Philosophiques

Identités psychophysiques et inférence à la meilleure explication

Filipe Drapeau Vieira Contim and Pascal Ludwig

Volume 40, Number 1, Spring 2013

URI: https://id.erudit.org/iderudit/1018382ar
DOI: https://doi.org/10.7202/1018382ar

See table of contents

Publisher(s)

Société de philosophie du Québec

ISSN

0316-2923 (print)
1492-1391 (digital)

Explore this journal

Cite this article

Drapeau Vieira Contim, F. & Ludwig, P. (2013). Identités psychophysiques et inférence à la meilleure explication. *Philosophiques*, *40*(1), 171–195. https://doi.org/10.7202/1018382ar

Article abstract

Most type identity theorists endorse a posteriori physicalism about phenomenal properties. On such a view, mind/brain identity statements can be justified by an inference to the best explanation (IBE) starting from the empirical premise of mind/brain correlations. We contend that Identity Theory cannot be based on such an abductive methodology. First, we argue that mind/brain identity statements cannot be justified by an IBE from correlations as it leads to the following dilemma : either correlations are trivial facts that cannot pretend to the status of *explananda*, or mind/brain identities must be admitted as brute facts, just like dualist laws. Second, we show that these identity statements can still be justified by an IBE starting from the premise of mental causation. However, this strategy is of no help for the a posteriori brand of Identity Theory here challenged since the same premise forces one to embrace a priori physicalism.

Identités psychophysiques et inférence à la meilleure explication

FILIPE DRAPEAU VIEIRA CONTIM
Université de Rennes 1
filipe.drapeauvieiracontim@univ-rennes1.fr

PASCAL LUDWIG
Université de Paris-Sorbonne
pascal.ludwig@paris-sorbonne.fr

RÉSUMÉ. — La plupart des théoriciens de l'identité des types souscrivent à un physicalisme a posteriori à l'égard des propriétés phénoménales. Selon cette conception, les énoncés d'identité esprit/cerveau peuvent être justifiés par une inférence à la meilleure explication (IME) partant du fait empirique des corrélations esprit/cerveau. Nous soutenons que la théorie de l'identité ne peut pas s'appuyer sur cette méthodologie abductive. Nous montrons tout d'abord que l'on ne peut pas justifier les énoncés d'identité esprit/cerveau au moyen d'une IME partant des corrélations esprit/cerveau car cette stratégie conduit à un dilemme : ou bien les corrélations ne sont pas des *explananda*, ou bien les identités esprit/cerveau doivent être admises comme des faits bruts au même titre que les lois dualistes. Nous montrons ensuite que ces énoncés d'identité peuvent être néanmoins justifiés par une IME partant du pouvoir causal des états phénoménaux. Cette seconde stratégie n'est toutefois d'aucune aide pour la forme a posteriori de théorie de l'identité ici discutée dans la mesure où elle débouche sur un physicalisme de type a priori.

ABSTRACT. — Most type identity theorists endorse a posteriori physicalism about phenomenal properties. On such a view, mind/brain identity statements can be justified by an inference to the best explanation (IBE) starting from the empirical premise of mind/brain correlations. We contend that Identity Theory cannot be based on such an abductive methodology. First, we argue that mind/brain identity statements cannot be justified by an IBE from correlations as it leads to the following dilemma : either correlations are trivial facts that cannot pretend to the status of *explananda*, or mind/brain identities must be admitted as brute facts, just like dualist laws. Second, we show that these identity statements can still be justified by an IBE starting from the premise of mental causation. However, this strategy is of no help for the a posteriori brand of Identity Theory here challenged since the same premise forces one to embrace a priori physicalism.

1. Introduction

Selon la thèse de l'identité des types, toutes les propriétés mentales, y compris les propriétés phénoménales — celles pour lesquelles on peut dire que cela fait un certain effet de les posséder, comme avoir mal à une dent ou ressentir une peur soudaine —, sont numériquement identiques à des propriétés

neuronales. Cette thèse implique donc qu'un certain nombre d'énoncés du type de (1) sont vrais :

(1) La sensation de douleur n'est rien d'autre que telle oscillation cortico-thalamique.

Notre objectif dans cet article n'est pas d'argumenter en faveur de la thèse de l'identité des types. Nous la présupposerons. Le problème qui nous intéresse n'est en effet pas métaphysique, mais épistémologique. Il ne suffit pas que les énoncés comme (1) soient vrais pour qu'ils puissent être défendus par un scientifique ou par un philosophe physicaliste ; encore faut-il également qu'ils puissent être *justifiés*, sans quoi nous ne pourrions pas même savoir qu'ils sont vrais. Or quel type de justification peut-on apporter en faveur de ces énoncés d'identité ? Quelle(s) raison(s) un physicaliste peut-il avancer pour nous convaincre d'adopter sa position ?

Ces questions sont cruciales dans le contexte des discussions actuelles concernant l'existence d'un fossé explicatif entre le domaine des propriétés phénoménales et celui des propriétés physiques. En effet, bon nombre de philosophes physicalistes, et notamment les défenseurs de la thèse de l'identité des types, admettent qu'il existe, entre les deux domaines, une lacune dans l'explication qu'il est impossible de combler. D'où cette question pressante : si l'on ne peut pas expliquer les identités esprit/cerveau, peut-on au moins les justifier, et si tel est le cas, par quelle(s) méthode(s) ?

Nous commencerons par présenter l'influente typologie des positions physicalistes proposée par David Chalmers. Cette typologie est particulièrement importante pour notre propos car elle repose précisément sur des critères épistémologiques. Selon Chalmers, les physicalistes, et notamment les physicalistes de l'identité, se rangent en deux types principaux, ceux de type A et ceux de type B. Les premiers soutiennent que les énoncés d'identité esprit/cerveau peuvent être à la fois justifiés et expliqués par la méthode de réduction fonctionnelle. Les seconds, en revanche, considèrent que ces énoncés ne sont pas explicables, mais qu'ils peuvent être justifiés de façon indirecte, à l'aide d'une *inférence à la meilleure explication* (IME) appuyée sur la découverte de corrélations esprit/cerveau.

Ensuite, nous nous proposons d'examiner le bien-fondé de la méthode abductive adoptée par le physicalisme de type B. Notre objectif est de montrer que le physicalisme ne peut pas s'appuyer sur l'IME. Notre objection procèdera en deux temps.

Tout d'abord, partant d'une objection de Jaegwon Kim, nous soutiendrons que les énoncés d'identité esprit/cerveau ne peuvent pas être justifiés par IME à partir des *corrélations* esprit/cerveau. Nous montrerons que le physicaliste de type B est enfermé dans un dilemme : ou bien les identités esprit/cerveau sont inexplicables car ce sont des faits triviaux qui ne demandent pas d'explication, auquel cas le même raisonnement montre que les corrélations esprit/cerveau n'ont pas elles-mêmes à être expliquées, privant

ainsi le physicalisme de type B de sa base abductive. Ou bien les corrélations peuvent être expliquées par les identités esprit/cerveau, auquel cas celles-ci sont des faits bruts, c'est-à-dire des faits qui ne peuvent pas être expliqués alors qu'ils le devraient ; dans cette seconde branche du dilemme, l'hypothèse de l'identité esprit/cerveau n'est pas un meilleur candidat à l'explication des corrélations que l'hypothèse rivale dualiste.

Nous montrerons ensuite que si les identités esprit/cerveau ne peuvent pas être inférées par IME à partir des corrélations, elles peuvent l'être néanmoins en partant du *pouvoir causal* des états phénoménaux. Nous soutiendrons cependant que cette nouvelle base abductive ne sort pas le physicaliste de type B de l'impasse dans la mesure où la prémisse causale requise pour l'IME fournit aussi le point de départ d'une réduction fonctionnelle des identités esprit/cerveau, ce qui débouche sur le physicalisme de type A. L'IME ne peut donc fournir l'épistémologie attendue par le physicalisme de type B : il ne semble pas y avoir de méthode qui permettrait de justifier les énoncés d'identité esprit/cerveau (à la différence du dualisme) tout en les tenant pour inexplicables (à la différence du physicalisme de type A).

2. Physicalisme de type A et réduction fonctionnelle

Selon Chalmers[1], on peut distinguer deux versions de la théorie de l'identité des types, qui diffèrent relativement au statut épistémique accordé au conditionnel :

(PTQ) Si PT, alors Q

où « P » est la conjonction de toutes les vérités physiques, « T » est une clause de clôture stipulant qu'il n'y a pas d'autres faits que ceux décrits par « P », et « Q » est la conjonction de toutes les vérités, à l'inclusion des vérités psychologiques, y compris celles décrites à l'aide d'un vocabulaire phénoménal (« sensation de douleur », « sensation visuelle », etc.).

Selon la thèse de l'implication *a priori*, aussi appelée « physicalisme de type A » par Chalmers, le conditionnel (PTQ) est à la fois nécessaire et *a priori*.

La nécessité de ce conditionnel est une conséquence du physicalisme en général. Supposons en effet que (PTQ) soit vrai mais de façon contingente. Cela signifie qu'il existe un monde métaphysiquement possible w qui est tel que : (i) tous les énoncés physiques qui sont vrais du monde actuel sont vrais de w, w est donc un monde physiquement indiscernable du nôtre ; (ii) w ne contient que des faits physiques ; (iii) il y a au moins un énoncé relevant du vocabulaire phénoménal, appelons-le « q », qui est vrai de notre monde et qui est faux de w. Il s'ensuit que le fait actuellement réalisé décrit par « q » constitue un fait « supplémentaire » ou « quelque chose de plus » que les faits

1. Chalmers, 1996, p. 166.

décrits par la conjonction P d'énoncés physiques, autrement dit : le fait que q est non physique, ce qui contredit l'hypothèse physicaliste. Par conséquent, (PTQ) est nécessairement vrai si le physicalisme est vrai.

Le physicalisme de type A a ceci de spécifique qu'il soutient que (PTQ) est non seulement nécessaire mais également *a priori*, ce qui ne va pas de soi même lorsqu'on est physicaliste. Cela revient en effet à dire que n'importe quel énoncé vrai contenant un terme phénoménal peut être *déduit a priori* à partir d'une conjonction arbitrairement grande de vérités physiques[2]. Il ne s'agit cependant pas d'un réductionnisme linguistique : les physicalistes de type A ne considèrent pas qu'il est possible de *traduire* directement les énoncés psychologiques dans le vocabulaire neuro-scientifique. Ils adhèrent plutôt à la méthode de *réduction par fonctionnalisation* défendue dans de nombreuses publications par David Lewis[3], et récemment popularisée par David Chalmers[4] et Frank Jackson[5].

La méthode de réduction fonctionnelle procède en trois grandes étapes. Afin de les illustrer, nous commencerons par un exemple de réduction moins controversé que celui des états phénoménaux : celui de l'eau.

(i) *L'étape de la régimentation*
Il s'agit de l'étape de l'analyse conceptuelle à proprement parler. Le but de l'analyse n'est pas de parvenir à une définition analytique du terme « eau » (procédant par exemple par conditions nécessaires et suffisantes) mais de recueillir les truismes contenant le terme qui sont communément acceptés. On peut ainsi facilement extraire une théorie naïve de l'eau de nos comportements linguistiques et de nos dispositions inférentielles, et cette théorie implique un certain nombre de truismes : l'eau est habituellement un liquide, on la trouve dans la mer, les lacs, les océans et les rivières ; lorsqu'elle est pure, l'eau est transparente ; l'eau désaltère ; c'est un solvant, etc.

(ii) *L'étape de la ramseyfication*
Il s'agit du processus logique d'abstraction qui permet de définir fonctionnellement une expression à partir des truismes dans lesquels elle figure. Cela consiste à forger une conjonction de ces truismes puis à remplacer toutes les occurrences de l'expression par une même variable liée par un quantificateur existentiel. Pour le terme « eau », nous obtenons :

2. Il importe de rappeler que le physicalisme de type A, ainsi défini par la thèse du caractère nécessaire et a priori de (PTQ), inclut non seulement la théorie de l'identité mais également le fonctionnalisme analytique, qui rejette les identités esprit/cerveau en arguant de la réalisabilité multiple du mental. Notre objet étant la justification des identités esprit/cerveau, notre présentation du physicalisme de type A s'en tiendra dans la suite à la théorie de l'identité.

3. Lewis, 1966, 1970 et 1972.

4. Chalmers, 1996.

5. Jackson, 1998.

(2) Il existe un x tel que x est un liquide, x coule dans les rivières, les fleuves, les lacs, les océans, x est transparent à l'état pur, x désaltère, x est un solvant, etc.

En présumant qu'il y a un unique objet satisfaisant ce prédicat conjonctif, on peut forger à partir de (2) une description définie fonctionnelle[6] :

(3) L'unique x qui est tel que x est un liquide, x coule dans les rivières, les fleuves, les lacs, les océans, x est transparent à l'état pur, x désaltère, x est un solvant, etc.

Pour faire court, abrégeons le prédicat conjonctif au moyen du prédicat « x est une substance aqueuse ».

(iii) *L'étape de l'identification*
Il s'agit de l'étape proprement *empirique* du processus de réduction, qui laisse place à l'investigation scientifique. Celle-ci a pour charge d'identifier une certaine entité, décrite dans un vocabulaire jugé plus fondamental, qui répond le mieux au rôle fonctionnel extrait par l'analyse conceptuelle. Dans le cas de l'eau, il s'agit par exemple d'établir que c'est la substance composée de molécules de H_2O qui répond le mieux à la description fonctionnelle « la substance aqueuse » attachée au terme « eau »[7].

Ce n'est qu'une fois ces trois étapes accomplies que la réduction fonctionnelle peut être menée à bien. Celle-ci prend la forme d'une *déduction* : l'information empirique formulée dans le vocabulaire de la théorie réductrice (étape (iii)), combinée avec l'analyse conceptuelle du terme relevant de la théorie réduite (étape (ii)), permet de dériver de façon purement a priori tous les énoncés vrais engageant ce terme. Dans le cas de l'eau, la réduction procède comme suit :

> **Prémisse** 1 (justifiée a posteriori à partir de l'investigation scientifique) : la substance aqueuse = la substance composée de molécules de H_2O.

---

6. Comme le rappellent Chalmers et Jackson, 2001, cela ne suppose nullement qu'un locuteur compétent pour ce qui est du terme « eau » puisse l'analyser explicitement au moyen d'une description définie fonctionnelle, ni que les réductions en science prennent littéralement la forme de ramseyfication. Il s'agit là d'une reconstruction rationnelle de connaissances implicites que les locuteurs manifestent en appliquant le terme à des scénarios décrits dans un vocabulaire plus fondamental.

7. L'étape (iii) présuppose que l'extension de certains termes figurant dans des théories scientifiques (ici « H2O ») peut recouper celle de termes issus de théories naïves (« eau »). Cette correspondance a toutefois été contestée pour ce qui est des termes de sortes biologiques (Dupré 1993) et de substances chimiques (Needham 2000, LaPorte 2004, Weisberg 2006), la critique visant notamment l'exemple de l'identification théorique eau/H2O mis en avant par les partisans de la réduction fonctionnelle. Pour une défense de l'identification théorique eau/H2O et de la correspondance entre termes vernaculaires et termes scientifiques de substances chimiques, voir la réponse du philosophe de la chimie Robert Findlay Hendry dans (Hendry 2006) et (Hendry 2010).

Prémisse 2 (justifiée a priori par analyse conceptuelle) : eau = la substance aqueuse.

Conclusion (déduite a priori des prémisses 1 et 2) : eau = la substance composée de molécules de H_2O.

La conclusion prend ici la forme d'un énoncé d'identité, et plus exactement celle d'un énoncé d'identification théorique. Il s'ensuit que tout énoncé contenant le terme « eau » peut être déduit a priori d'énoncés concernant la substance H_2O. Le programme réductionniste est donc rempli puisqu'il fournit une explication formulée dans un vocabulaire microphysique de toutes les propriétés macrophysiques de l'eau.

L'exemple de l'eau est particulièrement éclairant pour la question qui nous occupe, car il permet de dégager le rôle crucial que joue la réduction fonctionnelle dans l'épistémologie des énoncés d'identités : la réduction fonctionnelle du terme « eau » permet à la fois d'*expliquer* la vérité de l'énoncé d'identité eau/H_2O et de le *justifier*. Autrement dit, elle répond à la question « comment peut-il se faire que l'eau soit identique à un ensemble de molécules de H_2O ? », et, ce faisant, elle justifie ce jugement d'identité. Cela ne signifie pas qu'il faille attendre que l'explication réductive soit achevée, pour être justifié à croire que l'eau = la substance H_2O. Ainsi, ce n'est que très récemment que l'on est enfin parvenu à fournir une explication microphysique de la transparence de l'eau[8], et l'explication de sa fluidité fait toujours débat à l'heure actuelle. L'explication réductive des propriétés de l'eau n'est donc pas encore achevée, mais cela ne nous empêche pas de *savoir* que l'eau = la substance H_2O dans la mesure où un nombre suffisant de ses propriétés macrophysiques (son inertie thermique, son caractère désaltérant, la solidité de la glace, etc.) ont déjà reçu une explication réductive, nous permettant de justifier l'identification théorique eau/H_2O, par anticipation pourrait-on dire de l'explication réductive complète. Cela montre encore une fois le lien étroit entre l'explication réductive des identités et leur justification : nous ne serions pas dès à présent justifiés à croire que l'eau = la substance H_2O, si nous avions des raisons de penser qu'il est impossible par principe de fournir une explication réductive complète des propriétés de l'eau. Comme le souligne Malaterre[9], les scientifiques qui ont soutenu (à tort) qu'aucune explication microphysique de la transparence de l'eau ne pouvait être donnée sont aussi ceux qui ont tenu l'eau liquide ou solide pour un phénomène

8. La transparence de l'eau, c'est-à-dire le fait qu'une faible épaisseur d'eau laisse passer environ 98 % du spectre du visible, est désormais expliquée de façon réductive par le fait que les longueurs d'onde du visible ne correspondent que de façon très marginale aux modes vibratoires mécaniques de la molécule de H2O. Sur ce point, voir Chaplin, 2007, et Malaterre, 2012.

9. Malaterre, *ibid.*

émergent, numériquement distinct d'un réseau de molécules H_2O connectées par des liaisons hydrogène.

Les théoriciens de l'identité des types qui souscrivent au physicalisme de type A, comme par exemple Lewis[10] et Armstrong[11], soutiennent que les identités esprit/cerveau peuvent recevoir une explication réductive, au même titre que n'importe quelle autre identification théorique (eau/H_2O, température d'un gaz/énergie cinétique moyenne des molécules composant ce gaz, etc.).

Considérons par exemple la propriété phénoménale de la douleur. L'explication réductive de la douleur suppose tout d'abord que l'on procède à une analyse fonctionnelle du terme « douleur ». Cela consiste à extraire le rôle causal-fonctionnel de l'état de douleur à partir des énoncés de la psychologie naïve dans lesquels le terme « douleur » figure. Ces énoncés rapportent par exemple que l'état de douleur est causé dans des conditions normales par des coupures, des brûlures ou des pressions potentiellement dommageables au corps, que l'état cause à son tour tel comportement typique d'évitement, etc. Par commodité, abrégeons « le R », la longue description définie obtenue par « ramseyfication » des énoncés rapportant les causes et les effets typiques de la douleur.

Dans un second temps, les neurosciences identifient l'état neural, éventuellement très complexe, qui répond le mieux au rôle causal-fonctionnel extrait par l'analyse conceptuelle du terme « douleur ». Admettons par exemple qu'une certaine oscillation cortico-thalamique satisfasse la condition exprimée par la description définie « le R », et appelons « O » cet état d'activation neuronale[12].

---

10. Lewis, 1966, 1972 et 1994.

11. Armstrong, 1968.

12. Comme dans le cas de la réduction fonctionnelle de l'eau, cette étape suppose que les catégories phénoménales de la psychologie naïve puissent recouper en extension certaines catégories des neurosciences. S'il est douteux que les états cognitifs (croyances, désirs, etc.) de la psychologie naïve correspondent à des réalités neurofonctionnelles, la recherche sur les corrélats neuronaux des états phénoménaux plaide en revanche pour une concordance des taxinomies naïve et scientifique. Ainsi, les différents aspects des dimensions sensorielle et affective qui structurent la phénoménologie de la douleur et au moyen desquels les sujets classent leurs expériences, comme le type d'une douleur, son intensité, sa dynamique spatio-temporelle ou son caractère déplaisant, correspondent à des processus neuroanatomiques et neurophysiologiques bien spécifiques, qu'il s'agisse des récepteurs (fibres C pour les douleurs primaires, aiguës et localisées, fibres A-delta pour les douleurs secondaires, sourdes et diffuses) ou des aires de traitement de l'information nociceptive (cortex cingulaire antérieur pour l'aspect déplaisant d'une douleur, cortex somatosensoriel primaire pour son intensité). Voir Polger et Sufka, 2005. Ajoutons enfin que les neurosciences utilisent le terme « douleur » en continuité avec son usage ordinaire, comme en témoigne la définition qu'en donne l'IASP (International Association for the Study of Pain), qui reprend les principales caractéristiques attribuées par la psychologie naïve (Aydede, 2005, p. 1-5).

L'explication réductive de la douleur (notée ici « D ») prend la forme suivante :

Prémisse 1 (justifiée a posteriori à partir des neurosciences) : le R = O.

Prémisse 2 (justifiée a priori par analyse conceptuelle) : D = le R.

Conclusion (déduite a priori des prémisses 1 et 2) : D = O.

Cette déduction permet à la fois d'expliquer la vérité de l'énoncé d'identité D = O, et, ce faisant, de le justifier. Le physicaliste de type A soutient que tous les états phénoménaux peuvent recevoir une telle explication réductive. Plus encore : selon lui, nous ne pouvons pas être justifiés à accepter un énoncé d'identité esprit/cerveau si nous avons des raisons de penser qu'il est impossible d'en donner une explication réductive. C'est précisément ce lien étroit entre la justification des énoncés d'identité esprit/cerveau et leur explication que le physicalisme de type B prétend rejeter.

3. La thèse de l'implication a posteriori : le physicalisme de type B

Les physicalistes de type B, tels que Block et Stalnaker[13], et Papineau[14], sont sceptiques vis-à-vis de l'idée selon laquelle l'analyse conceptuelle permettrait d'associer des rôles causaux aux concepts phénoménaux[15]. Ils s'accordent avec les philosophes dualistes pour considérer que le problème de la conscience phénoménale ne pourra pas être résolu par la méthode de réduction fonctionnelle proposée par les physicalistes de type A : même une connaissance complète du monde physique ne permettrait pas, selon eux, de dériver a priori les énoncés d'identité esprit/cerveau. Ils soutiennent cependant qu'il existe non pas un gouffre ontologique entre le domaine physique et le domaine phénoménal, mais un gouffre conceptuel. Selon eux, la maîtrise de concepts phénoménaux les concepts des sensations de couleurs par exemple, ou des sensations corporelles comme la douleur ne permet pas de leur faire correspondre des rôles causaux à l'aide d'une analyse conceptuelle. Autrement dit, le physicalisme de type B accepte les trois thèses suivantes :

(T_1) Il existe un fossé explicatif entre le domaine physique et le domaine phénoménal : on ne peut pas dériver a priori les vérités phénoménales à partir des vérités physiques. Le conditionnel (PTQ) est nécessaire mais connaissable *seulement empiriquement*.

(T_2) Ce fossé explicatif ne vient pas d'un dualisme ontologique des propriétés (selon lequel les propriétés physiques et les propriétés phénoménales sont numériquement distinctes) mais d'un dualisme conceptuel :

13. Block et Stalnaker, 1999.

14. Papineau, 2002.

15. Un concept phénoménal est un concept dénotant une propriété phénoménale, et dont la maîtrise repose, de façon constitutive, sur le fait d'avoir vécu en première personne les expériences dénotées.

les mêmes propriétés cérébrales sont conçues en première personne à l'aide de concepts phénoménaux — et l'on parle alors de propriétés phénoménales, mais ce ne sont que des propriétés cérébrales conçues à l'aide de concepts spéciaux — et en troisième personne, à l'aide de descriptions physico-chimiques.

(T_3) L'analyse fonctionnelle ne permet pas d'associer des rôles causaux aux concepts phénoménaux ; pour cette raison, le fossé explicatif ne peut être comblé par une réduction fonctionnelle.

La thèse (T_3) est fondamentale, puisque c'est elle qui distingue le physicalisme de type A du physicalisme de type B[16]. Un physicaliste de type A peut en effet reconnaître qu'il existe, à un certain moment, une lacune explicative entre le domaine physique et le domaine phénoménal. En revanche, il rejettera (T_3) : 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 physique, doit aussi permettre de combler le fossé explicatif.

Le physicaliste de type B est donc engagé sur une voie étroite. D'un côté, il rejette fermement l'idée que les identités esprit/cerveau puissent recevoir une explication réductive, à la différence de toutes les autres identifications théoriques ; elles font donc figure d'exceptions, ce dont la théorie *ad hoc* des concepts phénoménaux est censée rendre compte. De l'autre, il rejette aussi le dualisme des propriétés : si les propriétés phénoménales sont identiques à des propriétés physiques, on ne peut pas considérer que les premières *émergent* à partir des secondes. Il faut donc soutenir à la fois que les énoncés d'identités esprit/cerveau peuvent être justifiés a posteriori, et qu'ils décrivent des *faits bruts*, c'est-à-dire des faits inexplicables ou irréductibles. Mais, si l'on ne dispose plus d'explication réductive pour justifier ces énoncés d'identité, par quelle méthode empirique pourra-t-on alors les justifier ?

4. L'argument abductif en faveur des identités esprit/cerveau

Les physicalistes de type B font valoir que l'impossibilité d'une réduction explicative des identités esprit/cerveau n'empêche nullement de les connaître car il existe une méthode permettant de justifier ces énoncés : l'*inférence à la meilleure explication* (IME) à partir des corrélations esprit/cerveau. L'argument est ancien, on le retrouve notamment chez Smart[17], l'un des pères de la

16. Chalmers distingue en outre un physicalisme de type C, mais nous n'en parlerons pas ici dans la mesure où cette forme de physicalisme s'accorde avec le physicalisme de type A sur la possibilité d'une réduction explicative des propriétés phénoménales. Voir Chalmers, 2002, p. 257-260.

17. Smart, 1959.

théorie de l'identité des types, mais il a été remis au goût du jour par des physicalistes de type B tels que Hill[18] et McLaughlin[19].

L'argument par IME en faveur des identités esprit/cerveau part d'un fait qui n'est contesté par personne, pas même par les philosophes dualistes : il est possible de découvrir des corrélats neuronaux des propriétés phénoménales[20]. Il s'agit donc de partir d'une prémisse empirique entièrement non problématique, la « thèse de la corrélation » :

> **Thèse de la corrélation** : pour tout type d'état de conscience phénoménal Q, il existe un type d'état neuronal N tel que pour tout organisme x, x est dans l'état Q si et seulement si x est dans l'état N.

Les physicalistes de type B soutiennent ensuite que les énoncés d'identité esprit/cerveau de la forme :

$$Q = N$$

fournissent la meilleure explication possible des corrélations esprit/cerveau de la forme :

$$\text{Pour tout organisme x, (Qx ssi Nx)}$$

L'argument abductif en faveur du physicalisme de type B repose donc sur la thèse centrale suivante :

> **Thèse de la meilleure explication (ME) :** la vérité des énoncés d'identité esprit/cerveau constitue la meilleure explication disponible de la vérité de la thèse de la corrélation.

L'hypothèse rivale est celle du dualisme des propriétés, hypothèse selon laquelle les corrélations esprit/cerveau s'expliquent par l'existence de lois psychophysiques *ad hoc*. Selon Hill et McLaughlin, (ME) est vraie car les identités esprit/cerveau fournissent une explication des corrélations qui est plus *simple* que celle offerte par les lois dualistes, à un double titre : premièrement, et c'est l'argument de simplicité considéré comme ayant le plus de poids, la théorie de l'identité des types est bien plus parcimonieuse en matière de lois de la nature fondamentales que sa rivale. Le dualisme des propriétés

18. Hill, 1991.

19. McLaughlin, 2001 et 2010.

20. Bien que les philosophes s'accordent sur l'existence de telles corrélations, ce n'est que très récemment, sous l'impulsion de travaux pionniers comme ceux de Crick et Koch sur l'expérience visuelle consciente (Crick et Koch 1990), que les neurosciences ont vu fleurir des programmes de recherche dédiés à l'identification des corrélats neuronaux des différents types de conscience phénoménale (voir Metzinger, 2000). L'une des difficultés vient de ce que l'étude neuroscientifique des états phénoménaux, à la différence de celle de la cognition, doit faire appel au jugement introspectif des sujets d'expérience. Pour une discussion des problèmes méthodologiques posés par le recours à l'introspection dans le cas de l'étude de la douleur, voir notamment les essais 13 à 18 réunis dans Aydede, 2005.

implique en effet de postuler un grand nombre de lois fondamentales *ad hoc* reliant les deux domaines pour rendre compte de chaque type de corrélation constaté, là où la théorie de l'identité ne rajoute rien au petit nombre de lois fondamentales déjà admises par les sciences de la nature. Deuxièmement, la théorie de l'identité surpasse sa rivale en termes non seulement de simplicité nomologique mais aussi de simplicité ontologique : elle n'admet dans son ontologie qu'un seul type de propriétés, les propriétés physiques (dont certaines, les propriétés cérébrales, sont aussi désignées au moyen de termes mentaux) tandis que le dualisme postule deux types hétérogènes de propriétés, les propriétés physiques *et* les propriétés mentales.

Concédons provisoirement que (ME) est vraie, et admettons aussi que l'inférence à la meilleure explication constitue une authentique méthode de justification épistémique au sens suivant : si l'on est justifié à croire qu'une hypothèse H explique mieux certaines données que les autres hypothèses rivales disponibles, alors on est justifié à croire que H est vraie, en un sens de « justification » suffisamment fort pour procurer de la connaissance. Il semble donc que le physicalisme de type B ait trouvé dans IME une méthode qui, partant du fait empirique des corrélations, permet de justifier a posteriori les identités esprit/cerveau, tout en maintenant la ligne dure d'après laquelle ces identités ne peuvent pas elles-mêmes être expliquées.

Dans la suite de notre discussion, nous allons nous concentrer en premier lieu sur le débat qui a récemment opposé Kim, d'un côté, à McLaughlin, Hill et Bates de l'autre. Kim remet radicalement en question la méthode abductive adoptée par les physicalistes de type B[21]. Kim avance plusieurs objections, mais nous ne retiendrons que l'argument le plus fort et sans doute aussi le plus discuté. Il consiste à dire que les identités esprit/cerveau ne peuvent pas être inférées à la meilleure explication à partir des corrélations pour la simple raison que ces identités ne peuvent pas les expliquer *tout court*. Selon Kim, cela ne tient pas spécifiquement aux identités esprit/cerveau, mais aux identités en général : les énoncés d'identité sont dénués de pouvoir explicatif, ils ne peuvent pas même prétendre au rôle d'*explanantia*.

5. L'argument de Kim à l'encontre du pouvoir explicatif des identités

Dans son ouvrage de 2005, Kim soutient que les identités esprit/cerveau ne peuvent pas bénéficier d'une IME partant des corrélations dans la mesure où, en tant qu'identités, elles sont dépourvues de pouvoir explicatif. Kim admet bien sûr que les énoncés d'identité peuvent figurer à titre de prémisses dans des explications, mais selon lui ces énoncés ne sont pas des *prémisses explicatives* : ils interviennent uniquement comme des règles de réécriture permettant de *transférer* des explications déjà disponibles en réécrivant certains faits. Kim écrit ainsi :

21. Kim, 2005, p. 121-148.

> Les identités semblent jouer le rôle de règles de réécriture dans les contextes inférentiels ; elles n'engendrent aucune connexion explicative entre l'*explanandum* et les phénomènes qui sont mentionnés dans l'*explanans*. Elles semblent ne posséder aucun pouvoir explicatif propre[22].

> Les identités ne semblent pas capables d'engendrer par elles-mêmes des explications ; au mieux, elles permettent de « transférer » des explications déjà achevées de les transférer non pas d'un phénomène à un autre phénomène, mais bien plutôt de la description d'un phénomène à une autre description du même phénomène[23].

Afin de mieux comprendre la position de Kim, prenons un exemple d'explication moins controversée que celle des corrélations esprit/cerveau. Supposons ainsi qu'on dispose déjà d'une explication scientifique parfaitement convaincante, par exemple issue de considérations génétiques et développementales complexes, du fait que Jean Giraud mesure 1,75 m. Supposons aussi qu'on apprenne par la suite que Jean Giraud = Moebius, alors qu'on cherche justement à expliquer le fait que Moebius mesure 1,75 m. L'explication de la taille de Moebius prend la forme de la déduction suivante (ci-dessous « H » abrège la théorie scientifique et les énoncés de conditions initiales dont on a besoin pour expliquer que Jean Giraud mesure 1,75 m) :

(i) H,
(ii) Jean Giraud mesure 1,75 m, par (i)
(iii) Or, Moebius = Jean Giraud,
(iv) Donc, Moebius mesure 1,75 m. par (ii) et (iii)

L'énoncé d'identité (iii) figure comme prémisse dans la déduction qui permet d'expliquer la taille de Moebius ; il joue donc un rôle dans l'argument, mais ce rôle n'est pas explicatif selon Kim. La fonction de (iii) est de re-décrire le fait relatif à la taille de Jean Giraud au moyen du nom « Moebius », de telle sorte que l'*explanans* H, qui porte la vraie charge de l'explication, puisse être transféré de l'énoncé (ii) vers l'énoncé (iv). En effet, en affirmant l'identité de Moebius et de Jean Giraud, nous nous apercevons qu'il n'y a pas deux faits demandant des explications distinctes, mais un même fait décrit de deux façons différentes. Or, s'il n'y a pas de fait supplémentaire, il n'y a pas non plus de nouvelle explication à donner. Le seul *explanans* ici est donc la théorie H qui explique tout à la fois le fait décrit par (ii) et celui — le même — décrit par (iv). L'énoncé d'identité joue un rôle comparable à celui des règles logiques et des tautologies : ce n'est pas parce que nous les utilisons lors d'un argument explicatif qu'elles font pour autant partie de l'*explanans*.

22. Kim, 2005, p. 132.
23. Kim, 2005, p. 146.

Si Kim voit juste, les identités n'expliquent rien, tout au plus transfèrent-elles des explications. Cette position nous semble motivée par deux principes, que nous allons discuter successivement.

Le premier est le principe de transparence :

> **Transparence des faits** : Si a = b, alors le fait que Fa = le fait que Fb (où « a » et « b » sont des termes authentiquement référentiels et « F » un prédicat ne contenant aucune expression hyperintensionnelle).

Selon ce principe, il suffit que les contextes « p » et « q » soient transparents pour que le contexte « le fait que p = le fait que q » le soit également[24]. En un mot, des faits sont identiques lorsque les mêmes propriétés ou relations sont attribuées aux mêmes objets[25]. Il s'ensuit que si Moebius = Jean Giraud, le fait que Moebius mesure 1,75 m = le fait que Jean Giraud mesure 1,75 m.

Le second principe auquel le raisonnement de Kim nous semble faire appel vise à empêcher toute circularité dans une explication, en stipulant que si celle-ci relie des faits, cette relation doit être irréflexive :

> **Irréflexivité de l'explication** : Un fait donné ne peut pas être invoqué dans sa propre explication.

Afin de vérifier que la thèse de Kim selon laquelle les identités n'ont pas un rôle explicatif découle des deux principes que nous avons présentés, supposons que ces principes soient vrais, mais que l'identité mentionnée dans l'étape (iii) de l'explication de la taille de Moebius possède néanmoins une fonction explicative propre. Si c'est le cas, on peut décomposer en deux parties l'explication de la taille de Moebius :

> *Sous-explication* 1 :
> H,
> Donc, Jean Giraud mesure 1,75 m.
>
> *Sous-explication* 2 :
> Jean Giraud mesure 1,75 m,
> Or, Jean Giraud = Moebius,
> Donc, Moebius mesure 1,75 m.

D'après le principe de transparence, le fait que Jean Giraud mesure 1,75 m = le fait que Moebius mesure 1,75 m. La sous-explication 2 constitue donc une infraction au principe d'irréflexivité, puisque le même fait figure

---

24. Un contexte est transparent lorsque deux expressions nécessairement co-extensionnelles peuvent y être substituées l'une par l'autre *salva veritate*. Le principe de transparence n'autorise pas une telle substitution lorsque la description d'un fait contient un verbe hyperintensionnel, comme dans « le fait que Marie croit que Jean Giraud mesure 1,75 m ».

25. Voir Wilson, 1974.

dans une des prémisses et dans la conclusion. On peut conclure, par *reductio*, que si l'on accepte les principes de transparence et d'irréflexivité, l'énoncé d'identité (iii) n'est pas une prémisse explicative dans l'explication du fait que Moebius mesure 1,75 m.

Appliquons à présent la thèse de Kim à l'explication des phénomènes mentaux. Soient par exemple Q_1 l'expérience auditive d'un son bref et intense, et Q_2 une émotion (consciente) de peur. Supposons que nous sachions que c'est une loi psychologique que Q_1 cause Q_2. Supposons également que les neurosciences aient établi l'existence d'une corrélation entre N_1 (un certain type d'activité neuronale dans le cortex auditif) et Q_1 d'une part, et entre N_2 (un certain type d'activité neuronale dans le cortex orbitofrontal) et Q_2 d'autre part. Enfin, supposons que les neurosciences aient fourni une explication physico-chimique du fait que N_1 cause N_2[26]. L'explication du fait que Q_1 cause Q_2 prendra la forme suivante :

(i) Explication neuroscientifique,
(ii) N_1 cause N_2, par (i)
(iii) $N_1 = Q_1$,
(iv) $N_2 = Q_2$,
(v) Donc, Q_1 cause Q_2. par (ii), (iii) et (iv)

Si Kim a raison, les prémisses (iii) et (iv) ne jouent pas réellement de rôle explicatif par elles-mêmes. Elles permettent simplement de réaliser que l'énoncé « Q1 cause Q2 » n'est qu'une re-description psychologique du fait décrit par l'énoncé « N1 cause N2 » dans le vocabulaire des neurosciences, ce qui permet du coup de transférer l'explication neuroscientifique (i) du fait décrit par le second énoncé à celui décrit par le premier. Mais ce sont les neurosciences qui expliquent le phénomène de causalité mentale observé, et non, en eux-mêmes, les énoncés d'identité (iii) et (iv).

Selon Kim, cette conclusion s'étend à tous les contextes d'explication : les identités esprit/cerveau ne jouent aucun rôle explicatif ; elles ne peuvent donc pas *a fortiori* faire l'objet d'une IME. L'argument abductif avancé par le physicalisme de type B constitue une erreur méthodologique.

6. La controverse Kim/Bates concernant l'explication des corrélations esprit/cerveau

Dans un article récent qui prend la défense de l'argument abductif[27], Jared Bates répond à l'objection de Kim en arguant que les identités esprit/cerveau

26. Pour la commodité de l'exposé, nous simplifions ce qui n'a ici qu'une valeur illustrative. En l'état actuel des connaissances, on ne peut pas exclure que N1, c.-à-d. le corrélat de l'audition consciente d'un son bref et intense, et N2, c.-à-d. le corrélat de l'émotion consciente de peur, soient en réalité deux effets concomitants d'une même cause, par exemple un circuit en boucle reliant l'amygdale et le cortex auditif.

27. Bates, 2009.

jouent bien un rôle explicatif dans le cas particulier de l'explication des corrélations esprit/cerveau. Bates montre ainsi qu'il existe une dérivation partant des identités esprit/cerveau et aboutissant aux corrélations :

Dérivation de Bates :

(i) Q = N	(Prémisse)
(ii) Qa	(Hypothèse)
(iii) Na	(i) × (ii) × (Loi de Leibniz)
(iv) Qa → Na	(ii) × (iii) × (Introduction du conditionnel)
(v) Na	(Hypothèse)
(vi) Qa	(i) × (v) × (Loi de Leibniz)
(vii) Na → Qa	(v) × (vi) × (Introduction du conditionnel)
(viii) Qa ↔ Na	(iv) × (vii) × (Introduction du biconditionnel)
(ix) (∀x)(Qx ↔ Nx)	(viii) × (Généralisation universelle)

De prime abord, cette dérivation constitue une réfutation de la position de Kim dans la mesure où elle satisfait toutes les conditions d'une explication *bona fide* : tout d'abord, l'*explanandum*, c.-à-d. la corrélation (∀x) (Qx ↔ Nx), est déduit de l'*explanans*, c.-à-d. l'identité Q = N. Ensuite, même si l'on accepte à la suite de Kim le principe de transparence, le fait d'identité est numériquement distinct du fait de corrélation, autrement dit l'explication satisfait le principe d'irréflexivité. Enfin, l'énoncé d'identité est la *seule* prémisse d'où est tiré l'énoncé de corrélation. On ne peut donc pas, semble-t-il, parler ici de « transfert d'explication », comme le voudrait Kim. Il semble bien, au bout du compte, que l'énoncé d'identité « Q = N » puisse jouer le rôle d'une prémisse explicative, et non pas celui de simple règle de réécriture.

La position de Kim n'est cependant peut-être pas si désespérée qu'il pourrait paraître. Considérons la conclusion de l'argument de Bates, à savoir la corrélation (∀x)(Qx ↔ Nx). Puisque cet énoncé est dérivé de la prémisse « Q = N », on peut aussi décrire le fait qu'il rapporte au moyen de l'énoncé « (∀x)(Nx ↔ Nx) ». Il existe donc au moins une description sous laquelle le fait figurant en conclusion est complètement trivial, et qui peut être dérivée à partir d'un ensemble vide de prémisses, à l'aide des seules règles logiques. Ce point peut être rendu manifeste si l'on présente d'une façon différente l'argument permettant de dériver l'énoncé de corrélation :

(i) Na	(Hypothèse)
(ii) Na	(i) × (Répétition)
(iii) Na → Na	(i) × (ii) × (Introduction du conditionnel)
(iv) Na → Na	(iii) × (Répétition)
(v) Na ↔ Na	(iii) × (iv) × (Introduction du biconditionnel)
(vi) Q = N	(Prémisse)
(vii) Qa ↔ Na	(v) × (vi) × (Loi de Leibniz)
(viii) (∀x) (Qx ↔ Nx)	(vii) × (Généralisation universelle)

Cette nouvelle formulation montre de nouveau que, *pace* Bates, l'identité ne joue pas un rôle explicatif mais se contente de transférer l'explication. Le cas qui nous occupe a ceci de particulier que l'*explanans* de (v) transféré à (vii) par l'identité (vi) est *nul* : il s'agit d'un cas limite où il n'y a rien à transférer, et où l'identité permet de re-décrire un fait de corrélation complètement trivial (c.-à-d. le fait que Na ssi Na), qui n'appelle aucune explication substantielle, sous un énoncé où le même fait *semblait* requérir une explication (c.-à-d. le fait que Qa ssi Na). En ce sens, accepter un énoncé d'identité esprit/cerveau ne permet pas d'apporter une explication substantielle à un fait de corrélation, mais bien plutôt de *dissiper le besoin* même de l'expliquer. Une fois que l'on a compris que la corrélation que l'on décrivait au départ comme *Qa ssi Na* peut être décrite également comme *Na ssi Na*, la question de savoir pourquoi cette corrélation existe disparaît complètement. Comme le soulignent Block et Stalnaker, qui défendent pourtant le physicalisme de type B :

> Si l'on considère que la chaleur est corrélée avec l'énergie cinétique moyenne des molécules, mais qu'elle ne lui est pas identique, on doit considérer comme légitime la question qui est de savoir pourquoi cette corrélation existe et quel est son mécanisme sous-jacent. Mais une fois que l'on réalise que la chaleur n'est rien d'autre que l'énergie cinétique moyenne des molécules, les questions de ce type apparaissent absurdes[28].

7. Le problème des « faits bruts »

Arrivé à ce stade de l'objection, le seul moyen dont dispose le physicaliste de type B pour défendre le pouvoir explicatif des identités esprit/cerveau est de neutraliser l'objection de Kim dès son point de départ, en rejetant l'un des deux principes sur lesquels elle s'appuie, à savoir ou bien la transparence des faits ou bien l'irréflexivité de la relation d'explication[29].

Une première option, sans doute la plus plausible, consisterait à accepter la transparence mais à rejeter l'irréflexivité. Selon ce point de vue, le fait que Jean Giraud mesure 1,75 m est certes le même que le fait que Moebius mesure 1,75 m puisque la même propriété est instanciée par la même personne ; mais on peut fort bien expliquer le second par le premier sans créer de circularité dans l'explication dans la mesure où les contextes d'explication sont hyperintensionnels au même titre que n'importe quel contexte épistémique : ils sont sensibles non seulement à l'identité des faits décrits, mais également aux énoncés sous lesquels on les rapporte[30].

Ainsi, il est généralement admis qu'on peut savoir que Jean Giraud mesure 1,75 m sans savoir que Moebius mesure 1,75 m, quand bien même

28. Block et Stalnaker, 1999, p. 24.

29. Nous ne discutons pas ici de l'option qui consisterait à rejeter les deux principes, car aucune vue cohérente ne nous semble pouvoir y correspondre.

30. Voir McLaughlin, 2010, p. 298.

les faits connus sont identiques. De même, on peut considérer qu'une explication du fait que Jean Giraud mesure 1,75 m n'est pas une explication du fait que Moebius mesure 1,75 m, quand bien même il n'y a ici qu'un seul fait rapporté sous deux énoncés. On peut en effet arguer de façon plausible que les *explanantia* et les *explananda* ne sont pas des faits *simpliciter*, comme semble le penser Kim, mais des faits-considérés-sous-certains-énoncés, ou des paires constituées d'un fait et d'un énoncé. Si tel est le cas, alors le fait que Jean Giraud mesure 1,75 m peut figurer dans l'*explanans* du fait que Moebius mesure 1,75 m sans que cela crée une circularité explicative : le même fait est certes décrit dans l'*explanandum* et l'*explanans* mais ceux-ci demeurent distincts dans la mesure où ils rapportent ce fait au moyen d'énoncés qui n'ont aucun lien *a priori* entre eux. L'objection de la circularité étant levée, plus rien ne s'oppose à considérer l'énoncé d'identité « Jean Giraud = Moebius » comme une authentique prémisse explicative. Cette conclusion s'étend à tous les contextes d'explication, et notamment à l'explication des corrélations par les identités esprit/cerveau.

La deuxième option qui s'offre au physicaliste de type B est de conserver l'irréflexivité de la relation d'explication mais de rejeter la transparence des faits. Selon ce point de vue, les faits ne peuvent pas figurer dans leur propre explication mais ils doivent en revanche être individués selon une plus grande finesse de grain que celle proposée jusqu'ici : leurs conditions d'identité doivent mentionner non seulement l'identité des objets et des propriétés concernés, mais également les expressions linguistiques ou les concepts — les « modes de présentation » en termes frégéens — sous lesquels ces objets et ces propriétés sont considérés[31]. Nous ne discuterons pas plus avant cette option, car elle nous semble n'être qu'une variante notationnelle de la première, à cette différence près qu'elle localise l'hyperintensionalité dans les *relata* de la relation d'explication là où la première option l'injecte dans la relation d'explication elle-même.

Pour les besoins de l'argumentation, concédons provisoirement au physicaliste de type B qu'il peut emprunter l'une des deux voies décrites ci-dessus. Il s'ensuit que les identités en général ont bel et bien un pouvoir explicatif. Les identités esprit/cerveau peuvent donc prétendre à expliquer les corrélations esprit/cerveau. Il importe cependant de garder à l'esprit que cela ne suffit pas pour mener à bien l'argument abductif en faveur du physicalisme de type B : encore faut-il établir que les identités esprit/cerveau expliquent les corrélations *mieux* que ne le font les lois dualistes. Est-ce le cas ?

Il y a une bonne raison de penser que les deux compétiteurs se valent en matière d'explication. Rappelons en effet le principal argument invoqué par les physicalistes de type B en faveur de la supériorité de leur théorie : le dualisme explique les corrélations en recourant à des lois psychophysiques qui doivent être admises comme des *faits bruts*, c'est-à-dire comme des faits

31. Voir Ruben, 1990, p. 174-177.

qui ne peuvent pas être expliqués par des lois plus fondamentales, notamment celles de la physique, tandis que le physicalisme en est précisément dispensé. Le problème que pose cet argument de simplicité est que les identités esprit/cerveau auxquelles le physicaliste de type B fait appel pour expliquer les corrélations doivent elles aussi être acceptées comme des faits bruts. Souvenons-nous que, contrairement au physicaliste de type A, le physicaliste de type B reconnaît l'existence d'un fossé explicatif, et soutient pour cette raison que les identités esprit/cerveau sont *inexplicables*, elles ne peuvent recevoir aucune explication réductive. Dès lors, en quoi ces identités brutes expliqueraient-elles les corrélations mieux que ne le font les lois brutes du dualisme ? Comme le soulignent Chalmers et Jackson, le fait qu'il s'agisse d'identités et non de lois de la nature ne change rien au coût explicatif que représente le recours à des faits bruts :

> Ontologiquement, ces identités diffèrent des lois [dualistes]. Mais du point de vue épistémique, elles se comportent exactement comme ces lois. Ce sont en effet des « principes-ponts » psychophysiques tenus pour épistémiquement primitifs, qui ne sont pas eux-mêmes expliqués mais qui, combinés aux vérités physiques, expliquent les vérités phénoménales. Une explication du domaine phénoménal comportera donc deux composants irréductibles : un composant physique et un composant psychophysique. En appelant ces principes-ponts des « identités » plutôt que des « lois », on préserve certes la structure ontologique du matérialisme, mais la structure explicative du matérialisme est en tout point comparable à celle du dualisme des propriétés[32].

S'il veut utiliser les corrélations comme point de départ d'une IME en faveur des identités esprit/cerveau, le physicaliste de type B doit donc nous donner une raison de penser que l'explication par les identités brutes conserve néanmoins un avantage décisif sur l'explication par les lois brutes du dualisme. Dans la section qui suit, nous présenterons et rejetterons les arguments censés rétablir la supériorité explicative du physicalisme de type B.

8. Le dilemme du physicaliste de type B

Les physicalistes de type B rétorquent qu'il existe une asymétrie fondamentale entre l'explication des corrélations par les identités physicalistes et leur explication par les lois dualistes, qui confère à la première un avantage décisif sur la seconde : les lois dualistes ne *peuvent* pas être expliquées alors qu'elles *doivent* l'être, ce pour quoi on peut parler d'un défaut explicatif. De leur côté, les identités esprit/cerveau sont tout aussi inexplicables que les lois dualistes mais, à la différence de ces dernières, elles *n'ont pas besoin* d'être expliquées. En un mot, ce ne sont pas des *explananda*. Dans la mesure où elles ne requièrent pas d'explication, le fait qu'elles ne puissent pas être expliquées ne doit pas être compté comme un défaut du physicalisme de type B.

32. Chalmers et Jackson, 2001, p. 353-354.

Il est donc trompeur de les traiter comme des « faits bruts » ou des « faits épistémiquement primitifs », comme s'il s'agissait de faits demandant une explication qui serait introuvable, à la manière des lois dualistes.

Pourquoi le physicaliste de type B devrait-il être exempté d'expliquer les identités esprit/cerveau ? Le principal argument motivant cette position est formulé de la façon suivante par Papineau[33] :

> Les identités authentiques n'ont besoin d'aucune explication. Si « deux » entités ne font qu'une, alors l'une n'« accompagne » ni ne « produit » l'autre — elle *est* l'autre. Et si c'est le cas, alors il n'y a rien à expliquer. On peut expliquer pourquoi *une* chose en accompagne une *autre*, ou pourquoi elle la produit. Mais l'on ne peut pas expliquer pourquoi une chose est identique à elle-même[34].

Cet argument peut être reconstruit de la façon suivante :

(i) Supposons que Q = N, (physicalisme des types)
(ii) Dans ce cas, le fait que Q = N n'est autre que le fait que N = N, (principe de transparence)
(iii) Or, le fait que N = N n'a pas besoin d'explication,
(iv) Conclusion : le fait que Q = N n'a pas besoin d'explication.

À la lumière de la discussion qui précède, on voit immédiatement que cet argument soulève un problème majeur. En effet, si on l'accepte, alors, par parité de raisonnement, on doit également accepter l'argument suivant :

(i) Supposons que Q = N, (physicalisme des types)
(ii) Dans ce cas, le fait que $(\forall x)\ (Qx \leftrightarrow Nx)$ n'est autre que le fait que $(\forall x)\ (Nx \leftrightarrow Nx)$, (principe de transparence)
(iii) Or, le fait que $(\forall x)\ (Nx \leftrightarrow Nx)$ n'a pas besoin d'explication,
(iv) Conclusion : le fait que $(\forall x)\ (Qx \leftrightarrow Nx)$ n'a pas besoin d'explication.

Cette conclusion est dévastatrice pour le physicaliste de type B : selon lui, la seule justification de l'identité (Q = N) provient justement du fait que nous en avons besoin afin d'expliquer la corrélation $(\forall x)\ (Qx \leftrightarrow Nx)$, ce qui permet d'utiliser une IME. Mais nous voyons que l'argument censé nous dispenser d'expliquer l'identité (Q = N) nous conduit aussi à reconnaître que la corrélation $(\forall x)\ (Qx \leftrightarrow Nx)$ ne requiert pas d'explication. Comment peut-on opérer une IME à partir d'une corrélation dont on reconnaît par ailleurs que, sous l'hypothèse que l'on veut défendre, elle ne nécessite justement aucune explication ? Le problème de l'argument de Papineau vient de ce que

33. Voir également Block et Stalnaker : « Remarquons que cela a un sens d'exiger une explication pour l'existence d'une corrélation entre deux ensembles d'événements. En revanche, cela n'a pas de sens d'exiger une explication pour une identité. Les identités n'ont pas d'explication », Block & Stalnaker, 1999, p. 24.

34. Papineau, 2002, p. 144.

le principe de transparence qu'il présuppose pour nous dispenser d'expliquer les identités est précisément celui qui permet à Kim de dissiper la demande d'explication des corrélations, privant ainsi le physicalisme de type B de sa base abductive.

La question cruciale est donc : existe-t-il un argument permettant de dissiper le besoin d'expliquer les identités esprit/cerveau sans affecter le statut d'*explananda* des corrélations ? Hill formule un argument qui semble de prime abord moins vulnérable que celui de Papineau dans la mesure où il ne présuppose pas le principe de transparence :

> La question « pourquoi *être une expérience consciente de type F* est-elle identique à *être un processus cérébral de type Y* ? » n'a pas de sens. « *Être une expérience de type F* » et « *être un processus cérébral de type Y* » sont des noms de propriétés. Si un énoncé est construit seulement à partir du prédicat d'identité et de noms, il sera vrai dans tous les mondes possibles, et il ne sera ni nécessaire ni même possible d'expliquer pourquoi l'énoncé est vrai. (Comme le dit Robert Causey dans un contexte semblable, des énoncés de ce type requièrent une justification, mais pas d'explication)[35].

Le principe invoqué ici est qu'un énoncé ne requiert pas d'explication s'il est métaphysiquement nécessaire, c'est-à-dire vrai dans tous les mondes possibles. Or, sous l'hypothèse ici défendue que (Q = N), il est métaphysiquement nécessaire que (Q = N) car l'identité en général est nécessaire[36]. Par conséquent, il n'est pas requis d'expliquer (Q = N). Malheureusement Hill ne semble pas remarquer que le même raisonnement nous dispense d'expliquer les corrélations : sous l'hypothèse que (Q = N), il est en effet métaphysiquement nécessaire que $(\forall x)(Qx \leftrightarrow Nx)$[37], ce qui revient à dire que les corrélations n'ont pas non plus à être expliquées. De nouveau, le physicalisme de type B perd sa base abductive.

Nous conjecturons que le physicalisme de type B est confronté ici à un problème général : tout argument qui dispense d'expliquer les identités esprit/cerveau retire du même coup le statut d'*explananda* aux corrélations esprit/cerveau, coupant ainsi le physicalisme de type B de sa base abductive. Il incombe bien sûr au physicaliste de type B de prouver qu'il existe un argu-

---

35. Hill, 1991, p. 24-25.

36. Hill utilise ici le principe largement admis de la nécessité de l'identité : $(a = b) \rightarrow \Box (a = b)$, où « a » et « b » sont des désignateurs rigides, et où « $\Box \Phi$ » signifie « il est métaphysiquement nécessaire que Φ ».

37. Plus précisément, $\Box(\forall x)\ (Qx \leftrightarrow Nx)$ est obtenu à partir de $\Box(N = Q)$ et de $\Box(\forall x)\ (Nx \leftrightarrow Nx)$. Il importe de rappeler que cette déduction ne repose pas sur le principe de transparence des faits mais sur celui, non controversé, d'après lequel deux expressions nécessairement co-extensionnelles peuvent être substituées *salva veritate* dans un contexte modal. L'argument de Hill est donc compatible avec la thèse selon laquelle le fait que $(\forall x)\ (Qx \leftrightarrow Nx)$ et le fait que $(\forall x)\ (Nx \leftrightarrow Nx)$ comptent pour deux faits distincts, c'est là ce qui le distingue de l'argument de Papineau.

ment permettant d'être dispensé de l'explication des premières et de ne pas l'être en ce qui concerne les secondes.

Il s'ensuit que le physicaliste de type B est enfermé dans un dilemme :

(i) Ou bien il considère que les identités esprit/cerveau n'ont pas besoin d'être expliquées parce que ce sont des faits triviaux ; dans cette première branche du dilemme, le même raisonnement conduit à dire que, sous l'hypothèse qu'il défend, les corrélations esprit/cerveau sont également des faits triviaux qui n'appellent pas d'explication, et l'on ne peut alors plus justifier les énoncés d'identité par une IME à partir des corrélations.

(ii) Ou bien il considère que même sous l'hypothèse de l'identité esprit/cerveau, les corrélations sont des faits substantiels qui requièrent une explication. Le même raisonnement conduit alors à dire que les identités esprit/cerveau sont des faits substantiels qui demandent eux aussi une explication. Toutefois, le physicaliste de type B reconnaît que ces identités ne peuvent pas être expliquées alors qu'elles devraient l'être, ce qui revient à en faire des faits bruts, au même titre que les lois dualistes. Dans cette seconde branche du dilemme, les identités esprit/cerveau ne sont pas mieux placées pour expliquer les corrélations que les lois dualistes, et elles ne peuvent donc pas être justifiées par une IME à partir des corrélations.

La conclusion est la même quelle que soit la branche choisie : le physicaliste de type B ne peut pas justifier ses assertions d'identité par une IME à partir des corrélations.

9. Explication et causalité : l'argument causal

De ce que les énoncés d'identité esprit/cerveau ne peuvent pas être justifiés par IME à partir des corrélations, il ne s'ensuit pas toutefois qu'il n'existe aucune autre façon de les justifier par IME. Peut-être existe-t-il une autre base abductive que les corrélations permettant de réaliser une telle inférence ?

Ainsi nous n'avons pour le moment rien dit des *pouvoirs causaux* des états phénoménaux, car les physicalistes de type B prennent généralement comme point de départ de leur argument abductif l'explication des corrélations, et non celle de la causalité mentale. Ces données portant sur les pouvoirs causaux nous semblent cependant fondamentales pour justifier les énoncés d'identité par IME. Il y a en effet un accord entre Kim et ses détracteurs pour considérer que les énoncés d'identité esprit/cerveau permettent tout au moins de transférer les explications causales du domaine des neurosciences vers le domaine psychologique. Si l'on admet que les états phénoménaux ont bien des pouvoirs causaux, que ces pouvoirs sont mentionnés par les explications de notre psychologie naïve, et si l'on admet par ailleurs le

principe de clôture causale du monde physique[38], on peut donc en tirer une justification indirecte particulièrement convaincante des identités. Papineau propose ainsi une version déductive de cet argument, qu'il nomme « l'argument causal »[39] :

(i) Les occurrences d'états mentaux conscients ont des effets physiques typiques (p. ex. le comportement).

(ii) Tous les effets physiques sont causés par des causes physiques qui suffisent à les expliquer (principe de clôture causale du monde physique).

(iii) Les effets physiques des occurrences d'états mentaux conscients ne sont pas sur-déterminés par des causes différentes de leurs causes physiques (autrement dit, leurs causes physiques sont leurs seules causes).

(iv) Conclusion : Les énoncés d'identité psychophysique sont vrais.

Même si cette présentation de l'argument causal est déductive, Bates fait remarquer qu'on peut aisément le reformuler sous une forme abductive[40] :

(i) Les occurrences d'états mentaux conscients ont des effets physiques.

(ii) La meilleure explication du fait que ces occurrences ont des effets physiques, c'est que les énoncés d'identité esprit/cerveau sont vrais.

(iii) Conclusion : Les énoncés d'identité esprit/cerveau sont vrais.

Cet argument par IME est bien plus prometteur que celui partant des corrélations, car il ne prête pas flanc aux objections qui minent ce dernier. Premièrement, à la différence des corrélations, les faits causaux qui fournissent l'*explanandum* de cette IME n'encourent pas le risque d'être « trivialisés », quand bien même le physicaliste de type B adopterait la ligne de Papineau selon laquelle les faits d'identité sont triviaux et n'ont pas besoin de ce fait d'être expliqués (1re branche du dilemme examinée plus haut). Ici, le physicaliste peut parfaitement se dispenser d'expliquer l'*explanans* (c.-à-d. les identités esprit/cerveau) sans pour autant faire disparaître l'*explanandum* (la causalité psychophysique), préservant ainsi la base abductive. Deuxièmement, les identités esprit/cerveau fournissent une explication de la causalité psychophysique qui est indéniablement meilleure que celle apportée par les lois dualistes. Le dualiste ne peut en effet expliquer ces faits causaux qu'au prix de grandes difficultés, c'est-à-dire ou bien en abandonnant le principe de clôture causale ou la prémisse de non-surdétermination causale, ou bien tout simplement en considérant ces phénomènes causaux comme illusoires — c'est la position épiphénoméniste.

38. Selon le principe de clôture causale du monde physique, tous les effets physiques ont une cause physique qui suffit à les expliquer. On considère aujourd'hui que ce principe ne peut avoir de justification qu'inductive.

39. Papineau, 2002, p. 17-18.

40. Bates, 2009, p. 320.

La force de conviction de l'argument causal n'est cependant pas une bonne nouvelle pour le physicalisme de type B. Selon sa première prémisse, « les occurrences d'états mentaux conscients ont des effets physiques typiques » ; autrement dit, ces états ont des pouvoirs causaux, des rôles causaux. D'où cette connaissance des pouvoirs causaux des états phénoménaux peut-elle provenir ? Sans doute de la psychologie naïve, qui mentionne les états phénoménaux comme les sensations, les émotions, les humeurs, etc., dans ses explications. Mais cette hypothèse, selon laquelle notre connaissance des pouvoirs causaux des états phénoménaux proviendrait de la psychologie naïve, entre directement en conflit avec la thèse centrale permettant de distinguer le physicalisme de type B du physicalisme de type A. Rappelons que les physicalistes de type B considèrent qu'une analyse fonctionnelle « à la Lewis » des concepts phénoménaux est impossible, car ces concepts sont censés ne pas pouvoir être associés à des rôles causaux. Ainsi Papineau écrit-il :

> Les concepts phénoménaux ne sont pas liés à des rôles causaux. Lorsque nous pensons à la douleur de façon pré-théorique, en utilisant un concept phénoménal, nous pensons à la douleur en termes de « ce que cela fait » [*what it is like*] d'avoir mal, pas en termes des causes et effets caractéristiques de la douleur[41].

Comment peut-on cependant à la fois recourir à l'argument causal pour justifier indirectement les énoncés d'identité esprit/cerveau, donc accepter la prémisse selon laquelle les états phénoménaux ont des effets physiques, et soutenir que lorsque nous raisonnons sur un état phénoménal, ce n'est jamais en termes « de causes et effets caractéristiques » de cet état ? Si nous avons une connaissance des effets typiques d'un état phénoménal Q, nous pouvons exprimer cette connaissance de la façon suivante :

(i) Dans les circonstances C1, l'occurrence de Q cause P1.
(ii) Dans les circonstances C2, l'occurrence de Q cause P2.
...
(n) Dans les circonstances Cn, l'occurrence de Q cause Pn.

Qu'est-ce qui nous empêche, dès lors, de former l'énoncé de Ramsey décrivant le rôle causal de Q, de la façon suivante ?

> (R) Il existe un unique état X tel que l'occurrence de X cause P1 dans les circonstances C1 et... l'occurrence de X cause Pn dans les circonstances Cn.

Nous obtenons alors une description fonctionnelle qui permet de procéder à une réduction par fonctionnalisation, selon la méthode mise en avant par les physicalistes de type A.

41. Papineau, 2002, p. 149.

Cette discussion montre qu'il y a une tension entre vouloir d'une part justifier les identités esprit/cerveau par une IME à l'aide de l'argument causal, et soutenir d'autre part que les pouvoirs causaux des états phénoménaux nous sont inaccessibles lorsque nous les concevons à l'aide des concepts phénoménaux. Si ces pouvoirs causaux nous sont inaccessibles, il n'y a rien à expliquer, et l'argument causal s'effondre ; mais s'ils nous sont accessibles, une réduction fonctionnelle est envisageable, ce qui revient à donner raison au physicalisme de type A.

10. Conclusion

Nous avons montré que, contrairement à la stratégie abductive mise en avant par les physicalistes de type B, il est impossible de justifier les énoncés d'identité esprit/cerveau par une IME qui partirait des corrélations. Le dilemme dans lequel s'enferre le physicaliste de type B vient de ce qu'il est incohérent de soutenir à la fois que les identités esprit/cerveau n'ont pas à être expliquées du fait de leur trivialité, et qu'elles peuvent néanmoins expliquer mieux que leur rival dualiste les faits de corrélation. La seule stratégie abductive qui s'offre au physicaliste consiste à partir non pas des corrélations, mais des pouvoirs causaux des états phénoménaux. Cette stratégie permet non seulement de justifier indirectement les énoncés d'identité, mais, au moins à terme, de les expliquer, en montrant comment les rôles causaux associés aux concepts phénoménaux peuvent être implémentés neurologiquement. Elle donnerait évidemment raison, cependant, au physicalisme de type A, qui nous paraît donc la seule véritable solution de remplacement au dualisme.

Références

Armstrong, David Malet. *A Materialist Theory of Mind*, London, Routledge & Kegan Paul, 1968.

Aydede, Murat (dir.). *Pain : New Essays on its Nature and the Methodology of its Study*, Cambridge (Mass.), MIT Press, 2005.

Bates, Jared. « A Defence of the Explanatory Argument for Physicalism », *Philosophical Quarterly*, vol. 59, n° 235, 2009, p. 315-324.

Block, Ned et Robert Stalnaker. « Conceptual Analysis, Dualism and the Explanatory Gap », *Philosophical Review*, vol. 108, n° 1, 1999, p. 1-46.

Chalmers, David. *The Conscious Mind*, New York, Oxford University Press, 1996.

————. « Consciousness and its Place in Nature », in *Philosophy of Mind*, David Chalmers (dir.), New York, Oxford University Press, 2002, p. 247-272.

Chalmers, David et Frank Jackson. « Conceptual Analysis and Reductive Explanation », *Philosophical Review*, vol. 110, n° 3, 2001, p. 315-361.

Chaplin, Martin. *Water Structure and Science*, 2007 (site web : http://www.lsbu.ac.uk/water).

Crick, Francis et Christof Koch. « Towards a Neurobiological Theory of Consciousness », *Seminars in the Neurosciences*, vol. 2, 1990, p. 263-275.

Dupré, John. *The Disorder of Things : Metaphysical Foundations of the Disunity of Science*, Cambridge (Mass.), Harvard University Press, 1993.

Hendry, Robert Findley. « Elements, Compounds, and Other Chemical Kinds », *Philosophy of Science*, vol. 73, 2006, p. 864-875.
————. « Science and Everyday Life : Water vs H_2O », *Insights*, vol. 3, n° 23, 2010, p. 2-10.
Hill, Christopher. *Sensations*, Cambridge, Cambridge University Press, 1991.
Jackson, Frank. *From Metaphysics to Ethics,* Oxford, Oxford University Press, 1998.
Kim, Jaegwon. *Physicalism, or Something Near Enough*, Princeton, Princeton University Press, 2005.
Kripke, Saul. *Naming and Necessity*, Oxford, Blackwell, 1980.
LaPorte, Joseph. *Natural Kinds and Conceptual Change*, Cambridge, Cambridge University Press, 2004.
Lewis, David. « An argument for the identity theory », *Journal of Philosophy*, vol. 63, n° 1, 1966, p. 17-25.
————. « How to Define Theoretical Terms », *Journal of Philosophy*, vol. 67, n° 13, 1970, p. 427-446.
————. « Psychophysical and Theoretical Identifications », *Australasian Journal of Philosophy*, vol. 50, n° 3, 1972, p. 249-258.
————. « Reduction of Mind », in *A Companion to the Philosophy of Mind*, Samuel Guttenplan (dir.), Oxford, Blackwell, 1994, p. 412-431.
Malaterre, Christophe. *Les origines de la vie*, Paris, Hermann, 2012.
Metzinger, Thomas. *Neural Correlates of Consciousness*, Cambridge (Mass.), MIT Press, 2000.
McLaughlin, Peter. « In Defense of New Wave Materialism : a Response to Horgan and Tienson », in *Physicalism and its Discontents*, Carl Gillett et Barry Loewer (dir.), Cambridge, Cambridge University Press, 2001, p. 319-330.
————. « Consciousness, Type Physicalism, and Inference to the Best Explanation », *Philosophical Issues*, vol. 20, 2010, p. 266-304.
Needham, Paul. « What is water ? », *Analysis*, vol. 60, 2000, p. 13-21.
Papineau, David. *Thinking About Consciousness*, Oxford, Clarendon Press, 2002.
Polger, Thomas et Kenneth Sufka. « Closing the Gap on Pain : Mechanism, Theory, and Fit », in Aydede, 2005, p. 325-350.
Ruben, David-Hillel. *Explaining Explanation*, New York, Routledge, 1990.
Smart, John Jamieson Carswell. « Sensations and Brain Processes », *Philosophical Review*, vol. 68, n° 2, 1959, p. 141-156.
Weisberg, Michael. « Water is *not* H_2O », in *Philosophy of Chemistry : Synthesis of a New Discipline*, Davis Baird, Eric Scerri et Lee McIntyre (dir.), Dordrecht, Springer, 2006, p. 337-345.
Wilson, N. L. « Facts, Events and Their Identity Conditions », *Philosophical Studies*, vol. 25, n° 5, 1974, p. 303-321.