Réfutation de l'*Argument du graphe filmé* de Bruno Marchal

Jean-Paul Delahaye
Université des Sciences et Technologies de Lille
LIFL Bât M3-ext
59655 Villeneuve d'Ascq cedex France
delahaye@lifl.fr

12 janvier 2011

==========

Résumé : Ce texte présente une analyse critique de *l'Argument du graphe filmé* proposé par Bruno Marchal au chapitre 4 de sa thèse "Calculabilité, Physique et Cognition" (Université de Lille, 1998). Quatre erreurs sont identifiées et localisées précisément dans le texte de Marchal. L'une de ces erreurs est au cœur même de l'argument, lorsque considérant la substitution d'une action opérée par un élément du graphe filmé par un rayon fortuit extérieur, Marchal admet l'équivalence des dispositifs avant et après la substitution, alors qu'ils ne le sont ni fonctionnellement ni causalement.

==========

Remarkable claims require remarkable proof.

Carl Sagan

One day in Naples the reverend Galiana saw a man from the Basilicata who, shaking three dice in a cup, wagered to throw three sixes; and, in fact, he got three sixes right away. Such luck is possible, you say. Yet the man succeeded a second time, and the bet was repeated. He put back the dice in the cup, three, four, five times, and each time he produced three sixes. "Sangue di Bacco," exclaimed the reverend, "the dice are loaded!" And they were.

George Pólya,

Introduction

L'*Argument du graphe filmé* est le point central des raisonnements du mémoire de thèse de Bruno Marchal soutenu en 1998 à l'Université de Lille. On trouvera le texte de la thèse en :

http://iridia.ulb.ac.be/~marchal/lillethesis/CPC.pdf

Marchal s'appuie par exemple sur cet argument dans la partie finale de son *Argument du déployeur universel* (UDA) pour lequel j'ai déjà proposé une analyse générale en :

http://www2.lifl.fr/~delahaye/dnalor/UDA2010.pdf

L'*Argument du graphe filmé* est présenté au chapitre 4 de la thèse de la page 19 à la page 31. Nous allons montrer qu'il comporte quatre erreurs passées inaperçues jusqu'à présent. Je dois avouer que, bien sûr, je n'avais pas vu ces erreurs au moment de la soutenance de la thèse Bruno Marchal en 1998, et que ce n'est que depuis qu'elles me sont apparues à force de réfléchir aux problèmes et aux raisonnements exposés dans la thèse. Ces erreurs ne sont devenues précises et parfaitement claires pour moi que depuis quelques mois, en reprenant l'étude du texte de la thèse à la demande Bruno Marchal.

Marchal, du fait de mon manque d'enthousiasme à soutenir ses conclusions, m'a mis au défi de préciser ma réfutation de ses raisonnements. Voilà donc, pour son *Argument du graphe filmé*.

Je sais bien qu'une seule réfutation suffit à invalider un raisonnement, et qu'en proposer quatre peut sembler inutile et même surprenant : je n'y peux rien, si examinant avec attention le déroulement de l'Argument du graphe filmé, ce n'est pas un point mais quatre qui sont apparu fautifs. Il en résulte que c'est chacun d'eux qu'il faudra traiter si Marchal souhaite remettre sur pied son argument.

Découvrant ces erreurs, je ne peux que regretter ma propre erreur d'avoir fait soutenir une thèse à Bruno Marchal en 1998. Je lui présente mes excuses pour mon aveuglement de l'époque, même si du fait de la nature principalement philosophique de l'Argument du graphe filmé, manquer de voir qu'il était faux n'a pas le même sens qu'à propos d'une démonstration mathématique (discipline où pourtant la chose se produit parfois aussi !). De même, je présente mes excuses à ceux qui ont lu cette thèse et qui ont pu en tirer des conclusions erronées en croyant qu'il s'agissait d'un travail solide et définitif.

*****

Faute 1 page 21, lignes 21-22

«Pour éliminer HE (et HU), il suffit donc de démontrer que COMP entraîne SUP-COMP.»

Cette phrase, remise dans le bon sens, signifie :

«*Si on démontre que COMP entraîne SUP-COMP, alors on peut éliminer HE (et HU)*».

Or, cela est faux. Il n'est pas vrai que l'on pourra éliminer HE (*Hypothèse Extravagante* que le déployeur universel physique est possible) et HU (l'*Hypothèse que l'Univers concret existe*) parce qu'on aura démontré que COMP entraîne SUP-COMP.

Si on démontrait que la lune est lumineuse d'elle-même (ce qui donnerait une explication à la lumière qu'on en perçoit la nuit ne s'appuyant pas sur l'idée que la lumière de la lune est due à celle du soleil qui l'éclaire), on n'aurait pas pour autant démontré que le soleil ne produit pas de lumière, et encore moins que le soleil n'existe pas.

Il y a confusion dans la phrase de Marchal entre le fait qu'une hypothèse n'est plus utile, et la possibilité de supprimer l'hypothèse définitivement.

Une telle suppression n'est évidemment pas possible lorsque l'hypothèse en question est utilisée pour d'autres choses. C'est encore moins le cas lorsque l'hypothèse sert de base à la conception générale que chacun a de l'organisation du monde.

Je suis un peu navré de devoir expliquer des choses aussi élémentaires et ridicules, mais tout le monde pense que l'hypothèse «le monde concret existe» est vraie et l'utilise à chaque instant de sa vie ; le fait que cette hypothèse pourrait ne plus être utile dans l'établissement de SUP-COMP à partir de COMP (qui ne préoccupe qu'un petit nombre de personnes sur terre) n'est évidemment pas suffisant pour renoncer à croire que le monde concret existe !

Je me répète car c'est important : il ne faut pas confondre «ne plus être nécessaire pour aboutir à une conclusion particulière» et «ne plus exister du tout». Prouver SUP-COMP (si on y arrive) à partir de COMP ne donne pas la preuve que le monde concret n'existe pas !!!

Il y a là une grossière erreur de logique qui rend inutile de lire la suite de l'Argument du graphe filmé. Nous acceptons cependant d'en poursuivre l'examen.

*****

Faute 2 page 21. La proposition 3, ligne 28.

«Proposition 3. COMP + non SUP-PHYS => SUP-COMP »

Bien sûr le '+' vaut pour un 'et'.

La proposition 3 est logiquement équivalente à :

COMP => (SUP-PHYS ou SUP-COMP)

La "démonstration" de cette proposition 3 est rapide (8 lignes) mais peu claire. Comme la proposition est fausse, on en déduit que la confusion de la démonstration sert à masquer sa fausseté.

Il est faux en effet qu'un computationnaliste n'a le choix qu'entre SUP-PHYS et SUP-COMP, et donc que si on lui démontre que SUP-PHYS est faux, alors il est obligé d'admettre SUP-COMP, du moins avec le sens que Marchal donne aux hypothèses de "supervénience physique" et de "supervénience computationnelle".

Le computationnaliste peut défendre une conception —et c'est ce que je ferais— où les contrefactuels ont de l'importance pour qu'on puisse parler de conscience ou d'esprit. Mais il peut le faire sans pour autant devoir accepter l'infinité des contrefactuels associés au déployeur universel et sans être obligé de croire que la conscience ne naît que de mécanismes universels de calcul. Il est faux qu'il n'y a que deux options : une supervénience *réduite* liée aux tokens individuels —ce que Marchal nomme SUP-PHYS— et une supervénience *extrême* liée au déployeur universel ou a des mécanismes universels de calcul, SUP-COMP.

La supervénience physicaliste et computationnaliste que je défendrais se formule en considérant la capacité d'un dispositif à faire survenir l'esprit. Elle est liée à sa capacité à réagir convenablement à certaines variantes des situations auxquelles il est soumis (contrefactualité limitée). Cependant, il n'y a besoin que d'un nombre fini de variantes pour fonder cette capacité. Dit autrement : le «programme» doit non seulement suivre un certain chemin de calcul précis, mais il doit avoir la capacité de suivre d'autres chemins de calculs proches, en réponse à des situations légèrement différentes auxquelles il n'est pas exposé. L'esprit ou la conscience surviennent pour moi dans les dispositifs computationnels ayant certaines capacités d'interactions avec le monde extérieur, ces capacités allant au-delà de ce que l'observation directe permet de constater. La supervénience est liée aux contrefactuels (comme Maudlin le montre) mais seule une contrefactualité finie est nécessaire.

Prenons un exemple informatique précis qui aidera à comprendre cette supervénience. Un programme dont je dis «il teste si un nombre est premier» doit être capable d'en tester plusieurs. Il doit même être capable d'en tester plus que ce que je ne peux vraiment lui en soumettre (sinon il est inutile !). Le fait d'être *un test de primalité* est une propriété que possèdent certains dispositifs matériels, c'est une propriété abstraite du programme qui ne change pas quand on change l'ordinateur utilisé, ou même le système d'exploitation. C'est une propriété qui fait intervenir des contrefactuels, c'est-à-dire des éventualités de tests qu'on n'effectue pas. Il y a plusieurs façons de s'assurer que cette propriété *d'être un test de primalité* est bien satisfaite : par l'expérimentation, par le raisonnement mathématique (si je dispose du programme, en partie ou en totalité), etc. Il est difficile d'arriver à une certitude —d'ailleurs l'équivalence des programmes est un problème indécidable—, mais le concept est clair et dans bien des cas particuliers on réussira à atteindre une quasi-certitude (positive ou négative) concernant la possession ou non de la propriété d'*être un test de primalité*.

La propriété de *posséder un esprit ou une conscience* est tout à fait comparable. Elle survient dans les dispositifs ayant certaines propriétés computationnelles faisant intervenir des contrefactuels (ce qui est lié à l'idée que ce sont des propriétés *dispositionnelles*, du moins si on s'écarte de l'idée naïve que ce sont des propriétés *tout ou rien* qui surgissent ou disparaissent instantanément comme lorsque Dieu dépose ou retire une âme). Il est

difficile d'avoir des certitudes absolues concernant ce type de propriétés, mais on peut comme pour la propriété *être un test de primalité* avoir des quasi-certitudes, et c'est d'ailleurs ce que nous expérimentons tous les jours avec nos proches, nos amis et les objets qui nous entourent : certains possèdent un esprit, d'autres non.

Un dispositif sera conscient ou possèdera un esprit non seulement s'il réagit correctement à une situation que je lui présente, mais, si par un moyen ou un autre, je peux me persuader qu'il réagira correctement à d'autres situations que je ne lui présente pas. Cette capacité contrefactuelle doit avoir une certaine amplitude, mais elle n'a pas besoin d'être infinie.

Entre les "tokens" et le "déployeur universel", il y a la place pour une conception de la supervénience physicaliste, mécaniste et computationnelle qui soit ouverte et limitée en même temps que parfaitement cohérente ! Une telle théorie *modérée* est la solution au problème du corps et de l'esprit, et même si sa formulation détaillée est difficile (mais c'est vrai aussi de la formulation détaillée d'une théorie qui dirait quand est-ce qu'un dispositif est *un test de primalité*) les principes d'une telle théorie sont clairs. Je ne comprends pas pourquoi Marchal ne semble pas en envisager la possibilité et *de facto* l'exclut sans l'examiner.

La «démonstration» que Marchal donne de sa Proposition 3 est d'une extrême confusion et insuffisante. Il semble y mélanger la forme faible de supervénience-physique (ne mentionnant que les tokens) et d'autres formes dont il écrit que ce sont des cas particuliers de la supervénience computationnelle. Or ce n'est pas vrai, les formes intermédiaires de supervénience computationnelle que j'ai mentionnées à l'instant ne sont pas des *cas particuliers* de la supervénience computationnelle (dans le sens de Marchal), mais des formes différentes et qui ne sont pas logiquement liées de manière simple à la supervénience computationnelle de Marchal (elles ne l'impliquent pas, et n'en sont pas conséquence). Il semble y avoir une grave lacune dans la conception qu'a Marchal des relations entre les différentes thèses de supervénience, et en tout cas, rien qui permettrait de démontrer la proposition 3 (l'erreur centrale provenant semble-t-il de l'hypothèse implicite par Marchal qu'il n'y a que deux possibilités : SUP-PHYS dans le sens étroit qu'il envisage, et SUP-COMP lié au pouvoir computationnel universel)

La supervénience computationnelle qu'il envisage semble n'avoir de sens que dans le cadre d'un déployeur universel (ou en référence à des mécanismes de calcul universels), or la supervénience physique que je défends (qui est aussi une forme de supervénience computationnelle) n'exige pas du tout de déployeur universel, ni même pour les dispositifs considérés d'être des dispositifs computationnelement universels. L'esprit peut très bien survenir dans des dispositifs de calcul n'ayant pas la propriété d'être équivalents à des machines universelles. D'ailleurs, notre caractère fini, et notre mémoire très limitée nous forcent à ne pas nous identifier à des machines universelles : de toute évidence nous n'en sommes pas ! Il existe

sur le sujet un très beau texte de Scott Aaronson : *Consciousness is finite (but I don't mind*). (http://www.scottaaronson.com/writings/finite.html)

Bien sûr dans la suite du raisonnement quand Marchal utilise la proposition 3, ce qu'il en tire est faux. Par exemple, page 30, ligne 11 et ligne 17, la formulation est incorrecte. Marchal argumente comme s'il était établi qu'il n'y avait que les deux possibilités (token ou déployeur universel) alors que cela n'a pas été établi. L'argumentation finale du chapitre est donc incorrecte du fait de cette erreur. Dit autrement, Marchal présuppose des résultats (faux) qu'il utilise comme s'ils avaient été établis.

*****

Faute 3 pages 24-28.

La modification apportée au dispositif (quand le film "supplée" à la défectuosité de la panne, page 24) n'en fait plus un dispositif physiquement équivalent car les relations causales et les dépendances fonctionnelles entre les pièces du dispositif sont changées. La structure causale et fonctionnelle du dispositif est modifiée d'une manière essentielle. La capacité à réagir correctement à des variations du dispositif —*robustesse contrefactuelle*— n'est pas préservée par le type de modifications que Marchal envisage.

Dit avec les mots de Marchal, je ne crois pas que la thèse de la supervénience physique que j'envisage entraîne la *thèse de la supervénience accidentelle active*, ni la *thèse de la supervénience accidentelle passive*. (page 25).

Le piège du raisonnement que Marchal propose consiste à entraîner l'approbation du lecteur du fait que la modification est faible *causalement et fonctionnellement*, puis par répétition de la modification faible à aboutir à une modification très importante. Le piège est exactement le même que dans le *paradoxe du tas* :

- si j'ai un tas de pierres et que j'enlève une pierre, cela reste un tas de pierre (tout le monde approuve), donc en itérant l'opération de retrait, une seule pierre (ou même aucune pierre) constitue un tas de pierres.

Il faut refuser de considérer comme négligeable les modifications de la structure causale et fonctionnelle introduite par l'évolution du dispositif dans l'expérience de pensée du graphe filmé, même quand elles ne concernent qu'un seul nœud du graphe.

Il n'est pas vrai, ***au sens strict***, que la thèse de la supervénience physique entraîne la *thèse de la supervénience accidentelle active*, ni la *thèse de la supervénience accidentelle passive*, comme il n'est pas vrai, ***au sens strict***, qu'un tas de cailloux auquel on enlève un caillou reste toujours un tas de cailloux.

C'est ici que la possibilité d'une réponse par les *fading-qualia* (disparition progressive et continue des sensations intimes, ou de la conscience) s'introduit. Du fait de la modification de la structure causale du dispositif (peu importante si un seul nœud du graphe est touché, mais de plus en plus importante si de nombreux nœuds sont concernés) défendre qu'il y a effacement progressif des qualia est parfaitement légitime et j'affirme, va de soi.

Contrairement au raisonnement de Chalmers (quand il argumente contre les *fading qualia* dans le cas d'un autre dispositif dont les modifications qu'il envisage préservent la structure causale) ici la structure causale est petit à petit complètement détruite. Quand on aura fait **n** modifications, il n'y aura plus aucune raison de croire que le dispositif obtenu sera, du point de vue de sa capacité à faire survenir l'esprit ou la conscience, resté équivalent au dispositif de départ.

Bien sûr tout le raisonnement de Marchal, qui s'appuie sur l'équivalence des diverses variantes de son dispositif quand on en modifie les éléments un à un, s'en trouve invalidé.

*****

Faute 4 pages 28-29.

Le "ce qui est absurde" (page 28, ligne 31) n'est pas si clair et la défense qui en est donnée page 29 n'est pas convaincante.

Le computationnalisme pourrait être plus abstrait que ce que Marchal croit. Le film pourrait faire survenir la conscience. On ne pourrait pas interagir avec elle, mais elle pourrait être là dans le film. Ce qui compte pour un computationnaliste, c'est la "structure" qui peut très bien être enregistrée. Le computationnaliste qui soutiendrait que la conscience survient dans le film n'est pas *évidemment faux*.

Il est un peu fort pour quelqu'un qui va ensuite défendre que le monde arithmétique (qui n'a rien de dynamique) de réfuter l'idée que le film puisse faire survenir la conscience avec un argument (implicite) de non-dynamicité. Cela me semble incohérent.

Marchal refuse au computationnaliste physicaliste, ce qu'il s'accorde à lui-même ! Pour moi, —en oubliant le problème des contrefactuels—, si on accepte qu'un déployeur arithmétique puisse faire survenir l'esprit, alors on peut tout autant admettre qu'un film lui aussi puisse faire survenir l'esprit.

À vrai dire, je considère que la question de savoir si la conscience survient sur le film n'a pas de sens autre qu'éventuellement *conventionnel*. Puisque les qualia ne se traitent pas à la troisième personne (il n'y a pas de test permettant de répondre avec certitude aux questions qu'on se pose à leur sujet), dès qu'on est dans des domaines éloignés de ce que le langage traite implicitement par les règles convenues par ceux qui l'utilisent, il n'y a plus de *vérité des choses*, mais uniquement des *conventions*. Savoir si la conscience survient quand on s'interroge sur des objets inertes (un film par exemple) devient une question de convention : rien ne permet de contrôler ce qu'on en dit. On peut trouver un accord entre nous, mais cela n'a pas de sens objectif, ce sera juste une *façon de parler*. Le plus raisonnable est sans doute de décider de ne pas répondre, comme je ne réponds pas à la question : est-ce que le *bleu* que je ressens est le même que celui que toi tu ressens et n'est pas plutôt ce que toi tu ressens *rouge* ? Pas de procédure pour répondre, donc pas de réponse, et sans doute, pas vraiment de sens à la question !

La question des qualia pour le film inerte ou pour les objets arithmétiques contrairement à ce que semble croire Marchal (qui ici oublie, je crois, ce que signifie *vérité à la première personne*) n'a pas de sens autre que celui que donnent les règles du langage qui à propos d'une telle question ne répondent pas, ou répondent plutôt qu'il n'y en a pas. Ce qui se passe avec le déployeur arithmétique abstrait, de même, ne concerne pas la conscience, ou alors, c'est une pure convention qu'on peut fixer dans un sens ou dans un autre.

L'idée de l'émergence de la conscience dans l'arithmétique est philosophiquement une illusion. Les conditions pour donner un sens réel à ce genre de considérations ne sont pas réunies. Non seulement Marchal ne prouve rien (car il commet des erreurs de logique), mais les choses qu'il prétend obtenir en conclusion de ses raisonnements n'ont pas de sens car il oublie les contraintes qu'imposent la distinction *première personne/troisième personne*.

*****

Conclusion

Tout d'abord, le détail de l'*Argument du graphe filmé* est faux (faute 3) du fait de la non équivalence causale des différents dispositifs envisagés. Ensuite l'utilisation qui est faite de la partie centrale du raisonnement est aussi

erronée (faute 2) puisqu'elle présuppose un résultat non établi —la proposition 3— qui n'envisage que deux types de supervénience alors qu'on peut en imaginer bien d'autres (et justement celles qu'on souhaiterait défendre !). Celles prenant en compte les contrefactuels de manière finie et n'exigeant pas de mécanismes computationnels universels semblent des positions assez raisonnables, et bien plus raisonnables en tout cas que la supervénience (infinitaire) computationnnelle envisagée avec les déployeurs universels.

Il se trouve, de plus, que Marchal s'attribue le droit de défendre une théorie de la supervénience sur des structures non temporelles (le déployeur universel est un objet arithmétique, hors du temps, situé, si cela a un sens, dans le monde immuable des entités mathématiques), droit qu'il refuse à un physicaliste (qui pourrait lui aussi défendre que le film, en un certain sens, continue de faire survenir l'esprit). Cette attitude de Marchal est logiquement incohérente. Cette incohérence utilisée de manière décisive dans son raisonnement (pour conclure le raisonnement par l'absurde à propos de SUP-PHYS et pour considérer comme acceptable sa version de SUP-COMP) à elle seule invalide le raisonnement de Marchal.

Le tout (pour celui qui n'en n'aurait pas vu les erreurs) n'aboutit à la conclusion que l'on peut se débarrasser de l'hypothèse d'un monde concret, qu'en commettant une quatrième erreur, celle de croire que ce qui n'est pas indispensable est nécessairement inexistant. Cette confusion entre les modalités *possible* et *nécessaire* est un peu étonnante chez un spécialiste de logique modale !

Le graphe filmé est de toute évidence totalement irrécupérable !

Bibliographie

• Bruno Marchal, *Calculabilité, physique et cognition,* Thèse soutenue en 1998 à l'Université des sciences et technologies de Lille. On trouve le document en format pdf en :

http://iridia.ulb.ac.be/~marchal/lillethesis/CPC.pdf

La thèse est aussi disponible par impression à la commande sous le titre : *Calculabilité, Physique et Cognition : une thèse originale*, Editions universitaires européennes, 2010.

• David Chalmers, *Absent Qualia, Fading Qualia, Dancing Qualia*, in *Conscious Experience*, edited by Thomas Metzinger. Imprint Academic, 1995. Voir en : http://cogprints.org/318/1/qualia.html

• Tim Maudlin. Computation and Consciousness. *The Journal of Philosophy*, 407-432, 1989.

• Jean-Paul Delahaye. *L'argument du déployeur universel de Bruno Marchal,* 2010 : http://www2.lifl.fr/~delahaye/dnalor/UDA2010.pdf