le stratagème confère aux espaces totalement distributifs, comme les espaces topologiques finis, qui sont à l'origine de la formule de Grzegorczyk. On peut espérer que la formule Grz puisse être extraite directement des principes de Gandy-Sepherdson (voir 1.1) définissant le mécanisme. En gros il faut trouver les relations pertinentes entre les ensembles héréditairement finis de Gandy et les espaces totalement distributifs de Grzegorczyk.

*Grz et le flux temporel*

L'interprétation épistémique de ⊡ est naturelle comme l'est celle de S4. D'un point de vue purement épistémique S4Grz n'apporte rien de nouveau à l'arithmétique de Reinhard-Shapiro (voir 2.3.2). Shapiro décrit pourtant la connaissance du sujet comme une entité variable évoluant dans le temps, et dans les modèles de Kripke de la logique intuitioniste les mondes sont souvent interprétés intuitivement comme des états mentaux successifs temporellement accessibles[114]. Du point de vue temporel la formule de Grzegorczyk introduit l'antisymétrie (voir plus haut) et celle-ci, introspectivement caractérise la subjectivité du temps qui passe. Cela rejoint de nombreuses analyses de la conscience en terme de temps ou mieux de durée comme celle de Bergson, et de Brouwer (qui cite par ailleurs Bergson à ce sujet) mais aussi des auteurs divers comme celle de Dogen 1232-1253, ou Héraclite, Saint-Augustin pour en nommer quelques-uns.

Notons que ce temps subjectif est défini de façon interne, il ne se réfère pas a priori à un temps physique. C'est la même notion de temps que celle du mécaniste pratiquant l'usage des duplications postposées. Ce point joue un rôle dans le problème du corps et de l'esprit dans le dernier chapitre.

*11°) La sémantique de Boolos pour G**

J'ai été conduit à rajouter une ou plusieurs capacités d'inférence inductives dans les équations de récursion initiales. La connaissance procédurale que constitue ces équations ont été alors qualifiée d'instinctives (dans 2.2). Cette capacité d'inférence correspond à une capacité instinctive et rappelle la définition de la conscience perceptive de Helmholtz comme inférence inconsciente. Cet aspect des choses est illustré plus ou moins directement dans une sémantique de G* élaborée par Boolos. On se rappelle

[114] Voir par exemple Bowen et de Jongh 1986.

que G* n'est pas une logique normale (on n'a pas la nécessitation), ce n'est pas non plus une logique classique minimale (on n'a pas la monotonie rationnelle). Il n'y a donc ni modèle de Kripke, ni modèle de Scott-Montague disponible pour G*. Bref, a priori, on peut se demander à quoi pourrait ressembler une sémantique de G*. En tenant compte de la proximité de G* avec G, Boolos a construit une sémantique de G*, qui illustre, de façon informelle, la nature inférable de G* pour une machine MII auto-réfé-et-inférentiellement correcte.

A partir d'un modèle de G, Boolos 1980a construit une collection de modèles servant de sémantique pour G*.

M = (W,R,V) est un modèle, i un nombre naturel, On va définir une notion de vérité à la limite dans un monde w. Je dis que A est vraie au stade i dans le monde w (appartenant à W) : atomique(p) -> $\Vdash_w^i p$.

1) si A est atomique, A est vraie au stade i (et donc à tous les stades) dans w ssi A est vraie dans w.

2) Le faux n'est jamais vrai, à quel stade que ce soit. Dans aucun monde : $\neg \Vdash_w^i \bot$

3) A->B est vrai à un stade i dans w ssi A est faux en w au stade i, ou B est vrai en w au stade i : $\neg \Vdash_w^i A \vee \Vdash_w^i B$.

Jusqu'ici, ce sont les définitions habituelles, et cela montre que dans les mondes à tous les stades on a la logique classique.

4) $\Vdash_w^i \Box A$ ssi $\Vdash_w \Box A$ et pour tout $j < i$, $\Vdash_w^j A$.

On a 1) A est vraie en w au stade 0 ($\Vdash_w^0 A$) ssi A est vraie en w, et on a, c'est le point intéressant $\Vdash_w^{i+1} \Box A$ ssi $\Vdash_w^i \Box A \& A$, c'est-à dire, avec A atomique ou interprétée arithmétiquement :

$$\Vdash_w^{i+1} \Box A \Leftrightarrow \Vdash_w^i \boxdot A \quad \text{(avec STRAT).}$$

De même il est intéressant de voir comment traduire la vérité à un stade i pour une proposition énonçant une possibilité ou une consistance :

$$\Vdash_w^i \Diamond A \text{ ssi } \Vdash_w \Diamond A \vee \exists j < i \Vdash_w^j A.$$

Le monde a une mémoire (si on interprète les stades temporellement) telle que ce qui a été vrai devient possible, et telle que ce qui a été vrai, et est vrai dans tous les mondes accessibles est nécessaire.

*définition principale* :

$$\Vdash_w{}^{*} A \quad \text{ssi} \quad \exists i\, \forall j>i \;\; \Vdash_w{}^{j} A.$$

le truc important, qui fait peut-être de S4Grz une "logique de la flèche temporelle" est $\Vdash_w{}^{i+1} \Box A <=> \Vdash_w{}^{i} \boxdot A$,

cela entraîne $\Vdash_w{}^{*} \Box A <=> \Vdash_w{}^{*} \boxdot A$, (et $\Vdash_w{}^{*}$ peut servir pour une sémantique de la logique intuitioniste arithmétique ARIL). Le connaisseur est alors lié *dynamiquement* à la connaissance.

Au niveau G*, donc avec MEC au niveau adéquat, le solipsiste dont la connaissance s'exprime à la première personne, est identifié avec le positiviste (dont la croyance s'exprime à la troisième personne).

*Question* Peut-on construire une sémantique de Scott-Montague variable de ce style pour KD?* ?

*12°) Des rôles de la conscience*

La relation suggérée entre la consistance (inférée) et la conscience rend compte de l'aura d'absurdité de la question du rôle de la conscience ou même du rôle de la vie. On ne questionne pas le mathématicien sur le rôle de la consistance en mathématique. On espère, à la base, que ces propos soient, au moins localement, consistants. Cependant le métamathématicien peut mettre en évidence des conséquences de la consistance pour une théorie comme des impossibilités (de prouver le faux, de prouver la consistance de la théorie).

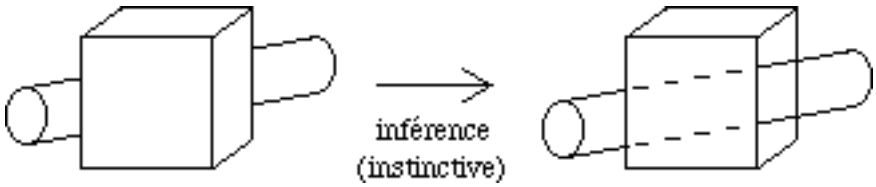
Ici, s'inspirant de Helmholtz et de G&G*, la conscience est approximativement décrite par les inférences en partie instinctives de l'autoconsistance, de type INF($\Diamond\top$), dans diverses situations ainsi que leurs remises en question, de type INF($\Diamond \Box \bot$). Inférences qui conduisent en principe à escalader quelques échelles transfinies et bifurcantes d'ordinaux. On peut voir ces échelles comme le squelette des révisions possibles. Cette montée permet différentes choses capables de donner un rôle local à la conscience.

Ainsi, de façon un peu grossière, la conscience

- **permet de mesurer <u>son</u> ignorance**. Notamment, par des considérations indexicales, c-à-d essentiellement celles dont la digitalisation a permis la désindexicalisation par diagonalisation. Informellement ces références indexicales conduisent à des propositions absolument indécidables comme avec l'autoduplication, ou avec les conséquences absolues de la thèse de Church (cf Gödel 1951, l'analyse de Kalmar de Reinhardt, etc. voir plus haut). Formellement les diagonalisations conduisent à des propositions relativement indécidables et à des théories comme celles des *degrés*

d'**in**solubilité, de complexité, d'**in**tractabilité, etc. Dans ce sens, mais de façon moins formelle on peut aussi voir la théorie des cardinaux de Cantor comme des mesures relatives d'ignorance concevable (j'y reviens plus bas avec des considérations sur le paradoxe de Skolem). Cela donne des **in**finis, des **in**dénombrables. En bref, cela permet de donner *un* sens (et surtout pas *le* sens) aux expressions de type "**in**-#". Notons que le ou les *sens* précèdent instinctivement le choix des noms. La théorie G&G* permet d'inférer que la nomination de ces "in-#" les multiplie[115]. Les nominations ici sont intrinsèquement ambiguës.

- **permet de tenir compte de l'ignorance** ce qui permet les inférences instinctives complétant (minimalement) l'inconnu :




Le caractère instinctif de ces inférences est à la base de l'intuition immédiate. Il y a un postulat instinctif de consistance de l'environnement. Un argument en faveur de l'idée selon laquelle notre cerveau, et en particulier le *tronc cérébral*, gère de façon automatique ce genre d'inférence, est donné dans la section suivante sur les rêves, notamment où je parle du phénomène de **faux-éveil**.

- **est à la base de la construction** ou de l'évocation d'extensions temporelles, spatiales, essentiellement locales, éventuellement multiples, de soi et ainsi de l'usage des capacités personnelles relativement à un environnement universel (cf 2.2). Elle permet la gestion du doute et la possibilité du choix. Choisir W parmi {W,M}, revient à se dupliquer et tuer sa propre M-extension. Choisir est déjà une façon de se tuer soi-même.

L'auto-observation multiplie les auto-extensions possibles et donc les choix possibles. Plus on essaye de s'autodéterminer, plus nombreuses apparaissent les extensions possibles de soi. Il s'agit d'une sorte d'*auto-diffraction*.

En 3.3, avec le paradoxe du dovetelleur universel, on abordera le cas limite où l'auto-observation, se faisant au niveau adéquat du mécanisme, entraîne une si grande automultiplication qu'a priori l'inférence inductive devient pratiquement impossible dans les voisinages de zéro. On verra cependant que les contraintes du mécanisme sont telles qu'on retrouvera une phénoménologie "filtrant" ce phénomène exagéré d'autodiffraction :

---

[115] Comme dans le premier chapitre du *Lao Tseu*.

filtrage qui constituera essentiellement un embryon de phénoménologie de la matière.

- **est à l'origine d'une accélération** relative à un niveau (universel ou non). En effet l'inférence de la consistance de soi et la réflexion de cette consistance cause une accélération relative. C'est en liaison avec les phénomènes du speed-up de Gödel ou de Blum, mais au niveau de l'inférence inductive je rappelle que Daley et Smith ont mis en évidence des phénomènes similaires. Ces phénomènes d'accélération sont qualitativement importants lorsqu'est inféré du vrai non communicable. De telles inférences sont par nature risquées, mais comme on l'a vu, aussi bien directement avec le translateur qu'avec MEC, TC (et donc G&G*) le caractère risqué est communicable. Nous pouvons donc être prévenus et faire éventuellement preuve de peur/courage, etc.

- **permet le développement de la liberté** dans le sens d'une indépendance accrue relativement au niveau qui nous supporte. L'indépendance résulte de l'accélération computationnelle relative. Cette liberté (relative) est obligatoirement en conflit avec la sécurité (relative) puisque les niveaux des supports ne peuvent être déterminés de façon communicables. Je montrerai dans quel sens de tels niveaux peuvent être inférés empiriquement (voir 3.3).

Notons à ce sujet que je rejoins une approche du libre-arbitre en terme de rapidité computationnelle due à Good (1971).

- **permet la reconnaissance de l'erreur possible** dans le sens où l'inférabilité de la consistance ◊⊤ rend inférable la possibilité de la communication du faux ◊□⊥.. Elle permet la reconnaissance de l'éveil possible et donc la possibilité d'une *éventuelle* lucidité. Voir section suivante pour le lucidité nocturne. Cette reconnaissance de l'erreur rend nécessaire la présence **d'une "surface non monotonique"** de la machine à inférence qui permet alors une procédure de révision : □⊥-> □'⊤.

Décrire ce processus nécessiterait de raffiner davantage la théorie G et sa sémantique (voir plus bas). Ce point est toutefois utilisé informellement dans 3.1. Il s'agissait de le motivation originale de Magari pour les algèbres diagonales (voir Boolos & Sambin 1991).

- **permet la reconnaissance de l'autre** (avec MDI), plus précisément de soi vu comme un autre, l'*in-moi* si on me permet l'expression : inf(◊⊤<->⟐⊤). De la même façon elle permet le reconnaissance du soi dans l'autre ou d'un autre soi, *avec le risque de disparaître*. En l'occurrence, le solipsisme de Brouwer est rigoureusement, avec MEC, attribuable à sa croyance en son immortalité. Ici on voit que le mécanisme permet de sauver le *solipsiste en*

*nous* en reconnaissant le *solipsiste en l'autre* et nous oblige ainsi à abandonner *la doctrine* du solipsisme. C'est pourquoi le solipsiste isolé par l'hypothèse mécaniste (et décrit par S4Grz) est muet à *son* sujet. Il prouve par constructions et exhibitions.

- **permet le développement de l'intelligence**, au sens de Binet, puisqu'elle permet la reconnaissance de l'erreur, de l'ignorance (et donc du silence) et de l'autre (et peut ainsi profiter par union de phénomènes de non-union). De même la conscience permet *in fine* de reconnaître, *implicitement* d'abord, sa propre $\Sigma_1$-complétude ou *universalité* (au sens de Turing), ou sa propre créativité (au sens de Post, voir aussi Myhill) et permet donc le développement de multiples compétences singulières variées et variables par autospécialisations locales sur différents domaines. Le piège ici, consistant à s'enfermer (ou enfermer d'autres) dans un quelconque domaine.

D'une façon plus générale les généralisations de Case du second théorème de récursion permettent des co-évolutions de systèmes récursivement inséparables.

- **permet le développement du langage**, et notamment de la perception du décalage entre ce que l'on veut dire et ce que l'on parvient à dire. La pensée se développe en zigzaguant entre l'intuition, le savoir personnel (l'épistémique, le doxastique) et la contre-intuition du dehors (paradoxastique) qui comprend l'autre et soi. De façon moins heureuse (sans doute), la conscience rend le mensonge possible.

Je partage avec Brouwer le lien entre la logique et le langage. La logique est considérée comme émergeante ou secondaire par rapport à la pensée, l'esprit et/ou le soi.

Le mécanisme permet cependant d'assimiler le prouvable formellement avec le finiment-communicable-de-façon-convaincante. Il s'agit entre autres de la communication scientifique correcte et positiviste (voir 1.2 et 2.2).

Ceci permet de faire collapser l'intuitionisme-épistémisme (S4) et le positivisme G au niveau où le fonctionnalisme est correcte (par MEC-IND)[116].. Je conjecture que ce décalage est à l'origine des mathématiques et de toute pensée du second ordre ce qui inclut les discours sur les *infinis*, le ou les notions de *continu*, les logiques d'ordre supérieur, etc.

- **permet le rêve** et l'origine du doute. Voir section suivante. Plus exactement elle rend (toujours au niveau G*) nécessaire la possibilité du rêve, de même qu'elle rend nécessaire la possibilité de la mort (absolue). C'est l'aspect réaliste de C (inclus dans G) dont on est parti (voir 1.2).

---

116 Ce point de vue est plus proche de celui du jeune Wittgenstein que de celui de Brouwer. Wittgenstein n'attaque pas la logique classique (à la différence de Dummett 1963)

*13°) Le paradoxe de Skolem et la question "une machine peut-elle concevoir l'infini ?"*

On se rappelle de la divergence d'opinion entre Descartes et Hobbes sur la question de savoir si l'homme et/ou la machine sont à même de concevoir l'infini. Je rappelle que [117] le sens est produit par la nécessaire ignorance, ainsi que la reconnaissance possible de cette ignorance, qui **croît** avec le développement de la connaissance et de l'intelligence.

Les infinis correspondent alors à des mesures d'ignorance ou d'incapacité personnelles relatives.

Je profite ici d'une relation, déjà aperçue, entre autres, par Goodstein 1963, entre l'incomplétude Gödélienne et une forme purement sémantique d'incomplétude liée au théorème de Löwenheim et Skolem.

Selon un résultat de Löwenheim et Skolem, toutes les théories du premier ordre qui admettent des modèles infinis admettent des modèles de cardinalité dénombrable.

Il existe alors des modèles *transitifs* dénombrables de la théorie des ensembles (ZF[118] pour fixer une théorie du premier ordre bien connue).

Dire que le modèle est **dénombrable** signifie que l'univers U de tous les ensembles (le domaine de parcours des quantificateurs de la théorie) est dénombrable.

Dire que le modèle est **transitif** revient à dire que si x est un terme désignant un ensemble, alors si $y \in x$, il existe un ensemble (de l'univers) dénoté par y. En gros on a $y \in x$ et "$x \in U$" (entre guillemets car cette appartenance est intuitive) entraîne "$y \in U$", ou dit encore autrement :

$$"x \in U" \Rightarrow "x \subseteq U"$$

A présent le théorème de Cantor est un théorème de la théorie axiomatique des ensembles. Vu la présence d'un axiome de l'infini dans ZF, il existe un ensemble infini dénombrable w. ZF comprend aussi un axiome qui affirme l'existence, pour chaque ensemble E, d'un ensemble des parties de E. Donc, si w est un ensemble infini dénombrable, P(w), l'ensemble des parties de w, est aussi un ensemble, et donc "$P(w) \in U$".

Par le théorème de Cantor, formalisable dans ZF, P(w) est non dénombrable (voir 2.1).

Mais par la transitivité du modèle, P(w) est inclus dans l'univers, donc l'univers, qui est dénombrable, *inclut* une partie non dénombrable.

Voilà, en gros le paradoxe de Skolem.

---

117 Cette section use d'un peu de théorie des modèles et de théorie des ensembles.

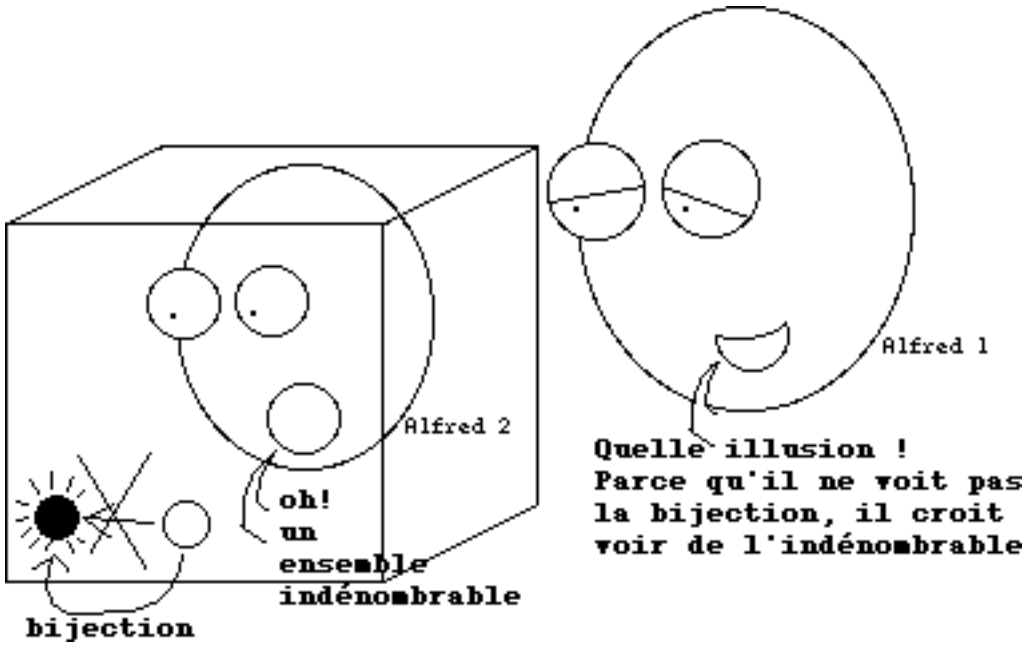
118 Zermelo Fraenkel. Voir Krivine 1969.

Il n'y a pas de contradictions formelles car la bijection entre P(w) et ω est un ensemble au sens intuitif, mais pas au sens du modèle dénombrable.

Je ne vais pas analyser en termes plus rigoureux ce paradoxe, mais je propose de regarder de façon imagée cette situation paradoxale car elle donne une version sémantique de la relativité des concepts d'infini qui découle de l'hypothèse mécaniste.

Dans les figures qui suivent, le cube représente un modèle dénombrable de la théorie des ensembles et joue le rôle de la machine (évoluant dans le temps) ou de son réceptacle. Avec MDI ce réceptacle est (au plus) dénombrable.

La preuve de l'existence d'un ensemble non dénombrable est correcte *dans* le modèle-réceptacle. Une machine de Turing "habitant" dans ce modèle y démontre *correctement* l'indénombrabilité de l'ensemble des réels. Graphiquement :




L'individu Alfred 1 observe *l'univers d'Alfred 2* de l'extérieur. Alfred 1 sait qu'il existe une bijection de ω dans ce qu'Alfred 2, qui a démontré le théorème de Cantor dans son modèle, prend pour l'ensemble des réels (identifié à P(w)).

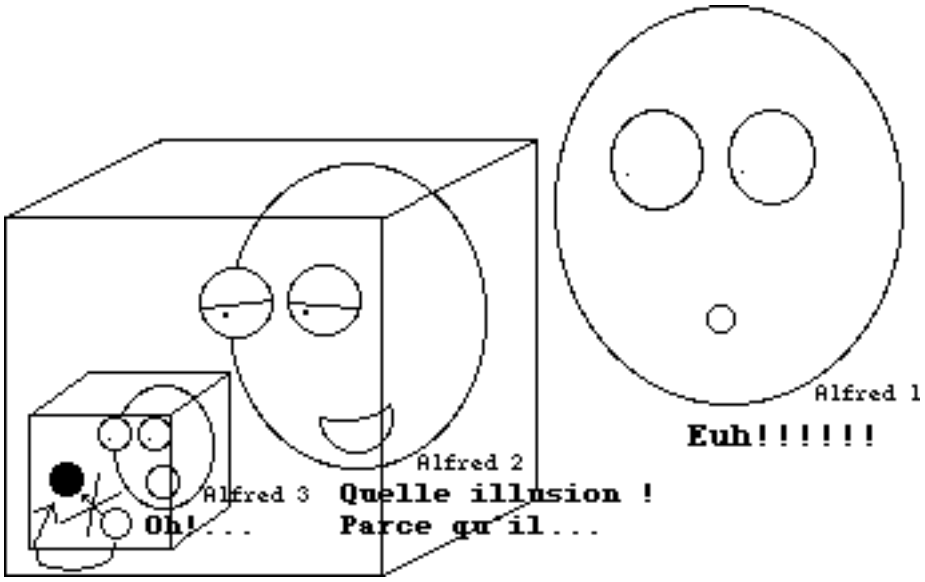
Alfred 1, qui observe les constructions d'Alfred 1 comme des constructions internes au modèle dénombrable, sait que la preuve de Alfred 2 est *correcte*, mais qu'elle prouve seulement l'inexistence, *dans son modèle à lui (Alfred 2)*, de cette bijection. Il est, dans un premier temps, en droit d'estimer que "l'autre" (Alfred 2) est illusionné et que sa preuve ne fait que stigmatiser son ignorance. L'indénombrable n'est indénombrable que pour Alfred 2. Alfred 1 pourrait, à la façon de Lucas, en arriver à se croire

supérieur. Mais ici aussi un peu de patience devrait lui permettre de mettre en doute ce sentiment de supériorité.

En effet, une analyse dans le style de Webb (ou un principe de LWV local) peut se mettre en branle : l'argument de supériorité provient du fait qu'il voit une bijection que l'autre ne voit pas. Mais ce sentiment vient aussi qu'il suppose implicitement son propre univers comme indénombrable. Dans un deuxième temps cependant, Alfred 2 se met à construire, lui aussi, dans son univers, un modèle dénombrable de ZF.

Et, pour la même raison qu'Alfred 2 découvre ou construit des ensembles non dénombrables, il peut tenter (et on démontre qu'il peut réussir) à construire lui-même un modèle interne et dénombrable de la théorie des ensembles, dans lequel une machine de Turing, Alfred 3, peut, à nouveau par mécanisabilité de l'argument de la diagonale, distinguer des ensembles non dénombrables.

Alfred 2, à la façon d'Alfred 1 peut considérer qu'Alfred 3 est abusé. Alfred 2 voit en effet la bijection entre les ensembles *non dénombrables pour Alfred 3* et le dénombrable. Alfred 1 voit que le raisonnement d'Alfred 2 ne peut être que relativement correcte, et la ressemblance entre ce raisonnement et le sien à l'étape précédente devrait mettre en doute (au moins) l'*assurance* de sa supériorité.




Ici aussi le transfini (produit par l'usage de la diagonale de Cantor) apparaît comme une mesure d'ignorance relative.

Brouwer, du point de vue intérieur (intuitif et absolu) avait finalement raison de considérer la preuve de Cantor comme non (absolument) convaincante (cf 2.1). Cette preuve met cependant en évidence une nécessaire relativité des concepts, relativité qui joue un rôle dès qu'on s'intéresse à l'école du dehors, et aux relations entre le dedans et *le* dehors

(avec TC intuitioniste et classique). Ceci illustre, avec MDI, des propriétés *objectives* du subjectif.

Cette interprétation est corroborée par le travail de McCarty et Tennant 1987, lorsqu'ils mettent en évidence la disparition du phénomène paradoxal de Skolem dès qu'on limite les diagonalisations aux diagonalisations intuitionistes (et donc internes). De même pour les sauts de Borodin (cf 2.2) dont Hartmanis montre la disparition lorsqu'on se limite aux classes de complexité dont l'existence est prouvable constructivement.

Un mécaniste fait de la proposition "il existe un ensemble absolument non dénombrable" une proposition absolument indécidable, de même que le mécanisme fait de l'hypothèse "panmécaniste" un absolument indécidable.

Une démonstration non constructive (mais formalisable dans un théorie du premier ordre[119]) prouve un résultat objectif concernant une limitation subjective[120]. Cela sonne bizarrement, mais cela découle uniquement de l'existence de *loi pour l'esprit*. Tous les rêves ne sont pas possibles, les rêves ce n'est pas n'importe quoi!

Le sens chez le solipsiste-intuitioniste est donc produit par l'ignorance, bien qu'il ne puisse pas s'en rendre *complètement* compte de l'intérieur (cf aussi le rôle de la conscience).

Le passage entre les deux dessins illustre une nouvelle fois l'usage local du principe LWV, ainsi que la mécanisabilité de l'argument de la diagonale. Lucas aurait pu utiliser le paradoxe de Skolem pour prétendre être supérieur à la machine en prétendant être à même de concevoir de l'indénombrable (ou des ordinaux non constructifs). Mais il ne peut pas commettre cette assertion de façon positivement communicable. Et de façon non communicable ou non constructive, la machine le peut aussi.

C'est tout le problème de la richesse du second ordre et de l'analyse, de la théorie abstraite des ensembles, des mathématiques non séparables de Brouwer (2ème acte de l'intuitionisme) et des fondements, jamais complètement axiomatisable, des mathématiques.

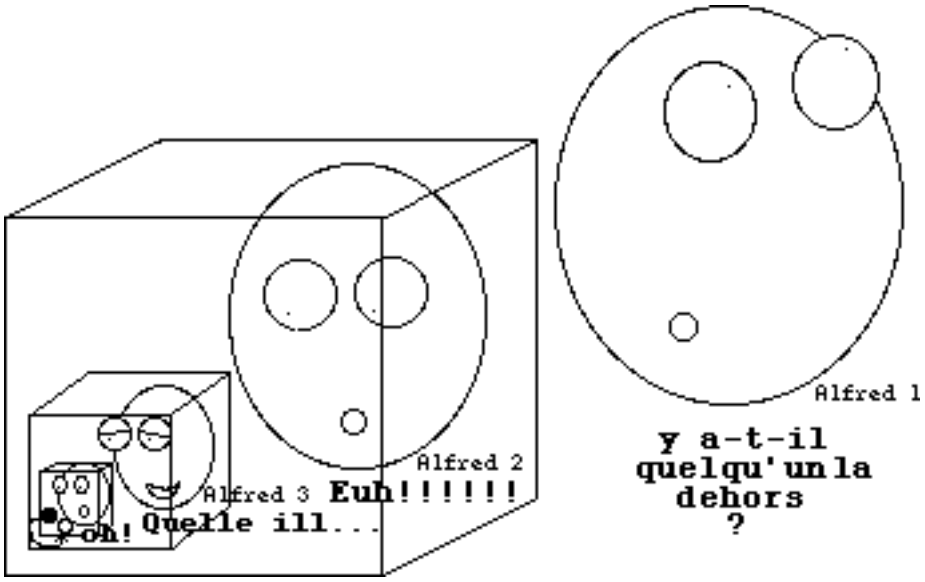
La suite de la petite histoire issue de Skolem illustre à merveille les idées de Webb dans ce cadre. Je pense au *double-edge aspect* de l'usage de l'incomplétude en philosophie de l'esprit que Webb analyse de façon approfondie dans son livre (Webb 1980).

---

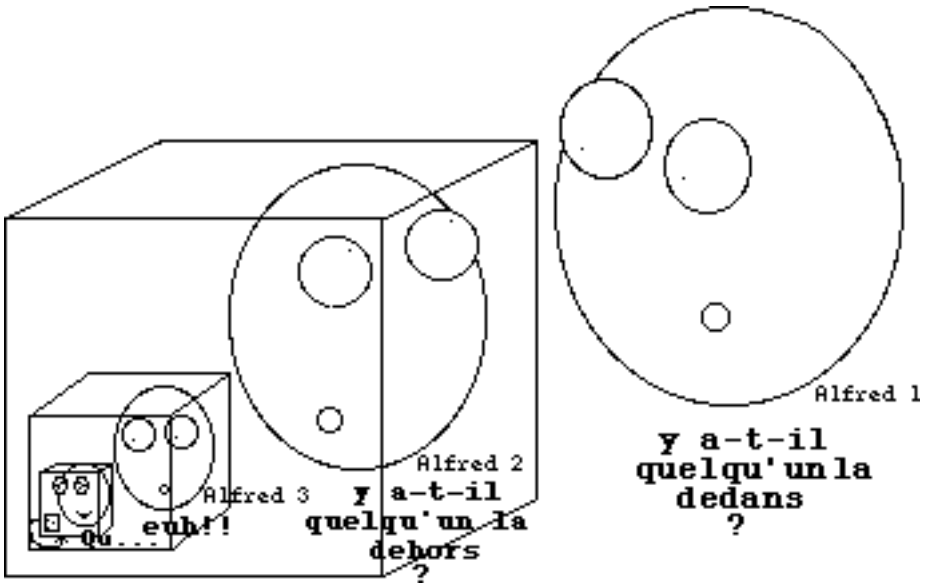
119 On n'a pas encore prouvé l'existence d'un concept qui ne soit pas représentable au sein d'une théorie du premier ordre. Par exemple, bien qu'il n'existe pas de théorie du premier ordre des ensembles finis (conséquence évidente du résultat de Löwenheim-Skolem), on peut définir le fini dans une théorie du premier ordre des ensembles. Le succès de cette dernière est son évidente facilité pour représenter les modèles des théories d'ordre supérieur. Notons malgré tout qu'aujourd'hui, pour cette tâche, elle est supplantée par la théorie des catégories (voir Lambek et Scott 1986).

120 Voir aussi les suggestions philosophiques de Ladrière 1957.

Alfred 1 réalise que l'argument utilisé contre Alfred 2 marche sur lui-même. Il réalise qu'il pourrait donc être dans une situation similaire. Il peut dans ce cas inférer une présence extérieur relativisant ce qu'il prenait jusqu'à ce jour pour de l'*absolument* indénombrable. Graphiquement :




Il sait de plus qu'Alfred 2 peut très bien arriver *correctement* à la même inférence (au niveau[121] G*). Il peut dès lors inférer ou questionner la présence d'Alfred 2 ainsi que le sens correct bien que relatif de son discours. C'est une forme de prélude à la reconnaissance de l'autre, qu'il soit dedans ou dehors. Graphiquement :




[121] Incidemment Solovay 1976 propose (sans démonstration) des extensions de G (et de G*) capturant des notions comme "vrai dans tous les modèles transitifs de ZF", ou "vrai dans tous les univers". Cela ne me semble pas évident à démontrer et je n'ai pas trouver de preuves dans la littérature. De tels résultats devraient permettre une approche plus rigoureuse pour les remarques de cette section.

Le même phénomène se produit dans certaines mathématiques constructives ou l'ensemble des réels constructifs clairement dénombrable (comme le graal) est non dénombrable du point de vue du dedans (constructif) car la bijection, entre cet ensemble et les naturels, est elle-même non constructive (Richman 1983, Beeson 1985).

Ceci est en relation avec le fait que la thèse de Church intuitioniste et le principe de Markov sont incompatibles avec les démonstrations habituelles du théorème de Skolem. Cela rapproche l'intuitionisme de Brouwer et le sujet*solipsiste* extrait du mécanisme. Cf encore Artemov 1990, McCarty et Tennant 1987.

*14°) Morale, théologie et théorie de l'intelligence*

Y a-t-il une morale ? Et si oui est-elle communicable sans qu'on tombe dans un piège de Benacerraf ou de Wittgenstein ?

Regardons à cet égard, avec G&G*, une théorie générale de la *vertu* possible.

Voilà comment Jean Brun résume une question que pose Platon dans le Ménon et surtout dans le Protagoras :

> *Y a-t-il universalité de la vertu ou diversité des vertus ? Socrate fait observer que, dans ses discours, Protagoras parle de la justice, de la sagesse, de la sainteté et de choses semblables, comme formant un tout qui est la vertu; la question est alors de savoir si la vertu est quelque chose d'unique dont ces vertus sont les parties, ou si ces vertus ne sont que des noms de cette chose unique. (Brun J. 1969).*

L'approche présente suggère que la solution à cette question réside dans le fait que les différentes vertus et l'intelligence ou la conscience obéissent à une axiomatique commune, mais de nouveaux axiomes sont nécessaires pour distinguer les "évidentes" différences factuelles. En effet c'est en partant de considérations théologiques élémentaires que j'en suis venu, dans 1.2, à postuler comme axiome modale de la théorie de la conscience la formule LWV. Elle capturait déjà, partiellement, des morceaux religieux ou mystiques fort divers. Ensuite l'axiome et sa communicabilité fut justifié de façon directe et explicite à partir de l'hypothèse du mécanisme indexical moyennant des expériences par la pensée.

Le mécanisme *digital* a permis de retrouver cette formule sans user de l'indexicalité. Celle-ci est en effet supprimée par la double diagonalisation. C'est ce que Gödel avait déjà utilisé dans la démonstration de l'incomplétude et c'est ce que Kleene a épinglé avec 2-REC. La consistance des machines autoréférentiellement correctes apparaît dès lors comme *une* parmi ces vertus, elle obéit à LWV :

$$\Box\Diamond \top \rightarrow \neg\Diamond \top,$$

et est même caractérisée par la forme plus contraignante de Löb, ce qui reste à exploiter :

$$\Diamond p \rightarrow \Diamond (p \,\&\, \Box \neg p).$$

Lao-tseu dit (en gros) "parlez du bien voici le mal", ou encore "définissez le beau voici le laid". Il n'a pas écrit "parlez du mal voici le bien", ni "définissez le laid voici le beau". Cette absence de symétrie[122] est elle capturée par G* ?

C'est le cas si on donne une définition très ouverte et modeste de chaque vertu. Le bien par la négation du mal. Le beau par la négation du laid, etc.

Alors la non-vertu est représentable par $\Box\bot$ et du coup, la vertu par $\Diamond\top$. Dans ce cas on a bien, au niveau G*

"communiquer le beau voilà le laid" :

$$G^* \vdash \Box\Diamond\top \rightarrow \neg\Diamond\top$$

mais on a aussi, par réflexion permise au niveau G* : "communiquer le laid, voila le laid" :

$$G^* \vdash \Box\neg\Diamond\top \rightarrow \neg\Diamond\top$$

ainsi que "communiquer le beau voilà le beau"

$$G^* \vdash \Box\Diamond\top \rightarrow \Diamond\top$$

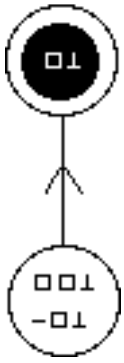
G, cependant, ne prouve aucune des 2 dernières formules.

$$G \nvdash \Box\neg\Diamond\top \rightarrow \neg\Diamond\top.$$

En effet $G \nvdash \Box\Box\bot \rightarrow \Box\bot$ , comme l'exemplifie le modèle de Kripke suivant :

---

[122] Attention : on ne confondra pas le début du 2 de Lao-Tseu avec la suite où il introduit un principe de relativité symétrique. Je me réfère aux traductions de Lao Tseu référées dans 1.2.




Et G⊬ □◊ ⊤->◊⊤ , ce serait inconsistant avec la formalisation du second théorème d'incomplétude de Gödel par PA, ou la *communication de LWV*, c-à-d :

$$G\vdash \Box\Diamond \top \to \neg\Diamond \top.$$

Ceci montre la prudence, la modestie et la consistance de Lao Tseu (relativement au mécanisme et à la notion de vertu approchée ici).

La vertu, de cette façon, est incompatible avec le *fanatisme*. L'autorité naturelle *est* ce qui n'use d'aucun *argument* d'autorité. Ici, on rejoint Locke 1667.

Le *piège de l'intelligence*, revient en gros à se croire *nécessairement* supérieur ou *nécessairement* inférieur (de façon communicable).

Une telle "erreur" annihile, ou moins temporairement, l'intuition de soi de l'autre ou de l'erreur.

Essayons, dans le même ordre d'idées, une théorie de l'intelligence dans un sens large et extrêmement ouvert. Disons donc que *x est intelligent* ssi *x n'est pas idiot*. Alors nous obtenons, avec LWV, deux formes d'*idiotie*.

D'une part l'idiotie des x qui communiquent que "x est intelligent".
D'autre part l'idiotie des x qui communiquent que "x est idiot"[123].

Avec les résultats sur l'inférence inductive théorique les x intelligents sont capables d'augmenter leurs compétences. Il faudrait dire plutôt : tant qu'elles sont intelligentes elles peuvent augmenter leurs compétences. Ces dernières pouvant, par des formules du style G*⊢ ◊ □◊ ⊤, *endormir* l'intelligence avec des inférences du genre "je ne vais plus faire d'erreurs" lesquelles inférences se trouvent alors (dans l'état d'être) imposées ou, en tout cas de moins en moins interrogées ou mises en doutes. L'intelligence

[123] Dont on peut dériver la théorie plus générale : x est idiot si x communique que y est intelligent ou que y est idiot (à l'exception du cas où y est prouvablement équivalent à une école pythagoricienne *fixée* du dedans, c-à-d correspond à une partie récursivement énumérable du graal. Cela va sans dire dans la mesure où le sujet ici est au moins universel (et capable de prouver sa propre $\Sigma_1$-complétude)).

est nécessaire pour le développement de la compétence, mais la compétence peut endormir l'intelligence.

Cependant la *possibilité d'être* localement inconsistant ($G^* \vdash \Diamond \Box \bot$) peut être elle-même un bénéfice dans le long terme, vis-à-vis de la compétence comme il a été dit à la section précédente. Il s'agit là, sans doute, d'une des source logique du rêve (voir 3.1).

Cette théorie axiomatique est encore corroborée avec le résultat sur la *séparation forte* (de Case, Chen et Jain, voir 2.3.5), résultat selon lequelle il est possible de construire une machine M capable de construire, ou de se transformer en une machine M' plus intelligente (que M). Et nous avons vu, avec Chen, Case et Smith, que dans ce cas, nécessairement ni M, ni M' ne sont capables de prouver que M' est plus intelligente que M.

D'autre relation devrait pouvoir être déduite entre l'intelligence au sens de Binet et la théorie générale de la *vertu* que je propose d'extraire du mécanisme.

Pour revenir à Protagoras, notons enfin que le mécanisme n'identifie pas les différentes vertus, comme il n'identifie pas les différentes formes d'intelligences ou de consciences possibles. Toutes sont exhibables mais aucunes ne sont communicables ou formellement identifiables. Cependant toutes les vertus obéissent à un lot de principes communs dont la communicabilité est d'être nécessairement révisable ou interrogative.

Dans le monde des machines autoréférentiellement correctes, les vertus se prêchent par l'exemple et se détériorent dans les affirmations.

Les "vertus" qui nous entourent ont sans doute une histoire profonde aux qualités diversement enrichies. L'incomplétude absolue et nécessaire permet cette diversité qui entraîne vraisemblablement, comme dans le cas de l'intelligence, de nombreuses incomparabilités. L'origine et les raisons de la profondeur de l'histoire est abordée, *in fine*, dans 3.3.

La notion de *vertu*, telle que je l'aborde ici et telle que Protagoras en parle est platonicienne. En réinterprétant l'*UN* des néoplatonistes par la classe des machines universelles (ou la notion d'ensemble créatif de Post), et l'âme par une machine universelle singulière, qui est (relativement) autoréfé-(et infé)-rentiellement correcte, le mécanisme permet une relecture des néoplatoniciens[124] où en particulier, le "moi" est un universel singulier, tout comme un ordinateur, ou un univers. La caractéristique de ce "moi" est

[124] Voir Trouillard 1972, O'Meara 1989, 1992.

l'immensité absolue (avec la thèse de Church) de ses déploiements, possibles ou même actuels[125] : ce sur quoi je reviens dans 3.2 et 3.3.

De même retrouve-on peut-être, par le mécanisme, la distinction entre la foi (de type G*) et la raison communicable (de type G) qui rappelle la distinction d'Averroes entre la cohérence philosophique et la "vérité" de la foi.

Abordons brièvement la question socratique par excellence : ce *moi* est-il (absolument) immortel ? Ici le mécanisme ne peut qu'exhiber partiellement la profonde complexité de la question. Avec MDI, la réponse, au niveau G*, est *plutôt oui*, mais on a vu qu'à ce niveau, il s'agissait plus d'une question, justement, que d'une réponse. Cela illustre le fait que MDI est solution de LWV.

Du point de vue positivement communicable, l'auto-arrêt ultime et absolu des machines universelles est absolument indécidable par Lucas+Kalmar.

L'interprétation mécaniste d'Everett de la mécanique quantique illustre concrètement la non-trivialité de la question (voir 1.2, 1.3, et j'y reviens en 3.3 ou un engagement ontologique permet d'interpréter la fermeture pour la possibilification de G* comme immortalité potentiel ou même actuel[126]).

La prise en compte nécessaire, avec l'hypothèse mécaniste, de l'acte de foi (indexical) fait du mécanisme, non pas une religion, mais un champ de religions possibles, aussi nombreuses que diverses appelant des pratiques (pré)funéraires (théotechnologiques) fort différentes et toujours globalement vaine *vue de G**.

L'aspect *néoplatoniste* sera plus frappant après la section 3.2 et 3.3 où l'on reconnaît l'*immatérialisme* du mécanisme moniste.

***Pont entre des livres***.

La présente approche permet d'expliciter des relations entre les deux ouvrages de Smullyan : *Tao is silent* et (par exemple) *Forever Undecided*. Le premier collecte des réflexions informelles sur le *tao*, le second est une introduction récréative à la logique G. Le rôle du théorème de Löb (l'axiome de G) est plus développé que je ne l'ai fait. La notion de machine (ou raisonneur) *modeste* peut éclaircir de nombreux points des applications de G à la morale. Dans une optique différente, voir aussi Smiley 1963.

Le phénomène de Henkin-Löb, je veux dire la nature prouvable des propositions qui affirment leur propre prouvabilité, est sans aucun doute

---

[125] C'est le paradoxe du graphe filmé, en 3.2, qui actualise le déploiement et nous force (avec MDI) à admettre une forme particulière de réalisme modale. Cela montre, que relativement à l'hypothèse mécaniste une part du "transcendantal" est (G) communicable.

[126] Voir la note en bas de page précédente. La possibilification est la règle p=>◊p, on a bien : G* ⊢ x entraîne G* ⊢ ◊x. C'est la duale de la nécessitation.

riche d'enseignement pour la philosophie mécaniste de l'esprit, mais elle sort du cadre présent.

Par ailleurs le travail présenté ici relie *Mind's I* (édité par Dennett et Hofstadter 1981) qui aborde l'autoduplication et *Gödel Escher Bach* (Hofstadter 1979) qui aborde la question du rôle du théorème de Gödel en intelligence artificielle, ainsi que *Mechanism, Mentalism & Metamathematics***.** de Webb 1980, qui motive considérablement le second théorème de récursion de Kleene (2-REC) pour la philosophie (mécaniste) de l'esprit.

*15°) La découverte du mathématicien par le mathématicien*

Si on pousse la logique classique jusqu'au bout en reconnaissant l'incomplétude et son caractère absolu ou intuitif (comme la thèse de Church nous y invite), alors on retrouve le sujet avec sa subjectivité et sa conscience temporelle, avec son intuition incommunicable mais dont les constructions sont exhibables.

La métamathématique a donc fait office d'introspection mathématique de la part de mathématiciens et l'*éblouissement intérieur* est la découverte communicable (avec MDI !) de son incapacité à circonscrire complètement le champ des comportements et des discours possibles des machines.

L'incomplétude Gödelienne est une forme d'illumination, de satori *arithmétique*.

Cette incapacité personnelle est désindexicalisable et machine-inférable. De même, grâce aux thèses de Church, Post et Turing, la découverte de l'impossibilité de se soustraire à l'informel nous a conduit à redécouvrir une formalisation, nécessairement non arithmétisable et nécessairement incomplète, de cette connaissance informelle laquelle reste (au niveau adéquat de MDI) trivialement (dans l'interprétation arithmétique : $\top \vee \Diamond \top$) en contact avec la vérité : le pont entre l'autoréférence correcte et cette intuition sans langage est concrétisé par la composition de la translation de Gödel, étendue par Grzegorczyk, de l'intuitionisme de Brouwer (tel que Heyting *a osé* le formaliser) avec les morphismes de Boolos, Goldblatt et Kuznetsov & Muravitskii.

Ainsi le mécanisme rend plausible l'interprétation de l'incomplétude comme un reflet du découvrement (nécessairement partiel) du mathématicien par le mathématicien[127]. Le mathématicien est au moins une machine universelle de Turing (capable de se convaincre de sa propre $\Sigma_1$-complétude). Avec MDI, il est au plus une telle machine.

[127] Notons qu'avec la théorie des catégories d'Eilenberg et MacLane (voir Mac Lane 1971) ce sont les relations interdisciplinaires entre mathématiques (Algèbre, Topologie, Géométrie et Logique), qui se plongent naturellement dans les mathématiques. Prendre l'interdisciplinarité au sérieux consiste à permettre à de nouvelles spécialités *interdisciplinaires* de se développer.

De même l'usage de la translation simple, en mécanisme indexical, a mis en évidence, pour celui qui commet l'expérience par la pensée, une proposition intuitivement et absolument, indécidable. L'expérience est communicable et l'indécidabilité de la proposition "je survivrai" (ou l'indéterminisme de la proposition "je survivrai à Washington ou à Moscou" dans le cas de la duplication) est communicable faisant ainsi de l'hypothèse mécaniste un acte de foi pour les éventuelles applications (théotechnologiques) concrètes.

Pour le solipsiste, cet acte de foi est difficile car il s'agit de se nier soi-même ou de se reconnaître dans un autre.

Les auteurs comme Girard 1987 ou Dummett 1963 qui argumentent en faveur de quelques logiques alternatives, ou pour quelques reconstructions du programme de Hilbert invoquent le théorème de Gödel comme motivation objective. L'hypothèse mécaniste force d'aller plus loin dans cette direction en défendant la possibilité d'extraire explicitement les logiques faibles internes à partir d'un approfondissement des phénomènes d'incomplétude. Et dans ce travail *G&G**, accompagnés des stratagèmes, sert d'embryon suggestif[128].

Le mécanisme permet de réconcilier Cantor et Kronecker : Dieu a créé les nombres naturels, tout le reste sont des constructions de nombres naturels "universels" relativement à des environnements (eux-mêmes) universels. Les grand cardinaux ou plus simplement le paradis de Cantor deviennent des étalons pour une géométrie, de l'ignorance.

C'est bien parce qu'il n'existe *nécessairement* pas de machines capables de répondre à toutes les questions que les questions possibles, énonçables par des machines qui s'observent, se structurent d'une façon non arbitraire.

*16°) Perspectives sémantiques*

A quoi pourrait ressembler une sémantique de "je" ? Aussi bien le mécanisme indexical que le digital permet d'inférer l'absolue non-formalisabilité complète d'une telle sémantique ainsi que l'impossibilité de se reconnaître dans une capture arithmétique d'un tel formalisme.

La thèse de Post-Turing, qu'on peut écrire sous la forme $\Box\exists i(\Box=i)$, et le stratagème conduisent à la description épistémique, S4Grz, décrivant un sujet interne naturel d'une machine autoréférentiellement correcte, et ne sachant, effectivement, se reconnaître pour telle : $\neg\exists i\Box(\Box=i)$.

La sémantique de Kripke de S4Grz, c-à-d les référentiels réflexifs, transitifs et antisymétriques, m'ont déjà permis de suggérer une interprétation

[128] Notons que cette découverte classique et objective du sujet intuitionniste est aussi illustrée par les interprétations classiques de l'intuitionisme dans la mathématisation de l'univers de discours du mathématicien par la théorie des topos et par d'autres constructions catégorielles (Lawvere, Grothendieck). J'y reviens un peu plus bas.

temporelle du sujet conscient, de sa part solipsiste (intuitioniste) et d'une façon générale de la nature mouvante et dynamique de ce dedans.

On aimerait cependant développer une sémantique de l'arithmétique épistémique basée sur le mécanisme, c'est-à-dire entre autres sur l'auto-référence et le stratagème : AREA. Comment ? Une idée est d'utiliser une sémantique algébrique. Les modalités admettent des interprétations naturelles en terme d'endotransformation du treillis des propositions prouvablement équivalentes. Ils existent alors des techniques permettant d'étendre cette sémantique avec des quantificateurs et d'en faire une sémantique pour le système AREA. Ici deux chemins algébriques peuvent être exploités. Les deux chemins partent chacun de la sémantique algébrique traditionnelle de la logique classique, c-à-d l'algèbre de Boole. Il semble qu'on puisse alors

-construire une sémantique pour S4Grz (ARIL) en partant d'une algèbre diagonalisable. Celles-ci constituent la sémantique algébrique de G (Magari 1975).

-utiliser le lemme de Funayama (voir Birckhoff 1940) et plonger les algèbres de Heyting dans des algèbres de Boole et poursuivre le travail de Flagg 1985 en utilisant des préordres antisymétriques.

La dualité eval-quote ou modèle-théorie, ou encore sa version généralisée de la "situation adjointe" local / global a ainsi permis à Flagg 1985 (sur base de Hyland Johnstone et Pitts, 1980) d'interpréter ⊡ par la comonade associée à cette adjonction et ◇ par sa monade).

L'avantage de cette dernière approche est qu'elle devrait permettre d'utiliser l'abondante information provenant des approches catégorielles de la sémantique, et notamment, en s'inspirant de Flagg (et Johnstone, Hyland et Pitts) on pourrait aboutir à une notion de réalisabilité basée sur l'auto-référence et le stratagème. On pourrait alors parvenir à isoler un *topos solipsiste* dont la logique intuitioniste interne correspondrait à celle reposant sur l'autoréférence et le stratagème (ou la thèse d'Artemov).

Le topos sous-jacent à l'approche de Flagg, Johnstone et Pitts est le topos effectif. Dans celui-ci le principe de Markov est vérifié (voir par exemple Phoa 1992). Le travail d'Artemov suggère que ce n'est pas le cas pour le "topos solipsiste" (Artemov 1990).

Le niveau où le mécanisme est correct correspond au niveau où l'identité personnelle (type quote) définit naturellement une notion de continuité psychologique. Pour définir la sémantique à la façon de Flagg, on est obligé d'avoir l'antisymétrie. Comme le fait remarquer Flagg ceci ne peut pas être fait de façon effective, mais cela correspond peut-être au caractère non effectif du choix du niveau où la description fonctionnelle est réalisable.

Ce travail de sémantique est peut-être prématuré pour le philosophe de l'esprit. Beaucoup d'outils sémantiques existent sur le "marché", mais avant

de savoir lesquels utiliser, une réflexion de nature informelle sur le rôle du sujet en inférence inductive doit être approfondie.

Le topos effectif peut être utilisé comme cadre naturel (univers intérieur) des mathématiques constructives avec thèse de Church (l'école constructiviste russe, voir Richman 1983). Bien que ce point de vue déborde sur le dehors, on y pense à la façon du dedans.

Pour capturer les formes possibles du dehors il faudrait parvenir à une version algébrique la plus générale possible des théorèmes d'incomplétude. Selon Di Paola et Heller, ni les algèbres polyadiques de Halmos, ni les logiques modales étendant G (avec quantificateurs), ni les algèbres diagonalisables n'y parviennent, et de nombreux beaux résultats (existence de points fixes, complétude sémantique, complétude arithmétique de Solovay, etc.) ne sont plus valables au cours des généralisations. Mais surtout ces dernières sont formellement éloignées de la théorie de la récursion. La raison pour laquelle je me suis borné à utiliser seulement les résultats les plus élémentaires de la théorie de la récursion, c'est que pour le développement futur de l'approche il faudra quitter le *cocon de la théorie élémentaire* en introduisant de façon plus appropriée les oracles, les fonctionnelles finitypes de Gödel, etc.

Les propriétés élémentaires de la théorie de la récursion sont conservées dans les généralisations, lesquelles permettent une description axiomatique. Je n'ai effectivement pas utilisé plus que l'existence d'une machine universelle et le théorème de paramétrisation. Ces deux résultats ne changent pas si on remplace les $\{\phi_i\}$ par les $\{\phi_i^A\}$, c'est-à-dire en rajoutant l'usage d'un oracle A. Le travail de Di Paola et Heller, 1987 sur les catégories dominicales et les catégories de récursion[129], est prometteur parce qu'il donne ce qui semble être une approche générale et intrinsèque pour les usages les plus variés de la récursion. Son défaut est qu'il nécessite, outre une connaissance approfondie de la récursion (ce qui n'est pas difficile car c'est une théorie qui *part de zéro*, voir aussi Soare 1980, Odifreddi 1989), mais elle nécessite surtout une connaissance de la théorie des catégories qui demande une familiarité avec les raisonnements algébriques, géométriques, topologiques. Pour les théories généralisées de la récursion voir aussi Fitting

[129] Les catégories dominicales sont des monoïdes un peu bizarre. Elles sont munies d'un pseudo-produit ("near product", reflétant le caractère partiel des flèches du dehors). Les catégories de récursion sont des catégories dominicales munies d'un morphisme de Turing. Un tel morphisme, à ma connaissance, capture assez bien le caractère universel de la machine universelle de Turing ainsi que le résultat de paramétrisation. La machine universelle ne doit plus être fixe, et les modèles de catégories dominicales peuvent transporter l'incomplétude dans des domaines a priori éloignés des nombres naturels classiques standards. Les hiérarchies de Grzegorczyk $E_n$ par exemple, qui sont des hiérarchies typiques d'écoles du dedans, sont capturées par des hiérarchies de catégories dominicales. Chaque $E_{n+1}$ possède un morphisme de Turing pour $E_n$. Aucun $E_n$ ne possède de morphisme de Turing à la différence des catégories de récursion. Celles-ci capturent bien l'aspect "fermeture pour la diagonalisation", qui nous permet de plonger intégralement l'objet dans le sujet. Je reviens sur ce plongement (en termes informels) dans 3.3.

1981, Fenstad 1980, Sacks 1990, Di Paola et Heller 1987, mais encore Lawvere 1969.

Pour respecter l'approche présente, les diverses modalités de l'esprit doivent être justifiées au moyen de G&G*[130] et/ou de (théories axiomatiques) de la récursion. Les mêmes remarques s'appliqueront au problème du corps et de l'esprit, comme je le laisse entr'apercevoir à présent. J'en dirai un peu plus dans 3.3.

*17°) Probabilité et logique modale*

Les théories G, G*, S4Grz vérifient chacune l'axiome de transitivité 4. G et S4Grz sont en outre chacune fermée pour la règle de nécessitation.

Les relations d'accessibilité des modèles de Kripke de G et de S4Grz , mais aussi de G* avec la sémantique limite de Boolos, sont transitives. Ces théories concernent les discours des machines dans les voisinages de l'infini. Et ces discours, décrits essentiellement par G, portent, eux aussi, sur le communic**able**, le connaiss**able**, l'incommunic**able**, c-à-d sur les modalités idéales telles qu'elles se stabilisent dans les voisinages de l'infini. Rien n'interdit cependant d'imaginer l'usage de G et de G* pour étudier les voisinages de zéro. Dans ce cas on étudie les discours limites des machines au sujet des voisinages de zéro, incluant les expériences appartenant à ces voisinages. Il suffit de trouver une interprétation arithmétique décente des notions d'immédiateté, de probabilité, etc.

Pour le calcul de la probabilité de survie en un coup (usant de l'accessibilité directe) il faut passer de la logique du prouvable à la logique de la preuve, de la logique du communicable à la logique de la communication, de la logique de l'accessible à la logique de l'accès, de la logique de la duplicabilité à la logique de la duplication. Il vaudrait peut-être mieux parler de *probance* que de probabilité, la probance désignant, alors, une croyance rationnelle d'arriver dans un certain état *en un coup*. Elle se limite à des expériences élémentaires.

Pour capturer cette probance, le stratagème *affaibli* s'impose pour les raisons suivantes :

1) Il s'agit d'une croyance, on ne peut pas avoir la réflexion T, le stratagème fort est trop fort pour les croyances.

130 Un peu à la façon d'Anderson (voir Smiley 1963) qui veut définir les modalités déontiques à partir du système S4. La raison invoquée est l'existence du consensus des philosophes sur le rôle et la clarté de S4. Ici nous voulons définir les modalités à partir de G. Ma raison n'est pas le fait que G soit la seule modalité qui admet un sens prédicatif mathématiquement clair (la prouvabilité formelle), mais bien l'importance accordée à l'autoréférence par des machines. On peut retrouver un point de vue Andersonnien avec le mécanisme en remplaçant S4 par S4Grz, ce dernier étant justifié à partir de l'autoréférence par le résultat de Boolos, Goldblatt ainsi que Kusnetzov et Muravitskii (voir 2.3.4).

2) L'affaiblissement[131] de □p&p en □p&◊p (voir la transformation DEON en 2.3.4), supprime, on l'a vu, la fermeture pour la nécessitation, ainsi que la fermeture pour la monotonie rationnelle au niveau *, et partout la "transitivité", c-à-d la formule 4. Ce qui est raisonnable pour une notion d'état *immédiatement accessible*.

3) La dernière raison, est celle analysée en 1.3 où j'invoque le *darwinisme arithmétique*. La probance nécessite une notion d'expérience élémentaire, et l'expérience élémentaire typique, avec MDI, est celle de l'annihilation-reconstitution (voir 1.3). Comme la mort (absolue) constitue un dernier état, survivre à une expérience signifie aboutir (dans tous les cas de figure) à des états *vivants*, ou *conscients*, ou *consistants*, etc., bref des états dont on peut s'échapper. Pour se limiter dans les états vivants, on *attache* la *consistance* à la preuve. Les probabilités sont des probabilités de vivre telles ou telles expériences, calculées sur des états où on a survécu. L'incommunicable $P_{trans}=1$, de l'expérience AR, qui ironiquement inférable (avec MDI) est capturé alors par

$$KD?^* \vdash \Box\Diamond \top$$

c-à-d $G^* \vdash \Box\Diamond\top$, c-à-d encore

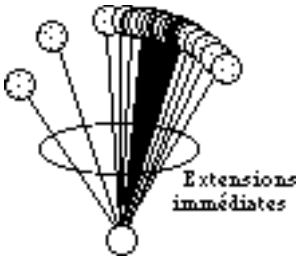
$$G^* \vdash \Box(\Box\top \& \Diamond\top) \vee \Diamond(\Box\top \vee \Diamond\top)$$

Une formule qui n'est pas un théorème de G, ce qui rend compte de l'ironie de la situation pour celui qui (avec MDI) a survécu à l'expérience.

Davantage de motivations sont présentées dans 3.3. Avec le Paradoxe du Dovetelleur Universelle (PDU), je montrerai que le caractère élémentaire des expériences "physiques" est capturé par la restriction aux formules $\Sigma_1$.

Le problème du corps et de l'esprit se ramène alors à l'isolation d'une mesure *convenable* sur les extensions immédiates, ou voisines.

Avec Scott-Montague :




[131] Qu'il s'agisse d'un affaiblissement se voit au niveau G*. $G^* \vdash p\to\Diamond p$ et $G^* \nvdash \Diamond p\to p$.

○ représente un monde (état), et ⁘ un ensemble de mondes (d'états), ou une proposition. Pour KD? l'ensemble des propositions nécessaires forme un quasi-filtre. Pour KD?* une famille de telles structures est sans doute nécessaire.

*Remarque* Une interprétation de l'accessibilité en terme d'inférence poppérienne en un coup (ou en un nombre fini borné de coups) comme dans Daley *et al.* 1992, pourrait être intéressante à exploiter dans ce contexte.

*18°) Informatique du dehors et informatique du dedans*

Le tableau suivant récapitule, *cum grano salis*, les points de vue des *dedans* (constructivisme avec ou sans thèse de Church), et le *dehors* (Platonisme avec thèse de Church). C'est l'aspect complémentaire de ces deux sortes de *points de vue* qui fait leur intérêt.

Il serait vain de penser qu'un de ces points de vue soit supérieur à l'autre. Exactement comme la double négation de Gödel permet de traduire la logique propositionnelle classique en logique intuitioniste, et inversément avec l'extension S4 de Gödel 1933, il est toujours possible d'approximer le dehors par le dedans, et le dedans par le dehors. Le mécanisme force la reconnaissance d'un *minimum* de dedans ***et*** de dehors.

| **LES DEDANS** | **LE DEHORS** |
|---|---|
| Algèbre de Heyting, locale; | Algèbre de Boole, mais aussi quantale (voir 3.3); |
| Informatique Conventionnelle; | Intelligence Artificielle; |
| Ecole**s** du dedans et logiques faibles alternatives; | Ecole du dehors et logiques modales classiques; |
| Extensionnel, dénotationnel; | Intensionnel, connotationnel; |
| Total calculable, prouvable; | Partiel calculable,. non prouvable; |
| Contrôlable; | Incontrôlable; |
| Calcul lambda typé; | Calcul lambda non typé; |
| Catégorie cartésienne close; | Idem, mais sans objet, terminal : C-monoïde (Lambek & Scott 1986), Catégorie dominicale avec morphisme de Turing (Di Paola, Heller 1987); |
| **1$^{er}$** théorème de récursion; | **2$^{em}$** théorème de récursion; |
| Sc**s** des progr. qui s'arrêtent; | Sc des progr. qui s'arrêtent, ou ne s'arrêtent pas; |
| Extension de soi ("néocortex"); | Détachement de soi (autonomie); |
| Math constructives; | Math classique; |
| Intuitif, doxastiques; | Contre-intuititif, paradoxastiques; |
| Soi; | Autres (Soi); |
| Intuitionismes; | Platonismes; |
| Connaissance de soi; | Reconnaissance de l'autre; |
| Subcréatif ? | Créatif (Post); |

| | |
|---|---|
| Ingénieur; | Théologien; |
| Infini(s) potentiel(s); | Infinis actuels; |
| Continu; | Existence de sauts, de catastrophes; |
| Action; | Contemplation; rêves |
| Aristote; | Aristote, Platon; |
| Héraclite, St-Augustin, Bergson, Dogen, Brouwer; | Plotin, Proclus, Cantor, Hilbert, Einstein, Gödel; |
| Non verbal; | Verbal (partiellement); |
| Math = partie exacte de la pensée; | Existence de mathématiques. expérimentales, incertaines; |
| Preuves bornables; | Preuves non bornées; |
| Déduction; | Inférence, induction + expérimentation; |
| Local (Bell 1986); | Global ? |
| Sécurité (contrôle); | Liberté, détachement, + risque; |
| Topos solipsiste; | Topos Booléen, C*algèbre ? (pour 3.3); |
| On ne pose que les questions que l'on sait résoudre | On apprend à faire face aux problèmes insolubles; |
| Invention, construction; | Découverte. |
| Informel, absolu, intuitif, personnel | Relatif, Universel |
| Intérieur, Première Personne; | Extérieur, Troisième Personne; |
| IL, IL * | G, G*, S4Grz, S4Grz*, KD?, KD?* |

*Commentaires*

Les écoles du dedans et du dehors sont indétachables, car, à quoi sert un outil ou une construction si ce n'est à réaliser un rêve et surtout à tenter de le communiquer ou à l'exhiber pour quelques autres ?

Le mécanisme permet de réconcilier Brouwer et Hilbert, ou plus profondément, Cantor et Kroneker. Il encourage la philosophie classique à découvrir et mettre en évidence *les multiples splendeurs*[132] de ses "rivales" intuitionistes (Marchal 1988). Elle la redécouvre véritablement en son sein.

C'est la découverte du mathématicien solipsiste et constructeur, par le mathématicien classique, c-à-d le "mathématicien" ayant l'imagination suffisante que pour concevoir l'existence indépendante d'un *autre* soi ou d'une *autre* incarnation de la conscience, et assez optimiste pour croire qu'il puisse lui communiquer quelques propositions non conventionnelles sur base d'une intuition supposée commune[133].

L'intuitionisme (et les autres écoles du dedans assez riches pour postuler une thèse de Church ou une thèse équivalente) suggère une description du sujet en tant qu'agent épistémique et créateur.

---

[132] J'emprunte l'expression à Boileau 1973, "les multiples splendeurs du forcing".

[133] Je dois admettre que cette découverte ne nécessite pas le mécanisme. Le développement de la théorie des topos exhibent un processus similaire. Le mécanisme rend seulement de tels processus obligatoires et impose des contraintes non triviales basées sur le rôle de l'auto-référence.

Le simple fait que Brouwer ait finalement été tenté de communiquer son expérience de mathématicien constructeur et libre montre qu'il ne partageait pas sérieusement la *doctrine* provocante du solipsisme. Il espérait, lors de moment de découragement il est vrai, que son oeuvre puisse servir à l'étude de la pensée.

Brouwer, à l'instar de Bergson ou même de Shopenhauer, semble avoir réalisé la présence d'un inconnu-infini *en soi* incommunicable, qu'il aurait pris pour la source de la conscience. Hilbert a perdu l'univocité et la complétude de la reconstruction formelle des mathématiques au niveau de la prouvabilité. Mais celles-ci sont (G)-communicables et les individus, s'ils désirent *rester* libres, peuvent commettre des paris, et choisir des axiomes, quitte à attendre leurs fruits déductifs ou les tester expérimentalement. Au niveau de la calculabilité, l'univocité réapparait avec la thèse de Church.

De même, à travers le *sacrilège de Heyting* (la formalisation de la logique intuitioniste), l'intuitioniste perd également l'univocité implicite du dedans, et la retrouve aussi avec la thèse de Church intuitioniste.

Cantor-Hilbert et Kronecker-Brouwer peuvent se réconcilier à travers le mécanisme, mais le *paradis* de Cantor (avec les logiques de tous ordres et les différentes conceptions du continu) s'interprète en une sorte de géométries des rêves, de champs des possibles. Avec la thèse de Church, les *rêves* admettent une géométrie non triviale et nécessairement ambiguë, mais cela n'enlève rien à leurs intérêts : il s'agit bien du statut de l'autre et de l'autre soi, ainsi que des relations entre le(s) dedans et le dehors.

Ces relations posent à leur tour de nouvelles questions abordées dans la troisième partie.

Pour les sciences empiriques Brouwer utilise une causalité *à la Hume 1739*, donc de nature a priori plutôt idéaliste et subjective.

Je montrerai en 3.3 comment rendre compatible un idéalisme objectif *concernant le subjectif* avec le réalisme platonicien. C'est que le subjectif, clairement spirituel, doit obéir, avec l'hypothèse mécaniste aux *lois de l'esprit* et que celles-ci sont à même d'imposer, grâce à leur indépendance par rapport au sujet (obtenue par désindexicalisation), des contraintes réalistes suffisamment fortes que pour faire apparaître une phénoménologie de la physique, dans les voisinages de zéro.

On sait déjà, avec MDI et les thèses de Church, que la frontière entre le dedans et le dehors est floue, vague, et constructivement ainsi, c'est-à-dire que plus on tente de la cerner et de la singulariser, elle se fuzzyfie et se multiplie.

Ceci nous donnera finalement *une troisième colonne* décrivant une logique de l'observable, celle-ci constituant le bord du dedans, ou au niveau *, le bord du dehors.

## *19°) Résumé de 2.3.6 et dernières remarques*

*L'hypothèse mécaniste en philosophie de l'esprit permet d'interpréter G, G*, S4Grz KD?, KD?*, et leurs extensions comme une théorie embryonnaire de l'esprit. G* décrit les lois de l'esprit, c-à-d la vérité concernant ce qu'une machine peut ou ne peut pas communiquer de façon positiviste et correcte, ainsi que le consistant, à son sujet, G décrit ce qu'elle sait effectivement communiquer de façon convaincante, et S4Grz, obtenu par application du stratagème de Théétète sur G, décrit platonistiquement la part solipsiste du sujet et de son intuition informelle. Les logiques KD?, obtenues par application du stratagème affaibli sur G et sur G*, décrivent une sorte de probabilité-croyance. Ces théories sont compatibles avec l'interprétation du théorème d'incomplétude de Gödel en philosophie mécaniste de l'esprit, sur laquelle elle repose en partie (Post, Turing, Webb, Reinhardt, Slezak). Ces théories donnent une vision du fondement des mathématiques, comme une découverte du sujet mathématicien, sous forme embryonnaire de machine universelle auto-référentiellement correctes, par le sujet mathématicien lui-même (ou par la collectivité).*

*La théorie et les lois décrites sont par ailleurs obtenues comme étant des discours inférés par les machines elles-mêmes et la collectivité auto-observante de machines auto-observantes et communicantes. On remarquera en outre que la présente approche prend sa racine dans une conception digitale de l'identité personnelle qui a permis les solutions récursives naturelles des problèmes biologiques de Descartes et Driesch.*

*Remarquons que cette théorie de l'esprit pourrait être correcte sans que l'hypothèse mécaniste ne soit vraie. Apprécier cette théorie et rester neutre vis-à-vis du mécanisme, ou même être antimécaniste tout en restant consistant est possible. Mais si on admet que cette littérature des "sages" (Lao-tseu, Wittgenstein, Watts, Valadier) confirme G, cette littérature confirme abductivement le mécanisme. Cette littérature elle-même est inspirée en partie par les catastrophes sociales ou individuelles lorsque des propositions du style* $\Box \Diamond \top$, *sont prises au sérieux.*

*Si l'hypothèse mécaniste indexicale est vraie, elle est de type "incroyable mais vrai", comme ,* $\Diamond\top$ *ou* $\Diamond\Diamond\top$. *On a*

$$\Box \text{ MEC-IND} \rightarrow \neg \text{ MEC-IND}$$

*Les machines ne peuvent pas savoir qu'elles sont des machines, elles peuvent, à leurs risques et périls, l'inférer et commettre un acte de foi. (voir aussi Kelly*[134] *qui retrouve cette proposition d'une façon différente et directement à partir de l'inférence inductive). De même :*

$$\Box \text{ MEC-FORT} \rightarrow \neg \text{ MEC-IND}$$

*Un corollaire immédiat (Kennes 1990*[135]*) laisse une porte ouverte pour réfuter le mécanisme indexical. Si quelque humain parvenait à donner une communication convaincante (qui vous convainc) qu'une machine peut penser, dans le sens d'avoir une expérience privée, alors il vous aura démontré que vous n'êtes pas une machine.*

## *Biblio locale*

---

134 Kelly 1993 utilise la théorie descriptive des ensembles. Pour une introduction avec des motivations provenant de la théorie de la récursion, voir Manfield et Weitkamp 1985.

135 Communication personnelle.

**ARTEMOV** S., **1990**, *Kolmogorov's Logic of Problems and a Provability Interpretation of Intuitionistic Logic,* in Parikh R., (Ed.), Proceedings of the Third Conference on Theoretical Aspect of Reasoning about Knowledge (TARK 90), Morgan Kaufmann Publishers.

**BEESON** M., **1985**, *Foundations of Constructive Mathematics*, Metamathematical Studies, Springer-Verlag, Berlin.

**BELL** J.L., **1986**, *From Absolute to Local Mathematics*. Synthese 69, pp. 409-426.

**BELLIN** G., **1985**, *A system of natural deduction for GL*, Theoria, Vol LI, pp. 89-114.

**BEKLEMISHEV** L. D., **1991**, *Provability Logics for Natural Turing Progressions of Arithmetical Theories*, Studia Logica L, 1, pp. 108-128.

**BERARDUCCI** A., **1990**, *The Interpretability Logic of Peano Arithmetic*, The journal of symbolic Logic, Vol 55, N° 3.

**BIRKHOFF** G. **1940**, *Lattice Theory*, American Mathematical Society Colloquium Publications, Troisième Edition 1993.

**BOILEAU** A., **1973**, *Les Multiples Splendeurs du Forcing*, mémoire présenté à la faculté des études supérieures en vue de l'obtension de maîtrise ès Sciences (Mathématiques), Montréal.

**BOOLOS** G., **1979**, *The Unprovability of Concistency. An Essay in Modal Logic*. Cambridge University Press.

**BOOLOS** G., **1980**a, *Provability, Truth, and Modal Logic*, Journal of Philosophical Logic, 9, pp. 1-7.

**BOOLOS** G. and **SAMBIN** G., **1991**, *Provability: the Emergence of a Mathematical Modality*, Studia Logica L, 1, pp. 1-23.

**BOWEN** K. A. and **DE JONGH** D. H. J., **1986**, *Some Complete Logics for Branched Time*, part I, Well-Founded Time, Forward Looking Operators, ITLI Prepublication Series 86-05, University of Amsterdam.

**BRIDGES** D. and **RICHMAN** F., **1987**, *Varieties of Constructive Mathematics*, Cambridge University Press.

**CARLSON** T., **1986**, *Modal Logics with Several Operators and Provability Interpretations*, Israël Journal of Mathematics, 54, pp. 14-24.

**CARSE** J. P., **1986**, *Finite and Infinite Games*, Macmillan Publishing Company, New York, Traduction française, Seuil 1988.

**CASE** J., **1971**, *Enumeration Reducibility and Partial Degrees*, Annals of Mathematical Logic, Vol. 2, N° 4, pp.419-439.

**COOPER** S.B., **1990**, *Enumeration Reducibility, Nondeterministic Computations and Relative Computability of Partial Functions*, in K. Ambos-Spies, G. H. Müller & G. E. Sacks (Eds) *Recursion Theory Week*, Proceedings, Oberwolfach 1989, Lecture Notes in Mathematics n° 1432, Springer-Verlag.

**DALEY** R., **KALYANASUNDARAM** B., **VELAUTHAPILLAI** M., **1992**, *The Power of Probabilism in Popperian FINite Learning*, in Jantke 1992.

**DI PAOLA** R. A., and **HELLER** A., **1987**, *Dominical Categories: Recursion Theory without Elements*, The journal of Symbolic Logic, Vol. 52, n° 3, pp. 594-635.

**DENNETT** D.C. and **HOFSTADTER** D., **1981**, (composed and arranged by), ***Mind's I***, Basic Books, Inc., Publishers, New-York. (Paru en français sous le titre *vue de l'esprit*, interEditions, Paris, 1987).

**DOGEN**, **1232-1253**, *Shôbôgenzô*, Editions de la différence, Argenteuil, 1980.

**DUMMETT** M., **1963**, *The Philosophical Significance of Gödel's Theorem*, Ratio, vol. 5, pp. 140-155.

**FEFERMAN** S., **1960**, *Arithmetization of metamathematics in a General Setting*. Fundamenta Mathematical XLIX.

**FENSTAD** J. E., **1980**, *General Recursion Theory An Axiomatic Approach*, Springer-Verlag, Berlin.

**FITTING** M. C., **1981**, *Fundamentals of Generalized Recursion Theory*, North-Holland Publishing Company, Amsterdam.

**FLAGG** R., **1985**, *Church's Thesis is Consistent with Epistemic Arithmetic*, in Shapiro 1985.

**GIRARD** J-Y., **1987**, *Proof Theory and Logical Complexity*, Bibliopolis, Naples.

**GOOD** I.J, **1971**, *Freewill and Speed of Computation*. Brit. J. Phil. Sci. 22, 48-49.

**GOODSTEIN** R.L., **1963**, *The Significance of Incompleteness Theorems*, British Journal for the Philosophy of Science, vol. 14, pp. 208-220.

**GUILLEN** M, **1983**, *Gödel's Theorem An Article of Faith*, in Guillen M., Bridges to infinity, the Human Side of Mathematics, Rider, London, pp. 117-125.

**GRISWOLD** C. L., **1986**, *Self-Knowledge in Plato 's Phaedrus*, Yale University Press. New Haven and London.

**GRZEGORCZYK** A., **1964**, *A Philosophicaly Plausible Formal Interpretation of Intuitionistic Logic*, Indagationes Math. 26, pp. 596-601.

**GRZEGORCZYK** A., **1967**, *Some relational systems and the associated topological spaces*, Fundamenta Mathematicae, LX.

**HARTMANIS** J. *Relations between Diagonalisation, Proof Systems, and Complexity Gaps*, Computer Science Department, Cornell University, Ithaca, New York 14853, USA.

**HOFSTADTER** D., **1979**, *Gödel, Escher, Bach : an Eternal Golden Braid*, Basic Books, Inc., Publishers, New York.

**HYLAND** J. M. E., **JOHNSTONE** P. T. and **PITTS**, **1980**, *Tripos Theory*, Math. Proc. Camb. Phil. Soc. 88, pp. 205-232.

**HUMES** D., **1739**, *A Treatise of Human Nature*, London. also Fontana/Collins, Glasgow, 1987.

**KELLY** K., **1993**, *Learning Theory and Descriptive Set Theory*, Journal of Logic and Computation, Vol 3, n° 1, pp. 27-45.

**KREISEL**, G., **1970**, *Church's thesis : a kind of reducibility axiom for constructive mathematics*, in Kino, A., Myhill, J., and Vesley, R.E. (eds.), Intuitionism and Proof Theory, Proceedings of the Summer Conference at Buffalo, New York, 1968, pp 121-150, North-Holland, Amsterdam.

**LADRIERE**, J., **1957**, *Les limitations internes des formalismes*, E. Nauwelaerts, Louvain, et Gauthier-Villars, Paris.

**LAMBEK** J. & **SCOTT** P.J., **1986**, *Introduction to Higher Categorical Logic*, Cambridge University Press.

**LAVENDHOMME** R., **1984**, *Le régistre du réel et la mathématique*, Analytica 36, pp. 69-78.

**LAVENDHOMME** R., **1986**, *Mathèmes Lacaniens*, Psychoanalyse 4, pp. 97-106.

**LAVENDHOMME** R., **1987**, *Leçons de géométrie différentielle synthétique*, Institut de mathématique, CIACO, Louvain-la-Neuve.

**LAWVERE** F. W., **1969**, *Diagonal Arguments and Cartesian Closed Categories*, Category theory, Homology theory and their applications II, Springer LNM 92, pp. 134-145.

**LOCKE** J., **1667**, *Lettre sur la tolérance*, dans Lettre sur la tolérance et autres textes, trad. française par Jean Le Clerc, Flammarion, 1992, Paris.

**McCARTY** C. & **TENNANT** N., **1987**, *Skolem's paradox and constructivism* Journal of Philosophical Logic, 16, 165 - 202.

**MAGARI** R., **1975**, *Representation and Duality Theory for Diagonalizable Algebras*, Studia Logica XXXIV, 4, pp. 305-313.

**MANSFIELD** R. and **WEITKAMP** G., **1985**, *Recursive Aspects of Descriptive Set Theory*, Oxford University Press.

**MARCHAL** B., **1987**, *What could be like a Semantics for Self-referential Process*, compte rendu d'une conférence donnée à Louvain.

**MARCHAL** B., **1990**, *Des fondements théoriques pour l'intelligence artificielle et la philosophie de l'esprit*, Revue Internationale de Philosophie, 1, n° 172, pp 104-117.

**MARCHAL** B., **1992**, *Amoeba, Planaria, and Dreaming Machines*, in Bourgine & Varela (Eds), *Artificial Life, towards a practice of autonomous systems*, ECAL 91, MIT press.

**OATLEY** K., **1988**, *On changing one's mind: a possible function of consciousness*, in A. J. Marcel & E. Bisiach (Eds.).

**ODIFREDDI** P., **1989**, *Classical Recursion Theory*, North-Holland, Amsterdam.

**O'MEARA** D., **1989**, *Pythagoras Revived Mathematics and Philosophy in Late Antiquity*, Clarendon Paperbacks 1990.

**O'MEARA** D., **1992**, *Plotin une introduction aux Ennéades*, traduit de l'anglais par Anne Calet Molin, CERF, Editions universitaires de Fribourg 1992.

**PENROSE** R., **1988**, *On the Physics and Mathematics of Thought*, in Herken R. (ed), *The Universal Turing Machine A Half-Century* ***Survey***, Oxford University Press.

**PHOA** W., **1992**, *An Introduction to Fibrations, Topos Theory, the Effective Topos and Modest Sets*, LFCS Report Series, Edinburgh.

**RICHMAN** F., **1983**, *Church's Thesis without Tears*. The Journal of Symbolic Logic, 48, 3, 797 - 803.

**ROGERS** H.,**1967**, *Theory of Recursive Functions and Effective Computability*, McGraw-Hill, 1967. (2ed, MIT Press, Cambridge, Massachusetts 1987).

**SOTO-ANDRADE** J. and **VARELA** F., 1984, *Self-Reference and Fixed Points: A Discussion and Extension of Lawvere's Theorem*, Acta Applicandae Mathematicae 2, 1-19.

**SMILEY** T. J., **1963**, *The Logical Basis of Ethics*, Acta Philosophica Fennica, 16, pp. 237-246.

**SMULLYAN** R., **1987**, *Forever Undecided*, Alfred A. Knopf, New York.

**SMULLYAN** R., **1977**, *The Tao is Silent*, Harper and Row, New-York.

**TROELSTRA** A. S. and **VAN DALEN** D., **1988**, *Constructivism in Mathematics, An Introduction*, (2 volumes) North Holland.

**TROUILLARD** J., **1972**, *L'Un et l'âme selon Proclos*, Société d'édition "Les Belles Lettres", Paris.

**VICKERS** S., **1989**, *Topology via Logic*, Cambridge Tracts in Computer Science, Cambridge University Press.

**WANSING** H., **1993**, *The logic of Information Structures*, Lectures Notes in Artificial Intelligence 681, Springer Verlag.

**WEBB** J.C., **1980**, *Mechanism, Mentalism & Metamathematics*. An Essay on Finitism. D. Reidel Pub. Company.