# L'ÉPISTÉMOLOGIE DES ÉNONCÉS D'IDENTITÉ CORPS/ESPRIT

Pascal LUDWIG
(Paris-Sorbonne)

Cet article porte sur la métaphysique de l'esprit, plus spécifiquement sur la métaphysique des états phénoménaux. Par « états phénoménaux », j'entendrai des états mentaux pour lesquels on peut dire que « cela fait un certain effet » d'être dans ces états. Les expériences perceptives — par exemple, la vision d'une rose rouge —, les sensations corporelles — par exemple, une vive douleur dans le bas du dos —, les émotions — par exemple, le sentiment accompagnant une violente colère —, sont tous des états phénoménaux en ce sens. Mon intention n'est cependant pas de défendre une thèse métaphysique. Je présupposerai, à vrai dire, que la position physicaliste réductionniste vis-à-vis des états phénoménaux est correcte, c'est-à-dire que l'on peut non seulement décrire leur nature et expliquer leurs pouvoirs causaux dans le vocabulaire des sciences de la nature, mais même identifier ces états à des états physico-chimiques. La question que je voudrais traiter est épistémologique : à supposer qu'une position physicaliste quant aux états phénoménaux soit vraie, comment cette position pourrait-elle être justifiée ? Quelles raisons pourrions-nous avancer, en tant que physicalistes, pour défendre notre position ? Il s'agit de questions épistémologiques, et même de questions méta-philosophiques : le but n'est en effet pas de défendre telle ou telle position dans l'espace des positions physicalistes possibles, mais de mieux comprendre la nature des justifications susceptibles d'être avancées en faveur de ces différentes positions. Cette perspective n'est pas entièrement nouvelle. La typologie la plus populaire pour classer les différentes positions physicalistes, que l'on doit à David Chalmers, repose en effet, comme je le rappellerai, sur des critères épistémologiques et non métaphysiques[1]. Il est cependant rare que l'épistémologie de la métaphysique de l'esprit soit abordée pour elle-même. Je pense pourtant que cette approche est susceptible d'apporter une perspective intéressante, en retour, sur le débat métaphysique lui-même. Le présent article a une ambition modeste : je veux avant tout convaincre le lecteur qu'une

[1]. Cf. (Chalmers, 2002).

perspective épistémologique sur la métaphysique de la conscience est pertinente.

## 1. L'intuition de la connaissance et les arguments épistémologiques contre le physicalisme

### *1.1 L'intuition de la connaissance et l'argument de la connaissance*

Selon les physicalistes réductionnistes, les états phénoménaux ne sont rien d'autre, métaphysiquement, que des états d'une substance physico-chimique. Il est donc possible, d'après cette position, de décrire la nature métaphysique de ces états dans le vocabulaire des science de la nature. Il n'y a cependant pas d'accord sur la nature de la relation logique existant entre ces descriptions scientifiques des états phénoménaux, et la description phénoménologique que nous pouvons en donner dans le vocabulaire de notre psychologie naïve. C'est une chose de soutenir que les états phénoménaux sont métaphysiquement identiques à des états physico-chimiques ; mais ces identités peuvent-elles être déduites à partir de l'ensemble de nos connaissances scientifiques ? Au moins depuis les écrits de Leibniz, de nombreux philosophes sont sceptiques quant à la possibilité d'expliquer la phénoménologie des états conscients à l'aide de théories formulées dans le vocabulaire des sciences de la nature. Ainsi J. W. Dunne, dans *An experience with Time*, soutient-il qu'il serait impossible de faire exactement comprendre la nature phénoménologique des expériences visuelles à un voyageur extra-terrestre provenant d'une planète dont tous les habitants seraient non-voyants. Certes, on pourrait décrire le mécanisme de la vision dans le vocabulaire des sciences physico-chimiques ; mais, affirme Dunne, « dans la vision, il y a bien plus que le simple enregistrement des contours des objets. Il y a, par exemple, la couleur ... Une description physique ne peut nullement procurer ici l'information qui serait apportée par l'expérience »[2].

Ce scepticisme quant au pouvoir explicatif des sciences naturelles dans le domaine de la phénoménologie repose sur l'intuition suivante, que nous nommerons l'intuition de la connaissance (IC) :

> (IC) Il existe un gouffre entre les concepts qui permettent de connaître le monde physique, et les concepts qui permettent de saisir la phénoménologie de l'expérience.

2. In (Ludlow, Nagasawa, Stoljar, 2004), Introduction.

Plusieurs arguments partent de cette intuition, mentionnée comme prémisse, pour tenter de conclure à la fausseté du physicalisme. On peut résumer ces arguments de façon lapidaire, en citant Frank Jackson : « rien de ce que vous pourriez me dire dans le vocabulaire de la physique ne traduit le parfum d'une rose. Donc le physicalisme est faux »[3]. L'argument de la connaissance, que cette citation de Jackson résume, développe l'intuition de la connaissance à l'aide d'une expérience de pensée. Imaginons qu'une personne, Mary, connaisse toutes les vérités physiques — et en particulier toutes les vérités formulables sur le système visuel humain et sur son fonctionnement. Saura-t-elle pour autant l'effet que cela fait de voir une rose rouge ? Selon Jackson, répondre « non » à cette question, comme nous y incite l'intuition de la connaissance, conduit logiquement à rejeter le physicalisme. Voici la structure de son argument :

> 1. Pour toute vérité *x*, ou bien *x* n'est pas une vérité exprimée dans le langage des sciences de la nature ou dérivable logiquement à partir de ces vérités, ou bien Mary connaît *x*.
> 2. Il existe une vérité *y* que Mary ne connaît pas, que l'on peut formuler ainsi, dans le langage de la psychologie populaire : « voir une rose rouge, cela fait cet effet » (où « cet effet » désigne une certaine sorte d'expérience vécue en première personne).
> 3. Il existe donc une vérité *y* qui n'est pas une vérité exprimée dans le langage des sciences de la nature, et qui n'est pas non plus dérivable logiquement à partir de ces vérités.

Il est un peu difficile de comprendre comment l'argument de la connaissance, formulé de la sorte, pourrait être interprété comme une réfutation directe et décisive du physicalisme. L'existence de vérités qui ne soient pas formulées dans le langage des sciences de la nature, et qui ne soient pas non plus dérivables à partir de ces vérités, n'est en effet pas directement en contradiction avec l'hypothèse physicaliste. Selon cette hypothèse, tous les faits qui existent dans le monde sont des faits physiques, des faits susceptibles d'être décrits et expliqués dans le vocabulaire des sciences naturelles. Mais cela n'implique pas *directement*

[3] Cf. (Jackson, 1982).

que toutes les vérités — c'est-à-dire toutes les propositions vraies décrivant ces faits — soient ou bien des vérités formulées dans le vocabulaire des sciences naturelles, ou bien des vérités dérivables logiquement à partir des théories scientifiques. L'hypothèse suivante n'est en effet pas contradictoire :

> (Dualisme conceptuel) Il existe au moins un fait physique F qui possède à la fois une description dans le vocabulaire des sciences de la nature, P, et une description dans le vocabulaire de la phénoménologie, Q. Par ailleurs, Q n'est pas dérivable logiquement à partir de l'ensemble des vérités physiques[4].

Selon l'hypothèse du dualisme conceptuel, Mary, lorsqu'elle voit une rose rouge pour la première fois, ne découvre pas un nouveau fait ; elle découvre plutôt une nouvelle façon de conceptualiser, et donc de décrire au moins mentalement, dans son for intérieur, un fait qu'elle connaissait déjà et qu'elle était déjà capable de décrire dans le langage des sciences de la nature. En vertu du cadre physicaliste que le dualisme conceptuel accepte, l'effet que cela fait de voir une rose rouge — une propriété phénoménale d'une expérience visuelle — n'est rien d'autre du point de vue métaphysique qu'une certaine propriété physico-chimique. Il est cependant concevable que cette unique propriété puisse être décrite de plusieurs façons différentes. C'est même non seulement concevable, mais très plausible : pourquoi donc nous représenterions-nous, en première personne et sur la base d'informations obtenues par introspection, de telles propriétés dans le langage des sciences physico-chimiques ?

La seule vraie question qui se pose est donc la suivante : à supposer que l'on dispose, comme c'est le cas dans l'expérience de pensée de Mary, d'une information exhaustive sur le monde formulée dans le langage des sciences de la nature, pourra-t-on déduire, à partir de cette information, l'ensemble des descriptions phénoménologiques des faits psychologiques ? En particulier, pourra-t-on déduire la description phénoménologique de l'effet que cela fait de voir du rouge à partir de la description scientifique de cette propriété ? Je reviendrai plus bas sur cette question, qui me semble au cœur du débat actuel sur la métaphysique des

4. Cette formulation du dualisme conceptuel m'est propre. Cette position a été défendue par de nombreux auteurs, en particulier Loar (1990), Levine (1993), Papineau (2002).

états conscients, mais je voudrais auparavant présenter un second argument contre le physicalisme, et le comparer à l'argument de la connaissance.

### *1.2 L'argument de la concevabilité*

Selon cet argument, on peut concevoir sans contradiction un organisme identique molécule par molécule à un organisme conscient, et néanmoins dénué de toute conscience : ce que l'on appelle, dans la littérature philosophique contemporaine, un « zombie »[5]. Dans la variante dite de «l'argument du spectre inversé», on considère qu'on peut concevoir un organisme identique molécule par molécule à un organisme conscient, mais dont les états qualitatifs diffèrent de façon systématique de ceux de l'organisme dont il est la réplique physique — un organisme percevant par exemple les choses rouges comme étant vertes, et les choses vertes comme étant rouges. On parle alors d'organisme « inversé » plutôt que de « zombie ». Quoiqu'il en soit, la forme générale de l'argument est la suivante :

> 1. Il est concevable qu'il existe un zombie (ou un organisme inversé).
> 2. S'il est concevable qu'il existe un zombie (ou un organisme inversé), il est métaphysiquement possible qu'il existe un zombie (ou un organisme inversé).
> 3. S'il est métaphysiquement possible qu'il existe un zombie (ou un organisme inversé), les états phénoménaux ne sont pas identiques à des états physiques, et le physicalisme est donc réfuté.

De prime abord, la structure de cet argument semble assez différente de celle de l'argument de la connaissance. Pourtant, les raisons avancées contre le physicalisme par l'argument de la connaissance et par l'argument de la concevabilité sont fondamentalement semblables. Dans ces deux arguments, la réfutation du physicalisme ne peut être atteinte que si l'on accepte de passer de considérations épistémologiques à des considérations métaphysiques. Dans l'argument de la connaissance, l'existence d'une vérité qui n'est ni formulée dans le vocabulaire des théories appartenant au sciences de la nature, ni dérivable à partir de ces théories est censée permettre de conclure à l'existence d'un fait non-physique. Dans

[5]. Cf. (Chalmers, 1996), (Chalmers, 2002).

l'argument de la concevabilité, la concevabilité de l'existence d'un certain organisme — donc une caractéristique épistémologique de la proposition selon laquelle cet organisme existe — est supposée justifier l'affirmation (métaphysique) selon laquelle l'existence de cet organisme est (métaphysiquement) possible. On voit donc qu'il existe une forte analogie entre les deux raisonnements.

Mais la proximité est encore bien plus grande. David Chalmers, l'un des principaux promoteurs de l'argument de la concevabilité, définit la concevabilité d'une proposition de la façon suivante :

> (Conc) La proposition P est concevable si, et seulement si la proposition que non P n'est pas *a priori*.

Par ailleurs, on peut présenter de façon plus générale et plus précise l'intuition de la concevabilité des zombies en soulignant qu'un monde où existe un zombie, en toute généralité, est un monde (i) dans lequel toutes les vérités physiques qui valent dans notre monde valent également — en particulier, la composition physico-chimique des organismes est exactement la même que dans notre monde — (ii) et dans lequel, cependant, au moins une vérité phénoménologique qui vaut dans notre monde n'est plus une vérité. On peut donc reformuler l'argument de la concevabilité de la façon suivante, en considérant que P est l'abréviation pour la conjonction de toutes les vérités physiques, et que Q est n'importe quelle vérité formulée dans un vocabulaire phénoménologique[6] :

> 1. La proposition selon laquelle ce n'est pas le cas que (P & non Q) n'est pas *a priori* ;
> 2. S'il n'est pas exclu *a priori* que (P & non Q), alors la proposition selon laquelle (P & non Q) est métaphysiquement possible.
> 3. Si la proposition selon laquelle (P & non Q) est métaphysiquement possible, le physicalisme est réfuté.

Cette reformulation permet de saisir immédiatement la très grande proximité entre l'argument de la concevabiltié et l'argument de la connaissance. En effet, « ce n'est pas le cas que (P & non Q) » est logiquement équivalent à « Si P, alors Q ». Ce qu'exprime la prémisse de l'argument de la concevabilité, c'est donc qu'il n'est pas exclu *a priori* que le conditionnel « Si P, alors Q » soit faux. Autrement dit, la conjonction de l'ensemble des vérités physiques n'implique pas *a priori* (au sens de

6. Voir en particulier (Chalmers, 2002). Voir également (Stoljar, 2006).

l'implication matérielle) toutes les propositions phénoménologiques, puisqu'au moins une de ces propositions peut être fausse même si l'antécédent du conditionnel « Si P, alors Q » est vrai. Or, soutenir cela, ce n'est rien d'autre que d'exprimer d'une façon un peu abstraite l'intuition de la connaissance (IC) : selon cette intuition en effet, on ne peut pas inférer, par la simple réflexion *a priori*, d'une connaissance de l'ensemble des vérités physiques — donc d'une connaissance de l'antécédent du conditionnel physico-psychique « Si P, alors Q », antécédent qui, rappelons-le, résume par une conjonction l'ensemble de nos connaissances physiques — à une connaissance de l'ensemble des vérités psychologiques. Pour le dire de façon imagée, en utilisant l'expérience de pensée de Jackson, la connaissance de l'ensemble des vérités physiques ne permet pas à Mary de savoir ce que cela fait de voir du rouge, et ce quelles que soient la profondeur et la durée de sa réflexion *a priori* sur ces vérités. Il est du coup possible de concevoir une situation où, comme dans l'expérience de pensée des zombies, toutes les propositions énoncées dans le vocabulaire de la physique sont vraies, mais où pourtant aucun organisme n'instancie une certaine propriété phénoménale — correspondant par exemple à l'effet que cela fait de voir du rouge.

On voit donc, comme Chalmers l'a souligné maintes fois[7], que l'argument de la connaissance et l'argument de la concevabilité se ramènent en fait à un argument plus fondamental, qui utilise le conditionnel physico-psychique « Si P, alors Q » :

> *Argument épistémologique fondamental* :
> 1. Le conditionnel « Si P, alors Q » n'est pas une vérité *a priori*.
> 2. Si le conditionnel « Si P, alors Q » n'est pas une vérité *a priori*, c'est une vérité contingente.
> 3. Si le conditionnel « Si P, alors Q » est une vérité contingente, le physicalisme est réfuté.

Les prémisses 1 et 3 semblent, de prime abord, les moins problématiques dans cet argument. La première prémisse peut être interprétée aussi bien comme énonçant l'intuition de la connaissance que comme énonçant la concevabilité d'un zombie. La prémisse 3 découle du lien entre physicalisme et survenance : si le physicalisme est vrai, alors il est au

7. Cf. (Chalmers, 1996), (Chalmers, 2002).

minimum impossible qu'un monde soit physiquement indiscernable du monde réel, mais psychologiquement discernable de ce monde (comme l'est justement un monde contenant un zombie). La vérité du physicalisme entraîne donc non seulement la vérité du conditionnel « Si P, alors Q », mais sa nécessité métaphysique. Dans une large mesure, la discussion philosophique s'est donc focalisée sur la prémisse 2 de l'argument.

## 2. Deux formes de physicalisme

David Chalmers propose une typologie des positions physicalistes réductionnistes qui part de la manière dont les tenants de ces positions réagissent aux arguments épistémologiques[8]. L'argument épistémologique fondamental est valide, et sa troisième prémisse ne semble pas discutable. Selon lui, deux grandes familles de réactions sont dès lors possibles :

(i) On peut d'abord rejeter la première prémisse, c'est-à-dire soutenir que le conditionnel physico-psychique « Si P, alors Q » est en fait une vérité a priori ; ce choix débouche sur ce que Chalmers nomme le physicalisme de type A.

(ii) On peut par ailleurs accepter la première prémisse, mais rejeter la seconde ; cela revient à soutenir que le conditionnel physico-psychique est une vérité métaphysiquement nécessaire, mais pour autant a posteriori. Chalmers nomme cette position le physicalisme de type B.

Selon David Chalmers, c'est fondamentalement une question épistémologique qui divise les physicalistes : est-il possible ou non de dériver, à partir d'un antécédent comportant la conjonction de toutes les vérités physiques, l'ensemble de toutes les vérités tout court, y compris donc les vérités phénoménologiques ? En particulier, pour reprendre la question fondamentale par laquelle j'avais conclu ma discussion de l'argument de la connaissance, peut-on déduire la description phénoménologique de l'effet que cela fait de voir du rouge à partir de la description scientifique de cette propriété ?

### *2.1. Le physicalisme de type A*

Répondre de façon positive à cette question revient à soutenir

[8.] Cf. (Chalmers, 2002). Voir également (Stoljar, 2006).

qu'il n'existe pas de gouffre explicatif ou épistémologique entre les concepts des sciences de la nature et les concepts à l'aide desquels nous décrivons en première personne nos états conscients. Chalmers fait remonter le physicalisme de type A aux travaux de David Lewis et de David Armstrong sur la conscience[9]. D'après ces auteurs, les concepts phénoménaux, comme le concept de douleur, peuvent en effet recevoir une *analyse fonctionnelle*, analyse qui permet de combler le gouffre explicatif supposé exister entre le domaine du physique et celui du phénoménal.

Commençons par illustrer le concept d'analyse fonctionnelle à l'aide d'un exemple tiré de la biologie, celui de la reproduction sexuée[10]. Il semble que l'on puisse donner une explication entièrement physicaliste du phénomène biologique de la reproduction. Il y a en effet reproduction sexuée lorsque deux organismes en produisent un (ou plusieurs) autre(s). Notons que ce dernier énoncé relève de l'analyse conceptuelle, non de la recherche empirique : il suffit de comprendre la signification de l'expression « reproduction sexuée » pour savoir que cet énoncé est nécessairement vrai. Cette analyse conceptuelle permet d'identifier la fonction de la reproduction, son rôle causal. Elle permet aussi, dans un second temps, d'identifier le mécanisme physique qui réalise cette fonction, puisque l'on peut supposer qu'il existe une suite de types d'événements physiques qui permettent à deux organismes d'en produire un (ou plusieurs) autre(s). On le voit, l'analyse fonctionnelle joue un rôle souvent inaperçu, mais central du point de vue conceptuel, dans la production d'explications réductrices au sein des sciences spéciales. Ici, cette analyse permet d'établir une connexion nécessaire entre un prédicat de la biologie, le prédicat « reproduction sexuée », et le vocabulaire de la physique et de la chimie : le mécanisme permettant à deux organismes d'en produire un troisième peut en effet être décrit dans ce vocabulaire.

La question est de savoir s'il est possible de formuler une telle analyse fonctionnelle pour un concept phénoménal, par exemple pour le concept phénoménal de douleur. Penser que c'est le cas revient à penser qu'il n'y a pas de gouffre explicatif, ou épistémologique, entre la douleur comprise et décrite dans le vocabulaire phénoménal et la douleur comprise et décrite dans le vocabulaire des sciences de la nature. Un malentendu doit cependant être immédiatement dissipé.

9. Cf. (Armstrong, 1964), (Lewis, 1970), (Polger, 2002). (Kim, 2005) défend le physicalisme de type A pour tous les états mentaux sauf les états phénoménaux.
10. Cf. (Chalmers, 1996), p. 44.

En premier lieu, un physicaliste de type A ne doit pas nécessairement soutenir que Mary saura, avant d'avoir sa première expérience visuelle d'une rose rouge, ce que cela fait de voir du rouge. Ce que le physicaliste de type A doit soutenir, c'est *qu'étant donnée une maîtrise préalable des concepts phénoménaux*, une analyse fonctionnelle de ces concepts doit permettre d'opérer une explication réductrice des propriétés phénoménales qu'ils désignent — par exemple, de la sensation de rouge. Or, il n'est pas évident que Mary, dans la situation décrite par Jackson, maîtrise le concept phénoménal de la sensation de rouge. Supposons que la possession d'un tel concept présuppose d'avoir déjà fait l'expérience correspondant au concept (en l'occurrence, d'avoir déjà fait l'expérience visuelle de la couleur rouge). Si c'est le cas, Mary ne possède pas le concept phénoménal pertinent avant d'avoir quitté son environnement en noir et blanc. Puisqu'elle ne maîtrise pas ce concept dans cet environnement, on voit mal comment elle pourrait le « fonctionnaliser » à l'aide d'une analyse fonctionnelle.

Considérons le conditionnel physico-psychique « Si P, alors Q ». Le physicaliste de type A soutient que ce conditionnel est *a priori*, puisqu'une analyse fonctionnelle des concepts phénoménaux doit permettre de dériver une explication physicaliste réductrice des états phénoménaux qu'ils dénotent. Il n'est *a priori*, cependant, que si tous les concepts figurant dans l'antécédent figurent aussi dans le conséquent. C'est la raison pour laquelle l'argument de la connaissance est difficile à interpréter. Certes, nous avons la forte intuition que Mary ne pourra pas dériver toutes les vérités phénoménales à partir d'une pure réflexion *a priori* sur l'ensemble des connaissances physiques. Mais, nous l'avons vu, il est plausible de soutenir que Mary ne possède pas, dans son environnement en noir et blanc, l'ensemble des concepts phénoménaux : puisqu'elle n'a jamais vu de rouge, elle ne possède pas le concept phénoménal de la sensation de rouge. Il y a donc deux façons d'interpréter l'intuition selon laquelle Mary ne parviendra pas à dériver *a priori* les vérités phénoménales à partir de la description physicaliste du monde dont elle dispose : on peut en déduire qu'il existe un gouffre épistémologique entre le domaine du physique et le domaine du phénoménal ; ou bien l'on peut en déduire qu'il manque certains concepts à Mary, et qu'en l'absence de ces concepts la dérivation est tout simplement impossible.

### *2. 2. Le physicalisme de type B*

Je présenterai plus rapidement le physicalisme de type B, qui est certainement actuellement la forme dominante du physicalisme réductionniste. Il s'agit d'une forme conceptuelle de dualisme, qui reconnaît l'existence d'un gouffre explicatif (ou épistémologique) entre le physique et le phénoménal, et qui cherche à expliquer ce gouffre à l'aide de la stratégie des concepts phénoménaux. L'idée centrale de cette stratégie est très simple : elle consiste à soutenir que l'explication de l'existence d'un gouffre explicatif ne réside pas dans l'existence d'un ensemble de propriétés métaphysiquement irréductibles, les propriétés phénoménales, mais plutôt dans l'existence de différentes façons de concevoir les mêmes propriétés cérébrales[11].

Le simple dualisme conceptuel — la thèse selon laquelle les propriétés cérébrales peuvent être conçues à la fois en première personne, à l'aide de concepts phénoménaux, et en troisième personne, à l'aide de descriptions physico-chimiques — ne caractérise cependant nullement le physicalisme de type B. Contrairement à ce que soutiennent certains auteurs, un physicaliste de type A peut en effet, comme je l'ai souligné plus haut, reconnaître l'existence d'une classe particulière de concepts phénoménaux. Un physicaliste de type B doit donc soutenir la conjonction des trois thèses suivantes :

(T1) il existe un gouffre explicatif entre le domaine du physique et celui du phénoménal ;

(T2) l'existence de ce gouffre explicatif est engendrée par un dualisme conceptuel : les mêmes propriétés cérébrales, les propriétés phénoménales, sont conçues en première personne à l'aide de concepts phénoménaux, et en troisième personne à l'aide de description physico-chimiques ;

(T3) les concepts phénoménaux ne peuvent être fonctionnalisés ; pour cette raison, le gouffre explicatif n'est pas destiné à être comblé.

La thèse (T3) est fondamentale, puisque c'est en fait elle qui distingue le physicalisme de type A du physicalisme de type B. Un physicaliste de type A peut en effet reconnaître qu'il existe, à un moment *t*, un gouffre explicatif entre le domaine du physique et celui du phénoménal,

11. Cf. (Block et Stalnaker, 1999), (Levine, 1983), (Levine, 1993), (Loar, 1990), (McLaughlin, 2001), (Papineau, 2002).

et même soutenir, comme nous l'avons vu, que ce gouffre s'explique par un dualisme des concepts. En revanche, il rejettera (T3) : à terme, pour le physicaliste de type A, une fonctionnalisation des concepts phénoménaux par analyse conceptuelle est possible, et cette fonctionnalisation, qui doit permettre de créer des liens conceptuels entre le vocabulaire phénoménal et le vocabulaire physicaliste, doit aussi permettre de combler le gouffre explicatif.

On considère en général, dans la littérature contemporaine, que la thèse (T3) constitue la force principale du physicalisme de type B, puisqu'elle permet de comprendre la persistance du gouffre explicatif. Selon le physicaliste de type B, en effet, le conditionnel physico-psychique « Si P, alors Q » est un énoncé *a posteriori*, mais nécessaire. Sa nécessité découle du physicalisme — nous avons vu que la contingence de ce conditionnel implique la fausseté du physicalisme —, et son caractère *a posteriori* s'explique par la thèse (T3) : même si l'antécédent et le conséquent du conditionnel contiennent les mêmes concepts, il est impossible d'analyser fonctionnellement les concepts phénoménaux, et l'on ne peut donc pas déduire *a priori* les vérités phénoménales à partir d'une description physique exhaustive du monde. On compare donc souvent l'analyse des énoncés d'identité psycho-physique par les physicalistes de type B à l'analyse des énoncés d'identification théorique issue des travaux de Saul Kripke et de Hilary Putnam[12]. L'énoncé psycho-physique (3) aurait fondamentalement les mêmes caractéristiques métaphysiques et épistémologiques que (1) et (2) :

(1) Eau = H2O

(2) Chaleur = énergie cinétique moyenne des molécules

(3) Douleur = stimulation des fibres-C

(4) M = P (où M est n'importe quelle propriété phénoménale, et P la propriété physique qui lui est identique)

Il faut cependant nuancer sérieusement l'analogie supposée exister entre (3) et (4) d'une part, et (1) et (2) de l'autre. Certes, tous ces énoncés sont à la fois nécessaires et *a posteriori* : une simple analyse du concept d'eau ne permet évidemment pas de dériver l'identité (1), par exemple. Mais, quoique la question soit controversée, les concepts « eau » et « chaleur » semblent pouvoir être fonctionnalisés[13]. Un locuteur maîtrisant le concept

12. Cf. (Kripke, 1980), (Putnam, 1985).

13. Voir le débat entre Chalmers et Jackson d'un côté, et Block et Stalnaker de l'autre : (Block et Stalnaker, 1999) et (Chalmers et Jackson, 2001).

d'eau est ainsi en principe capable de décrire le rôle qu'occupe la substance dénotée par le concept dans notre monde ; il sait que l'eau est un liquide transparent, qui occupe, dans un état impur, les lacs, les fleuves, les rivières, les mers et les océans. En principe, une description exhaustive du monde physique, formulée dans un langage scientifique, couplée à l'analyse fonctionnelle du concept d'eau que nous venons d'esquisser, doit donc permettre de dériver *a priori* l'énoncé (1). Nous avons cependant insisté sur le fait que (3) ne pouvait pas être dérivé *a priori* à partir de telles prémisses selon le physicalisme de type B.

Mon but dans le présent article n'est pas de remettre en question la thèse selon laquelle il existerait un contraste entre le concept phénoménal de douleur d'un côté, et le concept macroscopique d'eau de l'autre. Je voudrais simplement insister ici sur les conséquences épistémologiques de ce contraste. La question centrale à laquelle nous devons répondre est simple : s'il existe réellement un gouffre explicatif entre le physique et le phénoménal, empêchant d'établir des connexions entre les concepts phénoménaux et les concepts des sciences de la nature, quelle stratégie peut-on adopter pour justifier des énoncés d'identité psycho-physique comme (3) et (4) ?

## 3. De la justification des énoncés d'identification psycho-physiques.

Commençons par souligner que l'observation ne nous permettra pas de justifier *directement* un énoncé comme (3). Pour identifier un individu ou une propriété par l'observation, il faut en effet disposer d'une description préalable de cet individu ou de cette propriété, susceptible de permettre l'identification. Considérons ainsi que je veuille identifier Jaegwon Kim dans un colloque X, et que je ne l'ai jamais vu. On peut supposer que c'est ma connaissance de certaines propriétés de Kim qui me permettra de l'identifier — je sais par exemple quels sont ses thèmes de prédilection, je connais également le titre de l'article qu'il compte présenter au colloque. Tout cela me permet de formuler une description qui, dans le contexte du colloque, me permettra éventuellement de l'identifier. Dans une telle situation, la justification de l'énoncé d'identité (5) :

(5) Cet orateur est Jagwon Kim.
possède la forme suivante :
(i) Je sais que Kim va présenter un article sur l'argument d'exclusion causale.
(ii) Cet orateur est le seul, dans le colloque X, qui intervient sur l'argument d'exclusion causale.
(iii) Donc, cet orateur est Jaegwon Kim.

Une telle procédure directe de justification ne peut cependant pas être utilisée dans le cas d'une propriété phénoménale comme la douleur. Selon la thèse (T3), nous ne disposons pas d'une description fonctionnelle de la douleur, qui permettrait de l'identifier parmi l'ensemble des propriétés naturelles. Nous ne disposons que d'une description phénoménologique de la douleur. Or, d'après la thèse (T1), il existe un gouffre explicatif entre le domaine phénoménal et le domaine physique. Cela implique que notre connaissance phénoménale ne peut en aucun cas nous permettre, selon le physicalisme de type B, de formuler une description susceptible d'identifier la propriété physico-chimique que la douleur se trouve être.

Récapitulons. Ni un raisonnement *a priori*, ni l'observation assistée d'une connaissance par description de la propriété à identifier ne peuvent nous permettre de justifier un énoncé de la forme (4) dans le cadre du physicalisme de type B. Une telle identité ne peut donc être justifiée que de façon *indirecte*, à l'aide de raisonnements inductifs. Voyons comment exactement.

### *3.1. L'argument par la simplicité.*

L'argument par la simplicité a été défendu dès 1959 par J. J. C. Smart[14]. Selon Smart, les énoncés du type (4) sont à la fois *a posteriori* et contingents. Les physicalistes de type B les considèrent aujourd'hui, nous l'avons vu, comme nécessaires, mais cela ne change rien de fondamental à l'argumentation de Smart, qui reste intéressante dans le contexte théorique actuel. Selon Smart, nous disposons de données inductives qui indiquent qu'il existe des corrélations entre l'occurrence des propriétés phénoménales, et l'occurrence de propriétés neurales. Ainsi, nous disposons de données qui montrent par induction qu'à chaque occurrence de douleur correspond, dans les cerveaux des agents qui souffrent, une occurrence de stimulation de fibres-C. Nous pouvons donc accepter que « la

[14.] Cf. (Smart, 1959). Voir aussi la discussion de C. Daly, (Daly, 2010).

thèse de la corrélation corps-cerveau » vaut pour les états phénoménaux, comme pour tous les autres états mentaux (Kim 1996, p. 48) :

> (Thèse de la corrélation) Pour chaque type M d'état phénoménal ayant une occurrence pour un organisme O, il existe un état cérébral de type C tel que M a une occurrence pour O à *t* si et seulement si C a une occurrence pour O à *t*.

Remarquons qu'il n'est pas nécessaire, pour accepter cette thèse, de rejeter la thèse (T3) du physicalisme de type B. En effet, la thèse de la corrélation n'identifie pas les états phénoménaux aux états cérébraux ; elle se contente de souligner qu'il existe une corrélation entre ces deux sortes d'états. Pour établir une telle corrélation, il suffit de pouvoir identifier les états phénoménaux en tant que phénoménaux d'une part, par exemple par introspection, et d'autre part d'identifier les états cérébraux en tant que cérébraux. Il n'est à aucun moment nécessaire d'identifier ou de décrire les états phénoménaux dans le vocabulaire neural. Ainsi par exemple, je constate à l'aide de l'introspection qu'au moment *t* il y a une occurrence de douleur dans mon esprit ; et le chirurgien qui observe mon cerveau constate, grâce aux moyens d'investigation dont il dispose, que les fibres-C sont activées au même moment *t*. Ces constatations ne présupposent pas que je sois capable de trouver la moindre relation conceptuelle entre mon état de douleur et l'état neural décrit par mon chirurgien. A vrai dire, l'acceptation de la thèse de la corrélation est parfaitement compatible avec le dualisme des propriétés : il existe de très nombreuses théories dualistes — le dualisme interractionniste de Descartes, le parallélisme de Leibniz, l'occasionnalisme de Malebranche, pour ne prendre que quelques exemples — qui prétendent pouvoir fournir des explications causales de la thèse de la corrélation. Comment donc passer de la thèse de la corrélation à l'acceptation d'énoncés d'identification du type de (4) ? Selon Smart, en faisant appel au principe d'économie ontologique dit du «rasoir d'Occam». L'argument de Smart a donc la forme suivante :

> *Prémisse 1* : la thèse de la corrélation est vraie.
>
> *Prémisse 2* : toutes choses égales par ailleurs, la théorie la plus simple compatible avec la la thèse de la corrélation est la théorie de l'identité, qui soutient que les énoncés de type (4) sont vrais ; les autres théories sont plus complexes car elles multiplient les entités, ce qui constitue une

infraction au principe du rasoir d'Occam.

*Conclusion* : les énoncés de type (4) sont vrais.

Cet argument est malheureusement peu convaincant. Comme l'indique la prémisse 2 de l'argument, le principe du rasoir d'Occam ne donne qu'une raison défaisable, « toutes choses égales par ailleurs », d'adhérer à la conclusion. Or dans le cas précis des identités psycho-physiques, les adversaires du physicalisme prétendent avoir d'excellentes raisons de rejeter les énoncés d'identité du type de (4), à commencer par l'existence et la persistance d'un gouffre explicatif.

### *3.2. Une inférence à la meilleure explication partant des corrélations observées ?*

(McLaughlin, 2001) propose un argument qui diffère sensiblement de celui de Smart, tout en partageant la première prémisse — les données sur lesquelles se fonde l'argument sont les corrélations psycho-physiques — ainsi que son caractère indirect. Il s'agit d'une inférence à la meilleure explication, qui possède la structure suivante :

Prémisse 1 : la thèse de la corrélation est vraie.

Prémisse 2 : la vérité des énoncés de type (4) — des énoncés d'identité entre états phénoménaux et états cérébraux — constitue la meilleure explication disponible de la prémisse 1.

Conclusion : les énoncés de type (4) sont vrais.

Comme l'a souligné Kim à la suite de Block et Stalnaker[15], l'argument de McLaughlin possède une faiblesse : on ne voit pas en effet comment des énoncés d'identité pourraient, par eux-mêmes, posséder une force explicative, et en particulier expliquer des corrélations. L'idée même de corrélation semble en effet reposer sur la présupposition selon laquelle il existe une différence entre les événements corrélés. Si la propriété M est métaphysiquement identique à la propriété P, l'occurrence de M ne sera pas simplement corrélée à l'occurrence de P, mais bel et bien identique à cette occurrence : des occurrences de propriétés identiques sont évidemment des événements identiques et pas simplement des événements corrélés. Par ailleurs, il semble que les énoncés d'identité permettent le transfert d'explications déjà disponibles, mais qu'ils ne

15. Cf. (Kim, 2005), (Block et Stalnaker, 1999).

possèdent pas, en eux-mêmes, de force explicative. Considérons ainsi l'exemple suivant[16]. Supposons que nous disposions d'une explication biologique et génétique du fait que Superman mesure 1,90m, que nous abrégerons par la lettre E. Supposons par ailleurs qu'on veuille expliquer le fait que Clark Kent mesure également 1,90m. Voici une explication tout à fait satisfaisante de ce fait :

Prémisse 1 : E
Conclusion 1 : Superman mesure 1,90m (dérivée à partir de E)
Prémisse 2 : Superman = Clark Kent
Conclusion 2 : Clark Kent mesure 1,90m

Cependant, l'énoncé d'identité formulé dans la prémisse 2 ne contribue pas de façon positive à l'explication. Sa fonction est simplement d'opérer un *transfert d'explication* : accepter la prémisse 2 autorise en effet à appliquer l'explication E, formulée au départ à propos de Superman, à Clark Kent. En réalité, la prémisse 2 permet de transférer l'explication E en re-décrivant un fait d'une nouvelle manière : accepter l'énoncé « Superman = Clark Kent » nous autorise à décrire le fait « Superman mesure 1,90m » à l'aide de l'énoncé « Clark Kent mesure 1,90m ». Mais c'est un seul et unique fait qui est décrit par ces deux énoncés. Il est donc bien évident que la conclusion 1 et la prémisse 2 ne contribuent pas de façon substantielle à l'explication. Si c'était le cas, un fait — le fait que Clark Kent mesure 1,90m — serait expliqué par lui-même, puisque l'énoncé « Superman mesure 1,90m » n'est rien d'autre, comme nous venons de le voir, qu'une description différente du fait que Clark Kent mesure 1,90m.

Insistons bien : un énoncé d'identité ne peut pas constituer la prémisse explicative d'un raisonnement. Il ne permet pas d'expliquer de nouveaux faits, ni même d'étendre une explication déjà existante à de nouveaux faits. Sa seule fonction est de décrire d'une nouvelle façon un fait, et donc d'étendre une explication déjà disponible pour un fait sous une description donnée à une nouvelle description de ce fait.

On le voit, la stratégie proposée par (McLaughlin, 2001), qui consiste à soutenir que les énoncés d'identité psycho-physique sont justifiés car ils constitueraient la meilleure explication possible de la thèse de la corrélation, se heurte à une objection de taille : les énoncés d'identité

16. Je dois cet exemple, ainsi que l'idée de la discussion qui suit, à Filipe Drapeau-Contim, et je le remercie de m'autoriser à les reproduire ici.

ne peuvent tout simplement pas constituer des explications, et surtout pas une explication d'une corrélation puisque si ces énoncés sont vrais, il n'y a pas de corrélation psycho-physique, mais précisément une identité. Or s'ils ne peuvent constituer des explications à part entière, ils ne peuvent *a fortiori* pas non plus constituer de « meilleures explications », et la stratégie de justification indirecte de ces énoncés proposée par McLaughlin échoue donc.

Block et Stalnaker, qui ont perçu au moins en partie la difficulté que nous venons de discuter, ont néanmoins essayé de défendre la stratégie de justification indirecte par une inférence à la meilleure explication. Selon eux, cette inférence n'est pas fondée sur une explication des corrélations psycho-physiques (ou de toute autre corrélation entre les événements décrits par la théorie à réduire et ceux décrits par la théorie réductrice). Elle se fonde plutôt sur l'explication des phénomènes décrits par la science à réduire. Voici ce qu'écrivent Block et Stalnaker à ce propos :

> Pourquoi supposons-nous que la chaleur = l'énergie cinétique moyenne ? Considérons l'explication (...) de la raison pour laquelle chauffer de l'eau la fait bouillir. Supposons que la chaleur = l'énergie cinétique moyenne, que la pression = la quantité de mouvement des molécules, et que l'ébullition = une certaine sorte de mouvement moléculaire (...). Nous disposons alors d'une réponse à la question de savoir pourquoi chauffer de l'eau la met en ébullition. Si nous acceptions de simples corrélations à la place des identités, nous n'aurions qu'une réponse à la question de savoir pourquoi quelque chose qui se trouve être corrélé avec le chauffage de l'eau cause quelque chose qui se trouve être corrélé avec sa mise en ébullition. (...) Les identités permettent un transfert de la force explicative et de la causalité, que ne permettent pas de simples corrélations. [Les supposer] nous permet d'expliquer des faits que nous ne pourrions pas expliquer sinon. De la sorte, le principe de l'inférence à la meilleure explication nous justifie à inférer que ces identités sont vraies (1999, pp. 23-24).

Block et Stalnaker admettent que la fonction des énoncés d'identité n'est

pas à proprement parler d'expliquer, mais de transférer une explication : « les identités permettent un transfert de la force explicative et de la causalité, que ne permettent pas de simples corrélations ». En revanche, ils ne semblent pas conscients du fait que ces énoncés sont dénués de tout pouvoir explicatif propre, puisqu'ils soutiennent que les accepter « nous permet d'expliquer des faits que nous ne pourrions pas expliquer sinon ». Cette dernière phrase est en effet manifestement fausse : nous l'avons vu, accepter des énoncés d'identité ne permet pas d'augmenter le nombre de faits que l'on explique, mais simplement de décrire différemment des faits déjà connus, et donc d'appliquer à ces nouvelles descriptions d'anciennes explications également déjà connues. On peut admettre que la mécanique statistique nous fournisse la meilleure explication de la raison pour laquelle l'augmentation de l'énergie cinétique moyenne cause une certaine sorte de mouvement moléculaire — c'est-à-dire, si on accepte les énoncés d'identification théorique de la raison pour laquelle la chaleur cause la mise en ébullition de l'eau. Mais les énoncés d'identité ne jouent, en eux-mêmes, pas de rôle dans cette explication. Les accepter revient simplement à accepter que les phénomènes décrits par la mécanique statistique sont en fait exactement les mêmes que ceux qui sont décrits par la thermodynamique, et que les explications disponibles pour les premiers le sont également pour les seconds. Ne nous y trompons pas : la théorie physicaliste ne permet pas d'expliquer plus de faits que les théories dualistes. Un dualiste peut en effet accepter toutes les explications physicalistes disponibles des phénomènes cérébraux. Ce qu'il nie, c'est justement que l'on puisse transférer ces explications vers le domaine phénoménal, puisqu'il considère que les énoncés d'identité psycho-physique ne sont pas justifiés.

### *3.3 L'argument causal*

La dernière stratégie de justification indirecte des énoncés d'identité consiste à souligner que si la vérité de ces énoncés ne permet pas d'expliquer plus de faits, elle permet au moins de répondre à plus de questions. En particulier, supposer leur vérité permet de répondre à la question suivante : pourquoi les états phénoménaux possèdent-ils des pouvoirs causaux ? Nous savons déjà que les états cérébraux possèdent des

pouvoirs causaux ; accepter les énoncés d'identité psycho-physique conduit donc immédiatement à la conclusion selon laquelle les états phénoménaux possèdent exactement les mêmes pouvoirs : s'ils ne sont rien d'autre que des états cérébraux, ils possèdent aussi évidemment, d'après la loi d'indiscernabilité des identiques, tous les pouvoirs causaux de ces états. Voici comment David Papineau formule cet argument[17]:

> De nombreux effets que l'on attribue à des causes conscientes ont des causes physiques complètes. Mais il serait absurde de supposer que ces effets sont causés deux fois. En conséquence, les causes conscientes doivent être identiques à certaines parties de ces causes physiques. (Papineau, 2002, p. 17).

Le raisonnement de Papineau repose sur le principe de clôture causale du monde physique, selon lequel tout effet physique possède une cause physique qui suffit à l'expliquer. Si l'on accepte ce principe, puisque de nombreux états phénoménaux ont des effets physiques dans le comportement, la prémisse selon laquelle « de nombreux effets que l'on attribue à des causes conscientes ont des causes physiques complètes » doit être acceptée. Il repose aussi sur la thèse selon laquelle un même effet ne saurait avoir plusieurs causes différentes suffisantes à l'expliquer. Aucune de ces thèses ne peut être prouvée *a priori* : l'expression « absurde » employée par Papineau est sans doute trop forte. L'argument causal est donc, de nouveau, un argument abductif, une inférence à la meilleure explication : la vérité des énoncés d'identité psycho-physique constitue une explication du fait qu'ils aient des pouvoirs causaux que l'on peut reconnaître comme supérieure à l'explication que pourrait fournir un dualiste, qui consisterait ou bien à rejeter le principe de clôture causale du monde physique, ou bien à accepter qu'un même événement physique puisse avoir plusieurs causes susceptibles de l'expliquer.

17. Cf. (Papineau, 2002). Voir aussi, sur l'idée d'exclusion causale présentée ici par Papineau, (Kim, 1998) et (Kim, 2005).

## Conclusion

L'argument de Papineau me paraît acceptable : c'est en fait le seul argument indirect en faveur des identités psycho-physiques qui soit réellement convaincant. Il conduit semble-t-il cependant à sérieusement nuancer l'opposition entre physicalisme de type A et physicalisme de type B. Selon Papineau, les états phénoménaux ont des pouvoirs causaux, donc, il faut le supposer, un rôle causal : comment un état pourrait-il posséder des pouvoirs causaux, mais ne jouer aucun rôle causal, en effet ? Par ailleurs, la prise en considération de ces pouvoirs causaux constitue la seule explication des énoncés d'identité psycho-physique. Selon les physicalistes de type A, c'est également l'étude du rôle causal des états phénoménaux qui permet d'opérer une explication réductionniste des phénomènes mentaux liés à l'expérience consciente. S'il est possible d'étudier le rôle causal des états phénoménaux, on ne voit plus très bien comment l'on pourrait maintenir l'idée d'un gouffre explicatif.

## Références

Armstrong D. (1964) - *A Materialist Theory of Mind* (New York, Humanities Press)

Block N., Stalnaker R. (1999) - Conceptual Analysis, Dualism, and the Explanatory Gap (in *The Philosophical Review* 108, 1-46)

Chalmers D. (1996) - *The Conscious Mind* (Oxford, Oxford University Press)

Chalmers D. et Jackson F (2001) - Conceptual Analysis and Reductive Explanation (in *The Philosophical Review*, 110, 315-161)

Chalmers D. (2002) - Consciousness and its Place in Nature (in D. Chalmers (ed.) *Philosophy of Mind: Classical and Contemporary Readings*, New York, Oxford University Press)

Daly C. (2010) - *An Introduction to Philosophical Methods* (Buffalo, Broadview)

Hill C. (1991) - *Sensations* (Cambridge, Cambridge University Press)

Jackson F. (1982) - Epiphenomenal Qualia (in *Philosophical Quaterly* 32, 127-136)

Jackson F. (1998) - *From Metaphysics to Ethics: A Defense of Conceptual Analysis* (Oxford, Clarendon Press)

Kim J. (1998) - *Mind in a Physical World. An essay on the Mind-Body Problem and*

*Mental Causation* (Cambridge, Mass., MIT Press. Trad. fr. de F. Athané et E. Guinet (2006) - *L'esprit dans un monde physique. Essai sur le problème corps-esprit et la causalité mentale*, Paris, Syllepse)

Kim J. (2005) - *Physicalism, or Something Near Enough* (Princeton, Princeton University Press)

Kripke S. (1980) - *Naming and Necessity* (Cambridge, Mass., Harvard University Press)

Levine J. (1983) - Materialism and Qualia: the Explanatory Gap (in *Pacific Philosophical Quaterly* 64, 354-361)

Levine J. (1993) - On Leaving Out What It's Like (in M. Davies and G. W. Humphreys (eds.), *Consciousness* (Oxford, Blackwell))

Lewis D. (1970) - An Argument for the Identity Theory (in *The Journal of Philosophy* 67, 203-211)

Loar B. (1990) - Phenomenal States (in J. Tomberlin (ed.), *Philosophical Perspectives IV: Action Theory and the Philosophy of Mind*, 81-108 (Atascaredo, Ridgeview))

Ludlow P., Nagasawa Y., Stoljar D. (2004) - *There's Something About Mary. Essays on Phenomenal Consciousness and Frank Jackson's Knowledge Argument* (Cambridge, Mass., MIT Press)

McLaughlin B. (2001) - In Defense of New Wave Materialism: A Response to Horgan and Tienson (in C. Gillett et B. Loewer (eds.), *Physicalism and Its Discontents* (Cambridge, Cambridge University Press))

Papineau D. (2002) - *Thinking About Consciousness* (Oxford, Oxford University Press)

Polger T. W. (2002) - *Natural Minds* (Cambridge, Mass., MIT Press)

Putnam H. (1975) - *Mind, Language, and Reality: Philosophical Papers*, Vol. 2 (Cambridge, Cambridge University Press)

Smart, J.J.C. (1959) - Sensations and Brain Processes (in *The Philosophical Review* 68 : 141-156)

Stoljar D. (2006) - *Ignorance and Imagination. The Epistemic Origin of the Problem of Consciousness* (New York, Oxford University Press)