Rapport de lecture sur la thèse de M. Bruno Marchal, *Calculabilité, physique et cognition*

Paul Gochet, Séminaire de Logique et d'Epistémologie, Université de Liège, 32 Place du XX Août, 4000 Liège, Belgique, 7-3-1998, < pgochet@ulg.ac.be >, 78 Boulevard Louis Schmidt bte 35, 1040 Bruxelles, Belgique, TF 322-733 04 04

1. L'actualité du sujet

Dans un livre publié en janvier 1998 aux P.U.F. *la machine en logique*, Pierre Wagner écrit que les tenants de la conception "cogntiviste" de l'esprit soutiennent que "l'esprit est comparable à une machine de Turing dont le fonctionnement dépend à la fois des données initiales inscrites sur le ruban et des états internes (p.149)". B.Marchal prend cette analogie comme point de départ et déduit les conséquences qui en découlent lorqu'on la combine avec deux autres thèses : la thèse de Church selon laquelle toute fonction calculable au sens intuitif du terme "calculable" est programmable au sens technique du terme "programmable" et la thèse du réalisme arithmétique. Son travail s'inscrit dans un des courants de pensée dominants des sciences cognitives et de la philosophie de l'esprit de notre temps.

2. L'originalité du traitement du sujet

B. Marchal déduit des trois thèses précitées des conséquences qui renouvellent notre conception de l'identité personnelle, nos vues sur le déterminisme et notre conception des rapports de la psychologie et de la physique et qui mettent en question la validité des argments par lesquels certains auteurs ont cru pouvoir tirer du théorème de Gödel une réfutation du mécanisme . Si les solutions proposées par B.Marchal sont nouvelles, les problèmes qu'il aborde ont été formulés avant lui. Il nous paraît opportun d'évoquer brièvement ses prédécesseurs immédiats afin de mieux cerner son apport personnel.

3. La valeur de la méthode

Pour justifier ses thèses, B. Marchal fait uniquement appel à la *méthode hypothético-déductive* commune aux hommes de sciences et aux philosophes qui ont recours *exclusivement* à la méthode scientifique . Ces derniers comptent dans leurs rangs Russell, les Empiristes logiques et les philosophes de la mouvance analytique, par opposition aux philosophes qui font appel à une source *sui generis* de connaissance en philosophie (telle que l'intuition métaphysique chez Bergson ou la *Wesenschau* chez Husserl). M. Marchal s'impose également la contrainte de n'avancer que des thèses qui soient ou *empiriquement falsifiables* ou *logiquement réfutables*, ce qui leur confère le statut d'énoncés scientifiques, même si ces énoncés apportent des éléments de réponse à des questions traditionnellement cataloguées comme philosophiques tels que le problème des rapports de l'esprit et de la matière.

4. Du problème de l'identité personnelle

Le clonage ne pose pas un problème nouveau d'identité personnelle. L'individu produit par cette voie est simplement un *double,* une *copie conforme* de l'individu sur lequel une cellule a été prélevée. Il ne met donc pas en péril l'existence de l'individu dont on a prélevé une cellule pour le constituer. Pour qu'un problème intéressant touchant l'identité personnelle surgisse, il faut que l'individu reproduit ait des titres à faire valoir pour se présenter comme la *continuation* de l'individu dont il est issu. R. Chisholm a posé clairement le problème avant B. Marchal dans "Problems of Identity" édité dans l'ouvrage collectif *Identity and Individuation* de M. Munitz (N.Y. University Press 1971). Nous examinerons de près l'apport de ce précurseur pour localiser avec précision l'originalité de la thèse de B. Marchal.

Chisholm considère un individu AB qui subit une fission : son corps se dédouble en IJ et KL respectivement. Le premier corps issu de ce dédoublement conserve les caractéristiques physiques de AB telles que les empreintes digitales et les courants électriques qui parcourent le cerveau. Le deuxième corps est le produit d'une transplantation d'organes empruntés à un autre corps. En revanche, la mémoire est celle de AB. L'individu AB se pose alors la question: "Serai-je celui qui finit comme IJ ou celui qui finit comme KL?". Le propos de Chisholm est de donner une présentation imagée de la question "Quel est le critère d'identité d'une personne à travers le temps: son *corps* ou sa *mémoire* ?".

L'éventualité considérée par Chisholm anticipe l'une des éventualités envisagées par B.Marchal, celle où ce dernier évoque le cas d'une séparation du corps et du cerveau suivie de l'introduction d'un cerveau artificiel chez l'homme décérébré ainsi que la reconstitution d'un corps artificiel au profit du cerveau retiré du corps naturel. Posé en ces termes, le problème de l'identité personnelle donne au tenant du matérialisme et au tenant du mécanisme l'occasion d'affirmer leurs divergences, mais il ne permet pas de faire avancer la discussion.

Il en va tout autrement de l'éventualité imaginée par B.Marchal : celle où - le mécanisme étant postulé - le programme décrivant l'état computationnel du cerveau de l' individu initial est reconstitué simultanément de la même manière par deux machines situées dans des villes différentes. B.Marchal, comme nous allons le voir, s'est avisé des répercussions de ce nouveau type de duplication sur *l'emploi des pronoms personnels.* Cet emploi présente un grand intérêt philosophique. Songeons à la différence entre l'énoncé simplement faux: "l'auteur de ces lignes n'existe plus" et l'énoncé pragmatiquement contradictoire "je n'existe plus".

Même pour l'idée de duplication de la conscience, B. Marchal a un précurseur . Dans *Sameness and Substance* (Oxford, Blackwell 1980) , D. Wiggins écrit : "Si les lignes de conscience pouvaient se diviser", une nouveau concept de personne apparaîtrait "dont les consciences des membres seraient dérivées d'une point juqu'où elles partageraient la même biographie (p.165)". Et à la page suivante, il poursuit: "Chaque morceau de la personne éclatée entretient avec le clone-archétype la même relation que les pommiers de la sorte cox-orange...entretiennent avec le clone ou l'universel concret qu'ils perpétuent et constituent collectivement...(p.166)".

Ici encore, B. Marchal apporte quelque chose de neuf, cette fois par rapport à D.Wiggins. En effet, il met en lumière un effet de la duplication que Wiggins n'évoque pas. Considérant le cas où la description digitale de la personne initiale est reconstituée à deux endroits différents, Washington et Moscou en l'occurrence, B. Marchal fait remarquer que la personne ainsi dédoublée ne peut pas dire *avant son dédoublement* qu'elle se retrouvera *à la fois* à Washington et à Moscou . Comme l'écrit B. Marchal "la duplication...ne permet pas de se sentir [à la première personne] à deux endroits à la fois".

Il y a une bonne raison à cela que nous suggérons à B. Marchal de mentionner: selon le point d'arrivée, l'environnement sera différent et dès lors la *référence* du pronom "je" sera différente (le pronom personnel à la première personne attire l'attention sur un locuteur localisé dans la ville contenant la Maison Blanche dans le premier cas et sur un locuteur localisé dans la ville contenant le Kremlin dans le second cas). En revanche, le *sens* du pronom "je" reste le même. Dans les deux cas le pronom arraisonne linguistiquement le locuteur où qu'il soit.

Le dupliqué ne peut pas davantage dire qu'il ne sera qu'à un seul endroit . S'il le faisait, il nierait la possibilité même du télétransport dont la possibilité doit être admise pour que la question "Où meretrouverez-vous?" ait un sens. B.Marchal en déduit correctement une conclusion originale et importante : une forme nouvelle d'*indéterminisme*. Le sujet du télétransport sait qu'il ne se retrouvera pas à deux endroits et cependant ne peut dire à quel endroit il se retrouvera.

5. Du problème de la matière

Le physicalisme - qu'il ne faut pas confondre avec le matérialisme - domine assez largement la philosophie américaine contemporaine. Dans "Facts of the Matter" édité par R.W.Shahan et Ch.Swoyer , *Essays on the Philosophy of W.V.Quine,* The Harvard Press 1979, Quine écrit que le physicalisme qu'il professe ne doit pas se formuler par l'aphorisme "Pas de différence mentale sans différence physique", mais plutôt par l'énoncé plus précis: "il n'y a pas de différence dans le monde sans une différence dans le nombre ou l'arrangement des trajectoires de particules (p.165)". Contrairement au matérialisme éliminativiste qui nie l'existence d'états mentaux et contrairement au réductionnisme qui croit que les lois pyschologiques dont dérivables des lois physiques, la position physicaliste de Quine consiste à dire que les propriétés mentales requièrent l'existence de processus physiques. Une thèse rejetée , notons-le, par Ecclès.

Notons en passant que le computationnalisme et le matérialisme ne sont pas les seules théories possibles de l'esprit. La conception de l'esprit inspirée d'Aristote, de Saint Thomas et de Wittgenstein, n'est ni matérialiste, ni computationnaliste ni dualiste. Cette conception nous paraît plus apte à rendre compte de l'action rationnelle. Le passage suivant du livre de Roger Pouivet *Après Wittgenstein, saint Thomas* , Paris PUF 1997 illustre bien cette approche: "On ne décrira pas comme *choisie* l'action de quelqu'un dont on sait qu'il ne possède aucune faculté de discernement. L'action ne peut être dite

intentionnelle que si elle est aussi décrite comme celle d'un être qui reconnaît la fin de son action. La décrire comme choisie revient à considérer qu'un syllogisme pratique a porté sur les moyens (p.99)".

Moyennant l'hypothèse mécaniste combinée avec l'hypothèse dite "extravagante" du déployeur universel qui exécute tous les programmes possibles, le programme qui génère l'état de conscience dans lequel la craie ne tombe pas sur le sol quand on la lâche devrait recevoir une probabilité proche de zéro. Il est tenu pour négligeable. Au lieu de justifier la probabilité 1 pour la vie de tous les jours où la craie lâchée tombe sur le sol en invoquant les lois d'une substance, la matière, à laquelle on confère le statut de réalité, le computationnalisme que B.Marchal *étudie* exige qu'on fasse la démarche inverse. Pour "sauver les phénomènes", c'est-à-dire pour rendre compte que les craies lâchées tombent sur le sol au lieu de monter vers le ciel, on ne peut plus invoquer une *matière physique* régie par des lois physiques, mais "justifier sans invoquer de réalités substantielles ...que la collection ...des accidents lointains virtuels est négligeable ou de mesure nulle, pour une mesure définie sur les états ou sur les suites d'états computationnels apparaissant dans le déployement".

Il n'aurait pas été inutile que B.Marchal précise ce qu'il entend par "réalité substantielle". Dans la *Métaphysique* (1028a) , Aristote écrit que "La substance semble bien appartenir le plus manifestement aux corps. Ainsi disons-nous que sont des substances, non seulement les animaux, les plantes et leurs parties, mais aussi les corps naturels, tels que le Feu, l'Eau, la Terre, et chacun des autres éléments de ce genre, et encore toutes les choses qui sont des parties de ces éléments, ou des composés de ces éléments, c'est-à-dire l'Univers physique et ses parties, les astres, la Lune et le Soleil (Trad. Tricot Vrin 1948, pp.239-240)".

En accordant une priorité aux états de conscience reconstitués par le déployeur universel sur les muons, les gluons et autres "ondes-particules" qui font partie de l'ameublement de l'univers selon les physiciens contemporains, B.Marchal, impavide, tire une conséquence "renversante" de son hypothèse computationnaliste selon laquelle l'esprit est comparable à un programme d'ordinateur digital implémenté sur une machine dont *l'existence même* des matériaux n'ont pas d'importance. Cette conséquence renversante (p.15), s'énonce ainsi: "...avec le mécanisme toute solution au problème de la conscience nécessite une réduction de la physique à la psychologie".

Si B.Marchal en restait là, il aurait rendu non plausible son hypothèse de départ par les conséquences contre-intuitives qu'il en dérive plutôt qu'il n'aurait justifié ces dernières. Notre jugement sur les conséquences qu'il obtient redeviendra positif s'il parvient à trouver des *raisons indépendantes* de souscrire à ces vues, en d'autres termes s'il réussit à définir une *mesure de probabilité* sur les états computationnels apparaissant dans le déployement *qui soit "machine-indépendante"*. Le travail constitue ici un pas décisif vers la réalisation de cet objectif. Dans les quatre premiers chapitres, il démontre l'existence nécessaire de cette mesure. Dans le chapitre 5, il illustre une façon d'isoler mathématiquement cette mesure.

6. Le matérialisme face au computationnalisme

B.Marchal dans un article de 1988 et Tim Maudlin dans un article de 1989 s'accordent sur l'incompatibilité entre (1) la *conception matérialiste de l'esprit* selon laquelle les propriétés mentales "surviennent" sur une structure, des états et des processus physiques, chimiques, biologiques ou physiologiques et (2) la *conception computationnaliste de l'esprit* selon laquelle le fonctiàonnement de l'esprit est comparable à celui d'une machine de Turing, les propriétés physiques de l'ordinateur sur lequel la machine de Turing est implémentée n'ayant pas d'importance.

B.Marchal et T.Maudlin diffèrent par la position qu'ils adoptent face à cette incompatibilité:. Trois options sont en effet possibles : renoncer au computationnalisme et embrasser le matérialisme. C'est la position de Maudlin. Renoncer au matérialisme et embrasser le computationnalisme. C'est *l'hypothèse de travail* que Marchal choisit d'explorer. On peut, enfin, renoncer aux deux doctrines.

Epinglons au passage une *innovation philosophique importante* dont on peut créditer B.Marchal : il nous montre que le raisonnement par lequel il défend le computationnalisme et attaque le matérialisme ne dépend pas du niveau de substitution des parties.

Le computationnalisme offre une explication séduisante de certaines dispositions intellectuelles , telle que, par exemple, la compétence arithmétique que possède celui qui sait additionner les nombres naturels. Cette compétence réside dans la *disposition* à donner la réponse exacte à toute instance de la classe infinie de problèmes de la forme "Pour x et y donnés, quel est le z tel que $x + y = z$?". Une compétence peut rester à l'état latent. Avoir la capacité d'additionner deux nombres consiste à les additonner correctement *si* on nous demande de les additionner. Pour décrire une disposition, on utilise un énoncé de forme hypothétique.

La machine de Turing à additionner peut jouer un rôle dans la simulation de cette compétence humaine déjà présente chez les bébés de cinq mois (voir O. Houdé, *Rationalité, Développement et Inhibition* , Paris P.U.F. 1995, p.41). En effet la machine de Turing pour l'addition permet en un nombre fini de pas d'additionner n'importe quelle paire de nombres naturels. Mais il faut garder à l'esprit que les machines de Turing sont en fait des *programmes*. Selon la vue computationnaliste de l'esprit, l'esprit est comparable à un programme de machine de Turing universelle implémenté dans une machine concrète dont, comme nous l'avons vu plus haut, les propriétés physiques n'importent pas.

Pour identifier les conditions nécessaires et suffisantes que doit satisfaire une machine concrète pour simuler les opérations de l'esprit, T. Maudlin et B. Marchal se sont appliqués à transformer méthodiquement de telles machines. Nous ne pourrons juger de la valeur des enseignements qu'ils tirent de ces transformations qu'en entrant dans le détail de la description de celles-ci.

7. A propos des machines de T. Maudlin et de B. Marchal

T. Maudlin évoque le cas d'une machine matérielle, appelée Olympia I qui réalise concrètement une machine de Turing capable d'exécuter un programme π appliqué aux données figurant sur un certain ruban τ. Mais Olympia I est ainsi faite que si la question posée à la machine et formulée sur le ruban aurait

avait été banalement différente (si, pour fixer les idées, on lui avait demandé d'additionner 2 et 5 plutôt que 3 et 4), elle aurait été incapable d'exécuter la tâche demandée. Faute de disposer du *hardware* requis Olympia I n'est donc pas contrafactuellement correcte.

Maudlin construit alors Olympia II à partir de Olympia I en la dotant d'un hardware qui la rend cette fois *contrafactuellement correcte* dans le sens suivant: dans le cas non réalisé où elle devrait réagir à d'autres données sur le ruban, à savoir à un ruban τ' un dispositif (un loquet en l'occurrence) mettrait en action une machine de Turing universelle Klara et le calcul requis serait exécuté.

Olympia II manifeste une compétence qui fait défaut à Olympia I bien qu'elle soit aussi *causalement inerte* que cette dernière. Ce que l'expérience de pensée de T.Maudlin vise à montrer c'est que pour le computationnaliste, une *différence dans le hardware* qui n'a *aucun rôle causal* pendant l'exécution d'un calcul peut néanmoins être à la base d'une différence dans le mental : une compétence intellectuelle ou un état de conscience peut dépendre d'elle . Au contraire, pour le fonctionnaliste, la présence ou l'absence d'une pièce du hardware qui ne joue aucun rôle causal , "ne peut pas faire de différence pour l'intelligence, l'intentionalité ou la conscience du système durant l'exécution du calcul (Maudlin p.431)". Maudlin opte contre le computationnalisme et pour le fonctionnalisme . Il juge ce dernier beaucoup plus plausible en raison de la place qu'il accorde au rôle causal .

Un an avant Maudlin, B. Marchal avait proposé aussi une expérience de pensée dont nous allons évoquer la version la plus récente, celle de la thèse. Il décrit une machine de Neumann particulière, le graphe booléen, à l'intérieur duquel s'information se propage par l'intermédiaire de fibres optiques. Partant d'un graphe booléen contrafactuellement correct sur lequel on a implémenté le programme qui correspond au rêve fait par l'humain Hamlet. Appelons Macbeth la machine en question. Filmons l'exécution du programme par Macbeth . Le graphe filmé est *physiquement équivalent* au graphe booléen. "Ainsi, écrit B.Marchal, si on admet la thèse de la supervénience physique, qui concerne des exécutions particulières de machines, on doit reconnaître que le film véhicule ...les expériences subjectives correspondantes. Ce qui est absurde (P.28)".

On peut paraphraser ainsi le propos de B. Marchal : une machine accidentellement correcte (le graphe filmé) peut être *physiquement équivalente* à une machine contrafactuellement correcte (le graphe booléen), mais pour véhiculer un état de conscience, il faut à la fois davantage et moins. Il faut *davantage* en ce sens qu'il faut que la machine soit *computationnellement équivalente* à la machine contrafactuellement correcte. Il faut *moins* en ce sens que le substrat physique devient *indifférent* et peut-être même *superflu* . La conscience supervient à une exécution d'une computation effective par une *machine de Turing*, computation codifiable en nombres de Gödel (comme l'a montré Cutland dans *Computability* , Cambridge U.P. 1980). Les considérations de Maudlin sur l'opposition entre les parties de la machine qui sont physiquement actives et celles qui sont "inertes" cessent d'être pertinentes une fois qu'on a renoncé à la supervénience physique.

8.Contributions de la thèse à la logique

Lucas a présenté un essai de réfutation du mécanisme qui repose sur le théorème de Gödel. Il incombait dès lors à B. Marchal de fournir une réfutation du célèbre argument de Lucas repris par Penrose. La formalisation de cet argument à la p.41 et la localisation de l'erreur précisée p.42 sont exemplaires. B. Marchal a aussi le mérite d'avoir isolé la portion de l'argument de Lucas qui est valide et d'avoir précisé l'apport de cette partie correcte de l'argument (p.44). Lucas croit avoir établi que la conjonction de ces trois propositions est contradictoire: (1) La machine est saine (*Box* p -> p), (2) Je suis sain (*Box with a dot inside* p -> p), (3) Je suis une machine. En d'autres termes, tout ce que la machine prouve, je le prouve et réciproquement (*Box* p <-> *Box with a dot inside* p). Or tout ce que Lucas a réussi à prouver, c'est un résultat moins fort, à savoir que l'ensemble des trois propositions suivantes est contradictoire: (1), (2) et (3') où la proposition (3') est: *Je sais* que je suis la machine : *Box with a dot inside* (*Box* p <-> *Box with a dot inside* p).

Le carré modal introduit par B.Marchal, - carré modal qui reçoit une interprétation arithmétique , - combine la prouvabilité et la possibilité. La possibilité de p dont il s'agit ici, c'est la consistance de p, en d'autres termes l'impossibilité de démontrer Non p. Cet opérateur logique défini par B. Marchal permet de relier conceptuellement la *philosophie de l'esprit* à la *mécanique quantique*.:.Par surcroît, B. Marchal a réussi à relier la *logique de la prouvabilité* à la *logique quantique* par un théorème . Il a , en effet, démontré dans Z^*_1 la formule de Goldblatt: p -> *Box Diamond* p qui permet une interprétation modale de la logique quantique.

Le théorème 14 est loin d'être trivial. La démonstration élégante qu'en donne l'auteur n'utilise pas la nécessitation. Dès lors il a le droit d'utiliser le théorème de la déduction.

Il convient aussi de relever l'originalité de la démarche de l'auteur dans l'introduction de nouvelles logiques. Généralement on commence par donner une axiomatique pour laquelle on cherche ensuite une interprétation. On peut aussi commencer par formuler un langage et lui associer une interprétation, avant de se mettre à la recherche d'un système axiomatique. B. Marchal adopte une troisième et nouvelle manière de définir une logique. Il définit Z et Z* comme étant l'ensemble des formules produites par un démonstrateur automatique de théorèmes. Grâce à cette technique, Z et Z* reçoivent une définition précise, bien que leur axiomatisation ne soit que *partielle* (p.46).

9. Evaluation d'ensemble

La thèse de B.Marchal est une contribution profondément *originale* à un sujet interdisciplinaire , à l'intersection de la logique et de la philosophie de l'esprit. L'érudition de l'auteur est à la mesure de sa créativité. Il connaît tous les apports récents pertinents , apports qui sont très nombreux en raison précisément du caractère interdisciplinaire du sujet. Une publication rapide de cet admirable travail est hautement souhaitable, moyennant quelques améliorations de forme telles que : (1) remplacer "supervène" par "supervient", "possibilitation" par "possibilisation", (2) p.30, à la dernière ligne du premier paragraphe, remplacer "substituer la supervénience physique pour la supervénience cmputationnelle", par

"remplacer la supervénience physique par la supervénience computationnelle", (3) corriger la faute de frappe qui figure dans la démonstration du théorème 14, à savoir remplacer les occurrences de "q" par des occurrences de"p" dans les formules, (4) .rappeler dans une note que la distinction *Type /token* de Ch.S.Peirce s'appliquait initialement aux mots (*Collected Papers* , 4.537).(5) définir la règle de monotonie p. 37 où elle est utilisée pour la première fois et ne pas attendre la p.46.

Sous sa forme actuelle, la thèse est recevable et mérite d'être défendue .