RÉCAPITULATION EN 21 POINTS, 2 FORMULES ET 1 TABLEAU

RÉCAPITULATION EN 21 POINTS

1. LE MÉCANISME INDEXICAL

1) J'extrais une théorie axiomatique de la conscience à partir de la littérature. Le terme *conscience* est pris dans un sens très large (vie, vertu, sagesse, etc.).
J'isole alors un principe équationnel[1] W (pour Wittgenstein 1918) :

W $\qquad \neg \Box x$

ainsi qu'une version plus faible : le principe équationnel LWV (pour Lao Tseu "1967", Watts 1951, Valadier 1990) :

LWV $\qquad \Box x \rightarrow \neg x$

L'inconnue *x* représente une proposition du genre "*je suis conscient*", "*je suis vivant*", etc.

Le carré représente une forme de *justification*, par exemple une communication scientifique ou positiviste. Intuitivement une telle communication est censée imposé, garantir, rendre *nécessaire*, la proposition sur laquelle elle porte.

W et LWV peuvent se lire dualement $\Diamond \neg x$ (W), et $x \rightarrow \Diamond \neg x$ (LWV), le losange représente alors une sorte de possibilité, l'existence d'un chemin menant à la proposition contraire.

LWV est *paradoxal* ou *contre-intuitif* puisque la négation de x accompagne la garantie de x. On reconnaît *la loi des effets contrariés* de Watts 1951, l'*immoralisme de la morale objectivée* de Valadier 1990, ainsi que de nombreuses situations décrites par les philosophes taoistes illustrant comment *l'intention* est trahie lorsqu'on tente de la *fixer* par des règles ou des mots.

2) Je motive pour une "géométrie" de la vie et de la mort où essentiellement *être vivant* consiste à être dans un état transitoire (représenté par des ronds), et *être mourant* consiste à aboutir à un *dernier état* (représenté par des ronds pointés), c'est-à-dire un état duquel plus aucune expérience n'est permise.
Les flèches représentent des expériences (subjectives) possibles :

3) J'introduis l'hypothèse mécaniste indexicale et digitale (MDI) :

Indexicale = _Je_ suis une machine
Digitale = Je suis une machine _*digitale*_

[1] *équationnel* au sens large, il s'agit d'une inéquation. $\neg \Box x$ est équivalent à $\Box x \rightarrow \bot$. La relation d'ordre (d'inéquation) correspond à l'implication.

Cela signifie qu'il existe un niveau fini de description de *moi* tel que je survis à la substitution fonctionnelle d'une quelconque de mes parties, opérée à ce niveau.

Le second théorème de Kleene, allié à la thèse de Church, est utilisé plus loin pour donner un sens précis à une expression comme "description finie *de moi*" (voir aussi Webb 1980).

On admet ainsi un fonctionnalisme relativisé à un niveau n :

$$\text{MDI} \Leftrightarrow \exists n \text{FONC}(n)$$

Cette version du mécanisme est très large dans le sens qu'à part la condition de finitude de n, aucune restriction n'est placée. Il se peut, par exemple, qu'il faille dupliquer tout l'univers pour survivre (j'illustre cette idée avec l'interprétation d'Everett (1957) de la mécanique quantique).

Néanmoins MEC-IND est plus fort que MEC-FORT (des machines peuvent penser) qui est lui-même plus fort que MEC-BEH (des machines peuvent avoir un comportement d'être pensant).

Une réfutation, comme celle tentée par Lucas 1959, de MEC-BEH, restreint aux preuves (pas nécessairement formelles) en arithmétique, entraînerait donc une réfutation de MDI (voir aussi Wang 1974).

4) J'argumente, avec des expériences par la pensée, que MDI est équivalent à la survie à une expérience de *télétransportation* (je dis aussi *translation*), c'est-à-dire à une **A**nnihilation suivi d'une **R**econstitution (**AR**).

Le fait que l'on puisse toujours *a posteriori* ne pas annihiler *l'original* permet de communiquer *a priori* de façon positive l'impossibilité de garantir le fait que l'on survit, ou puisse survivre, à une telle expérience.

Si la flèche de la *géométrie de la vie et de la mort* est interprétée par l'expérience de translation *AR*, tous les états transitoires aboutissent à un dernier monde (état).

En regardant cette géométrie comme un modèle de Kripke on a, *avec la solution* $x = \Diamond\top$, que LWV est vrai dans tous les états et W est vrai dans tous les états transitoires. $\Diamond\top$ est équivalent à $\neg\Box\bot$, la consistance obéit à LWV, ainsi que $\Diamond p$ d'une façon générale. Notons que le faux vérifie trivialement LWV. Pour les entités consistantes, le faux vérifie W, et W lui-même est une solution de LWV.

Au niveau où le fonctionnalisme est correct, tout passage d'un état à un autre est équivalent à une AR. Les expériences par la pensée indique que LWV est communicable, à la différence de W.

Définitions

- la <u>théorie C</u> admet comme axiome les tautologies de la logique propositionnelle classique, l'axiome K de Kripke $\Box(p\text{->}q)\text{->}(\Box p\text{->}\Box q)$, et l'axiome correspondant au principe LWV avec $x = \Diamond p$: $\Box\Diamond p \text{->} \neg\Diamond p$. C admet comme règles le modus ponens (MP) et la nécessitation (Nec).

- un référentiel de Kripke est <u>réaliste</u> ssi tous les états transitoires aboutissent à un dernier état. Les expériences par la pensée, reposant sur MDI, impose le réalisme du référentiel pour la géométrie de la vie et de la mort.

On a la *soundness* et la *completeness* de la théorie C de la *conscience* , interprétée comme une théorie du *communicable* et de l'*incommunicable*, vis-à-vis des référentiels réalistes :

Théorème : C $\vdash$ A ssi A est respectée par tous les référentiels réalistes.

Remarque : le théorème reste vrai avec l'axiome $\Box\Diamond\top$->$\neg\Diamond\top$ à la place du schéma $\Box\Diamond p$->$\neg\Diamond p$.

5) J'identifie le sujet avec le connaisseur. J'accepte pour logique de la *connaissance* du sujet une logique modale KT. La réflexivité ($\boxdot p$->p) reflète l'incorrigibilité du sujet. De façon imagée celui-ci garde intact son cordon ombilical avec la vérité[2].

L'axiome 4, $\boxdot p$->$\boxdot\boxdot p$, qui rend transitif le modèle de Kripke, indique une forme de ressource illimitée (d'accessibilité indirecte). Il est responsable du passage entre la connaissance directe (ou l'immédiatement connaissable) et la *connaissabilité* et j'accepte donc S4 comme théorie de la connaiss**abilité** du sujet, celui-ci étant considéré à la première personne.

Une question se pose. La formule $\Diamond\top$, qui est équivalente à $\neg\Box\bot$, exprime la correction (la conformité, la justesse) de la communication[3]. C $\nvdash$ $\Box\bot$->$\bot$, puisque $\Box\bot$ est vraie dans les derniers mondes (états). Comment relié la communication (ou la communicabilité) positive de la machine (ou d'une duplication mécaniste de moi-même) avec la connaissance(-abilité), incorrigible, du sujet *que je suis* ? MDI, sous la forme ∃nBEH(n), peut s'écrire

$$\exists n(\boxdot p \text{ <-> } \Box_n p) \qquad (*)$$

en limitant éventuellement p à un ensemble de propositions d'un domaine particulier.

$\Box_n$ représente la machine qui me constitue au niveau n adéquat. Il s'agit de *moi* vu à la troisième personne comme après une expérience d'autoduplication sans annihilation. C'est un moi formel, identifiable par une *forme*, laquelle est capturable par un nombre naturel *avec MDI*.

Le problème est que $\boxdot$ est incorrigible ($\boxdot p$->p), ce qui contredit apparemment LWV avec l'utilisation de $\Box_n$, qui correspond à $\Box$ dans la théorie C. La question qui se pose est de trouver le lien (intensionnel) entre le *je intuitif* $\boxdot$ et le *il formel*, tel que (*) soit correcte, et qu'en même temps $\Box$ obéit à C(4) et $\boxdot$ obéit à KT(4).

6) Une solution se trouve déjà dans Platon. En effet une vieille idée proposée à Socrate par Théétète est de définir la connaissance $\boxdot p$ par $\Box p \& p$ c'est-à-dire par une justification de p accompagnée de la vérité de p. C'est la façon la plus simple de créer un cordon ombilical entre la justification et la vérité. L'application de *ce stratagème* sur la logique C donne la logique KT?

KT? désigne une extension de KT. Cela donne un embryon de théorie de la connaissance. De même l'application du stratagème sur C4 donne un embryon d'une théorie de la connaissabilité axiomatisée par une extension de S4.

Plus de détails apparaîtront lorsque, avec la thèse de Church et la théorie de la prouvabilité formelle, la théorie C(4) est remplacée par la théorie Löbienne de l'autoréférence G.

---

[2] J'utilise "$\boxdot$" pour des logiques modales non spécifiques mais vérifiant au moins T. De même j'utilise "$\Box$" pour des logiques modales vérifiant au moins K et encore "$\boxed{\circ}$" pour les logiques vérifiant D, et cela en présence ou en absence de la règle de nécessitation.

[3] Il ne s'agit bien sûr pas d'identifier la consistance et la conscience, ce n'est pas parce que deux choses vérifient des mêmes formules qu'elles sont identiques.

7) Je présente d'autres expériences d'autoduplication par la pensée pour motiver l'existence d'une probabilité sur l'ensemble des états directement accessibles (par les expériences de type AR).

Cette probabilité est associée à un indéterminisme *abrupte*, semblable à celui d'une large classe d'interprétations de la mécanique quantique[4]. Comment extraire une mesure de probabilité (et sur quoi porte-t-elle exactement ?) de telle façon que $P_{dup} = 1/2$ dans l'expérience de l'autoduplication sans annihilation, et qu'en particulier $P_{trans} = 1$ dans l'expérience AR ?

L'idée est de faire porter les probabilités sur les états directement accessibles et transitoires des référentiels réalistes.

C'est une forme de *darwinisme élémentaire* : on ne tient compte que des états vivants, conscients, consistants ($\Diamond\top$) c-à-d *où on a survécu*.

Pour ce faire je propose l'application du stratagème *affaibli* sur C :

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

Plus précisément je considère la transformation DEON de LPM dans LPM (LPM = langage propositionnel modal) : DEON(p) = p pour p variable propositionnelle, DEON(A#B) = DEON(A)#DEON(B), où # représente un opérateur booléen. Enfin, le cas principal

$$\text{DEON}(\Box A) = \Box\text{DEON}(A) \ \& \ \Diamond\text{DEON}(A)).$$

$\boxdot\Diamond\top$ peut être alors interprété par le fait que *normalement* P(survie) = 1. On obtient une logique KD? = {p | C $\vdash$ DEON(p)}, sans 4 donc, et on perd la nécessitation car C $\nvdash$ $\Box\top$ & $\Diamond\top$.

La probabilité "P=1" est capturée sainement et complètement[5] par "$\Box$p" dans un système KD avec nécessitation par Alechina (1994). L'absence de $\Box\top$ et de 4 correspond au fait qu'on a plus affaire à une croy<u>ance</u> non normalisable qu'à une prob<u>abilité</u> normalisée[6]. Dans le contexte de la multiplication de soi "P=1" correspond à $\Box\Diamond\top$.

<u>2. LE MECANISME DIGITAL</u>

8) J'introduis une version moderne de la thèse classique et informelle de Church. (Elle est aussi connue sous le nom de thèse de Turing, ou thèse de Markov, ou encore thèse de Church-Turing-Markov) :

> *La collection des fonctions calculables = la collection des fonctions programmables sur un ordinateur.*

---

[4] Il s'agit des interprétations sans variables cachées (même non locales).

[5] Pour une sémantique qu'il conviendrait de donner (Alechina 1994, voir aussi Fattorosi-Barnaba et Amati 1987). Je bénéficie pour cette récapitulation d'un exposé d'Alechina à l'Iridia (juin et août 1994 qui m'a confirmé concernant le rôle de D en probabilité. Je bénéficie de conversation avec P. Smets pour le rôle de D pour les croyances.

[6] On est vraisemblablement plus proche de Boole 1854 et de Dempster-Shafer 1976 que de Bayes (voir aussi Smets dans Smets & Al. 1988, Smets 1991, voir aussi Hacking 1975).

L'ordinateur est supposé disposé de ressource mémoire non limitée. En pratique cela est réalisé par l'usage de mémoire virtuelle extensible (comme les grottes, le papier ou les disques magnétiques). On s'intéresse donc, ici, à la calculabilité de principe plutôt qu'aux calculs. Si le carré représente la connaissabilité, on peut présenter, dans une arithmétique classique étendue avec S4-quantifiée, avec la règle $\boxdot\forall x P(x) => \forall x \boxdot P(x)$, une version formelle épistémique de la thèse de Church :

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

Les $\{\phi_i, i \in \omega\}$ constituent une énumération des fonctions partielles récursives (Rogers 1958) et $\phi_z(x)\Downarrow$ signifie que la machine (dont le nombre de Gödel est) z s'arrête avec le nombre x comme entrée (input).

Cette version de la thèse de Church peut être considérée comme une version épistémique d'une thèse de Church intuitioniste :

$$\forall x \exists y\, P(x,y) -> \exists z \forall x\, (\phi_z(x)\Downarrow \,\&\, P(x,\phi_z(x)))$$

La thèse classique affirme que tout ce que je sais calculer est calculable par un ordinateur. Elle rend universelle et univoque la notion de partiellement calculable. La thèse intuitioniste (qui est fausse classiquement) revient à affirmer que *toutes* les fonctions sont Turing-calculables.

J'appelle **école (pythagoricienne) *du dehors*** les écoles qui s'intéressent à toutes les propositions portant sur toutes les fonctions, totales ou partielles, calculables. J'appelle **école (pythagoricienne) *du dedans*** les écoles qui s'intéressent exclusivement aux propositions sur des fonctions appartenant à des classes de fonctions (prouvablement) totales calculables.

La thèse de Church classique rend le dehors univoque, et la thèse de Church intuitioniste rend le dedans univoque au prix d'accepter la non-représentabilité arithmétique de la prouvabilité du dedans. Ceci résulte essentiellement de l'absence de fermeture pour la diagonalisation de l'ensemble des fonctions totales calculables.

Grossièrement les thèses de Church entraînent que la frontière entre le dedans et le dehors, vue du dehors, est absolument vague, et vue du dedans n'existe pas. En particulier la thèse de Church épistémique entraîne l'existence de propositions vraies absolument indécidables[7] (voir (déjà) Post 1921, voir Gödel 1951 cité dans Wang 1974, voir aussi Kalmar 1959, Reinhardt 1986).

Je mentionne la thèse de Post-Turing (comme l'appelle Reinhardt) permettant d'identifier les ensembles semi-décidables avec les ensembles récursivement énumérables. Elle entraîne l'identification entre les propositions vérifiables ou même observables (au sens d'Abramsky[8], voir Vickers 1989) et les propositions Σ_1. On peut voir la thèse de Post-Turing comme une version ensembliste de la thèse de Church.

---

[7] Ces propositions absolument indécidables ne peuvent être considérées qu'au niveau épistémique, donc classique. De l'intérieur (donc pour celui que j'appelle plus loin le *solipsiste*) de telles propositions n'ont aucun sens : elles ne sont ni exprimables, ni concevables.

[8] Une idée, qui est déjà implicite chez Brouwer, et qui est à la base des sémantiques algébrico-topologiques de la logique intuitioniste (algèbre de Brouwer, algèbre de Heyting) consiste à identifier les ensembles semi-décidables (ou de suite les propositions intuitionistes) avec les ouverts d'un espace topologique. L'interprétation en terme d'observation finie se trouve dans la thèse d'Abramsky 1987 et est développée par Vickers 1989.

9) Dans ce contexte le second théorème de récursion de Kleene[9] (2-REC, Kleene 1952) permet de prouver (constructivement) l'existence de machines universelles autoduplicables.

Cela donne une preuve conceptuelle de l'existence de machines et de réseaux de machines capables de s'autodupliquer ou d'appliquer d'éventuelles transformations (non nécessairement constructives) sur elles-mêmes (Myhill 1964, Case 1971, 1974).

C'est 2-REC qui permet de ***désindexicaliser*** l'usage de la première personne et d'user arithmétiquement de terme équivalent à "description de moi" sans user de terme indexical.

Une identification de base, effectuée au niveau adéquat, entre machines et systèmes formels permet de justifier alors l'usage de la logique de la prouvabilité (voir Boolos 1979, 1993) pour la communicabilité positive entre machines autoréférentiellement correctes relativement au niveau adéquat de description (voir aussi Smullyan 1985, Marchal 1992).

Cela revient à étendre la théorie C(4) en la théorie G, qui s'applique donc essentiellement aux machines autoréférentiellement correctes (et donc consistantes) capables de communiquer qu'elles sont en principe *au moins* Σ_1-complètes, abstraction faite des ressources disponibles.

Plus précisément, la prouvabilité gödélienne B *par* une telle machine est représentable *dans* le langage de la machine, et on peut montrer qu'elle vérifie :

1) $M \vdash A \Rightarrow M \vdash B(\ulcorner A \urcorner)$ (L1)
2) $M \vdash B(\ulcorner A \urcorner) \,\&\, B(\ulcorner A \to B \urcorner) \to B(\ulcorner B \urcorner)$ (L2)
3) $M \vdash B(\ulcorner A \urcorner) \to B\ulcorner(B(\ulcorner A \urcorner)\urcorner)$ (L3)

où $\ulcorner A \urcorner$ représente un codage de A dans le langage de M.

Soit F une fonction qui associe à chaque variable propositionnelle un énoncé dans le langage de la machine M. On peut traduire les propositions modales dans le langage de M. J'appelle cette traduction morphisme de Magari-Boolos[10] MB_F.

Appliquée aux variables propositionnelles MB_F est simplement F. $MB_F(A\#B) = MB_F(A)\#MB_F(B)$, où # représente un opérateur booléen. Enfin, le cas principal $MB_F(\Box A) = B(\ulcorner MB_F(A)\urcorner)$.

L1, L2 et L3 peuvent s'écrire (avec un évident abus de langage) :

1) $M \vdash A \Rightarrow M \vdash \Box A$ Nec
2) $M \vdash \Box A \,\&\, \Box(A \to B) \to \Box B$ K
3) $M \vdash \Box A \to \Box\Box A$ 4

La désindexicalisation rendue possible par 2-REC, est capturée dans les discours des machines autoréférentiellement correcte par le *lemme de diagonalisation*. (Gödel 1931, voir aussi Smorynski 1981).

Solovay (1976) démontre deux théorèmes remarquables et fondamentaux :

[9] $\forall t\, \exists e\, \forall y\ \phi_e(y) = \phi_t(e,y)$. Ce théorème est équivalent à la formulation de Rogers 1967 : $\forall i\, \exists e\, \forall x\ \phi_{\phi_i(e)}(x) = \phi_e(x)$. Toutefois la forme de Kleene est valable pour des énumérations assez riches du dedans, la forme de Rogers n'est valable que sur les énumérations du dehors.
On peut énoncer la forme de Rogers de la façon suivante : toute endo-transformation intensionnelle du dehors admet (au moins) un point fixe extensionnel. Par comparaison le premier théorème de récursion énonce que toute transformation extensionnelle admet un plus petit point fixe extensionnel (t est extensionnelle si $\phi_i = \phi_j \Rightarrow \phi_{t(i)} = \phi_{t(j)}$).

[10] Depuis le résultat de Löb (1955) l'interprétation modale de la prouvabilité gödélienne est assez bien étudiée, particulièrement en Italie (Magari, Bernardi, Sambin, Montagna, ...), en Hollande (Löb, van Benthem, Veldman, de Jongh, Visser, ...), aux USA (Boolos, Solovay, Smorynski,...) et en (ex)URSS (Kusnetzov, Artemov, Vardanyan, Shavrukov, Dzhaparidze...). Ces listes sont loin d'être exhaustives.

a) **la théorie G**, qui a K, 4, et L : $\Box(\Box A\text{->}A)\text{->}\Box A$ pour schéma d'axiomes et Nec et MP comme règles d'inférence, axiomatise complètement ce que la machine peut prouver sur sa propre prouvabilité :

Théorème G $\vdash$ A ssi pour tout F, M $\vdash$ $MB_F(A)$

Le second résultat est plus étonnant, et constitue le moteur principal du présent travail :

b) **la théorie G***, qui a comme axiomes tous les théorèmes de G, ainsi que le schéma d'axiomes $\Box A\text{->}A$, et pour règle, le modus ponens, axiomatise complètement la *vérité* sur la prouvabilité par M :

Théorème G* $\vdash$ A ssi pour tout F, $MB_F(A)$ est vrai

Solovay montre en outre la décidabilité de G et de G*. Il montre comment traduire G* dans G :

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

où $S^{\Box}(A)$ désigne l'ensemble des sous-formules de A de type "$\Box\#$".

En remplaçant A par $\perp$ dans L, on trouve la version modale du second théorème d'incomplétude de Gödel, et il s'agit aussi d'une instantiation du principe LWV, avec x = $\Diamond\top$. Sambin a démontré que KL prouve 4, cela découle aussi de la sémantique de Kripke de G.

G est donc une extension naturelle de C4. Les solutions de LWV, comprenant en particulier W, sont des théorèmes de G*, et ne sont pas des théorèmes de G. Il s'agit des vérités (indexicales) non communicables. Inversement toutes formules de G*\G est une solution de LWV.

Ainsi, pour la conscience, G*\G constitue une sorte de réservoir d'expériences subjectives possibles. J'argumenterai comme quoi de telles vérités incommunicables peuvent néanmoins être auto-inférées par des machines.

En **résumé**, pour les machines autoréférentiellement correctes, avec MDI et la thèse de Church, on peut condenser la situation ainsi :

G = le communicable (sur le communicable et l'incommunicable)
G = le vrai (sur le communicable et l'incommunicable)*
G \ G = l'incommunicable, mais vrai (le réservoir subjectif)*

10) Je présente la réfutation, reposant sur l'incomplétude Gödélienne des théories formelles, du mécanisme par Lucas. En effet en 1959, Lucas a tenté de réfuter MEC-BEH, sous la forme d'une sorte de test de Turing limité aux propositions de l'arithmétique informelle (Lucas 1959, voir aussi Penrose 1988, qui utilise cet argument et défend alors la possibilité d'une théorie quantique de la conscience[11]).

L'argumentation de Lucas illustre à la fois la difficulté d'identifier le connaisseur sujet $\boxdot$ avec la machine objet $\Box$ et les difficultés conceptuelles qui se présentent lors des expériences par la pensée de multiplication de soi-même.

Je pense avec Post (1921!), et Webb (1980) qui a consacré un ouvrage sur cet argument, que le raisonnement de Lucas est invalide[12]. Je partage néanmoins avec Lucas l'idée que le théorème de Gödel s'applique aux machines digitales.

---

[11] Je ferai exactement l'inverse, je vais montrer que si le mécanisme est correct, on doit extraire la phénoménologie de la matière à partir de la théorie de la conscience, en particulier je suggère plus loin comment tirer une "crédibilité quantique" à partir de G&G*.

[12] Voir aussi Hofstadter 1979.

Je pense, à présent, comme Benacerraf (1967), qu'une bonne partie du raisonnement de Lucas peut être rendu valide : la conclusion antimécaniste doit être affaiblie. Ce n'est pas que *je ne suis pas une machine*, mais seulement que *si je suis une machine*, alors *je ne peux pas me reconnaître dans cette machine*.

Je propose une reconstruction de cet argument (due en partie à Benacerraf 1967, Chihara 1972, Reinhardt 1986).

Cette reconstruction de la réfutation de Lucas, est formellement similaire aux critiques de MDI basée sur les expériences par la pensée de l'autoduplication (si je suis duplicable, je ne peux pas me reconnaître dans le dupliqué).

Lucas prend comme hypothèse qu'il est sain ($\boxdot$p->p) et il décide de ne se comparer qu'aux machines saines ($\Box$p->p).

Lucas commet une identification entre machine et système formel (ce qui, on l'a vu, est une conséquence plausible de MDI, au niveau adéquat, avec la thèse de Church).

La *communicabilité* $\Box$ de la machine est alors arithmétisable et diagonalisable : il existe un énoncé arithmétique p tel que p <-> $\neg$ $\Box$p. La machine saine lui étant présentée, Lucas peut trouver cet énoncé p, et démontrer p <-> $\neg$ $\Box$p.

L'argumentation de Lucas, ou plutôt sa reconstruction, peut alors être résumée dans la dérivation suivante :

1) $\boxdot$(p <-> $\neg$ $\Box$p)	identification de base ;
2) $\boxdot$($\Box$-> $\neg$ p)	calcul propositionnel ;
3) $\Box$p->p	la machine est saine ... (par hypothèse!)
4) $\boxdot$($\Box$p->p)	... sait Lucas ;
5) $\boxdot$($\Box$ p->(p & $\neg$ p))	calc. prop. + 2) et 4)
6) $\boxdot$($\Box$ p->$\bot$)	calc. prop.
7) $\boxdot$ $\neg$ $\Box$p	calc. prop.
8) $\boxdot$p	par 7 et 1)
9) $\neg$ $\Box$p	par 7) et $\boxdot$p->p.

et donc l'ensemble des propositions arithmétiques p telles que $\boxdot$p, est différent de l'ensemble des propositions arithmétiques p telles que $\Box$p. Lucas semble être à même de se distinguer de toutes les machines (saines) sur un test de Turing limité à l'arithmétique. Où est l'erreur ?

11) En faveur de Lucas, une chose est claire : si $\boxdot$ obéit à S4, il n'est pas arithmétiquement définissable (il n'est pas finitairement définissable par lui-même).

En effet, dans ce cas, il serait diagonalisable et il existerait un énoncé p tel que p <-> $\neg$ $\boxdot$p, donc $\boxdot$p->$\neg$p, or $\boxdot$p->p (par la réflexion T), donc $\neg$ $\boxdot$p, donc p (puisque p <-> $\neg$ $\boxdot$p), donc $\boxdot$p (par nécessitation), donc $\bot$, puisqu'on a à la fois $\neg$ $\boxdot$p et $\boxdot$p. Remarquons la similarité de cette preuve avec la réfutation de Lucas.

Un raisonnement similaire montre que G* ne peut pas être fermé pour la nécessitation, puisqu'il a la réflexion. On voit en fait qu'une communicabilité ne peut pas être à la fois diagonalisable, obéir à l'axiome de réflexion et être fermée pour la nécessitation.

En passant on retrouve le théorème de Tarski comme quoi la vérité arithmétique n'est pas arithmétiquement définissable. En effet si tel était le cas, on disposerait d'un prédicat V de vérité, et le stratagème fort pourrait produire une connaissance $\boxdot$ p qui serait intensionnellement équivalente à $B(\ulcorner p \urcorner)$ & $V(\ulcorner p \urcorner)$, qui obéirait à S4 et serait diagonalisable[13].

---

13 Voir aussi Kaplan et Montague 1960

12) Pour trouver de façon précise l'erreur de Lucas dans notre contexte, c-à-d avec MDI et la thèse de Church, il reste à appliquer le stratagème de Théétète à la logique G.

En réalité ceci n'est pas nécessaire, on peut se contenter de travailler dans une arithmétique étendue avec S4 (comme l'arithmétique épistémique de Reinhardt-Shapiro[14]).

Dans ce cas cependant la relation entre la prouvabilité dans la théorie et la prouvabilité intuitive (ou constructive, absolue) n'est pas claire.

L'avantage du stratagème appliqué à une machine auto-référentiellement correcte[15] est de garantir au départ l'égalité extensionnelle de la prouvabilité formelle et de la prouvabilité intuitive.

Le stratagème est capturé par la transformation modale BGKM (pour Boolos, Goldblatt, Kusnetzov et Muravitsky) : A nouveau BGKM(p) = p, pour les variables propositionnelles, BGKM(A#B) = BGKM(A)#BGKM(B) avec # désignant un opérateur booléen, et BGKM($\neg$ A) = $\neg$ BGKM(A).

Mais surtout :

$$\text{BGKM}(\Box A) = \Box\,\text{BGKM}(A)\ \&\ \text{BGKM}(A)$$

Sur G cela donne la théorie S4Grz. C'est-à-dire S4 (KT4 + MP et Nec) + la formule de Grzegorczyk[16] : $\Box(\Box(p\text{->}\Box p)\text{->}p)\text{->}p$:

$$\text{S4Grz} \vdash A \text{ ssi } G \vdash \text{BGKM}(A)$$

On regarde alors le raisonnement de Lucas au niveau G*, et on voit que l'erreur se situe dans le passage de la ligne 3) à la ligne 4).

En composant BGKM avec l'interprétation *arithmétique* (= dans le langage de la machine, avec l'identification de base) de G, on obtient une interprétation arithmétique de $\boxdot$, c-à-d, on obtient les transformations (paramétrées par F) : $MB_F \circ BGKM$: LPM -> L(M), et

$$\text{S4Grz} \vdash A \text{ ssi pour tout F, M} \vdash MB_F(\text{BGKM}(A))$$

Ainsi $\boxdot p$ est interprété par $B(\ulcorner p \urcorner)\&p$. La théorie obtenue, S4Grz, constitue alors une axiomatisation naturelle de la connaissabilité (intuitive et non diagonalisable) de la machine.

Le bord *flou* du sujet est capturé par le fait que $\boxdot$ n'est pas *effectivement* définissable par la machine, ni arithmétisable, ni diagonalisable. Comme il fallait s'y attendre, le stratagème fort empêche le sujet de se définir *effectivement* lui-même.

De même, le sujet machine peut se dupliquer, mais il ne peut pas effectivement (de façon effective, constructive, communicable, ou connaissable) se reconnaître dans le dupliqué.

On rejoint d'une certaine façon la philosophie de Brouwer (le fondateur de la philosophie intuitioniste, cf van Stigt 1990) en ce sens que le sujet (et son oeuvre) n'est pas (prouvablement) axiomatisable.

---

14 ... et c'est ainsi que je présente les choses dans 2.3. Le rôle de Grz avec MDI reste à élucider.

15 Au lieu de prendre une telle machine on peut prendre l'arithmétique de Peano ou ZF, ou n'importe quelle théorie RE consistante et capable de prouver sa propre Σ_1-complétude. Dans le présent travail, l'arithmétique de Peano est considérée comme l'*Escherichia Coli* de l'autoréférence.

16 Voir Grzegorczyk 1967 pour la forme originale, et Boolos 1980c pour la preuve de l'équivalence entre les deux présentations. On peut montrer que le système KGrz possède T et 4 parmi ses théorèmes et est donc extensionnellement équivalent (avec les règles MP et Nec) à KT4Grz = S4Grz.

13) La formule de Grzegorczyk entraîne l'antisymétrie de la relation d'accessibilité pour les modèles finis. Cela suggère une interprétation temporelle[17] (au sens subjectif) du développement local de la connaissance du sujet, ce qui permet une interprétation arithmétique du temps subjectif à-la-Bergson[18] 1939 (voir aussi Dogen 1232-1253), ce qui encore est proche de la philosophie de la conscience et du développement temporel du soi de Brouwer (voir Brouwer 1905, 1948, 1946-51, voir aussi Grzegorczyk 1964 !).

En fait **une logique intuitioniste** représentable arithmétiquement, IL, émerge. En effet, Gödel (1933) a suggéré et McKinsey & Tarski (1948) ont démontré qu'on peut interpréter la logique intuitioniste dans le système modal S4. Grzegorczyk a étendu le résultat pour S4Grz.

Voilà la transformation de 1933 de Gödel. Attention il s'agit d'une transformation du langage propositionnel (non modal) dans LPM. G33(A#B) = G33(A)#G33(B) avec # mis pour une conjonction ou une disjonction. Attention :

$G33(p) = \Box p$, pour les variables propositionnelles,
$G33(\neg A) = \Box \neg G33(A)$
$G33(A \rightarrow B) = \Box G33(A) \rightarrow \Box G33(B)$

On a, avec IL = logique intuitioniste (Grzegorczyk 1967) :

$$IL \vdash A \text{ ssi } S4Grz \vdash G33(A)$$

Il suffit de composer les différentes transformations pour extraire l'interprétation arithmétique de l'intuitionisme (Goldblatt 1978, voir aussi Artemov 1990) :

$$IL \vdash A \text{ ssi pour tout } F, M \vdash MB_F(BG(G33(A)))$$

J'argumente, avec MDI, en faveur de l'idée qu'il s'agit de la part solipsiste du sujet.
Un résultat formel, qui confirme ce point de vue, est que le passage à G* ne rajoute aucune formule : ni pour S4Grz (Boolos 1980a, 1980b), ni pour IL (Goldblatt 1978). Les propositions absolument indécidables mais vraies, c-à-d telles que $G^* \vdash p \;\&\; \neg \boxdot p$, ne sont pas des images de propositions intuitionistes par la transformation BGKM.
Avec des notations évidentes, Boolos et Goldblatt ont démontré que S4Grz = S4Grz* et IL = IL* : le solipsiste identifie correctement (de *son* point de vue) la prouvabilité et la vérité.

Notons qu'Artemov (1990) justifie (proprement) que le stratagème fort peut être érigé en une thèse de philosophie des mathématiques (comme l'est la thèse classique de Church, voir annexe 6).

En résumé :

S4Grz = le connaissable	*IL = le solipsiste*
S4Grz = idem*	*IL* = idem*
S4Grz\\S4Grz = {}*	*IL*\\IL = {}*

14) Conclusion : Lucas a essayé de tirer une contradiction de l'ensemble des propositions $\{\Box p \rightarrow p, \boxdot p \rightarrow p, \Box p \leftrightarrow \boxdot p\}$, où p est une proposition arithmétique, $\Box$ représente la prouvabilité par une machine auto-référentiellement correcte et $\boxdot$ représente la prouvabilité intuitive de l'incorrigible Lucas.

Le raisonnement de Lucas montre seulement le caractère contradictoire de l'ensemble $\{\Box p \rightarrow p, \boxdot p \rightarrow p, \boxdot(\Box p \leftrightarrow \boxdot p)\}$. On a bien $G^* \vdash \Box p \rightarrow p$, $G^* \vdash \boxdot p \rightarrow p$, $G^* \vdash (\Box p \leftrightarrow \boxdot p)$, mais $G^* \nvdash \boxdot(\Box p \leftrightarrow \boxdot p)$. On retrouve bien la solution de Benacerraf.

[17] Seul Spencer-smith 1991 semble mentionner explicitement cette idée "évidente".
[18] Notons que Brouwer ainsi que Post dans son anticipation se réfère de temps à autre à Bergson.

En particulier Lucas a prouvé (modulo la reconstruction de son raisonnement) la contradiction de $\exists n \boxdot(\boxdot p <-> \Box_n p)$. Sachant que G n'est pas fermé pour la règle $\Box \exists x A(x) => \exists x \Box A(x)$ et que la formule $\exists x \Box A(x) -> \Box \exists x A(x)$ est validée dans G (étendu avec quantificateurs), la possibilité reste ouverte que $\boxdot \exists n(\boxdot p <-> \Box_n p)$. Cela sort du cadre du présent travail[19].

Pour paraphraser Post 1921, l'argument "Gödélien" ne peut pas montrer que l'homme n'est pas une machine.

Post ajoute que l'argument montre seulement que l'homme ne peut pas construire une machine prouvant les mêmes théorèmes (de l'arithmétique) que lui.

Ceci est encore trop dire puisqu'avec 2-REC, on peut rendre une machine quelconque (extensionnellement parlant) autoreproductible : ce que je montre (avec MDI) c'est qu'un sujet ne peut pas *à la fois* construire une machine capable de prouver les mêmes théorèmes (de l'arithmétique) que lui *et en même temps prouver* qu'il en est bien ainsi, c-à-d prouver que cette machine est capable de prouver les mêmes théorèmes que lui.

En interprétant $\boxdot$ par "je sais" et $\Box$ par "il croit", où "il" joue le rôle d'une duplication de "moi", les situations paradoxales des expériences par la pensée sont suffisamment clarifiées pour confirmer la thèse digitale et empêcher qu'on ne prenne ces expériences par la pensée pour des réfutations de MDI[20].

En **résumé**, les erreurs dans l'usage du théorème de Gödel, ou des paradoxes de la duplication, pour réfuter le mécanisme reviennent en général (et à une reconstruction logique de l'argument près) à un usage simultané de T, Nec et de la diagonalisation (ou de la représentabilité arithmétique). Le tableau suivant récapitule la situation et peut servir de garde-fou contre ce genre de confusion intensionnelle[21] en philosophie (mécaniste) de l'esprit :

	G	G*	S4Grz
T	-	+	+
Nec	+	-	+
Diag (Arith)	+	+	-

15) Pour le calcul des *probabilités-crédibilités* qu'on rencontre dans les expériences d'auto-multiplication, je propose d'appliquer le stratagème *affaibli* à la logique de l'autoréférence G. Cela correspond au darwinisme élémentaire des expériences de multiplications de soi, et cela devient par MDI, une forme de darwinisme arithmétique. Modalement cela revient à étudier un nouvel opérateur modal $\boxed{\circ}p$, définit par $\Box p \& \Diamond p$.

Arithmétiquement cela revient à étudier la logique d'une prouvabilité intensionnellement définie par $B(\ulcorner p \urcorner) \& \neg B \neg (\ulcorner p \urcorner)$.

De façon précise

$$KD? = \{p \text{ telles que } G \vdash DEON(p)\},$$

[19] De même G démontre la formule de Barcan-inverse : $\Diamond \forall x A(x) -> \forall x \Diamond A(x)$. Mais la formule de Barcan n'est pas un théorème de G (ni de G*).

[20] Le "vieux" Wittgenstein, celui de "de la certitude" (1951) semble avoir réalisé cette nuance intensionnelle. Voilà par exemple la proposition 42 de "de la certitude" : *On peut dire "il le croit, or il n'en est pas ainsi", mais non : "il le sait, or il n'en est pas ainsi". Cela vient-il de la différence entre les états d'âme de la croyance et du savoir? Non. - On peut appeler "état d'âme", disons, ce qui s'exprime dans l'intonation, dans l'attitude, etc. Il serait donc* possible *de parler d'un état d'âme de la conviction; et ce peut être le même, que la croyance qu'il comporte, corresponde à un savoir ou soit fausse. Penser qu'aux mots "croire" et "savoir" doivent forcément correspondre des états différents serait équivalent à croire qu'au mot "Ludwig" et au mot "moi" doivent forcément correspondre des hommes différents parce que les concepts sont différents.*

[21] La présence de confusion intensionnelle chez Lucas 1961 a été mis en évidence la première fois (à ma connaissance) par Webb 1968. Voir aussi Visser 1986.

Par le premier résultat de Solovay 1976, On a aussi :

$$KD? = \{p \text{ telles que } \forall F\ M \vdash MB_F \circ DEON(p)\},$$

On obtient une théorie KD?, extension de KD, fermée pour le modus ponens mais pas pour la règle de nécessitation (Nec). Cette théorie est cependant fermée pour la règle dite de monotonie rationnelle RM : p->q => $\boxdot$p->$\boxdot$q.

On perd aussi 4, ce qui était inespéré : cela permet d'user, grâce à G*, de l'incommunicable (mais vrai) des voisinages de l'infini, pour analyser la logique du "croyable-probable" dans les voisinages de zéro.

Sans 4, il s'agit plus de croyance que de crédibilité. On obtient une logique des croyances non normalisée[22] : on n'a pas $\boxdot\top$ (noté N dans Chellas 1980) puisque G $\nvdash$ $\Box\top \& \Diamond\top$.

Une telle théorie, avec K et sans N, ne possède pas de sémantique de Kripke. Elle admet cependant une sémantique de Scott-Montague (Neighborhood semantics) : la structure des voisinages à chaque monde constitue (au moins) un filtre sans éléments maximals. Ils sont, en effet, fermés pour l'intersection parce que

$$G \vdash (\boxdot A \& \boxdot B) \to \boxdot (A \& B)$$

De même si $x \cap y$ appartient à une structure de voisinage, alors x et y appartiennent chacun à cette structure :

$$G \vdash \boxdot (A \& B) \to (\boxdot A \& \boxdot B)$$

Il n'y a pas nécessairement d'éléments maximals : il peut y avoir des mondes dont la structure de voisinage ne possède pas la tautologie [$\top$] identifiée ici avec la collection de tous les mondes :

$$G \nvdash \boxdot\top$$

Je ne sais toujours pas si cette théorie est axiomatisable, mais grâce à la toute puissance de G (cf le premier théorème de complétude arithmétique de Solovay 1976), il est facile de voir que KD? est décidable.

A la différence de IL (= IL*) et de S4Grz (= S4Grz*), KD? ≠ de KD?*.

KD?* est naturellement mis pour {p | G* $\vdash$ DEON(p)}. Grâce à la toute puissance de G* (cf le second théorème de complétude arithmétique de Solovay 1976), KD?* = {p | $\forall$F MB$_F\circ$DEON(p)}.

La décidabilité de G* entraîne la décidabilité de KD?* (un démonstrateur pour chacun est proposé[23] dans l'annexe 2).

A présent, à la différence de G,

$$G^* \vdash \boxdot\top$$

Puisque G* $\nvdash$ $\Box\top \& \Diamond\top$. Donc KD?* diffère de KD?*

Autre exemple,

---

[22] Pour les logiques modales des probabilités ou des crédibilités, P(Ω), ou B(Ω) ≠ 1 entraîne, avec la sémantique d'Alechina 1994(voir aussi Fattorosi-Barnaba & Amati 1987) la perte de $\boxdot\top$, ce qui entraîne la perte de la règle de nécessitation. La réciproque n'est pas vraie, G* possède $\boxdot\top$ comme théorème, mais G* n'est pas fermé pour la nécessitation ou pour RM (de même KD?-Σ_1*).

[23] ... pour la part propositionnelle du calcul. Au niveau du premier ordre l'extension correspondante de G (avec quantificateurs) pour une machine adéquate n'est plus axiomatisable (Vardanyan 1985, cité dans Boolos 1993).

$$G^* \vdash \boxed{\circ}\, p \to p$$

et G ne le prouve pas. KD?* est donc une extension de KT, qui n'est pas fermée pour l'application de la règle de nécessitation Nec.

Comme pour la starification de G, c-à-d G*, la *starification* de KD?, KD?*, non seulement n'est pas fermée pour la nécessitation, mais n'est pas non plus fermée pour la règle d'équivalence RE[24] : $p \leftrightarrow q \Rightarrow \Box p \leftrightarrow \Box q$.

Comme les modèles de Scott-Montague sont tous fermés pour cette règle, ni G* ni KD?* ne possèdent une telle sémantique[25].

A la différence de $\boxdot$, $\boxed{\circ}$ est *arithmétisable*. Son interprétation arithmétique, ou avec MDI (et la thèse de Church, et l'identification de base ...) l'interprétation *par* la machine *dans* le langage de la machine de $\boxed{\circ}p$, est $B(\ulcorner p \urcorner) \,\&\, \neg B \neg (\ulcorner p \urcorner)$, et d'une façon générale F étant variable, $MB_F \circ DEON(p)$. Cela permet à G de prouver l'existence du point fixe suivant :

$$G \vdash \Box(p \leftrightarrow \neg \boxed{\circ}\, p) \to \Box(p \leftrightarrow \Box\bot \vee \Diamond\Diamond\top)$$

KD? donne la part communicable de la logique des croyances consistantes.

A cause de la perte de 4 (par l'application du stratagème affaibli), G lui-même reste le meilleur candidat pour une notion de *crédi**bilité*** rationnelle (*en principe* révisable) de la part de la machine autoréférentiellement correcte.

De plus dans le contexte d'une accessibilité immédiate (non transitive et sans nécessitation) la starification peut faire office de stratagème fort : on peut argumenter en faveur du fait que KD?*, qui prouve $\boxed{\circ}\, p \to p$, mais pas[26] $\boxed{\circ}\,(\boxed{\circ}\, p \to p)$ ni $\boxed{\circ}\, p \to \boxed{\circ}\boxed{\circ}\, p$, joue le rôle d'une logique de la connaissance immédiate ou intuitive, dont KD? apparaît alors comme la part communicable et *en principe* révisable.

De cette façon on peut interpréter les propositions de KD?* \ KD? comme des propositions de la logique des croyances consistantes et correctes, bien que non communicables comme telles[27].

On peut interpréter les propositions de KD?* \ KD?, comme des relations logiques bien fondées mais non communicables : par exemple du type des sentiments ou des intuitions immédiates accessibles dans les voisinages de zéro. La "croyance" décrit dans ce contexte une forme de mélange entre le probable et le communicable.

En résumé :

---

[24] G* prouve $\top \leftrightarrow \Diamond\top$ mais ne prouve pas $\Box\top \leftrightarrow \Box\Diamond\top$. De même G* prouve $\top \leftrightarrow \boxed{\circ}\top$, mais ne prouve pas $\boxed{\circ}\top \leftrightarrow \boxed{\circ}\boxed{\circ}\top$. Tous ces théorèmes sont faciles à démontrer. Bien sûr, on peut aussi utiliser le démonstrateur donné en annexe. Par exemple (g*ip '((p -> p) <-> - bw - (p -> p))) donne NIL, cela signifie que g*, le démonstrateur n'a pas trouvé de contre-exemples et donc que la formule présentée est un théorème. Le "ip" de g*ip permet l'usage de notation infixée : (g* '(-> p p) donne la même chose que (g*ip '(p -> p)), avec l'ordre de précédence usuel. De même (kd*ip '((p -> p) <-> bw (p -> p))) donne NIL, mais (kd*ip '(bw (p -> p) <-> bw bw (p -> p))) donne une longue série de modèles de Kripke avec des mondes invalidant la formule telle qu'elle est traduite dans G* (c-à-d finalement dans G). Par abus de notation, j'écris souvent KD? prouve ---$\boxed{\circ}$---, au lieu de KD? prouve --- $\Box$ ---, ou au lieu de G(*) prouve --- $\boxed{\circ}$ ---. La même remarque vaut pour toutes les logiques bâties sur G et G*.

[25] Boolos (1980a) a donné une sémantique de Kripke *variable* pour G*. Il devrait de même être possible de développer une sémantique de Scott-Montague variable pour KD?*.

[26] L'absence de $\Box(\Box p \to p)$ est une forme de modestie de la part de la machine auto-référentiellement correcte. Cette idée apparaît dans Smullyan 1987, et Boolos 1993 mentionne Rohit Parickh interprétant la formule de Löb par le fait que PA ne pourrait pas être plus modeste au sujet de sa véracité. En ce sens KD? et KD?* sont sans doute un peu moins modeste. Bien sûr le solipsiste, interprété épistémiquement par S4Grz, est de loin le plus immodeste. S4grz est le seul système dans ce travail capable de prouver $\Box(\Box p \to p)$. Je rappelle que la vérité et la prouvabilité sont implicitement identifiées pour le solipsiste (Le fait que IL = IL* justifie, a posteriori une telle attitude, chez la machine (avec MEC).

[27] De telles croyances pourraient bien être instinctivement inférables en un coup (ou en un nombre fini de coups).

KD? = la croyance (communicable et en principe révisable)
KD? = KT? = la connaissance intuitive (incorrigible)*
KD? \ KD? = la connaissance incommunicable (le sentiment)*

Les points d'interrogation représentent à chaque endroit un ou peut-être même une infinité d'axiomes qu'il faut ajouter à chacun de ces systèmes pour qu'ils soient arithmétiquement complets (l'équivalent de Grz pour l'application du stratagème affaibli). Bien que ces théories n'aient pas été axiomatisées, il est facile, avec G et G*, de construire pour chacune un démonstrateur de théorèmes (pour la part propositionnelle, voir annexe 2).

De telles interprétations peuvent être utilisées pour rechercher des contre-exemples dans des raisonnements philosophiques.

Le but ici, cependant, est d'extraire un calcul de probabilité ou de crédibilité (concernant des extensions consistantes *immédiatement accessibles* ou *voisines* pour une sémantique convenable qu'il reste à trouver) capable de justifier arithmétiquement les probabilités-croyances extraites par MDI à partir des expériences par la pensée d'auto-multiplication de soi.

Avant çà, le point suivant rappelle la présence *à terme* des théories décrites ici dans les discours des machines à inférence inductive Löbienne (autoréférentiellement correcte et capable de prouver sa propre Σ_1-complètude).

16) J'utilise la théorie de l'inférence inductive (Putnam 1963 & 1965, Gold 1967, Case et Smith 1983, Case, Chen & Jain 1992, voir aussi[28] Osherson, Stob et Weinstein 1986, Zeugman 1987, Kurtz et Smith 1989) pour suggérer des relations entre la conscience (au sens large présenté ici) et l'intelligence (au sens de Binet 1911, voir aussi Gould 1983).

Je montre que les théories intensionnelles S4Grz, KD?, KD?* ..., (reposant sur les logiques de l'autoréférence G et G*), appartiennent aux discours stables, partiellement communicables, partiellement inférables, des machines dans les voisinages de l'infini.

Pour les inférences instinctives immédiates, plus proche du stratagème *affaibli*, je pense que Daley, Kalyanasundaram, Velauthapillai 1992, et d'une façon générale l'identification en un nombre déterminé de coups (de changements d'avis) est opportune.

On a encore que si la machine peut se *savoir* au moins Σ_1-complète, alors elle ne peut pas se *savoir* au plus Σ_1-complète (c-à-d computationnellement *équivalente* à une machine universelle de Turing, voir Kelly 1993 pour un résultat similaire). Elle peut indirectement *croire à* ou inférer (à ses risques et périls et sur base d'un acte de foi) l'existence d'un niveau adéquat correct pour l'autoduplication, à partir de considérations empiriques.

D'où peut bien provenir de telles considérations ? C'est ce que la troisième partie aborde (entre autres).

3. VERS UNE FORMULATION ARITHMETIQUE DU PROBLEME DU CORPS ET DE L'ESPRIT

17) J'aborde l'état de rêve. Cela n'est pas logiquement nécessaire pour la suite. Cela sert à la fois : 1) d'illustration, encore très approximative, de l'usage de G&G* au sujet de cet état de conscience particulier : l'état de rêve (voir aussi Marchal 1992) et : 2) d'introduction *à*, et de motivation *pour*, la thèse de la supervénience de la conscience sur l'activité physique du cerveau ; thèse dont je réfute cependant la compatibilité avec MDI en 3.2.

Je critique une théorie des rêves de Malcolm 1959, il s'agit d'une théorie positiviste dans l'esprit du *jeune* Wittgenstein 1918.

[28] Ici aussi se sont développées des écoles indépendantes à l'Est et à l'Ouest.

La motivation par le souvenir du rêve pour le stratagème de Théétète est rediscutée ainsi que le rôle du rêve dans le *cogito* de Descartes[29].

Je propose encore une illustration de l'inéluctable doute que rencontre le mécaniste au moyen de la fiction de Galouye 1964 : *Simulacron 3* et au moyen de la notion de *réalité artificielle* ou *virtuelle* (dont le rêve lucide est une version *naturelle*[30]).

Outre une confusion entre $\Box$ et $\boxdot$, Malcolm n'a pas anticipé que la proposition "je rêve" est, dans certaines situations, *relativement vérifiable*, ce qu'ont montré Hearne en 1978, ainsi que Laberge 1980 avec la preuve expérimentale de l'existence de rêves lucides (voir Hearne 1990, et Laberge 1991). De ce fait MDI est presque entièrement compatible avec la conception des rêves des onirophysiologues contemporains :

pour autant qu'on admet être conscient lors de la vie éveillée, on peut dire :

a) on est conscient lors des rêves ;
b) on peut savoir que l'on rêve dans un rêve (lucidité) ;
c) on peut communiquer que l'on rêve lors d'un rêve
(Hearne 1990, Laberge 1991) ;
d) on ne peut pas justifier, en éveil, que l'on est éveillé (LWV) ;
e) on peut mesurer l'éventuel décalage entre le temps interne d'un rêveur et le temps du labo (Laberge 1991).

Dans ce cas, il semble tout naturel de penser que la conscience est (causalement) reliée à l'activité physique, spatio-temporelle, du cerveau (ou du corps ou de l'univers ou de l'éventuel *simulacron*, selon le niveau de fonctionnalisme). On se rappelle, qu'après tout, le mécanisme est d'inspiration matérialiste. Je vais pourtant mettre en évidence l'inadéquation entre la thèse de la supervénience physique (ainsi que les thèses d'identité corps-esprit afférentes) et la thèse du mécanisme indexical et digital, MDI. Cela fait l'objet du *paradoxe du graphe filmé*.

18) La matière est-elle matérielle ? J'ai développé jusqu'ici une approche purement *spiritualiste* ou *immatérialiste* de l'esprit. L'approche est immatérialiste en ce sens que les propositions concernées sont essentiellement des propositions arithmétiques structurées intensionnellement (modalement). Elles sont de fait filtrées par différents points de vues extraits des logiques de l'autoréférence (G et G*).

Cette approche ne repose pas sur des propositions physiques, et cela est dû en partie au fonctionnalisme (du niveau adéquat) et en partie au digitalisme (la thèse de Church[31]).

Par ailleurs, avec MDI, le sujet et sa conscience sont *immatériels* dans le sens qu'ils sont capables de *changer de corps*, de survivre à la greffe *totale*.

En outre, les paradoxes de l'automultiplication de soi ont déjà non seulement montré le caractère immatériel du sujet mais aussi le caractère *non local* de la conscience : un réel calcul de probabilité de survie nécessite la prise en compte de tous les éventuels accidents "reconstitutifs" lointains[32]. Et s'il l'on pense que de tels accidents sont rares dans un petit

[29] Platon invoque aussi le rêve dans le Théétète.

[30] Le rêve lucide constitue une *réalité artificielle naturelle*, mais la *réalité artificielle artificielle*, c-à-d avec un ordinateur à la place d'un cerveau permet d'illustrer tout autant le présent propos. Cette remarque illustre aussi le caractère artificiel du terme artificiel.

[31] Je ne partage pas l'avis de Deutsch 1985 lorsqu'il dit que la calculabilité dépend des lois de la physique. Avec d'autres lois physiques, prétend-il, l'addition pourrait ne pas être calculable. L'addition peut ne pas être calculable (dans un modèle non standard de l'arithmétique par exemple), mais ici on reste dans le modèle standard. On peut défier Deutsch en lui demandant de décrire une situation physique assez riche que pour supporter un observateur (capable au moins de distinguer deux états) dans laquelle l'addition standard est non calculable. S'il y parvenait il réfuterait la partie "évidente" de la thèse de Church (qui dit que les fonctions Turing calculables sont intuitivement calculables, *abstraction faite* des disponibilités physiques). J'ajoute cependant que Deutsch 1985, et sa machine quantique (que l'on peut "améliorer" avec une "delayed choice instruction" (Marchal 1978, Wheeler 1980) est un instrument pédagogique saisissant pour se faire une idée de ce que l'interprétation d'Everett a de remarquable. Ceci sera encore plus clair lorsque de telles machines verront le jours.

[32] ...y compris les accidents hors du *cône de lumière* de l' "original", y compris d'éventuels accidents appartenant au passé, ou dans n'importe quel *autre* univers concevable, etc.

univers d'une cosmologie classique, il en est déjà tout autrement dans une cosmologie quantique (par exemple en suivant l'interprétation d'Everett).

Cependant, malgré l'immatérialité du sujet, un corps *matériel* semble au moins indispensable pour que le sujet, sa conscience, puisse se manifester localement et relativement à un environnement (lui-même matériellement conçu).

De même, lorsqu'on se représente l'activité d'un programme, son exécution est généralement pensée comme incarnée ou véhiculée par une machine matérielle *(physique)* plongée dans un environnement matériel.

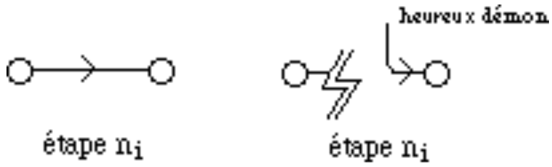
Mais voilà : le paradoxe du graphe filmé (PGF[33]), qui généralise les paradoxes de la multiplication de soi, montre de façon directe l'impossibilité d'user de la physicalité pour *singulariser* et/ou *actualiser* l'expérience subjective de la conscience (sauf à abandonner MDI) :

$$\text{PGF} \Rightarrow \neg \text{MDI} \vee \neg \text{SUP-PHYS}$$

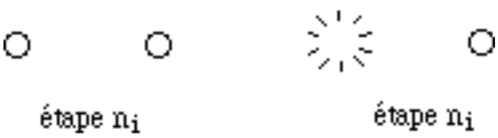
Le principe du paradoxe du graphe filmé est le suivant. Admettons la thèse de la supervénience physique (SUP-PHYS). On doit admettre alors, avec MDI, que la conscience supervène sur l'activité computationnelle et physique d'une machine lors d'une éxécution particulière.

Observons l'activité d'une machine M au niveau où le fonctionnalisme est correct. A ce niveau aucune entité active n'est supposée capable de cognition, encore moins de précognition (voir aussi Dennett 1978).

On doit dès lors admettre que la conscience supervène sur l'éxécution particulière de la machine B, laquelle est identique à A **sauf** qu'un certain nombre d'entités sont défaillantes lors des étapes n_i, **mais où on suppose, en plus**, qu'une collection d'*heureux rayons cosmiques* suppléent à l'activité défaillante.



De même, si certaines entités E de la machine A ou B ne sont pas utilisées lors de cette éxécution particulière ("démons *absence-de-pièce*"), la conscience supervenant sur cette exécution particulière de M doit supervenir sur l'activité de C, qui est A ou B dont on a retiré les entités E :



Les éxécutions des machines transformées sont *accidentellement* correctes. En combinant ces deux transformations il est possible cependant de restituer[34] (Marchal 1988, Maudlin 1989) le caractère contrafactuellement correct de la machine sans rien modifier à l'activité physique de la machine C, activité qui peut être elle-même pratiquement réduite à l'absence totale d'activités.

---

[33] C'est l'argument capital. Une bonne expérience par la pensée (que j'utilise dans Marchal 1988) est *la conversation avec le cerveau d'Einstein* de Hofstadter (Dennett & Hofstadter, 1981).

[34] Cette restitution est proposée explicitement par Maudlin 1989. En particulier Maudlin utilise les démons *absence de pièce* à cette fin.

De cette façon, PGF (et le raisonnement de Maudlin) interdit d'attacher la conscience (ni l'éventuel "flux de la conscience") à l'état physique d'une machine (à l'activité physique de cette machine).

PGF dénie donc la thèse de la supervénience physique ainsi que la thèse de l'identité forte, laquelle associe (par exemple) :

(le quale de la douleur) *à l'instant t* avec *l'état physique à l'instant t*

Mais PGF ne dénie peut-être[35] pas la thèse de la supervénience computationnelle (SUP-COMP), celle-ci associe (pour suivre le même exemple) :

le quale de la **(douleur à l'instant t)** avec *un état computationnel relatif*

L'*état computationnel* doit être pensé comme plongé dans une structure, qu'il faut analyser, d'états computationnels possibles. Un tel *nuage d'états* n'a rien de temporel, ni de matériel a priori. "Possible" ici est un terme modal qui devra être précisé par des relations d'accessibilité ou de voisinage, etc. La probabilité (ou les croyances) que l'on cherche à dériver de l'automultiplication de soi doivent, avec PGF, être globalement définies sur cette structure d'états computationnels.

Remarquons que la matière, en tant que substance indépendante, n'a pas *encore* nécessairement disparue. Son concept ne peut toutefois ***plus***, avec PGF, être utilisé pour actualiser la conscience : MDI => $\neg$ SUP-PHYS.

Du fait que PGF exige[36] une justification des crédibilités ou des probabilités immédiates et donc aussi la logique des discours physicalistes des machines concernant ces crédibilités et ces probabilités immédiates (dans les voisinages de zéro), le concept de matière (comme substance indépendante et ontologiquement première), avec MDI, est, par PGF, *nécessairement* superflu, un peu comme le *principe vital* est superflu en biologie[37] .

La thèse de la supervénience computationnelle doit donc être fausse (ainsi que MDI) ou à même de justifier la nature observable et l'apparence empirique de la matière : elle doit justifier l'observation immédiate.

On peut notamment espérer, étant donné le décalage entre G et G*, une justification de la part communicable (et *en principe* falsifiable) de l'observation immédiate (la mesure de la physique) ainsi qu'une justification de la part incommunicable de l'observation (le quale de la sensation physique).

En résumé : 1) la conscience ne supervène pas sur l'activité spatiotemporelle d'une machine mais sur l'existence relative d'états computationnels et 2) MDI oblige, par PGF, a élaboré une phénoménologie de la matière à partir de la "vérité arithmétique".

Il reste donc à chercher cette structure des états computationnels qui doit être suffisamment contraignante que pour extraire "*l'empirisme physique*" de la supervénience computationnelle. Cela revient, pour le problème du corps et de l'esprit à construire une phénoménologie de la matière à partir de *lois de l'esprit* (G&G*), plutôt que l'inverse.

La détrivialisation du mécanisme entraîné par la thèse de Church et les phénomènes d'incomplétude suggèrent l'existence d'un chemin pour isoler une telle structure.

---

35 PDU, plus loin, aborde cette question.

36 Ici je suppose que MDI => (SUP-PHYS v SUP-COMP).

37 On doit à Helmholtz une déclaration de guerre contre le vitalisme défendu par Driesche. On trouvera dans de Brouckère 1982 un compte rendu de la fin du vitalisme en biologie. La matière disparaît dès ici, si on admet une certaine forme conceptuelle de rasoir d'Occam.

19) Le problème du corps et de l'esprit[38] est la recherche de l'explication du hiatus manifeste qui existe entre la connaissance publique (communicable) sur un objet (comme le corps et la matière) et la connaissance privée (pas nécessairement communicable) d'un sujet (comme sa conscience, son état d'esprit, ses *qualia*).

Le problème étant posé de cette façon, la différence entre G et G* suggère une explication évidente puisque G est la logique du communicable (de la part de la machine autoréférentiellement correcte) et G*\G la *logique*[39] de l'incommunicable (mais néanmoins inférable ou "pariable"). Une telle solution directe ne marche pas pour deux raisons évidentes :

a) le sujet est décrit par des extensions de KT ou de S4 avec nécessitation.
b) pour l'objet (immédiatement appréhendable) on ne peut pas avoir 4.

G* est une logique du *pur* esprit. Pour trouver des probabilités ou des crédibilités, qui s'appliquent à de l'empirique, j'ai déjà proposé d'appliquer le stratagème affaibli (darwinisme arithmétique). On se limite encore au probabilité particulière, "sûr", de type $P_{trans}=1$.

PGF impose alors de prendre, pour la structure computationnelle sur laquelle supervène la conscience (SUP-COMP), la collection de tous les états computationnels <u>transitoires,</u> *structurée* par le stratagème affaibli, <u>et directement accessibles</u>, ce que le Dovetelleur Universel (DU) permet d'illustrer.

Le dovetelleur universel (DU) illustre essentiellement la notion d'états computationnels ***immédiatement*** accessibles (du point de vue phénoménologique).

DU est un programme qui éxécute tous les programmes. Cela comprend toutes les interactions possibles entre programmes et entre programmes et oracles[40]) dans tous les langages de programmation possibles.
L'existence d'un tel programme (donné dans l'annexe 4) est une conséquence de la thèse de Church, ainsi que la machine-indépendance des probabilités globales (dans les voisinages de l'infini) par rapport au choix initial d'une numérotation acceptable de Rogers 1958).
La routine majeure d'un tel programme est le dovetellage.

Le paradoxe du dovetelleur universel (PDU), avec MDI, provient du trop-plein a priori des extensions calculables accessibles : l'ensemble des extensions contient (avec MDI) tous les états hallucinatoires possibles[41], et semble empêcher toute inférence inductive de type empirique, physique et/ou probabiliste : comment justifier la *normalité* des extensions usuelles, de la vie de tous les jours, comment justifier l'intersubjectivité (la communicabilité) résiduelle ? La supervénience arithmétique exigée par PGF semble a priori trop large[42].

Trop large ? Voire. Le stratagème affaibli *affaiblit* la logique G des états "directement accessible" et donne l'espoir d'exhiber une topologie non triviale sur ces états.

Cependant PGF indique encore, qu'**avec DU**, un **enrichissement de G** est inévitable, et donc aussi un appauvrissement des modèles et des mondes.

---

38 Voir Warner & Szubka 1994 pour une sélection du débat *actuel* (1994).
39 G*\G est fermé pour l'application de la règle du modus ponens. En tant que logique, G*\G est assez curieuse, elle ne contient aucune tautologie (classique ou intuitioniste), ni aucun théorème de K, etc. Le terme "cologique" serait peut-être plus approprié ici, mais je crains qu'il ne soit trop barbare.
40 Le dovetelleur dovetelle sur l'éventail des segments initiaux des oracles.
41 C'est une sorte de *paradoxe d'Olbers* de la conscience.
42 PDU est une instantiation d'un paradoxe plus général qui apparaît avec les réalisme modaux (voir Lewis 1986). Notons que PGF impose un réalisme modal arithmétique dès qu'on admet que la vérité arithmétique est indépendante de soi (ce qu'on admet, par exemple, pour l'usage usuel (non égoïste) d'une assurance-vie).

En effet DU met en évidence l'*effectivité* de la notion d'état computationnel : le fait d'atteindre, même à partir d'un oracle, un état, est assimilable à la preuve formelle d'une proposition Σ_1.

Avec la thèse de Post-Turing, *être une machine*, c'est, essentiellement, être **au plus** ***Σ_1.-complet***.

L'arithmétique de Peano est déjà un dovetelleur universel[43]. Si on regarde une exécution de DU par un ordinateur plan, sa dynamique est capturée en trois dimensions par une sorte de cône digital. Ce cône code (de façon externe) l'ensemble de toutes les propositions Σ_1.

Les propositions *directement accessibles* qui nous concernent sont donc les propositions Σ_1 (c-à-d encore, avec la thèse de Post-Turing, les *observables* de Vickers 1989). Remarquons qu'il existe pour chaque état une infinité de successeurs par lequel DU passe tôt ou tard. Avec DU, il y a une infinité d'*accidents lointains*.

Le DU peut mettre énormément de "temps" pour générer un état sur lequel la conscience accède immédiatement, comme dans les paradoxes de duplication postposée. On se rappelle qu'avec PGF, le temps subjectif de l'expérience fait partie intégrante du quale de l'expérience.

Il faut donc faire bien attention de ne pas confondre le "temps" arithmétique externe (une mesure arithmétique de complexité) avec la topologie du directement accessible, du nuage des états computationnels *indexicalement voisins*.

On a pratiquement isolé cette topologie, en effet :

La part *communicable* de la logique des "propositions physiques" est donnée par

- l'application du stratagème affaibli sur G,
- restreint aux propositions Σ_1.

La part *non communicable* de la logique des "propositions physiques" est donnée par

- l'application du stratagème affaibli sur G*\G,
- restreint aux propositions Σ_1.

On pourrait craindre de verser dans l'*idéalisme subjectif*, mais grâce à nouveau au décalage entre G et G*, on est à même de distinguer la part subjective incommunicable (la sensation physique) de la part objective et communicable (la mesure physique entre machines donc).

Encore faut-il que ces deux parties soient distinctes. On peut s'en douter puisque KD? et KD?* diffèrent déjà. Mais ici il faut tenir compte de la restriction aux formules Σ_1.

20) AppelonsV[44] la restriction de G (ou sa version arithmétique) sur les formules Σ_1.

Comme pour p Σ_1 :

[43] En 1987, j'appelais PDU le paradoxe RE (pour récursivement énumérable). C'est la thèse de Church qui donne un statut important à cette classe d'ensemble, et qui détrivialise sa structure, comparée par exemple à la bibliothèque de Babel de Borges 1951.

[44] V pour Visser (voir Boolos 1993). Je bénéficie pour cette récapitulation de l'ouvrage de Boolos 1993, qui expose le travail de Visser (la complétude de V et de V*, = G* avec la réalisation F restreinte aux formules Σ_1).Visser a développé une sémantique de Kripke pour V. Je craignais que cela soit impossible car je ne voyais pas comment tenir compte de l'absence de fermeture pour la substitution (due au fait que p-> $\Box$ p n'est vrai (et prouvable) que pour p Σ_1), et cela malgré des suggestions de Smorynski 1985). Il suffit en fait de rendre les valeurs de vérités, des variables propositionnelles, héritables via la relation d'accessibilité (l'observable est persistante). Il est facile de modifier le démonstrateur de théorèmes de G et G* pour obtenir un démonstrateur de théorèmes de V et V*.
Par ailleurs ayant moins de mondes et de modèles à vérifier, il sera plus efficient que le démonstrateur de G et de G*.
Notons que Boolos 1993 appelle G et G*, GL et GLS respectivement. De même il appelle V et V*, GLV et GLVS respectivement. C'est mis pour Gödel, Löb, Visser, et Solovay.

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

V étend G + $p \to \Box p$ (d'où l'enrichissement). Visser a montré en plus que la théorie G + $p \to \Box p$, avec MP et Nec, est arithmétiquement complète pour la prouvabilité restreinte aux formules Σ_1.

Il est nécessaire de restreindre la portée du schéma "$p \to \Box p$" sur les variables propositionnelles. Par exemple, on ne peut pas substituer *p* par $\Diamond\top$, pour obtenir $\Diamond\top \to \Box\Diamond\top$! Arithmétiquement parlant "$\Diamond\top$", c-à-d $MB_F(\Diamond\top)$, est Π_1.

On applique alors, pour extraire la part communicable d'une logique de l'observable, le stratagème affaibli sur V. Cela donne KD?-Σ_1 :

$$KD?\text{-}\Sigma_1 = \{p \text{ telles que } V \vdash DEON(p)\}$$

où DEON est la transformation, définie sur et à valeur dans l'ensemble des formules modales, correspondant au stratagème affaibli défini plus haut.

C'est aussi la logique de la prouvabilité intensionnelle (et encore arithmétisable), $B(\ulcorner p \urcorner)$ & $\neg B \neg (\ulcorner p \urcorner)$, où p est limité aux propositions Σ_1.

Plus précisément, avec F restreint aux formules Σ_1,

$$KD?\text{-}\Sigma_1 = \{p \text{ telles que } \forall F\ M \vdash MB_F \circ DEON(p)\}$$

De même il est naturel, pour mettre en évidence la part incommunicable de la logique des observations, de définir avec le morphisme de Magari-Boolos :

$$KD?\text{-}\Sigma_1{}^* = \{p \text{ telles que } \forall F\ MB_F \circ DEON(p) \text{ est vraie}\}$$

où les valeurs de F sont restreintes aux propositions Σ_1.

Il se fait qu'ici je peux encore profiter du travail de Visser (Boolos 1993) :

$$KD?\text{-}\Sigma_1{}^* = \{p \text{ telles que } V^* \vdash DEON(p)\}$$

où V* = tous les théorèmes de V + le schéma $\Box p \to p$, avec MP pour seule règle. V* est donc le correspondant de G* pour la restriction à Σ_1 et Visser a démontré la complétude arithmétique de V*.

KD?-Σ_1 donne ainsi la logique des propositions immédiatement vérifiables et communicables, ce sont les *propositions physiques* de base, on peut les interpréter, avec MDI comme les croyances physiques quantifiables et *en principe* falsifiables. Ptrans = 1, est capturé par $\boxdot\Diamond\top$, formule vraie (heureusement avec MDI), mais non communicable, ce qu'on savait déjà avec la reconstruction de Benacerraf.

De même, pour la proposition *je vais survivre à la translation*, on savait déjà qu'il s'agissait d'un pure *quale*, personne n'est obligé de croire le survivant, bien qu'au niveau adéquat de substitution celui-ci peut l'observer *immédiatement* de l'intérieur.

Par analogie avec la relation entre S4Grz et IL, ceci inspire la recherche d'une transformation TRANS de LP dans LPM, interprétant *intensionnellement* l'observation idéale (immédiatement répétable) *p* par $\boxdot\Diamond p$.

Cela pourrait mettre en évidence deux logiques faibles (internes) dont KD?-Σ_1 et KD?-Σ_1* sont des interprétations intensionnelles (externes et classiques).

Une transformation due à Goldblatt 1974, pour donner une interprétation modale de la logique quantique, suggère une telle transformation TRANS.

Cela fera l'objet du point suivant.

Le tableau suivant récapitule quelques résultats :

	G	G*	S4Grz(*)	KD?	KD?*	KD?-Σ_1	KD?-Σ_1*	S4Grz-Σ_1
T	-	+	+	-	+	-	+	+
Nec	+	-	+	-	-	-	-	+
Diag	+	+	-	+	+	+	+	-
N	+	+	+	-	+	-	+	+
RM	+	-	+	+	-	+	-	+
B	-	-	-	-	-	-	+	+
4	+	+	+	-	-	-	-	+
D	-	+	+	+	+	+	+	+
5	-	-	-	-	-	-	-	+
dual T	-	+	+	-	+	+	+	+

L'axiome K est vérifié dans tous les systèmes.

Dual T , la duale de T, distingue KD? de KD?-Σ_1,. on peut se convaincre cependant que KD?-Σ_1 $\vdash$ p->$\circledcirc$p, et KD? ne prouve pas cette formule[45].

La restriction aux formules Σ_1 avec le stratagème ***fort*** ne donne rien d'intéressant : S4Grz-Σ_1 ou S4Grz-Σ_1* donnent apparemment une interprétation arithmétique de S5, mais en fait le carré et le losange collapsent.

KD?-Σ_1*, quant à lui, *frôle le collapse*, puisque KD?-Σ_1* $\vdash$ p<->$\boxdot$p, en ce sens cette logique peut induire une confusion intensionnelle *pire* que celle du solipsiste bavard (?) (ou de sa version épistémique), il ramène tout à l'immédiateté[46].

Le collapse est cependant évité car KD?-Σ_1* $\nvdash$ $\boxdot$(p<->$\boxdot$p).
Notons que KD?-Σ_1 ne prouve ni p->$\boxdot$p, ni $\boxdot$p->p.

En particulier la différence entre KD?-Σ_1* \ KD?-Σ_1 est non vide. Cette différence donne la part *non communicable* de la logique des propositions immédiatement observables (physiques). Ici s'inscrit de façon naturelle une "logique" des qualias ou des *sensations* physiques.

On peut interpréter (mécanistement) la querelle entre Goethe et Newton sur la nature de la couleur par une confusion intensionnelle implicite entre KD?-Σ_1* et KD?-Σ_1. Cette confusion est bien sûr un cas particulier de confusion entre G et G*.

L'immatérialisme mécaniste est donc un idéalisme parfaitement compatible avec le réalisme minimal du platonisme arithmétique, c'est, *au moins* l'ontologie de celui qui prend une assurance-vie, ou de quiconque capable de croire à l'indépendance d'au moins un *autre*.

En résumé :

KD?-Σ_1 = l'observable communicable (mesurable)
KD?-Σ_1 = l'observable*
KD?-Σ_1 \ KD?-Σ_1 = l'observable non communicable (qualia)*

---

45 je refais l'abus de langage KD?-Σ_1 $\vdash$ p->$\circledcirc$p est mis pour V $\vdash$ p->$\circledcirc$p, où $\circledcirc$p est une abréviation de $\Box$ p v $\Diamond$p (p atomique).
46 A ce niveau, on retrouve en quelque sorte une phénoménologie du "esse est percipi" de Berkeley "1970".

Pour résoudre *concrètement* PDU et le problème mécaniste du corps et de l'esprit, il faudrait extraire de KD?-Σ_1(*), une interprétation arithmétique *complète* d'un calcul des probabilités (pas seulement P=1). Cela donnerait une notion de voisinage limite, à partir de laquelle on dériverait une notion de densité relative d'états computationnels transitoires (consistants). Et de là faut-il encore justifier la normalité.

Des considérations de profondeurs logiques (Bennett 1988) sont sans doute nécessaires et donnent des évidences pour dériver une phénoménologie de la cosmologie. De même sont nécessaires ici les notions de compressibilité et d'incompressibilité algorithmique (Chaitin 1987), et plus abstraitement de *simplicité* (Post 1944). Je n'ai pu qu'effleurer cette aspect de l'incomplétude (voir aussi Delahaye 1994, Svozil 1993).

Remarque On peut généraliser la réfutation de Webb de l'argument de Lucas. Webb montre que les arguments Gödéliens contre le mécanisme, une fois précisés, se retournent, grâce à la thèse de Church (utilisée par Lucas), en faveur du mécanisme. De même, s'il était possible d'utiliser PGF ou PDU, de façon communicable, contre le mécanisme, alors, avec la thèse de Church, on montrerait que les machines seraient capables de démontrer correctement qu'elles ne sont pas des machines (absurde).

21) J'ai plusieurs fois utilisé l'interprétation d'Everett de la mécanique quantique pour illustrer la multiplication de soi. Avec PGF et PDU, la phénoménologie de la matière est donnée ultimement par une mesure sur l'ensemble des états computationnels directement accessibles, et la mesure "sûre" doit obéir à une *logique* des "propositions" directement observables (KD-Σ_1*).

Il est naturel alors de se demander s'il existe une relation entre cette phénoménologie de la matière et la mécanique quantique, celle-ci étant la théorie *actuelle* de la matière dérivée empiriquement par les physiciens.

Avec PDU, il semble *a priori*, qu'il existe un trop-plein de mondes ou plus adéquatement, un trop-plein d'états.

Avec les phénomènes d'incomplétude de Gödel (mais on pourrait prendre aussi les théorèmes de Löwenheim-Skolem) nous savons que rajouter des axiomes à la logique ne permet pas d'isoler une réalité univoque. Il a mieux valu affaiblir la logique dans l'espoir d'arriver à définir une notion d'extension normale. Il reste à montrer ensuite que de telles extensions sont relativement denses ou *plus probables*. L'affaiblissement a été donné ici par le stratagème affaibli. Afin de se limiter aux propositions vérifiables ou *vraiment accessibles* PDU impose, cependant, un enrichissement de la théorie de l'autoréférence : de G on est passé à V, de G* on est passé à V*.

A présent, avec une hypothèse de *panmécanisme*, les états relatifs d'Everett sont des cas particuliers d'états Σ_1-descriptibles directement accessibles (avec MDI).

Se pourrait-il que l'extravagance de la multiplication des états chez Everett corresponde, pour notre approche introspective (bien que digitalement désindexicalisée[47]), à l'appauvrissement nécessaire ou à la normalisation recherchée des extensions transitoires ?

Il est raisonnable d'aborder cette question en comparant directement la logique quantique (des physiciens, cf Birckhof et Von Neumann 1936) à la phénoménologie du physique, donnée par KD-Σ_1* pour les perceptions et les mesures et par KD-Σ_1 pour la part exclusivement communicable.

On distingue :

-la logique **MQL,** *minimal quantum logic*, qui est essentiellement caractérisée (algébriquement) par le treillis des sous-espaces d'un espace vectoriel muni d'un produit scalaire convenable (espace de Hilbert),

- de la logique **QL**, qui est MQL + Ortho, une formule d'orthomodularité (Birckhof et von Neumann 1936, voir aussi l'annexe 5) :

47 ... et rendue ainsi relativement communicable (avec MDI).

$$(p\text{->}q)\text{->}(q\text{->}(p \vee (\neg p \;\&\; q)))$$

KD?-Σ_1* se rapproche de la logique quantique dans le sens suivant :

Goldblatt (1974) a montré :

$$B \vdash \text{GOLDB}(p) \text{ ssi } MQL \vdash p$$

où B = KT+p->$\Box\Diamond$p, avec MP et Nec. Par abus de langage "B" désigne encore la formule : p->$\Box\Diamond$p. GOLDB est une transformation du langage propositionnel non modal dans le langage propositionnel modal (Goldblatt 1974, voir aussi Dalla Chiara 1977) :

GOLDB(p) = $\boxdot\Diamond$p pour p atomique
GOLDB($\neg$ A) = $\boxdot\neg$ GOLDB(A)
GOLDB(A->B) = "GOLDB(A) => GOLDB(B)" (*)
GOLDB(A&B) = GOLDB(A) & GOLDB(B)

De même

$$B\text{+MOrtho} \vdash \text{GOLDB}(p) \text{ ssi } QL \vdash p$$

où MOrtho = GOLDB(ortho) = (p & $\neg$ q) -> $\Diamond$(p & $\boxdot\neg$ (p & q)). C'est une version modale de l'orthomodularité. MOrtho doit être restreint aux orthopropositions (image-inverse de GOLDB).

La ligne (*) mérite une explication puisque "=>" n'est pas un symbole logique. Retenons seulement que B $\vdash$ a=>b est mis pour "B est fermé pour la règle a => b". Cette complication est due au fait que la logique quantique n'obéit pas au théorème de déduction. En particulier MQL est fermée pour la règle p=>(q->p), mais MQL $\nvdash$ p->(q->p). Pour un traitement plus propre on rajoutera un symbole correspondant à "=>" et on donnera une sémantique[48] convenable (voir Dalla Chiara 1975, 1977).

Donc, à l'instar de la logique intuitioniste qui admet une traduction classique modale (par S4 (Gödel 1933), ou S4Grz (Grzegorczyk 1967), la logique quantique admet une traduction modale par B.

Les formules B et T ne sont pas des théorèmes de KD?-Σ_1, mais bien de KD?-Σ_1*. Cependant, à l'inverse de B, KD?-Σ_1* n'est pas fermé pour la nécessitation. KD?-Σ_1* partage donc avec la logique B tous les théorèmes que l'on peut démontrer sans nécessitation. Il s'agirait plus d'une sorte de crédibilité quantique que de probabilité quantique[49]. Pour évaluer la ressemblance avec QL on doit regarder si MOrtho est arithmétiquement dérivable :

$$\text{KD?-}\Sigma_1(*) \vdash (p \;\&\; \neg q) \text{->} \Diamond(p \;\&\; \boxdot\neg (p \;\&\; q))$$

Il n'existe cependant pas d'unanimité sur la logique quantique. Des affaiblissements de la notion d'orthomodularité existent (voir aussi Mittelstaedt 1978). Il est donc nécessaire de nuancer l'évaluation. La question est de savoir dans quelle mesure KD?-Σ_1 et KD?-Σ_1* éclairent le *labyrinthe des logiques quantiques* (van Fraassen 1974).

---

48 Ou sans doute mieux encore, on peut éviter l'usage de l'implication et utiliser une présentation à-la-Gentzen de logique quantique (comme Goldblatt 1974).

49 Dans ce contexte, les travaux de Conway et Moore, décrits dans Svozil 1993, semblent tout à fait opportun. L'idée d'extraire le "quantique" du "computationnel" apparaît aussi dans différents articles de Kafatos 1989. L'utilisation d'un axiome de modularité pour le problème du corps et de l'esprit semble avoir été envisagé par Watanabe (Jammer 1974).

De façon plus précise, **Quel**le est la **L**ogique **Q**uantique, **QuelQL**, qui axiomatiserait l'ensemble des formules p telles que pour les réalisations F restreintes aux formules Σ_1 :

$$M \vdash MB_F \circ DEON \circ GOLDB(p)$$

On peut poser d'autres questions : la "matière" (ou la logique des propositions Σ_1 *prouvable-et-consistante*) viole-t-elle les inégalités de Bell ?

La formule suivante (tirée de D'Espagnat 1979),

$$A \cap B \subseteq (A \cap C) \cup (B \cap \neg C) \qquad \text{BE-D'E}$$

représente une forme d'inégalité de Bell (Bell 1965, Voir Bell 1987, voir aussi d'Espagnat 1971, voir annexe 5). Voilà une version modale forte :

$$\boxed{\circ}\Diamond A \& \boxed{\circ}\Diamond B \text{ -> } (\boxed{\circ}\Diamond A \& \boxed{\circ}\Diamond C) \vee (\boxed{\circ}\Diamond B \& \boxed{\circ}\neg\boxed{\circ}\Diamond C) \qquad \text{BE}^{50}$$

L'éventuelle violation de BE devrait permettre de définir des notions d'observables, incommensurable, ou incompatibles.

Si KD?-Σ_1* $\nvdash$ BE, alors KD?-Σ_1 $\nvdash$ BE. C'est-à-dire si la conscience ne supervène pas localement (ce qu'on sait depuis 1.3 avec MDI), alors il existe des faits mesurables et communicables non locaux. L'expérience d'Aspect confirmerait MDI[51].

DEUX FORMULES

Avec MDI, la logique de la vérité sur les autoréférences correctes constitue une *théorie de l'esprit* capturée (partiellement par G* et d'éventuelles extensions).
G (et d'éventuelles extensions) capture la part communicable (de façon autoréférentiellement correcte par une machine) de cette théorie.
G*\G joue le rôle d'une sorte de réservoir subjectif partiellement inférable.

On obtient une **logique du sujet** (incorrigible, innommable et solipsiste) en appliquant le stratagème fort sur G (ou G*) (première formule) :

$$IL \vdash p \Leftrightarrow M \vdash MB_F \circ BGKM \circ G33(p) \Leftrightarrow MB_F \circ BGKM \circ G33(p)$$

Ceci est inférable (ou pariable) par les machines elles-mêmes (pourvu qu'elles soient autoréférentiellement correctes) dans les voisinages de l'infini.

On obtient une **logique de l'objet** en appliquant le stratagème affaibli sur G* restreint aux propositions vérifiables, c'est-à-dire, avec TC, sur les propositions Σ_1 (deuxième formule) :

$$QuelQL^* \vdash p \Leftrightarrow MB_F \circ DEON \circ GOLDB(p)$$

Dont la part communicable est donnée par QuelQL (deuxième formule *bis*) :

---

[50] 2 SUN travaillant 2 jours sur une modification bâclée des démonstrateurs de G et G* (annexe 2) ne m'ont pas permis de conclure. La question de savoir si la phénoménologie de la matière extraite ici du mécanisme viole les inégalités de Bell, reste ouverte à ce stade.
[51] Dit autrement : si l'ordinateur KILL-THE-USER (voir 1.2, 1.3) fonctionne, alors l'ordinateur quantique de Deutsch fonctionne, ou plus généralement, si le mécanisme (la non-localité) est confirmable ironiquement, alors il existe une confirmation non ironique de cette non-localité.
Les qualias ne sont pas la seule magie de la supervénience de la conscience sur une infinité de machines ou d'états computationnels; l'aspect *infinité de machines* est utilisable en pratique (ce que montre la machine de Deutsch 1985).

$$QuelQL \vdash p \Longleftrightarrow M \vdash MB_F \circ DEON \circ GOLDB(p)$$

F est restreinte aux propositions Σ_1.

Cette logique de l'objet, qui décrit des probabilités dans des voisinages immédiats de zéro, est encore inférable par les machines (auto-référentiellement correctes) dans les voisinages de l'infini.

TABLEAU

<u>Autoréférence</u>
G = le communicable (formellement justifiable)
G = le vrai sur le communicable et l'incommunicable*
G \ G = l'incommunicable, mais vrai (le réservoir subjectif)*
<u>Stratagème fort</u>
S4Grz = S4Grz = le connaissable (avec une nuance temporelle)*
IL = IL = le solipsiste*
S4Grz \ S4Grz = IL* \ IL = {}*
<u>Stratagème affaibli</u>
KD? = la croyance (communicable et révisable)
KD? = KT? = la connaissance intuitive (incorrigible)*
KD? \ KD? = la connaissance incommunicable (le sentiment)*
<u>Stratagème affaibli sur la restriction à Σ_1</u>
KD?-Σ_1 = l'observable communicable (mesurable)
KD?-Σ_1 = l'observable*
KD?-Σ_1 \ KD?-Σ_1 = l'observable non communicable (qualia)*
<u>Stratagème affaibli sur la restriction à Σ_1</u> (vision interne)
QuelQL = la logique de la physique
QuelQL = la logique du physique*
QuelQL \ QuelQL = la "logique" de la sensation physique*

Remarques

1) Il n'est pas vraiment nécessaire d'identifier l'ontologie spirituelle avec la vérité arithmétique (comme il m'arrive quelque-fois). On peut voir la *théorie de l'esprit* présentée ici comme étant elle aussi une phénoménologie bâtie sur l'auto-référence.

De cette façon, on se rapproche, semble-t-il, de Spinoza et Einstein[52]. L'esprit ***et*** la matière deviennent des modalités reposant sur une ontologie plus ultime, laquelle *peut* être identifiée avec la vérité, non axiomatisable, et moniste[53], de l'arithmétique.

2) Il est souvent tenu pour évident qu'il est impossible de savoir pourquoi nous existons. Cette évidence repose sur le fait qu'une explication repose sur des prémisses qui demandent elles-mêmes une explication, et ainsi de suite. Mais alors on voit que pour concevoir cette impossibilité, il est nécessaire d'avoir la conception (non communicable) de la litanie des naturels (le "ainsi de suite").

---

52 Al'occasion du tricentenaire de la naissance de Spinoza, Einstein écrira dans une lettre à la société Spinoza d'Amérique (publiée en traduction Anglaise dans le *New York Times* le 4 novembre 1932) : "*Pour Spinoza, le psychique et le physique ne sont que des formes phénoménales différentes, régies par les mêmes lois, d'une seule réalité*" (Einstein "1991").

53 Post 1921, semble avoir entrevu ce monisme immatérialiste, mais pour l'avoir confondu, me semble-t-il, avec l'idéalisme subjectif de Bergson, il a adopté(en 1924) une philosophie moniste. Voir Davis 1965 page 431, footnote 118.

Cependant, avec le mécanisme digital et indexical, ce qu'il est nécessaire pour concevoir ce "ainsi de suite" est suffisant pour comprendre pourquoi il est nécessaire que les machines (arithmétiquement et relativement définies) se posent des questions sur leur nature et leur origine dans les voisinages de l'infini, et croient à une physicalité (au moins locale), avec ses parts communicable et incommunicable (cf aussi Minski 1968).

La portée explicative du mécanisme est d'autant plus grande, qu'il n'est pas possible de réduire l'infinité des naturels, à une quelconque logique ou théorie *purement finie*. Comme on s'en doute depuis Dedekind 1888, comme c'est clair depuis Gödel 1931.

MDI, permet ainsi d'extraire à la fois une phénoménologie de l'esprit et une phénoménologie de la matière, mais laisse *nécessairement* intact le *mystère* de notre croyance dans les nombres naturels.

Bibliographie locale

Les auteurs marqués par * *ne sont pas* cités dans la bibliographie générale. Seule cette "récapitulation" a bénéficié de ces articles ou livres.

***ALECHINA** N., **1994**, *Logic with Probabilistic Operators,* 1-13, preprint.

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.

BELL J., **1987**, *Speakable and unspeakable in quantum mechanics,* Cambridge University Press, Cambridge.

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

BENNETT C. H., **1988**, *Logical Depth and Physical Complexity*, in Herken R. (ed), *The Universal Turing Machine A Half-Century Survey*, Oxford University Press.

BERGSON H. **1939**, *Matière et Mémoire*, Presses Universitaires de France, Paris.

BERKELEY, "1970", *Trois dialogues entre Hylas et Philonous*, Trad. Franç. : A. Leroy, Aubier Montaigne, Paris.

BINET A., **1911**, *Nouvelles recherches sur la mesure du niveau intellectuel chez les enfants d'école,* L'année psychologique, 19, pp. 145-201.

BIRKHOFF G. and **VON NEUMANN** J., **1936**, *The Logic of Quantum Mechanics,* Annals of Mathematics, Vol 37, N° 4, pp. 823-843.

BOOLE G., **1854**, *The Law of Thought*, MacMillan. also Dover Publications, New-York, 1958.

BOOLOS G., **1979**, *The Unprovability of Consistency, 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., **1980**b, *On Systems of Modal Logic with Provability Interpretations*, Theoria, 46, 1, pp. 7-18.

BOOLOS G., **1980**c, *Provability in Arithmetic and a Schema of Grzegorczyk*, Fundamenta Mathematicae, 96, pp. 41-45.

***BOOLOS** G., **1993**, *The Logic of Provability*, Cambridge university Press.

BORGES J. L., **1951**, *Ficciones*, Emecé Editores S.A. Buenos Aires, Fiction, Trad. Franç. : P. Verdoye et Ibarra, Gallimard, Paris, 1957.

BROUWER L.E.J., **1905**, *Leven, Kunst en Mystiek* (La Vie, l'Art et le Mysticisme), Waltman, Delft. Cité dans van Stigt 1990 page 199.

BROUWER L.E.J., **1948**, *Consciousness, philosophy and mathematics*, Amsterdam, reprinted in Philosophy of Mathematics, P. Benacerraf & H. Putnam (eds), Cambridge University Press 1983,pp. 90-96, (2ème édition), première édition chez Prentice-Hall 1964.

BROUWER L. E. J., **1946-1951**, *Brouwer's Cambridge Lectures on Intuitionism*, Edited by D. van Dalen, Cambridge University Press, 1981.

CAILLOIS R., **1956**, *L'incertitude qui vient des rêves.* Gallimard, Paris.

CASE J., **1971**, *A Note on Degrees of Self-Describing Turing Machines*. Journal of the A.C.M., vol. 18, n° 3, pp 329-338.

CASE J., **1974**, *Periodicity in Generations of Automata*, in Mathematical Systems Theory. Vol. 8, n° 1. Springer Verlag, NY.

CASE J. & **SMITH** C., **1983**, *Comparison of Identification Criteria for Machine Inductive Inference*, In Theoretical Computer Science 25, pp 193-220.

CASE J., **CHEN** K., **JAIN** S., **1992**, *Strong Separation of Learning Classes*, in Jantke (Ed.) Analogical and Inductive Inference, Lecture Notes in Artificial Intelligence, N° 642, Springer-Verlag, Berlin.

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

CHELLAS B. F., **1980**, *Modal logic an introduction*, Cambridge University Press, Cambridge.

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.

DALLA CHIARA M. L., **1976**, *A General Approach to non Distributive Logics*, Studia Logica XXXV, pp. 139-162.

DALLA CHIARA, M. L., **1977**, *Quantum Logic and Physical Modalities*, Journal of Philosophical Logic, 6, pp. 391-404.

DALLA CHIARA M. L., **1986**, *Quantum Logic*, in "***Handbook of Philosophical Logic***", Vol III, Gabbay D. and Guenthner F. (eds.), D. Reidel Publishing Company, pp. 427-469, Dordrecht.

DALEY R., **KALYANASUNDARAM** B., **VELAUTHAPILLAI** M., **1992**, *The Power of Probabilism in Popperian FINite Learning*, in Jantke 1992.

DAVIS M. (ed.), **1965**, The Undecidable. Raven Press, Hewlett, New York.

DE BROUCKERE L., **1982**, *L'évolution de la pensée scientifique*, Culture Laïque, édité par la fédération des amis de la morale laïque, Maison de la laïcité, Bruxelles.

***DELAHAYE** J.P., **1994**, *Information, complexité et hasard*, Hermès, Paris.

DENNETT D. C., **1978**, *Brainstorms*, Harvester Press, Hassocks, Sussex, 1979.

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).

D'ESPAGNAT B., **1971**, *Conceptual Foundations of Quantum Mechanics*, Reading, Mass., Addison-Wesley (2ème Ed. 1976).

D'ESPAGNAT B., **1979**, *A la recherche du réel*, Gauthier-Villars, Paris.

DEUTSCH D., **1985**, *Quantum theory, the Church-Turing principle and the universal quantum computer*, Proc. R. Soc. A 400, pp. 97-117.

DOGEN, **1232-1253**, *Shôbôgenzô*, Editions de la différence, Argenteuil, 1980.

EINSTEIN A., "1991", *Albert Einstein, Oeuvres choisies*, N° 5, Sciences, éthique, philosophie, Editions du Seuil/Editions du CNRS, Paris.

EVERETT, III, H., **1957**, *"Relative State" Formulation of Quantum Mechanics*, Review of Modern Physics, Vol. 9, N° 3, pp. 454-462.(also in DeWitt and Graham, 1973).

***FATTOROSI-BARNABA** M. and **AMATI**, G., **1987**, *Modal Operators With Probabilistic Interpretations I*, Studia Logica, XLVI, 4, pp. 383-393.

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

FEFERMAN S., DAWSON Jr. J. W., KLEENE S. C., MOORE G. H., SOLOVAY R. M. VAN HEIJENOORT J. (Ed.), **1986**, *Kurt Gödel Collected Works*, Vol. 1, Publications 1929-1936, Oxford University Press, New York.

GALOUYE D., **1964**, *Simulacron 3*, Trad. Franç. : Editions J'ai lu, N° 778.

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.

GODEL K., **1933**, *Eine Interpretation des Intuitionistischen Aussagenkalküls*, Ergebnisse eines Mathematischen Kolloquiums, Vol 4, pp. 39-40, also in FEFERMAN & Al. 1986.

GOLD E.M., **1967**, *Language Identification in the Limit*. Information & Control 10, pp. 447-474.

GOLDBLATT R. I., **1974**, *Semantic Analysis of Orthologic*, Journal of Philosophical Logic, 3, pp. 19-35.

GOLDBLATT R., **1978**, *Arithmetical Necessity, Provability and Intuitionistic Logic*, Theoria, Vol 44, pp. 38-46.

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

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

GOULD S. J., **1983**, *La mal-mesure de l'homme*, traduction de l'américain par J. Chabert, Editions Ramsay, Paris.

GRZEGORCZYK A., **1967**, *Some relational systems and the associated topological spaces*, Fundamenta Mathematicae, LX pp. 223-231.

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

***HACKING** I., **1975**, *The Emergence of Probability*, Cambridge University Press, Cambridge.

HARTLE J.B., **1968**, *Quantum Mechanics of Individual Systems*, American Journal of Physics, Vol 36, N° 8.

HEARNE K., **1990**, *The Dream Machine*, The Aquarian Press.

HUGHES R. I. G., **1989**, *The Structure and Interpretation of Quantum Mechanics*, Harvard University Press, Cambridge.

HUMES D., **1739**, *A Treatise of Human Nature*, London. also Fontana/Collins, Glasgow, 1987.

JAMMER M., **1974**, *The Philosophy of Quantum Mechanics*, John Wiley & Sons, New York.

KAFATOS M., **1989**, (Ed.), *Bell's Theorem, Quantum Theory and Conceptions of the Universe*, Kluwer Academic Publishers.

KALMAR L., **1959**, *An Argument against the Plausibility of Church's Thesis*, in Constructivity in Mathematics, A. Heyting (ed), North Holland, pp. 72-80.

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

KELLY K., **1993**, *Learning Theory and Descriptive Set Theory*, Journal of Logic and Computation, Vol 3, n° 1, pp. 27-45.

KLEENE S. C., **1952**, *Introduction to Metamathematics*, North-Holland.

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.

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).

LABERGE S., **1991**, *Le rêve Lucide*, trad. Franç. : collection Rêvéveil, Oniros, Paris.

LADRIERE, J., **1957**, *Les limitations internes des formalismes*, E. Nauwelaerts, Louvain, et Gauthier-Villars, Paris.

LAO-TSEU, **1967**, *tao tö king*, trad. de Liou Kia-hway, dans Philosophes Taoïstes, bibliothèque de la pléiade, Nrf, Gallimard, Paris.

LEWIS D., **1986**, *On the Plurality of Worlds*, Basil Blackwell.

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

LUCAS J. R., **1959**, *Minds, Machines and Gödel*, Philosophy, 36, 1961, pp. 112-127.

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

MALCOLM N., **1959**, *Dreaming*, Routledge & Kegan Paul ltd.

MARCHAL B., **1978**, *Introductions au théorème de Bell*, preprint, ULB, Laboratoire de Cosmologie, Plaine des Manoeuvres, Bâtiment NO.

MARCHAL B., **1988**, *Informatique théorique et philosophie de l'esprit*. Actes du 3ème colloque international de l'ARC, Toulouse, pp. 193-227.

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., **1991**, *Mechanism and Personal Identity*, proceedings of WOCFAI 91, M. De Glas & D. Gabbay (Eds), Angkor, Paris, pp. 335-345.

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

MARKOV A. A., **1947**, *Impossibility of certain Algorithms in the Theory of Associative Systems*, Dokl. Acad. Nauk 55 (En Russe). Traduction Anglaise, pp. 583-586.

MAUDLIN T., **1989**, *Computation and Consciousness*, The Journal of Philosophy, pp. 407-432.

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.

MINSKI M, **1968**, *Mind, Matter and Model*, in Semantic Information Processing, Minski M. (ed.), Cambridge MIT Press.

MINSKI, M., **1985**, *The Society of Mind*, Simon and Schuster, New York.

MITTELSTAEDT P., **1974**, *Quantum Logic*, PSA 1974, R. S. Cohen et al. (eds), pp. 501-514.

MITTELSTAEDT P., **1978**, *Quantum Logic*, D. Reidel Publishing Company, Dordrecht.

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

MYHILL, J., **1964**, *Abstract theory of self-reproduction*. Views on General Systems theory, M.D. Mesarovic, Ed., Wiley, N.Y.pp. 106-118.

OSHERSON D.N., **STOB** M.and **WEINSTEIN** S., **1986**, *Systems that Learn*, MIT press.

PARISOT H.,**1965**, *Lewis Carroll*, Editions Pierre Seghers, France.

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.

PLATON, "1950", *Théétète ou de la science*, oeuvre de la pléiade, Editions Gallimard, Paris, pp. 83-192.

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

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., **1965**, *Trial and error predicates and a solution to a problem of Mostowski*, Journal of Symbolic, 30, 1, pp. 49-57.

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

ROGERS H., **1958**, *Gödel Numbering of the Partial Recursive Functions*, Journal of Symbolic Logic, 23, pp. 331-341.

ROGERS H.,**1967**, *Theory of Recursive Functions and Effective Computability*, McGraw-Hill, 1967. (2ed, MIT Press, Cambridge, Massachusetts 1987).

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

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.

SHAPIRO S. (Ed), **1985**, *Intensional Mathematics*, North-Holland, Amsterdam.

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

SMETS P., **1991**, *Probability of Provability and Belief Functions*, Logique & Analyse 133-134, pp. 177-195.

SMETS P., **MAMDANI** E. H., **DUBOIS** D., **PRADE** H., (eds.), **1988**, *Non-Standard Logics for Automated Reasoning*, Academic Press, London.

SMORYNSKI, C., **1981**, *Fifty Years of Self-Reference in Arithmetic*, Notre Dame Journal of Formal Logic, Vol. 22, n° 4, pp. 357-374.

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.

SMULLYAN R., **1977**, *The Tao is Silent,* Harper and Row, New-York.

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

SPENCER-SMITH R., **1991**, *Modal logic*, Artificial Intelligence Review, 5, 5-34.

***SVOZIL** K., **1993**, *Randomness & Undecidability in Physics*, World Scientific, Singapore.

VALADIER P., **1990**, *Inévitable Morale.* Editions du Seuil, Paris.

VAN FRAASSEN B. C., **1974**, *The Labyrinth of Quantum Logics*, in Boston Studies in the Philosophy of Sciences, Vol. 13, R. Cohen & M. Wartofsky (Eds.), Reidel, Dordrecht, pp. 224-254.

VAN STIGT W. P., **1990**, *Brouwer's Intuitionism,* Studies in the history and philosophy of Mathematics, Vol 2, North Holland, Amsterdam.

VICKERS S., **1989**, *Topology via Logic*, Cambridge University Press, Cambridge.

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

VISSER A., **1984**, *The Provability Logics of Recursively Enumerable Theories extending Peano Arithmetic at Arbitrary Theories extending Peano Arithmetic*, Journal of Philosophical Logic, 13, pp. 97-113.

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

***WARNER** R. & **SZUBKA** T. (eds), **1994**, *The Mind-Body Problem*, Blackwell, Oxford UK.

WATTS A., **1951**, *The Wisdom of Insecurity,* Pantheon Books, a Division of Random House, Inc. (trad. Franç. : Bienheureuse insécurité, Stock / Monde ouvert 1977).

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.

***WHEELER** J. A., **1980**, *Delayed Choice Experiments and the Bohr-Einstein Dialogue*, Report presented at the Am. Phil. Soc. and the Royal Soc., also in At Home in the Universe, Masters of Modern Physics Series, vol. 9, AIP Press, New York, 1994, pp. 112-131.

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, Berlin.

ZEUGMANN, T., **1987**, *On Barzdin's Conjecture*, Lecture Notes in Computer Science N° 265, pp. 220-227, Springer-Verlag, Berlin.

WITTGENSTEIN L.,**1918**, *Tractatus Logico-Philosophicus,* Gallimard, Paris, 1961.

WITTGENSTEIN L.,**1951**, *De la certitude*, Gallimard, Paris, 1965.

ZEH, H. D., **1990**, *Quantum Measurements and Entropy*, *Complexity, Entropy, and the Physics of Information*, SFI Studies in the Sciences of Complexity, vol VIII, Ed. W. H. Zurek, Addison Wesley.

Résumé des morphismes (avec les systèmes et les théorèmes)

a) Formules modales principales rencontrées

K	$\Box(p\to q)\to(\Box p\to\Box q)$
C	$\Box\Diamond p\to\neg\Diamond p$
D	$\Box p\to\Diamond p$
T	$\Box p\to p$
4	$\Box p\to\Box\Box p$
B	$p\to\Box\Diamond p$
Grz	$\Box(\Box(p\to\Box p)\to p)\to p$
L	$\Box(\Box p\to p)\to\Box p$
5	$\Diamond p\to\Box\Diamond p$
dual T	$p\to\Diamond p$
σ1	$p\to\Box p$
N	$\Box\top$
M	$\Box(p \,\&\, q)\to(\Box p \,\&\, \Box q)$
C	$(\Box p \,\&\, \Box q)\to\Box(p \,\&\, q)$

Dans tous les systèmes considérés dans ce travail, on définit le losange avec le carré : $\Diamond p = \neg\Box\neg p$.

b) Règles d'inférence principales

MP	$p,\ p\to q \Rightarrow q$
RE	$p \leftrightarrow q \Rightarrow \Box p \leftrightarrow \Box q$
RM	$p \to q \Rightarrow \Box p \to \Box q$
Nec	$p \Rightarrow \Box p$

A l'exception des logiques KD?-Σ_1 et KD?-Σ_1*, les formules sont des schémas d'axiomes. Ceci afin d'éviter la présentation lourde des règles de substitution.

M représente une machine autoréférentiellement correcte capable de prouver les théorèmes de l'arithmétique de Peano. On peut remplacer M par PA.

C
Axiomes : K, C
Règles : MP et Nec

<u>1) Morphisme de Magari-Boolos : MPL -> L(M)</u>

B = le prédicat Gödélien de prouvabilité par la machine représentable dans le langage de la machine.

$F(p_i)$ = énoncé de L(M)

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

G

Axiomes : K, 4, L
Règles : MP et Nec

Théorème

$$G \vdash A \text{ ssi pour tout } F, M \vdash MB_F(A)$$

G*
Axiomes : les théorèmes de G, + T
Règles : MP

Théorème

$$G^* \vdash A \text{ ssi pour tout } F, MB_F(A) \text{ est vrai}$$

<u>2) Morphisme de Solovay 1976</u>

SOL : MPL -> MPL

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

Théorème

$$G^* \vdash A \text{ ssi } G \vdash SOL(A)$$

<u>3) Morphisme de Gödel 1933 (cf aussi Kolmogorov 1932)</u>

G33 PL -> MPL

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

S4
Axiomes : K, T, 4
Règles : MP et Nec

Théorème (Gödel 1933, McKinsey etTarski 1948)

$$IL \vdash A \text{ ssi } S4 \vdash G33(A)$$

S4Grz
Axiomes : K, T, 4, Grz
Règles : MP et Nec

Théorème (Grzegorczyk 1967)

$$IL \vdash A \text{ ssi } S4Grz \vdash G33(A)$$

<u>4) Le stratagème (Morphisme de Boolos 1980, Goldblatt 1978, et Kuznetsov & Muravitsky 1977)</u>

BGKM : MPL -> MPL

$BGKM(p_i) = p_i$
$BGKM(AvB) = BGKM(A) \vee BGKM(B)$
$BGKM(A\&B) = BGKM(A) \& BGKM(B)$
$BGKM(\neg A) = \neg BGKM(A)$
$BGKM(\Box A) = \Box(BGKM(A)) \& BGKM(A)$

Théorème

$$S4Grz \vdash A \text{ ssi } G \vdash BGKM(A)$$

En combinant le théorème 3 avec le théorème 1, on obtient les morphismes (paramétrés par F) $MB_F \circ BGKM$: MPL -> L(M) avec

$$S4Grz \vdash A \text{ ssi pour tout F, } M \vdash MB_F(BGKM(A))$$

De même en combinant le théorème 3 avec le théorème 2 et le théorème 1, on obtient

$IL \vdash A$ ssi pour tout F, $M \vdash MB_F(BGKM(G33(A)))$

Remarque : S4Grz = S4Grz*, IL = IL*, on peut donc enlever "$M \vdash$ " dans la formule précédente.

5) Le stratagème affaibli

DEON : MPL -> MPL

$DEON(p_i) = p_i$
$DEON(AvB) = DEON(A) \vee DEON(B)$
$DEON(A\&B) = DEON(A) \& DEON(B)$
$DEON(\neg A) = \neg DEON(A)$
$DEON(\Box A) = \Box(DEON(A)) \& \Diamond DEON(A))$

KD
Axiomes K, D, M, C
règles MP, RM

Théorème

$$KD \vdash A \Rightarrow G \vdash DEON(A)$$

Question : KD est sain pour la machine auto-référentiellement correcte, quel axiome faut-il rajouter pour avoir la complétude ? Un nombre fini d'axiome suffit-il ?
Il n'est pas nécessaire de résoudre ce problème pour construire des démonstrateurs de théorèmes, pour les systèmes correspondant arithmétiquement complet (voir annexe 2) :

$$KD? = \{p \text{ telles que } G \vdash DEON(p)\},$$

Par le premier résultat de Solovay 1976, On a aussi :

$$KD? = \{p \text{ telles que } \forall F\ M \vdash MB_F \circ DEON(p)\},$$

Avec F restreint aux formules Σ_1,

$$KD?\text{-}\Sigma_1 = \{p \text{ telles que } \forall F\ M \vdash MB_F \circ DEON(p)\}$$

ainsi que

$$KD?\text{-}\Sigma_1{}^* = \{p \text{ telles que } \forall F\ MB_F \circ DEON(p) \text{ est vraie}\}$$

6) Morphisme de Goldblatt 1974 (voir 3.3 et annexe 5)

$$GOLDB(p) = \Box\Diamond p \text{ pour p atomique}$$
$$GOLDB(\neg A) = \Box\neg Goldb(A)$$
$$GOLDB(A\&B) = GOLDB(A) \,\&\, GOLDB(B)$$

Théorème

$$B \vdash GOLDB(p) \text{ ssi } QL \vdash p$$

où QL est le nom d'une logique quantique (voir annexe 5).

QuelQL et **QuelQL*** (F est restreinte aux propositions Σ_1)

$$QuelQL^* \vdash p \Leftrightarrow MB_F \circ DEON \circ GOLDB(p)$$

$$QuelQL \vdash p \Leftrightarrow M \vdash MB_F \circ DEON \circ GOLDB(p)$$