Analyse conceptuelle et connaissance phénoménale

Pascal Ludwig[*]
(Université de Paris IV)

I. Introduction

David Lewis est un physicaliste antécédent, c'est-à-dire qu'il adhère au matérialisme, à la thèse selon laquelle « la physique, ou du moins quelque chose d'assez proche de la physique actuelle, (...) constitue une théorie exhaustive du monde, complète aussi bien que correcte » (Lewis, 1983a in Lewis, 1999, p. 33-34). Sa contribution la plus remarquable à la philosophie de l'esprit consiste dans la clarification de cette position physicaliste. Nous disposons d'un riche vocabulaire d'expressions mentales ou psychologiques, qui nous permettent de décrire des états mentaux – les croyances, les désirs, les intentions, les expériences sensorielles, les émotions, … –, de les attribuer à nos semblables dans des circonstances pertinentes, et de prédire leurs comportements sur la base de ces attributions. Ce vocabulaire caractérise la psychologie populaire, l'ensemble des règles ou des préceptes, explicites ou implicites, qui nous permettent de donner un sens à nos actions. Quoique dans une certaine mesure rudimentaire lorsqu'on la compare à la psychologie scientifique, la psychologie populaire a un pouvoir prédictif important (Lewis, 1994 in Lewis, 1999, p. 298). Lewis prétend expliquer ce pouvoir prédictif en montrant comment les expressions psychologiques parviennent à nommer non seulement des propriétés naturelles, mais aussi des propriétés physiques – en l'occurrence, des états du cerveau. Il ne s'agit pourtant pas de défendre une conception linguistique du physicalisme, qui soutiendrait que l'on peut traduire les explications de la psychologie populaire dans le langage de la physique, ou au moins dans celui des neurosciences. Lewis défend une

---

[**] Pascal Ludwig est Maîtres de conférences à l'Université de Paris-Sorbonne et membre de l'équipe Rationalités Contemporaines. Ses recherches portent sur la philosophie du langage et de l'esprit, plus particulièrement sur la place de l'expérience consciente dans la monde naturel. Il s'intéresse également à l'épistémologie de la métaphysique. Il a récemment publié Kripke : référence et modalités (PUF, 2005), en collaboration avec Filipe Drapeau-Contim, L'individu (Vrin, 2008), un volume collectif co-dirigé avec Thomas Pradeu. Derniers articles publiés : « Réduction et émergence », in Barberousse, Bonnay, Cozic, Précis de Philosophie des sciences (Vuibert, 2011) et « De la relativité des jugements moraux », Dialogue, 51, 2012.

forme de réductionnisme, mais une forme entièrement différente de celle qui était en vigueur dans l'école du positivisme logique.

En premier lieu, il donne un sens précis à la thèse physicaliste appliquée au domaine du mental, que l'on peut résumer de la façon suivante : l'ensemble des propriétés qui existent dans le monde dépend de façon systématique d'un ensemble beaucoup plus restreint de propriétés, les propriétés fondamentales, ou « parfaitement naturelles »[1]. De plus, nous avons des raisons a posteriori de soutenir que les propriétés qui sont fondamentales dans le monde réel sont celles décrites par la physique. On peut donc avancer, pour utiliser le concept technique de survenance, que les états mentaux des personnes surviennent sur leurs états physiques : deux mondes possibles exactement semblables du point de vue de l'instanciation des propriétés physiques seront également exactement semblables du point de vue de l'instanciation des propriétés psychologiques[2].

Dans une perspective métaphysique, la thèse de la survenance permet de répondre de façon satisfaisante à la question de savoir quelle est la place des états mentaux dans le monde naturel. Nous pouvons en effet affirmer, si cette thèse est vraie, que l'instanciation d'une propriété mentale sera toujours conditionnée par l'instanciation d'un ensemble, peut-être complexe, de propriétés physiques. De façon générale, nous pouvons donc dire que la conjonction de toutes les vérités, à l'inclusion des vérités psychologiques, est nécessairement conditionnée par la conjonction des vérités physiques. Cela revient à considérer que l'énoncé suivant est nécessaire :

> (1) Si P, alors Q. (où P désigne la conjonction de toutes les vérités physiques décrivant le monde réel, et où Q désigne la conjonction de toutes les vérités, à l'inclusion des vérités psychologiques).

Néanmoins, la thèse de la survenance en elle-même ne nous dit pas si les propriétés mentales sont des propriétés naturelles : chaque propriété mentale pourrait en effet être identique à une disjonction extrêmement complexe de propriétés naturelles. Elle ne nous dit pas non plus si l'on peut ou non dériver les vérités mentales à partir d'une connaissance complète du monde physique. Certes, la survenance implique que le conditionnel (1) est une vérité nécessaire. Si ce conditionnel était contingent, il existerait en

[1] Cf. Lewis (1983a et 1994), et les discussions récentes et éclairantes de Sider (2011). Selon Lewis, la « naturalité » d'une propriété est une question de degré. Sur ce point, cf. Hall (2010).
[2] Cf. Lewis (1994, in Lewis, 1999) p. 292 pour une formulation plus précise.

effet un monde possible dans lequel l'ensemble des vérités physiques seraient vérifiées, mais dans lequel pourtant au moins une proposition psychologique, vraie dans notre monde réel, serait falsifiée. Mais cela n'implique pas que l'on puisse dériver a priori l'ensemble des vérités psychologiques à partir des vérités physiques. Le fait d'admettre la thèse de la survenance est compatible, autrement dit, avec la thèse selon laquelle le conditionnel (1) est une vérité a posteriori quoique nécessaire. Prendre position relativement au statut épistémologique a priori ou a posteriori de l'énoncé (1) revient ainsi à prendre position sur la nature du physicalisme. On peut adhérer à une forme a posteriori de physicalisme, selon laquelle l'établissement de certaines vérités empiriques sera requise pour pouvoir dériver le conséquent du conditionnel (1) à partir de son antécédent. Mais on peut également soutenir, au contraire, qu'une analyse des concepts figurant dans l'antécédent suffira pour permettre une telle dérivation. Cela revient à accepter une forme a priori de physicalisme, selon laquelle l'ensemble des vérités psychologiques peut être dérivé à partir de l'ensemble des vérités physiques grâce à une analyse conceptuelle.

C'est précisément à une version du physicalisme a priori que David Lewis adhère. Son physicalisme va en effet bien au-delà de la simple acceptation de la thèse de la survenance. Selon lui, l'impressionnant succès prédictif de la psychologie populaire implique que l'on peut identifier les propriétés psychologiques à des propriétés physiques, et que l'on peut également justifier de façon déductive une telle identification. Il défend, autrement dit, une version de la thèse de l'identité des types : chaque type psychologique est identique à un certain type neuronal[3]. Il soutient par ailleurs qu'une réduction du vocabulaire psychologique au vocabulaire neuroscientifique est possible, à l'aide de la méthode de l'analyse conceptuelle. Il s'en suit que toutes les vérités, y compris les vérités portant sur l'expérience consciente, doivent pouvoir être déduites a priori à partir d'une connaissance exhaustive du monde physique.

Cette position entre en conflit avec ce qu'on appelle parfois l'intuition de la connaissance[4], c'est-à-dire avec l'intuition selon laquelle il existe un gouffre épistémologique, et peut-être du coup également un gouffre ontologique, entre le domaine physique et le domaine phénoménal. On peut développer les conséquences de cette intuition à l'aide de

[3] Cf. Lewis (1966, 1972, 1994). Pour une approche semblable de la théorie de l'identité, cf. Armstrong (1968).

[4] Sur l'intuition de la connaissance et son histoire, voir l'introduction de Ludlow et al. (2005).

l'expérience de pensée célèbre de Franck Jackson (1982)[5]. 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, la réponse est « non » : il existe une vérité psychologique, qui porte sur l'effet que cela fait de voir du rouge, que l'on ne peut pas dériver logiquement à partir d'une connaissance même complète du monde physique. Voici la structure de l'argument dit «de la connaissance», de Jackson :

> 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.
> 4. Le physicalisme doit donc être rejeté sous toutes ses formes.

Lewis n'accepte pas la conclusion de l'argument de Jackson. Il est en effet convaincu que tous les faits peuvent être connus à partir de prémisses contenant uniquement les vérités physiques. Le physicalisme de Lewis apparaît, on le voit, très ambitieux. Il ne s'agit pas simplement de montrer que l'esprit peut avoir une place dans le monde naturel ; il s'agit de situer précisément cette place, en expliquant de façon réductive les processus mentaux, y compris les processus conscients, et donc de combler le gouffre épistémologique censé exister entre le domaine physique et le domaine phénoménal. Le but du présent article est d'abord de présenter et de défendre la méthode de réduction par fonctionnalisation de Lewis. Dans un second temps, je discuterai les difficultés spécifiques que soulèvent nos connaissances phénoménales, en attachant une attention particulière à l'argument de la connaissance.

[5] Pour un recueil des principales discussion que cet article a suscitées, Ludlow et al. (2005).

II. La méthode de réduction par fonctionnalisation

Dès sa première publication (Lewis, 1966), David Lewis met en place une méthode de réduction que j'appellerai la méthode de « réduction par fonctionnalisation ». Il détaille cette méthode dans Lewis (1970 et 1972), et la discute en la modifiant substantiellement dans Lewis (1980, 1994 et 1997). Comme la méthode de réduction par fonctionnalisation s'inspire de la proposition, faite par Carnap dans Les fondements philosophiques de la physiques, pour interpréter les termes théoriques comme « électron » ou « boson » figurant dans nos discours scientifiques de façon fonctionnelle, je commencerai par présenter cette proposition carnapienne, avant de discuter le parti qu'en tire Lewis[6].

A. Carnap, Ramsey, et la question du sens des termes théoriques

Remarquons d'abord que, dans le contexte dans lequel s'exprime Carnap, les expressions théoriques suscitent une question sémantique plutôt qu'une question métaphysique. Selon Carnap, les questions d'existence légitimes sont toujours relatives à des cadres conceptuels. A cet égard, les termes actuels du débat relatif à l'existence des entités théoriques lui sont étrangers : relativement au cadre conceptuel constitué par la physique des particules, il n'est pas du tout problématique d'affirmer que les électrons existent, puisque ce cadre conceptuel comporte des expressions dont l'usage dans des inférences implique l'existence d'électrons. Ce qui est plus difficile en revanche, c'est de savoir comment le sens de telles expressions se trouve déterminé, puisque l'on ne peut pas mettre leur usage en relation directe avec des conditions observables. Or il s'agit d'un problème sérieux pour un empiriste relativiste et conventionnaliste comme Carnap. Certes, pour Carnap, la question de l'existence des électrons peut être facilement tranchée, à partir du moment où cette existence est relativisée à un cadre linguistique : les physiciens sont entièrement libres d'introduire toutes les expressions théoriques qui peuvent leur être utiles afin d'expliquer au mieux les observations. Reste qu'une frontière précise doit pouvoir être tracée, dans un tel cadre théorique, entre les énoncés synthétiques, qui possèdent un contenu cognitif et qui parlent du monde, et les énoncés analytiques, qui ne possèdent pas de tel contenu mais ne font que refléter nos conventions

[6] Dans l'importante littérature secondaire sur cette question, on peut consulter Galinon (2009), Demopoulos (2007) et Uebel (2011). Voir également Brandon-Mitchell & Nola (2009) pour une présentation du programme dit de « Canberra », qui reprend dans une large mesure la méthode lewisienne de réduction par fonctionnalisation.

linguistiques. Carnap reconnaît qu'il n'est tout simplement pas possible de définir les concepts théoriques dans le vocabulaire de l'observation :

> « Il est tentant de croire qu'un scientifique doit être (...) capable de définir en termes familiers les notions théoriques qu'il emploie. Mais ce n'est pas possible. Un physicien ne peut pas nous montrer une image de l'électricité comme il montre à son enfant une image d'un éléphant » (Carnap, 1966, p. 228)

Pour préciser la signification d'une expression théorique, le physicien, nous dit Carnap, ne peut recourir uniquement au vocabulaire observationnel. En fait, le physicien devra faire appel aux lois théoriques elles-mêmes pour expliquer, par exemple, le comportement d'un électron : en renvoyant à ces lois, il pourra décrire « le champ produit par un électron, la réaction d'un électron placé dans un champ, etc. ». Mais notons que dans cet exemple, l'explication partielle de la signification du terme « électron » repose sur la compréhension d'un autre terme théorique, « champ électromagnétique ». Il existe aussi, bien entendu, des connexions inférentielles entre les termes observationnels et les termes théoriques, connexions qui se reflètent dans ce que Carnap appelle les « règles de correspondance », mais ces connexions ne permettent pas d'opérer à elles seules de véritables définitions[7].

C'est donc de l'ensemble du contexte créé par l'affirmation de toutes les thèses d'une théorie qu'il faut partir. La théorie a en bloc des conséquences observables testables, ce qui suffit à lui conférer un contenu empirique bien déterminé. Et si elle possède un contenu empirique bien déterminé, ce contenu doit pouvoir être exprimé à l'aide des seuls termes observationnels. A ce stade, Carnap fait appel à un article de Ramsey (1929), dans lequel celui-ci explique qu'une théorie peut, dans une certaine mesure, être remplacée par un unique énoncé ne contenant que des expressions observationnelles. Cet énoncé, que nous nommerons « l'énoncé de Ramsey » de la théorie, ne lui est pas équivalent, mais possède exactement le même contenu observationnel. Pour obtenir l'énoncé de Ramsey d'une théorie T, contenant des termes théoriques comme « électron », et incluant des règles de correspondance, des énoncés observationnels et des énoncés théoriques, il convient :

---

[7] Si c'était le cas, on devrait parler de définitions opérationnelles ; le problème réside cependant en ce qu'il existe de nombreuses façons différentes de « définir » un même concept de façon opérationnelle. Il n'est donc pas raisonnable, selon Carnap, de considérer les règles de correspondance comme des définitions.

(i) de prendre la conjonction de tous les énoncés de T, que nous nommerons le «postulat» de la théorie T ;
(ii) de remplacer tous les termes théoriques par des noms, que ce soient des noms de classes, de relations, etc. ..., ce qui conduit à l'énoncé $T(t_1, ..., tn, o_1, ..., o_n)$, où $t_1 ...t_n$ sont des noms théoriques, et $o_1 ...o_n$ des termes observationnels ;
(iii) remplacer tous les noms théoriques par des variables appropriées, ce qui conduit à une formule ouverte de type $T(x_1, ...x_n, o_1, ...o_n)$; tout n-tuple d'entités satisfaisant cette formule ouverte peut être nommé une « réalisation de la théorie »
(iv) enfin, former la clôture existentielle de la formule ouverte obtenue, soit l'énoncé de Ramsey RT : $\exists x_1...x_nT(x_1, ..., x_n, o_1,...o_n)$; cet énoncé affirme que la théorie T possède au moins une réalisation.

Comme le souligne Carnap, l'énoncé de Ramsey RT ne possède plus aucun terme théorique, tous ces termes ayant été remplacés par des variables quantifiées existentiellement. En revanche, il a exactement le même contenu empirique que T : les mêmes énoncés observationnels peuvent en être logiquement déduits. Peut-on considérer, du coup, que l'énoncé RT continue de parler, par exemples, des électrons ? Le terme d'électron ne figure plus dans la théorie, mais, insiste Carnap :

Cela ne veut nullement dire que les électrons eux-mêmes disparaissent, ou plus précisément que tout ce à quoi le terme « électron » se réfère dans le monde extérieur disparaisse aussi. L'énoncé de Ramsey continue à affirmer, par ses quantificateurs existentiels, qu'il existe quelque chose dans le monde extérieur dont les propriétés sont celles mêmes que les physiciens attribuent à l'électron. Il ne met pas en cause l'existence – la « réalité » – de ce quelque chose ; il propose seulement une autre manière d'en parler. (Carnap, 1966, tr. fr. p. 245).

Carnap ne s'arrête cependant pas là. Il remarque que l'on peut également formuler un énoncé, le conditionnel (v), que nous nommerons à la suite de Lewis « énoncé de Carnap », et qui reflète exactement les conventions déterminant la signification des expressions théoriques :

(v) $\exists x_1...x_nT(x_1, ..., x_n, o_1,...o_n) \rightarrow T(t_1, ..., t_n, o_1, ..., o_n)$

Contrairement à l'énoncé de Ramsey, l'énoncé de Carnap n'a pas de conséquences observables. Il y a en effet deux manières d'interpréter cet énoncé.

Si l'on suppose, en premier lieu, que les termes théoriques sont déjà bien compris, qu'ils ont déjà une signification, cet énoncé affirme que s'il existe au moins une réalisation de la théorie T, alors le postulat de cette

théorie est vrai, et que donc les entités théoriques constituent l'une de ces réalisations. On ne peut strictement rien en déduire quant aux entités observables, puisque l'énoncé n'est que conditionnel. L'énoncé de Carnap n'a donc pas de contenu empirique.

Puisque le postulat de la théorie est logiquement équivalent à la conjonction de l'énoncé de Ramsey de l'énoncé de Carnap, et puisque l'énoncé de Ramsey a exactement les mêmes conséquences observationnelles que le postulat alors que l'énoncé de Carnap n'a pour sa part aucun contenu empirique, on peut considérer cet énoncé comme un pur postulat de signification, vrai de façon analytique, c'est-à-dire comme un énoncé exprimant une contrainte portant sur la signification des expressions théoriques qui y figurent. A tout le moins, la contrainte exprimée par l'énoncé (v) est selon Carnap exactement la même que celle qui se trouve exprimée par l'ensemble de la théorie T.

Cela mène à une seconde interprétation possible de l'énoncé (v), dans laquelle celui-ci est vu comme une convention permettant de spécifier, au moins partiellement, la signification des expressions théoriques qui y figurent. Dans cette lecture, une expression comme « électron », figurant dans le conséquent de (v), est considérée comme non-interprétée. C'est l'interprétation qui est privilégiée par Carnap : celui-ci considère que l'on peut voir (v) comme un postulat de signification, exprimant une contrainte sur le sens des expressions théoriques : cet énoncé « affirme seulement que si l'énoncé de Ramsey est vrai, alors nous devons comprendre les termes théoriques de façon que la théorie entière soit vraie », c'est-à-dire que si l'énoncé de Ramsey est vrai, et donc si les conditions observables sont telles que décrites par la théorie, alors les désignations des termes théoriques sont les entités qui satisfont la théorie (Carnap, 1966, trad. fr. p. 262). Puisque l'énoncé de Ramsey a exactement les mêmes conséquences observables que la théorie elle-même, nous pouvons considérer la théorie, sous l'hypothèse que cet énoncé est vrai, comme une manière plus ramassée d'exprimer exactement le contenu cognitif de l'énoncé de Ramsey.

Il n'y a de fait que trois possibilités relativement à l'énoncé de Carnap, qu'il est important de discuter à tour de rôle (Lewis, 1970, p. 82) :

(i) Première possibilité : l'énoncé de Ramsey d'une théorie T est faux, car il n'existe aucune réalisation de T. Dans ce cas l'énoncé de Carnap est trivialement vrai, son antécédent étant faux, mais il ne peut pas fonctionner comme un postulat de signification, et il échoue donc à déterminer la signification des termes théoriques. C'est ce qui se passe, souligne Lewis, dans le cas de la théorie du phlogistique : comme il n'y a aucun x qui corresponde au rôle du principe phlogistique tel que décrit par

cette théorie, il n'existe aucune réalisation de la théorie, et il n'est donc pas possible d'attribuer une signification à l'expression « phlogistique ». Il faut cependant nuancer un peu ces remarques : ce qui importe n'est pas tant, au fond, de savoir s'il existe ou non des entités réalisant exactement la théorie T, mais plutôt de savoir s'il existe des entités réalisant approximativement T. Il peut arriver en effet que certains détails d'une théorie soient faux, et qu'il n'existe donc à strictement parler aucune réalisation de la théorie, mais qu'il y ait néanmoins une réalisation de la théorie légèrement modifiée. Dans cette situation, on peut considérer que la théorie possède bien une réalisation. Je reviendrai longuement sur ce point plus bas, car il est très important pour comprendre l'usage précis que fait Lewis de la méthode de ramseyfication.

(ii) Deuxième possibilité : l'énoncé de Ramsey de T est vrai, et il existe une unique réalisation de la théorie. On peut dans ce cas considérer que chaque expression théorique désigne l'entité jouant, dans la réalité, le rôle qui lui correspond. Supposons par exemple que la première variable x_1 dans l'énoncé de Ramsey de T corresponde à l'expression « électron », et qu'il existe bien dans la réalité une entité E_1 satisfaisant la formule ouverte produite à partir de l'énoncé de Ramsey. Si c'est bien le cas, on peut considérer que « électron » désigne E_1 ; c'est exactement ce que l'on peut inférer, selon Carnap, du conditionnel (v) dans ce cas précis à propos de l'expression « électron ».

(iii) Troisième possibilité : l'énoncé de Ramsey de T est vrai, mais il existe plusieurs réalisations distinctes de la théorie. Dans ce cas, le postulat de signification exprimé par l'énoncé de Carnap dit simplement que chaque expression théorique dénote l'un des composants d'une réalisation possible, sans préciser laquelle. Le fait que la possibilité d'une réalisation multiple de la théorie T soit toujours ouverte conduit à une certaine indétermination du sens des termes théoriques.

B. De l'élimination des termes théoriques à la justification des identités : le programme de Lewis

Lewis reprend l'essentiel des résultats de Carnap, en opérant cependant quelques inflexions importantes.

En premier lieu, la distinction entre termes théoriques et termes d'observation n'a pas de grande pertinence dans le cadre de son projet, qui

est avant tout métaphysique et non épistémologique[8]. L'intérêt de la méthode de Carnap réside en ceci qu'elle permet d'éliminer n'importe quel ensemble de termes d'une théorie scientifique donnée. Supposons que nous disposions d'une théorie portant sur un ensemble de phénomènes, et que cette théorie comporte des expressions au statut jugé problématique. À titre d'illustration, considérons une théorie comportant toutes les expressions issues des sciences de la nature, mais également des termes psychologiques, comme « croyance », « désir », « intention », « préférence », « émotion », « douleur », etc. Nous supposerons que ces expressions figurent dans des énoncés généraux exprimant des lois, auxquelles nous faisons appel pour prédire et pour expliquer les phénomènes psychologiques. Ainsi, nous pouvons expliquer le fait que Pierre soit allé chercher une bière au frigo par le fait qu'il croyait qu'il y avait une bière dans le frigo, qu'il avait envie de se désaltérer, et que de façon générale, lorsque X désire que la condition C se réalise, et qu'il croit qu'accomplir l'action A permettra de réaliser la condition C, il accomplit l'action A. Mais quoique nous supposions, dans nos pratiques habituelles d'attribution d'états mentaux, que ces états possèdent des pouvoirs causaux, nous ne sommes certainement pas au clair sur leur statut ontologique. Pourrait-on en principe, par exemple, prédire les croyances possédées par un agent à partir d'une description physique exhaustive du fonctionnement de son cerveau ? Est-il possible, à partir de notre compréhension du vocabulaire mental, de déterminer la nature métaphysique des états dénotés par les expressions de ce vocabulaire ? Ce sont à ces questions ontologiques que Lewis veut avant tout répondre.

Cela nous amène à la seconde différence importante entre la démarche de Carnap et celle de Lewis. Le but ultime de Carnap, lorsqu'il reprend la méthode de Ramsey, est de trouver une frontière nette entre les énoncés analytiques d'une théorie T et ses énoncés synthétiques[9]. Ce qui est déterminant pour lui n'est pas d'identifier les référents réels des expressions théoriques, mais bien plutôt de parvenir à exprimer le postulat de signification susceptible de contraindre la détermination de la référence. Nous l'avons vu, Carnap laisse donc ouverte la possibilité selon laquelle les expressions théoriques auraient plusieurs réalisations distinctes. Dans ses premiers écrits au moins, Lewis exprime un désaccord avec Carnap sur ce

[8] Il faut de plus souligner que Lewis ne partage pas la thèse de Carnap selon laquelle il serait possible de tracer une frontière nette entre le vocabulaire théorique et le vocabulaire observationnel.

[9] Sur ce point, voir Demopoulos (2007) et Uebel (2011).

point précis[10]. Selon lui, une théorie est, normalement, suffisamment informative pour qu'il soit possible d'identifier les référents des expressions problématiques de façon unique. Ainsi l'expression « douleur » de notre psychologie populaire dénote-t-elle l'unique propriété cérébrale réalisant le rôle causal attribué aux états de douleur par le postulat de cette théorie, si cette unique propriété existe, et rien du tout si aucune propriété ne correspond à ce rôle causal, ou si plusieurs propriétés différentes lui correspondent.

Cette seconde divergence avec Carnap est importante, car elle débouche sur une conception descriptiviste des concepts de propriétés « problématiques » – dans le cas qui nous intéresse ici, des concepts psychologiques. Si l'on part du présupposé selon lequel une théorie vise à ne posséder que des réalisations uniques de ses expressions problématiques, on peut en effet formuler, à partir de l'énoncé de Carnap de la théorie, des descriptions définies de chacune des dénotations des expressions problématiques. Voici par exemple la description définie que l'on peut produire pour l'expression problématique t_1 de notre théorie (Lewis, 1970, p. 87) :

> (2) t_1 = l'unique y_1 tel que : $\exists\, y_2 ... y_n \;\forall\, x_1 \; ... x_n$ ($Tx_1 ... x_n$ si et seulement si ($y_1 = x_1$ et $y_2 = x_2$ et ... $y_n = x_n$).

Pour le dire plus simplement : t_1 est l'unique entité qui satisfait, au moins typiquement[11], le rôle fonctionnel associé par la théorie T à l'expression « t_1 ». Soulignons que selon cette conception, les différents rôles attribués aux expressions problématiques ne pourront probablement être définis que les uns relativement aux autres. Cela tient au fait que l'énoncé de Ramsey d'une théorie décrit les rôles causaux des entités désignées par les expressions problématiques, et que ces rôles causaux sont interdépendants les uns des autres. Ainsi la définition de la douleur sera-t-elle la suivante :

> (3) avoir mal = l'unique propriété P jouant le rôle causal de la douleur.

---

[10] Cf. Lewis, (1970, p. 83 et 1972), et Galinon (2009) pour une comparaison entre les approches respectives de Carnap et Lewis.

[11] Pour une explication de la précision « au moins typiquement », voir le paragraphe « Douleur de fou et douleur de martien » plus bas.

Dans cette description, l'expression « jouant le rôle causal de la douleur » est l'abréviation de l'énoncé quantifié dont (2) donne la forme logique. Il est clair que la description du rôle causal de la douleur ne peut être donnée qu'en mentionnant les relations causales existant, dans la cognition, entre les occurrences d'expériences de douleur et les occurrences d'autres états mentaux. Ainsi une occurrence de douleur cause-t-elle typiquement une croyance portant sur cette occurrence, ainsi qu'un désir de voir l'occurrence cesser. Mais les expressions « croyance » et « désir » sont elles-mêmes problématiques, et doivent donc elles-mêmes être définies fonctionnellement. On voit donc bien que les expressions problématiques forment un réseau, et que les définitions sont des inter-définitions (Lewis, 1970, 1972 et 1994).

Selon Lewis, les définitions fonctionnelles des expressions problématiques expriment les intensions de ces expressions[12]. Elles permettent en effet d'identifier leur dénotation dans chaque monde possible : dans chaque monde, la description fonctionnelle correspondant à une expression détermine l'unique réalisation du rôle causal associé à l'expression dans le monde en question. Ces définitions sont donc bien censées refléter le sens des expressions problématiques, sens qui se trouve décrit uniquement à l'aide des expressions non-problématiques. A cet égard, l'identification des référents des expressions problématiques relève de l'analyse conceptuelle : la connaissance des rôles causaux des référents provient en effet de la connaissance du sens de ces expressions. Ce point est crucial pour le projet de Lewis, car il lui permet de défendre une conception tout à faire originale du physicalisme.

Considérons en effet l'ensemble des vérités formulées dans le vocabulaire des sciences physiques d'un côté, et de l'autre un ensemble de vérités formulées dans un vocabulaire au statut plus problématique, par exemple l'ensemble des vérités psychologiques. Selon le point de vue physicaliste, les propriétés psychologiques, étant des propriétés naturelles, peuvent être décrites dans le vocabulaire des sciences physiques, ce qui implique la vérité d'un ensemble d'énoncés d'identités, du type de (4) et (5), qui permettent d'établir un pont entre le vocabulaire physicaliste et le vocabulaire psychologique :

(4) Douleur = Activation des cellules X_i dans le Y_i

(5) Colère = Activation des cellules X_j dans le Y_j

... etc.

[12] Rappelons que dans la tradition sémantique issue de Carnap, l'intension d'une expression nominale est une fonction faisant correspondre à chaque monde possible la dénotation de l'expression dans ce monde.

La question qui se pose cependant, comme le souligne Lewis de façon particulièrement lucide dès ses premiers écrits, est de savoir comment ces énoncés peuvent être justifiés[13]. Or il ne semble y avoir que deux grandes méthodes de justification disponibles pour de tels énoncés, qui sont les suivantes :

(i) La méthode de justification directe par déduction, en premier lieu, d'après laquelle les énoncés d'identité du type de (4) et (5) doivent être déduits à partir de l'ensemble des vérités exprimées dans le vocabulaire physicaliste.

(ii) La méthode de justification indirecte, par inférence à la meilleure explication.

Selon la méthode de justification indirecte, il n'est pas possible de dériver logiquement les énoncés d'identité à partir de l'ensemble des vérités physiques. On peut, néanmoins, les justifier indirectement. À la suite de Smart (1959), on pourrait en effet penser que les énoncés d'identité doivent être acceptés essentiellement en raison de l'économie ontologique qu'ils permettent de réaliser. Si l'on considère que la douleur n'est autre qu'un état d'activation de certaines cellules, le concept de douleur n'est au fond qu'une façon commode de décrire des phénomènes qui, dans leur nature, sont des phénomènes neuronaux.

La méthode de justification directe doit néanmoins, toutes choses égales par ailleurs, être préférée à la méthode indirecte. On sait que les considérations de simplicité, auxquelles la méthode de justification indirecte fait essentiellement référence, n'ont qu'une valeur épistémologique assez limitée : si le seul argument avancé par un philosophe physicaliste pour justifier la supériorité de sa position face à une position métaphysique pluraliste réside dans sa plus grande simplicité ontologique, on peut parier que cette justification indirecte ne convaincra pas son adversaire[14].

Or les énoncés d'identification des entités ontologiquement problématiques, du type de (4) et de (5), peuvent être logiquement déduits à partir de l'ensemble de la théorie. Cette déduction aura la forme suivante :

1. La douleur = l'unique propriété jouant typiquement le rôle causal attribué à la douleur par T (analyse conceptuelle) ;

[13] Voir par exemple la discussion de Lewis (1970, p. 91-92).

[14] Cf. sur ce point, Daly (2010, chap. 4) qui est tout entier consacré à la discussion de l'usage des arguments de simplicité en philosophie. Voir également la critique par Kim (2005) de l'usage de l'inférence à la meilleure explication pour justifier les énoncés d'identité.

2. L'unique propriété jouant typiquement le rôle causal attribué à la douleur par T = l'activation des cellules X_i dans Y_i (d'après la partie purement physicaliste de T) ;
3. Conclusion : La douleur = l'activation des cellules X_i dans Y_i.

Soulignons qu'une telle inférence constitue à la fois une justification déductive directe de l'identité psychophysique qui figure en conclusion, mais également une explication de cette identité. Elle nous permet en effet de comprendre pourquoi il est légitime d'affirmer que la douleur n'est autre que l'activation d'un certain ensembles de neurones : c'est parce que cet état d'activation est l'unique entité qui, dans notre monde, joue le rôle causal attribué à la douleur par notre théorie totale de la réalité.

III. La place de l'analyse conceptuelle dans la méthode de réduction fonctionnelle

La démarche réductionniste de Lewis accorde une place centrale à l'analyse conceptuelle[15]. C'est en effet par une analyse de l'usage théorique qui est fait des expressions psychologiques, analyse qui conduit à la ramseyfication de ces expressions, que l'on peut justifier les descriptions définies des différents rôles causaux qui leur sont associés. Ces descriptions permettent à leur tour, dans un second temps, à la fois de justifier et d'expliquer les énoncés d'identité psychophysique. Il nous semble important de discuter quelques problèmes que l'on pourrait soulever à l'encontre de cette démarche. En premier lieu, quelle place l'analyse conceptuelle joue-t-elle exactement dans la réduction fonctionnelle ?

A. Carnap et Lewis sur la signification des expressions problématiques

Afin de répondre à cette question, je voudrais d'abord revenir sur la différence fondamentale qui existe entre l'approche de Carnap et celle de Lewis. Carnap, à l'époque où il reprend les idées de Ramsey pour les développer, c'est-à-dire après la publication de l'article « Empirisme, sémantique et ontologie » en 1950, adhère à une conception relativiste de l'ontologie. Il convient selon lui de distinguer entre deux sortes de questions d'existence, les questions internes et les questions externes. Pour tout prédicat F appartenant au vocabulaire d'une théorie T, on ne peut

[15] Il a été récemment rejoint, sur ce point, par Kim (2005), Chalmers & Jackson (2011), et les auteurs se réclamant du programme de Canberra – sur ce programme voir Brandon-Mitchell & Nola (2009).

s'interroger sur la question : « y a-t-il des F ? » que relativement au cadre constitué par T. Par exemple, étant donnée une théorie employant le concept d'électron, on peut se demander s'il y a des électrons dans le monde, et répondre positivement à la question. Il s'agit là d'une question interne. On peut s'interroger également sur la question de savoir si les électrons existent absolument, indépendamment du cadre théorique introduisant le terme « électron ». Mais d'après Carnap, cette seconde question, qu'il nomme « question externe », ne peut pas recevoir de réponse factuelle. Tout au plus peut-on lui donner une réponse pragmatique, si l'on considère qu'il est commode d'employer un langage introduisant le terme « électron ». L'approche développée par Carnap n'est donc ni réaliste, ni anti-réaliste, mais bien relativiste : les questions factuelles d'existence présupposent l'acceptation préalable d'un cadre théorique, c'est-à-dire d'un langage dont toutes les expressions doivent posséder une signification. On ne peut donc bien évidemment pas considérer que les référents des expressions théoriques existent préalablement aux conventions qui permettent de définir, au moins partiellement, les significations des termes observationnels et des termes théoriques.

C'est sans doute pour cette raison que Carnap, de façon prudente, formule l'énoncé de Ramsey sans lui associer de condition d'unicité[16] : il suffit, pour que l'énoncé de Ramsey de T tel que Carnap le construit soit vrai, qu'il existe une réalisation, quelle qu'elle soit, de la théorie T. Ce n'est en effet pas la relation de dénotation entre le terme théorique et l'éventuelle unique entité qu'il dénote qui détermine la signification de ce terme, mais bien plutôt l'ensemble des relations qui existent, d'une part, entre l'usage du terme et des observations possibles, et, d'autre part, entre les usages du termes et ceux des autres termes théoriques. Rappelons que selon Carnap, les théories comportent, en plus de lois théoriques où figurent exclusivement des termes théoriques, des règles de correspondance, qui relient les termes théoriques aux termes d'observation. En voici un exemple : « la température d'un gaz (mesurée par un thermomètre) est proportionnelle à l'énergie cinétique moyenne de ses molécules » (Carnap, 1966, tr. fr. p. 227). Grâce à ces règles, les expressions théoriques entrent en contact avec la réalité observable. Néanmoins, elles n'épuisent pas à elles seules la signification de ces expressions. D'abord parce que les énoncés purement théoriques contraignent également leur signification. Mais également et surtout parce que les scientifiques peuvent à tout moment, selon Carnap, introduire de nouvelles règles de correspondance. Cette

[16] Cf. Galinon (2009) et Uebel (2011).

liberté dans l'introduction de nouvelles règles est même selon lui « un processus qui n'a pas de fin » (Carnap, 1966, tr. fr. p. 231), d'une part parce que « rien dans l'histoire de la physique ne permet, jusqu'à maintenant, de supposer que la physique sera un jour achevée » (ibid.), et d'autre parce que si l'on atteignait un jour le terme auquel la signification d'une expression théorique ne pourrait plus être précisée par l'introduction de nouvelles règles de correspondance, cette expression « cesserait du même coup d'être théorique » et « s'intégrerait au langage d'observation » (ibid., p. 231). On voit donc pourquoi il n'est pas souhaitable selon Carnap d'essayer de formuler une définition explicite du sens de chacune des expressions théoriques : le sens de chacune de ces expressions, à un moment du stade de développement d'une théorie donnée, est indéterminé, et il est destiné à rester indéterminé tout au long du développement historique de la théorie[17].

La perspective de Lewis est complètement différente, puisque ce dernier adhère au réalisme métaphysique : il considère que lorsqu'une théorie est vraie – ou du moins approximativement vraie – les termes scientifiques qui y figurent possèdent des dénotations réelles, qui existent de façon absolue et indépendamment du cadre de la théorie. Ces dénotations sont des entités naturelles, et ni leur existence ni leur nature ne dépendent de nos schèmes conceptuels. Il est donc raisonnable, dans ce contexte réaliste, de proposer des descriptions définies des entités « problématiques », formulées uniquement dans le vocabulaire non-problématique – le fait que ce vocabulaire soit purement observationnel n'ayant guère d'importance pour un réaliste qui récuse la distinction même entre ce qui est observable et ce qui ne l'est pas.

La question méthodologique de la place de l'analyse conceptuelle ne se pose cependant que de façon plus pressante dans ce nouveau contexte. Dans le cadre carnapien, ce sont tous les énoncés de la théorie pris ensemble qui permettent de déterminer, de façon holiste et à un moment donné t de l'histoire des sciences, le postulat de signification exprimé par l'énoncé de Carnap. Il s'agit au fond d'une conception très modeste et très prudente de la manière dont le sens des expressions théoriques se trouve contraint par leurs usages dans des énoncés.

Lewis va beaucoup plus loin, et son ambition pourrait sembler, du coup, exagérée : l'analyse vise en effet selon lui à dégager une description du référent de chaque expression problématique de la théorie, référent que la

[17] Ce point est bien établi par Uebel (2011).

description nous permet d'identifier dans chaque monde possible[18]. Ainsi, la pure réflexion portant sur le rôle théorique du concept de douleur est censée nous apprendre que la douleur est l'unique état x jouant typiquement le rôle causal de la douleur dans T, et cette description désigne la réalisation de la douleur dans chaque monde possible – réalisation qui peut être différente dans chacun des mondes.

B. La douleur du fou et la douleur du martien

Avant d'aborder les difficultés qu'elle suscite, remarquons que cette analyse possède un avantage, puisqu'elle permet de distinguer deux éléments dans la signification d'une expression problématique : son intension d'une part, c'est-à-dire son sens descriptif tel que dégagé par l'analyse conceptuelle, et d'autre part, relativement à un monde possible donné, sa dénotation dans ce monde possible. C'est sur cette distinction que s'appuie Lewis, dans l'article célèbre « Mad pain and martian pain », pour résoudre les problèmes liés à l'existence de réalisations multiples des concepts mentaux définis fonctionnellement (Lewis, 1980). Rappelons que Lewis décrit la « douleur du fou » comme une douleur possédant un rôle fonctionnel très différent de nos douleurs habituelles – éprouver de la douleur aide le fou à se concentrer et l'incite à se livrer à des réflexions mathématiques –, mais pourtant indiscernable de nos douleurs du point de vue phénoménal. Avoir mal fait au fou exactement le même effet en première personne qu'à nous, bien que cet état n'ait ni les mêmes causes typiques, ni les mêmes effets typiques, que notre douleur. La douleur du martien, en revanche, est un état physique qui possède la même description causale/fonctionnelle que nos douleurs typiques. Ainsi, un martien qui ressent de la douleur en ce sens est motivé à la faire cesser. Mais la réalisation physique de la douleur du martien n'a rien à voir avec la réalisation physique de notre douleur : le rôle causal de la douleur est en effet implémenté, chez les martiens, par « le gonflement de nombreuses petites cavités dans les pieds ».

Une théorie crédible de la douleur, soutient Lewis, « doit faire une place tant à la douleur du fou qu'à la douleur du martien » (Lewis, 1980, in Lewis, 1983b p. 123), aussi difficile que cela puisse sembler de prime

---

[18] Lewis (1970) soutient en effet que les définitions des expressions problématiques décrivent le sens de ces expressions précisément parce qu'elles permettent d'identifier leur référent dans tous les mondes possibles : « Qu'en est-il de leurs sens ? Nous avons déjà spécifié leurs sens. Car nous avons spécifié leurs dénotations dans chaque monde possible, et pas seulement ici, dans le monde réel » (p. 85).

abord. Or l'approche à la fois fonctionnaliste et réductionniste qu'il revendique permet de comprendre pourquoi nous pouvons parler, en deux sens différents, de douleur à propos du fou et du martien. Si l'on considère la douleur comme l'état physique dénoté, dans notre monde réel, par notre concept habituel de douleur, le fou peut être dit avoir mal : il instancie en effet cet état physique, même si dans son système cognitif particulier l'état en question ne possède ni les causes typiques ni les effets typiques de la douleur. En ce sens du concept de douleur, en revanche, le martien n'a pas mal, puisqu'il n'instancie pas l'état qui, dans le monde réel, réalise notre concept de douleur. Si l'on privilégie au contraire non le référent actuel du concept de douleur, mais son intension, alors, comme cette intension ne s'applique qu'aux états ayant certaines causes et effets typiques, on devra dire que le fou n'a pas mal, mais que c'est le cas, en revanche, du martien.

Présentée ainsi, la réponse de Lewis n'est pas entièrement satisfaisante. Un concept ne peut en effet avoir qu'une seule signification, qu'il serait en l'occurrence plausible d'identifier à son intension. Il est peu satisfaisant de considérer le concept de douleur comme ambigu, ou du moins une telle hypothèse appelle une justification qui ne soit pas ad hoc. Par ailleurs, nous y avons insisté plus haut, la description fonctionnelle issue de l'analyse conceptuelle des usages de nos concepts dans la psychologie populaire est censée permettre d'identifier un unique référent dans chaque monde possible. Or, comme Lewis le remarque lui-même, le martien dont il est question dans l'expérience de pensée que nous discutions pourrait exister dans notre monde possible. En général, il semble parfaitement concevable que le rôle causal correspondant à notre concept naïf de douleur soit réalisé par des états physiques différents dans des organismes appartenant à des espèces ayant évolué différemment au cours de leur histoire. Mais si c'est le cas, c'est-à-dire s'il y a dans un monde possible plusieurs réalisations de l'énoncé de Ramsey de la théorie naïve de la douleur, la description définie exprimant l'intension du terme « douleur » ne pourra être uniquement satisfaite.

Pour cette raison, Lewis propose à partir de 1980 de relativiser les descriptions fonctionnelles à certains éléments contextuels[19]. Ainsi peut-on s'interroger sur l'unique réalisation pour une espèce donnée d'une description multi-réalisable. On pourra alors bien parler de la douleur de l'être humain – c'est-à-dire de la propriété naturelle qui réalise le concept de douleur dans l'espèce humaine –, de la douleur du martien, etc. Cette idée a

[19] Cette relativisation, introduite dans Lewis (1980), sera en effet reprise dans Lewis (1994).

eu une importante descendance dans la philosophie contemporaine de l'esprit, puisqu'elle a été reprise par Jaegwon Kim (1998, chap. 4, trad. fr. p. 135-137), qui propose de remplacer la notion de réduction globale par une notion de réduction locale. Considérons ainsi un concept C défini de façon fonctionnelle, par exemple le concept du gène responsable de la couleur des yeux. Supposons que le rôle fonctionnel correspondant à C soit occupé dans une première espèce E_1 par la propriété physico-chimique P_1, et qu'il soit réalisé dans une seconde espèce E_2 par la propriété physico-chimique P_2, dans une troisième espèce par P_3, ... etc. On ne peut pas définir globalement le concept C à l'aide du vocabulaire physico-chimique permettant de décrire P_1, puisque l'extension de C comprend également les propriété P_2, P_3, ...etc. Il est cependant possible d'appliquer localement l'idée de réduction, c'est-à-dire de soutenir que le concept C peut être considéré, relativement à l'espèce E_1, comme un concept dénotant la propriété P_1. Ainsi, pour les membres de l'espèce E_1, on peut considérer que le concept C a la même extension que la description physico-chimique de P_1 ; pour les membres de l'espèce E_2, il aura en revanche l'extension de la description physico-chimique de P_2, ...etc.[20].

Qu'en est-il cependant de la douleur du fou ? Remarquons en premier lieu que la situation imaginée par Lewis est très particulière, puisque la propriété naturelle D, dénotée par notre concept habituel de douleur, joue un rôle causal différent du rôle causal de la douleur. Il faut donc supposer que dans l'architecture cognitive particulière du fou, les occurrences de D sont causées de façon inhabituelle, et qu'elles ont également des effets inhabituels. Néanmoins, la propriété naturelle D posséderait les causes et les effets typiques de la douleur si elle avait des occurrences au sein d'un système cognitif normal. Pour cette raison, Lewis soutient que nous pouvons appliquer notre concept de douleur à la douleur du fou. Ce concept vise en effet à dénoter une propriété naturelle : comme nous l'avons expliqué, le but d'une réduction fonctionnelle est d'identifier les propriétés désignées par des expressions problématiques à des propriétés décrites dans le vocabulaire des sciences naturelles. En ce sens, Lewis défend une théorie de l'identité des types : avoir mal n'est rien d'autre métaphysiquement qu'être dans un certain état d'activation neuronale. Nous décrivons cet état neuronal grâce au rôle causal que nous pouvons extraire

[20] L'idée de réduction locale soulève cependant des difficultés, dont la discussion a débouché sur un débat complexe. Le principal problème est sans doute de savoir si cette idée ne débouche pas sur une forme d'éliminativisme. Cf. Polger (2004) pour une défense de la position de Lewis et de Kim ; cf. Esfeld & Sachse (2007) pour une critique de cette position et pour une proposition alternative.

de notre connaissance de la psychologie naïve de la douleur. Mais ce rôle causal ne nous dit rien de la nature de la douleur, de son essence, à laquelle seule la description physico-chimique de cette propriété peut nous donner accès. Nous voyons donc pourquoi il est important d'ajouter le qualificatif « typiquement » à la description du rôle causal d'une propriété problématique : ces rôles causaux ne se manifestent pas toujours, mais uniquement dans des contextes typiques. Reste que la notion de typicité est relative[21] : l'accent typique d'un locuteur du français, par exemple, dépend de son origine géographique. Il y a donc là un second argument qui doit nous conduire à relativiser les descriptions fonctionnelles à des éléments contextuels. Ainsi le rôle conceptuel de la douleur doit-il être considéré comme le rôle typiquement possédé par une certaine propriété, elle-même susceptible d'être instanciée au sein d'une population à la fois contextuellement saillante et aux frontières naturelles – par exemple une espèce biologique.

La conception descriptiviste des concepts « problématiques » tels que les concepts psychologiques, défendue par Lewis, a, on le voit, de nombreux avantages. Elle permet de concilier les points forts du fonctionnalisme avec ceux de la théorie de l'identité. Comme dans les approches strictement fonctionnalistes[22], Lewis peut en effet expliquer l'intuition selon laquelle les propriétés mentales possèdent – ou au moins, pourraient posséder – des implémentations physiques multiples. Cela tient au fait que les rôles fonctionnels sont relativisés à des éléments contextuels : dans le cas des concepts psychologiques, à des espèces biologiques. De la sorte, notre concept de douleur est analysé comme une expression ambiguë, qui se subdivise en fait en un concept de douleur-pour-les-humains, de douleur-pour-les-martiens, etc. Mais la propriété de douleur est une propriété physico-chimique de premier ordre, et non un rôle fonctionnel – donc une propriété de second ordre – comme le soutient par exemple Hilary Putnam (1967). On peut être assuré, pour cette raison, qu'elle possède bien des pouvoirs causaux[23].

[21] Cf. sur ce point la discussion éclairante de Daniel Nolan (2005 et 2009).
[22] Cf. Putnam (1967) et Fodor (1974).
[23] A cet égard, le fonctionnalisme de Lewis échappe à l'argument de l'exclusion causale opposé par Kim aux tenants du fonctionnalisme métaphysique anti-réductionniste. Cf. Kim (1998 et 2005).

C. Analyse conceptuelle et magnétisme référentiel

David Lewis, bien qu'il soit parti de l'analyse carnapienne des termes théoriques, l'a modifiée en profondeur. Dans sa métaphysique de l'esprit, l'élimination des expressions mentales problématiques joue un rôle pivot, puisqu'elle permet de s'assurer qu'il est en principe possible de dériver logiquement les énoncés d'identité psychophysique à partir d'une connaissance complète du monde physique.

Il me semble cependant que dans le cadre réaliste accepté par Lewis, son approche de la réduction fonctionnelle pose des problèmes épistémologiques dont la conception carnapienne de l'élimination des termes théoriques était exempte. Dans ce cadre en effet, les entités « problématiques » ont une existence indépendante, en un sens absolu. Par ailleurs, ces entités problématiques ne sont pas par nature inobservables. Peut-on, dés lors, maintenir la thèse selon laquelle la découverte des descriptions permettant de les identifier, par l'intermédiaire de leur rôles causaux, repose entièrement sur une analyse conceptuelle ? Cette conception de la réduction ne présuppose-t-elle pas, du coup, une approche descriptiviste du sens et de la référence discréditée depuis longtemps ? Je vais soutenir que les choses sont en réalité plus complexes qu'on pourrait le supposer de prime abord. Certes, la méthode de Lewis est descriptiviste : il s'agit bien d'utiliser la ramseyfication pour obtenir des descriptions des entités problématiques. Mais nous allons voir qu'elle comporte en fait un élément externaliste absolument essentiel. Du coup, je soutiendrai que le rôle de l'analyse conceptuelle dans la démarche de Lewis peut être interprété de façon assez modeste, ce qui renforce me semble-t-il la plausibilité de son approche.

Commençons par souligner qu'une interprétation purement descriptiviste ne permettrait pas d'attribuer des significations satisfaisantes aux expressions problématiques, en raison du problème de l'indétermination de la référence mis en évidence par Quine, et développé par Putnam (1981). Si l'argument de Putnam est complexe dans son détail, l'idée générale peut être formulée de façon extrêmement simple (Lewis, 1984). Nous savons – il s'agit là d'un résultat de logique assez élémentaire – qu'un ensemble cohérent de formules possède au moins un modèle, c'est-à-dire qu'il est possible de trouver un ensemble d'entités et une fonction d'interprétation tels que toutes les formules de la théorie soient vraies dans l'ensemble en question, relativement à la fonction d'interprétation. Par ailleurs, nous savons que si un ensemble de formules a un modèle, l'image de ce modèle par un isomorphisme rendra également toutes les formules de l'ensemble

vraies. Cela implique que le monde, s'il contient suffisamment d'objets, peut rendre vraies n'importe quel ensemble non-contradictoire de formules logiques. Afin d'illustrer ce point[24], considérons une théorie T, ainsi que l'énoncé « les cochons ont des ailes », qu'on supposera non-contradictoire relativement aux énoncés de T. On nommera T' la théorie obtenue en ajoutant cet énoncé à T. Y a-t-il une interprétation du vocabulaire théorique de T' qui soit telle que T' soit vraie dans le monde réel selon cette interprétation ? D'après le raisonnement qui précède, c'est forcément le cas. Et l'on peut effectivement facilement trouver une telle interprétation : par exemple, on pourrait stipuler que « cochon » désigne l'union de l'ensemble des cochons et de l'ensemble des libellules, et que « a des ailes » possède son interprétation standard. Evidemment, ainsi que le notait Quine, cela implique aussi de compenser le caractère non-standard de l'interprétation de « cochon » par d'autres interprétations non-standard des prédicats de la théorie. Par exemple, on voudra que l'énoncé « les cochons sont roses » soit vrai, ce qui pourrait sembler difficile si l'on donne à « rose » son sens standard, puisque les libellules ne sont pas toutes roses. Mais il suffit d'attribuer à « rose », par exemple, la dénotation suivante : l'union des objets roses et des insectes. Dans cette l'interprétation étrange que nous avons stipulée, « les cochons sont roses » est bien vrai, puisque l'union des cochons et des libellules est incluse dans l'union des objets roses et des insectes. On voit donc qu'on peut trouver une interprétation non-standard qui rend vrais à la fois l'énoncé « les cochons ont des ailes » et l'énoncé « les cochons sont roses ». En fait, on peut montrer qu'une théorie quelconque peut, dans au moins une interprétation non-standard, être rendue vraie par le monde, à partir du moment où celui-ci contient suffisamment d'objets[25]. Les contraintes strictement théoriques n'aboutissent au bout du compte que sur des contraintes qui concernent la cardinalité du domaine des entités existantes dans le monde.

Ces remarques pourraient sembler dévastatrices pour le programme de Lewis, puisqu'elles montrent qu'une définition fonctionnelle d'une expression problématique peut toujours être trivialement satisfaite par des interprétations non-standard tout à fait fantaisistes. Il s'agirait d'une réfutation décisive du programme d'analyse conceptuelle proposé par Lewis si ce programme était uniquement descriptiviste, c'est-à-dire si les

[24] Cf. Sider (2011, p. 24).

[25] Et par « suffisamment », on peut entendre au maximum une infinité dénombrable d'objets, comme le montre le théorème de Löwenheim-Skolem. En effet, selon ce théorème, si une théorie finiement axiomatisable possède un modèle infini, elle possède un modèle dénombrable.

descriptions théoriques extraites par ramseyfication devaient à elle seules permettre d'identifier les référents des expressions problématiques[26]. Ce n'est heureusement pas le cas, comme Lewis le souligne d'ailleurs lui-même dans sa remarquable analyse de l'argument modèle-théorique de Putnam[27]. Certes, on ne peut formuler de l'intérieur d'une théorie des contraintes suffisamment forte pour parvenir à une définition implicite satisfaisante des expressions nouvelles qu'elle introduit ; mais le monde lui-même peut nous fournir le « supplément de contrainte » nécessaire, si l'on suppose qu'il possède une structure métaphysique indépendamment de notre pensée, de notre langage et de nos théories. L'idée centrale est la suivante : toutes les classes d'entités ne sont pas disponibles pour servir d'interprétations à nos prédicats ; seules les classes relativement naturelles le sont. Du coup, la plupart des interprétations non-standards des prédicats peuvent être laissées de côté :

> « parmi les innombrables choses et classes qui existent, la plupart sont hétéroclites, regroupées n'importe comment, leurs contours sont mal définis. Seule une petite minorité, l'élite, peuvent être découpées selon leurs articulations, de sorte que leurs frontières soient établies à l'aide d'identités et de différences objectives fondées en nature. Seules ces choses et ces classes d'élite sont éligibles pour servir de référents. » (Lewis, 1984 in Lewis, 1999, p. 65)

Même si l'expression n'est pas de David Lewis, on peut parler de « magnétisme référentiel »[28] pour décrire cette hypothèse : nous ne devons pas chercher, dans notre entreprise de réduction, à identifier n'importe quels référents possibles pour nos expressions problématiques, mais uniquement des référents naturels, voire même les plus naturels, puisque la naturalité est pour Lewis une question de degré[29]. Seuls de tels référents naturels peuvent jouer le rôle « d'aimants » référentiels, ce qui suffit à bloquer l'argument de Putnam[30].

---

[26] L'impact de ce type d'argument pour la méthode de ramseyfication a fait l'objet de nombreuses discussions. Voir des mises au point récentes Demopoulos & Friedman (1985), Ketland (2009) et Uebel (2011).
[27] Cf. Lewis (1984), et pour une discussion de cette article, Sider (2011).
[28] Sur l'idée du magnétisme référentiel, voir Sider (2011, p. 26 sq).
[29] Cf. Lewis (1983a) et Hall (2010).
[30] Ketland (2009, p. 44), parvient à une conclusion semblable : la méthode de ramseyfication ne peut permettre de définir de façon non-triviale les expressions problématiques d'une théorie qu'à condition que l'on présuppose que ces expressions dénotent des classes naturelles.

On voit, en conséquence, que Lewis adopte une version nuancée du descriptivisme. Certes, la dénotation des expressions problématiques analysées est donnée par des descriptions, issues du processus de ramseyfication. Mais ces descriptions doivent avant tout être conçues comme des guides permettant de repérer des classes naturelles. Lewis admet donc parfaitement que certaines parties d'une description ne soient pas satisfaites par le référent de la description, si ce référent constitue par ailleurs une bonne approximation de la description. Autrement dit, le fait que certains énoncés inclus dans le postulat d'une théorie soient faux ne doit pas nécessairement nous empêcher de parvenir à une identification des entités problématiques dénotées par les termes théoriques.

Ce point, nous le verrons plus bas, est important lorsqu'on s'intéresse à la métaphysique des états phénoménaux : Lewis considère que certaines des conditions que notre psychologie populaire impose à ces états ne peuvent tout simplement pas être satisfaites par des propriétés naturelles[31]. On voit donc que le statut des conditions descriptives associées par ramseyfication aux expressions problématiques est complexe. Il ne suffit pas qu'une description associe la condition C à la dénotation d pour que l'énoncé « Si d existe, Cd » soit analytiquement vrai : la dénotation est l'entité naturelle qui satisfait au mieux la description, et si cette entité naturelle ne satisfait pas la condition C, tant pis pour C que l'on laissera alors de côté. On peut donner de nombreuses illustrations, plus ou moins radicales, de telles situations. Ainsi la découverte que les baleines ne sont pas des poissons, mais des mammifères, ne nous a-t-elle pas conduits à penser que les baleines n'existent pas : plutôt que de considérer que la définition issue de notre conception naïve des baleines ne peut être satisfaite, nous intégrons l'information issue de l'observation à cette définition, en opérant un changement dans le sens du terme « baleine » (Nolan, 2009). Dans certains cas, il est difficile de dire si l'on choisirait de modifier la signification du terme, ou de considérer qu'il ne peut pas être associé à une dénotation réelle.

Considérons ainsi l'expérience de pensée suivante, développée par M. Tye (2009, p. 58 sq.) mais inspirée d'expériences de pensée similaires imaginées initialement par H. Putnam. Supposons que l'eau soit en fait non un liquide transparent, mais une matière granuleuse, composée de toutes petites particules roses, chacune de la taille d'un grain de sable. Ces particules composent également les nuages dans l'atmosphère. Supposons également que cette matière granuleuse nous apparaisse comme un liquide à

[31] Voir Lewis (1995) et la discussion de cet article ci-dessous.

cause d'une manipulation cognitive massive opérée par des martiens. Que dire d'une telle situation ? Selon Tye, on doit maintenir que l'eau n'est pas ce que l'on croit ; autrement dit, l'énoncé « l'eau est le liquide transparent présent dans les rivières » devrait être considéré non comme analytiquement vrai, mais comme empiriquement faux. Je ne partage pas, pour ma part, les intuitions de Tye. Il me semble que si notre théorie naïve de l'eau s'avérait aussi radicalement fausse, nous considérerions que l'eau n'existe en fait pas. Mais ce que je voudrais surtout souligner, c'est que l'approche de Lewis permet de donner un sens à toutes nos intuitions concernant de telles situations. Si nous avons l'intuition que nous aurions fait dans ce cas une découverte empirique radicale sur l'eau, c'est en raison du magnétisme référentiel : il y a bien une propriété naturelle correspondant au moins à certains aspects importants du rôle causal associé à « eau », et cette propriété est en un sens éligible pour servir de dénotation au terme « eau ». Si nous éprouvons au contraire, comme c'est mon cas, une forte réticence à admettre qu'on continuerait à parler d'« eau » à propos d'une matière granuleuse non-liquide, c'est que nous privilégions les conditions descriptives associées à cette expression par l'analyse. Dans les deux cas, la conception à la fois descriptiviste et externaliste que Lewis propose peut fournir une explication satisfaisante : identification de la propriété naturelle correspondant à une partie du rôle fonctionnel associé à « eau » accompagnée d'une modification radicale de la définition fonctionnelle associée dans le premier cas ; élimination de l'eau de notre ontologie, suivi d'un remplacement de cette substance par la matière granuleuse découverte dans le second.

Toute cette discussion débouche sur une question fondamentale : dans quel cas doit-on considérer une condition descriptive C comme appartenant de manière essentielle à la définition associée à un terme d, de sorte que « Si d existe, Cd » soit analytiquement vrai ? Dans quel cas l'attribution de C à d peut-elle être en revanche considérée comme négociable, de sorte que la découverte empirique que d ne possède en fait pas C ne conduise pas à éliminer d de l'ontologie, mais bien plutôt à réviser notre théorie sur d, et donc la description par ramseyfication ? Lewis ne répond pas précisément à cette question, même si, comme je l'ai souligné, on peut inférer de l'esprit de son approche qu'à peu près n'importe quelle condition peut être considérée comme négociable[32]. A la suite de Daniel

[32] On trouve des prises de position assez variées sur ce sujet dans l'oeuvre de Lewis. Dans Lewis (1995) il écrit qu'il n'y a « pas de réponse bien déterminée » à la question de savoir à partir de quel moment on peut considérer que des conditions descriptives littéralement non satisfaites peuvent néanmoins être considérées comme ayant une dénotation approximative.

Nolan (2005, p. 129), je veux surtout insister sur un point qui me paraît crucial, et qui me semble souvent négligé dans les discussions actuelles de la méthode de l'analyse conceptuelle en philosophie de l'esprit. Même si l'on rejetait complètement, comme le fait par exemple Michael Tye, la thèse selon laquelle nos théories naïves permettent de construire des définitions au moins en partie analytiques des expressions problématiques – par exemple des termes mentaux –, l'approche de Lewis garderait une partie importante de sa pertinence. Le rôle que jouent les descriptions obtenues par ramseyfication dans les explications réductives ne dépend en effet pas du fait que ces descriptions soient connues de façon a priori. Ce qui est important, au regard de leur statut épistémique, c'est que ces descriptions soient correctes au moins dans une large mesure, suffisamment en tour cas pour identifier les propriétés naturelles correspondantes. Ici encore, l'idée de magnétisme référentiel est cruciale. Seule une théorie philosophique des concepts permettrait de déterminer s'il existe des conditions descriptives canoniquement associées aux usages des expressions problématiques, et de préciser éventuellement la nature de ces conditions. Or, il est indubitable que Lewis n'a jamais essayé de développer une telle théorie. Mais est-ce si important relativement au projet réductionniste ? Quelle que soit la façon dont nous concevons naïvement une entité problématique, on peut supposer que si nous possédons réellement un concept de cette entité, c'est-à-dire si nous disposons d'un ensemble de capacités cognitives nous permettant d'acquérir des connaissances sur celle-ci, nous devons également posséder des connaissances, éventuellement tacites, sur son rôle causal. Que ces connaissances puissent être justifiées a priori ou non ne me semble pas essentiel pour l'utilisation de la méthode de réduction par fonctionnalisation, du moins dans la plupart des domaines théoriques[33].

IV. Peut-on réduire le vocabulaire phénoménal ?

Il y a au moins un domaine de vérités dans lequel il ne va pas de soi, au moins pour la plupart des philosophes, que notre maîtrise de concepts implique une connaissance de rôles causaux : c'est le domaine de la connaissance phénoménale, c'est-à-dire de nos connaissances portant sur l'effet que cela fait de vivre, en première personne, telle ou telle sorte

---

Dans d'autres écrits en revanche, par exemple dans Lewis (1997), il semble considérer que la non-satisfaction d'une des conditions exprimées par la théorie naïve des couleurs suffit à éliminer un candidat possible pour la dénotation.

[33] Pour un point de vue radicalement différent du mien, qui conduit donc à douter de la pertinence de la méthode de réduction par fonctionnalisation, voir Tye (2009).

d'expérience. De nombreux philosophes physicalistes seraient en effet certainement prêts à soutenir que l'on peut savoir précisément l'effet que cela fait de voir du rouge sans être pour autant en position de déterminer le rôle causal de l'état cérébral correspondant à cette sensation de rouge. Il nous faut donc présenter maintenant la façon dont Lewis conçoit l'application de sa méthode de réduction par fonctionnalisation au domaine phénoménal. Selon cette méthode, nous devons partir de notre psychologie naïve, c'est-à-dire des connaissances qui nous permettent de rationaliser, de prédire et d'expliquer les comportements des êtres humains en leur attribuant des états mentaux. Ces connaissances, comme le souligne Lewis (1999, p. 298), peuvent être tacites, et donc s'apparenter à nos connaissances grammaticales telles que Chomsky les conçoit. Elles doivent cependant appartenir à l'ensemble de nos connaissances communes – au sens où, pour toute connaissance C appartenant à cette théorie naïve, chaque agent doit non seulement posséder C, mais être connu des autres agents comme possédant C, savoir lui-même que les autres agents savent qu'il possède C, etc. Cette condition exclut que la psychologie scientifique soit utilisée pour déterminer la nature des rôles causaux associés au vocabulaire mental en général, et phénoménal en particulier, du moins si l'on considère que la psychologie scientifique va au-delà de la simple explicitation de nos connaissances naïves. Néanmoins, comme je l'ai souligné plus haut, ce point n'est pas essentiel : la méthode de réduction par fonctionnalisation peut être appliquée à la psychologie scientifique comme à la psychologie naïve. Nous allons discuter deux obstacles qui pourraient sembler s'opposer à la démarche de Lewis dans le cas spécifique des états phénoménaux. Les deux problèmes que nous allons discuter sont de nature épistémologique : le premier obstacle provient de la façon dont nous concevons naïvement ces états ; le second est issu de l'argument de la connaissance de Franck Jackson.

A. La thèse de l'identification

Commençons par remarquer que l'analyse fonctionnelle du vocabulaire mental ne nous donne accès qu'aux rôles causaux des états mentaux, mais non à leur nature intrinsèque. Considérons ainsi le terme « douleur ». Selon Lewis, tout agent qui maîtrise la psychologie populaire sait, du fait simplement de cette maîtrise, que les états de douleur ont un rôle causal spécifique – c'est-à-dire certaines causes typiques et certains effets typiques. Mais la connaissance de ces causes et de ces effets typiques ne nous dit rien, en elle-même, de la nature de la douleur. Pour connaître cette

dernière, il convient selon Lewis de s'enquérir des propriétés naturelles qui, selon ce que nous apprend notre connaissance scientifique du monde, peuvent réaliser dans notre espèce le rôle causal de la douleur.

Il découle donc de l'approche réductionniste que préconise Lewis que nous ne pouvons accéder à la nature intrinsèque de la douleur ou de tout autre état phénoménal qu'à partir d'une perspective en troisième personne, supposant une identification des rôles causaux issue de l'analyse conceptuelle avec les états cérébraux réalisant ces rôles causaux. Or cette conséquence entre directement conflit, selon lui, avec la manière dont nous concevons naïvement les états phénoménaux, et en particulier avec ce qu'il nomme la thèse de l'identification :

> (Thèse de l'identification) : avoir une expérience consciente E dotée d'un caractère phénoménal C implique de pouvoir connaître la nature intrinsèque de C. Pour le dire autrement : le fait d'avoir l'expérience consciente E révèle l'essence de C.

L'idée de Lewis est très simple : selon lui, la psychologie populaire de la douleur implique qu'il suffit d'avoir mal, ou en tout cas d'avoir eu mal, pour savoir tout ce qu'il y a à savoir sur la nature profonde de la douleur – et il en va de même de tous les autres états conscients. La thèse de l'identification est solidaire d'une conception mettant la perspective en première personne au premier plan dans l'acquisition de connaissances phénoménales : identifier, en première personne, un état phénoménal par la propriété phénoménale qu'il instancie permet, selon cette conception, de savoir tout ce qu'il y a à savoir sur l'état en question.

Il est clair que la thèse de l'identification est incompatible avec l'approche réductionniste de Lewis. En fait, cette thèse entre en conflit avec toutes les formes de physicalisme. Quelle que soit exactement la nature des propriétés phénoménales – qu'elles soient identiques à des propriétés cérébrales, ou à des propriétés fonctionnelles, ou encore à toute autre propriété physique que l'on pourrait concevoir – il n'est guère plausible de soutenir que l'introspection nous révèle cette nature. Comme le souligne Lewis : « Faire des découvertes en neurophysiologie n'est pas si facile ! » (Lewis, 1995, in Lewis, 1999, p. 329).

L'analyste se trouve donc placé face à un dilemme. Il peut ou bien intégrer la thèse de l'identification à la théorie naïve de l'esprit, et donc considérer que le fait que l'essence des propriétés phénoménales se révèle dans l'introspection appartient essentiellement à la façon dont nous les concevons. Mais alors, il sera impossible de trouver des propriétés

naturelles susceptibles de jouer le rôle attribué aux propriétés phénoménales par notre théorie naïve. Ou alors il doit considérer que la thèse de l'identification n'est pas essentielle à la manière dont nous concevons les propriétés phénoménales. Si l'on accepte cette seconde branche du dilemme, on devra reconnaître qu'il n'existe aucune propriété naturelle méritant exactement le titre de « propriété phénoménale », mais qu'il existe cependant des propriétés méritant presque ce titre : des propriétés qui certes ne satisfont pas la thèse de l'identification, mais qui satisfont néanmoins toutes les autres conditions que notre théorie naïve impose aux états phénoménaux. C'est l'option que Lewis (1995, in Lewis, 1999, p. 329) préconise : se contenter des « candidats imparfaits, mais acceptables (imperfect but good enough deservers) » susceptibles de correspondre à notre vocabulaire phénoménal.

Quelques précisions doivent, me semble-t-il, être apportées sur la thèse de l'identification. La première, c'est qu'il n'est pas complètement clair que quiconque ait réellement soutenu cette thèse, et surtout pas la psychologie populaire. Les concepts d'essence ou de nature sont des concepts philosophiques techniques, et il ne me semble pas plausible de considérer que notre théorie naïve de l'esprit implique quelque engagement que ce soit vis-à-vis de la connaissance de l'essence des propriétés phénoménales.

Mais ce n'est pas le plus important, et j'en viens à ma seconde remarque. Il semble que Lewis propose, au travers de la thèse de l'identification, une interprétation très forte – sans doute même trop forte – de la manière particulière dont la psychologie populaire conçoit la connaissance phénoménale. Il s'agit, au travers de cette thèse, de donner un sens à l'idée, commune au moins en philosophie de l'esprit, qu'il y aurait une dissymétrie entre les connaissances des propriétés phénoménales acquises « en première personne », et les connaissances acquises « en troisième personne »[34]. Or il semble qu'on puisse reconnaître cette dissymétrie sans adopter pour autant la thèse de l'identification. Il est ainsi plausible de soutenir, à la suite par exemple de Christopher Peacocke (1999), qu'il existe un lien a priori entre le concept d'expérience et certains jugements d'observation[35]. Prenons l'exemple des expériences visuelles de

[34] Lewis renvoie, pour justifier sa formulation de la thèse de l'identité, à un article d'Adams dans lequel celui-ci parle d'un mode « en première personne » d'identification des propriétés phénoménales ; cf. Lewis (1995 in Lewis, 1999, p. 328).

[35] Pour une étude du lien conceptuel qui semble exister entre les pensées démonstratives et les auto-attributions d'expériences, voir aussi Ludwig (2005).

couleurs. Il semble qu'un agent susceptible de juger (6) sur la base d'une expérience visuelle puisse du même coup juger (7) de façon justifiée :

(6) Ceci est rouge.
(7) J'ai maintenant une expérience de rouge.

On peut par ailleurs noter que la connaissance exprimée par (7) est, dans ce cas précis, infaillible. Nous pourrions donc admettre que la psychologie populaire des sensations de couleur inclut, au moins tacitement, la clause suivante :

(8) Un agent autorisé à juger « ceci est de couleur P » sur la base de l'observation est par là même autorisé à s'auto-attribuer une expérience du même contenu que son jugement – une expérience de la couleur P.

Le but de la présente discussion n'est pas de défendre une telle hypothèse, mais simplement de faire remarquer que si elle était correcte, on pourrait parfaitement comprendre la dissymétrie entre la première et la troisième personne dans le domaine phénoménal, sans accepter pour autant la thèse de l'identification. La clause (8) n'implique pas, en effet, que nous ayons accès à l'essence de notre sensation lorsque nous sentons, mais simplement que nous possédons, en tant qu'agents, un privilège épistémique concernant l'auto-attribution des expériences que nous sommes en train d'éprouver. L'existence d'un tel privilège ne pose pas de difficulté particulière pour un physicaliste. Le fait qu'un agent puisse produire un jugement d'observation implique qu'il possède l'expérience correspondant à ce jugement ; si l'on imagine qu'il existe un mécanisme qui puisse lier de façon a priori le jugement d'observation et l'auto-attribution correspondante, l'auto-attribution sera infaillible, puisque produite précisément par l'occurrence de l'expérience du type attribué. Mais cela n'implique à aucun moment que le système cognitif ait un accès direct à l'expérience en question, ni qu'il soit capable de l'identifier par l'une ou l'autre de ses propriétés essentielles.

On peut s'étonner que Lewis considère la thèse de l'identification comme étant au coeur de la conception naïve des propriétés phénoménales. La conviction, sans doute largement partagée dans le cadre de cette conception naïve, qu'il existe un privilège de la première personne vis-à-vis des informations portant sur le domaine phénoménal nous est en effet apparue logiquement indépendante de la thèse de l'identification. Comme

nous allons le voir, Lewis adopte une position déflationniste vis-à-vis de la connaissance phénoménale : il considère que c'est une connaissance absolument comme les autres, si l'on excepte certains savoir-faire spécifiques qui lui sont associés.

B. L'argument de la connaissance et l'hypothèse du savoir-faire

Je l'ai signalé en introduction : l'argument de la connaissance soulève une importante difficulté pour l'approche réductionniste de Lewis. Selon cette approche, l'analyse fonctionnelle des expressions phénoménales doit permettre d'identifier les dénotations de ces expressions dans le vocabulaire de la physique. Cela revient à soutenir que toutes les vérités phénoménales – par exemple, les vérités concernant les propriétés phénoménales de la sensation de rouge – peuvent être déduites logiquement à partir de l'ensemble de toutes les vérités physiques. Or, selon l'argument de la connaissance, Mary apprend quelque chose lorsqu'elle sort de son environnement en noir et blanc et qu'elle découvre l'effet que cela fait de voir du rouge. Quel sens donner à cette intuition, dès lors que l'on considère qu'absolument toutes les vérités sont déjà connues par Mary dans son environnement confiné ?

La réponse de Lewis (1988) tient en une distinction : on peut en effet reconnaître l'existence d'une connaissance non-propositionnelle, relevant du savoir-faire plutôt que du savoir que[36]. Illustrons ce point à l'aide d'une analogie. Mary connaît tous les faits physiques. Elle peut donc décrire avec exactitude la séquence de mouvements qu'il convient d'accomplir pour nager le crawl de la manière la plus efficace possible. Par exemple, elle sait que dans la phase dite de « prise d'eau », il convient de garder le coude le plus haut possible. Mais rien de tout cela n'implique qu'elle puisse nager le crawl de façon efficace. Pour parvenir à garder le coude haut lors de la phase de prise d'eau, il faut en effet s'entraîner de nombreuses années, si possible dès l'enfance. La différence pertinente entre un crawleur efficace et un crawleur inefficace n'est donc le plus souvent pas une différence épistémique, mais bien une différence pratique : le premier possède un savoir-faire qui manque au second.

Considérons maintenant le cas de la sensation de rouge, qui est selon Lewis tout à fait analogue. En raison de sa connaissance totale du monde physique, et de son application, supposons-le, minutieuse de la méthode de

[36] L'idée centrale de Lewis est également défendue dans plusieurs publications de Lawrence Nemirow (1980, 1990 et 2007).

réduction par fonctionnalisation, Mary sait tout ce qu'il y a à savoir sur cette sensation. Elle peut identifier « ce que cela fait de voir du rouge », au sens où elle sait exactement quelle est la propriété physique qui joue le rôle fonctionnel que nous associons à l'expérience consciente du rouge. Elle connaît également toutes les caractéristiques de cette propriété, y compris celles de ses caractéristiques que nous nommons « phénoménales ». Par exemple, elle sait que la sensation de rouge « ressemble » plus à celle de rose plus qu'à celle de vert, au moins au sens où elle est capable de prédire et d'expliquer les jugements de ressemblance entre sensations que des locuteurs normaux seraient susceptibles d'exprimer. Pour un physicaliste comme Lewis, les propriétés phénoménales d'un état comme la sensation de rouge sont toutes identiques à des propriétés cérébrales. Puisque Mary possède une connaissance totale du monde physique, il est donc également inévitable qu'elle possède une connaissance totale des propriétés phénoménales de l'état en question.

Pourtant, il y a une chose qui différencie Mary dans son environnement confiné de Mary après qu'elle a pris contact avec les couleurs : la seconde, puisqu'elle a eu une expérience de rouge, a instancié la propriété cérébrale correspondant à la sensation de rouge. Selon Lewis, du simple fait qu'elle ait instancié cette propriété découle une nouveauté, qui se situe sur le plan pratique du savoir-faire : Mary peut désormais imaginer une (nouvelle) sensation de rouge, elle peut évoquer son souvenir épisodique de la sensation de rouge, elle peut enfin reconnaître de nouvelles sensations comme des sensations de rouge. Pour cette raison, on peut considérer selon lui que Mary apprend bien quelque chose de nouveau ; mais cet apprentissage ne correspond pas à la découverte d'un nouveau fait, il s'agit de l'acquisition d'une capacité pratique (ability), une capacité à se souvenir, à imaginer, et à reconnaître les expériences.

L'argument qu'avance Lewis en faveur de l'hypothèse des savoir-faire est simple : la seule alternative possible à cette hypothèse susceptible d'expliquer l'intuition selon laquelle Mary apprend quelque chose consiste à soutenir qu'il existe des informations spécifiquement phénoménales ; mais cette alternative n'est pas compatible avec le physicalisme, qui est bien justifié par ailleurs. Nous avons donc d'excellentes raisons d'adhérer à l'hypothèse des savoir-faire[37].

---

[37] Selon Lewis, (1988) on peut identifier la connaissance d'un agent à un moment donné à un ensemble de mondes possibles ; cf. aussi sur ce point Lewis (1986). Apprendre une nouvelle information revient toujours, selon lui, à se trouver dans une position épistémique permettant d'éliminer certaines des possibilités de cet ensemble. Soutenir que Mary acquiert une nouvelle information spécifiquement phénoménale reviendrait donc à soutenir

La suggestion de Lewis paraît plausible. Il semble indéniable que l'expérience consciente soit une condition au moins nécessaire, et sans doute souvent suffisante, de l'acquisition de certaines capacités cognitives spécifiques. Ainsi ne voit-on pas comment un agent qui n'aurait jamais fait l'expérience consciente de la couleur rouge pourrait se souvenir d'une expérience de rouge, former une image mentale de la couleur rouge, ou distinguer visuellement des fauteuils rouges de fauteuils verts. On peut donc avancer la thèse suivante :

> (Contrainte de l'expérience pour les savoir-faire cognitifs) : certaines capacités cognitives d'ordre pratique ne sont effectivement acquises, normalement, qu'à condition d'avoir vécu certaines expériences conscientes[38].

Admettons que cette thèse soit vraie. On peut aller un peu plus loin que Lewis et s'interroger sur la nature des capacités cognitives concernées. On ne peut en effet inférer, du fait que X constitue une capacité cognitive d'ordre pratique, à la conclusion selon laquelle l'usage de X ne jouerait pas de rôle essentiellement épistémique. Bien au contraire, même. La possession d'un concept est en effet toujours étroitement liée à la maîtrise de certains savoir-faire cognitifs, et ce quel que soit le concept considéré. Ainsi, on n'attribuera à un agent la maîtrise du concept de conjonction qu'à condition que cet agent possède un certain savoir-faire inférentiel : il faut qu'il soit capable de voir que l'acceptation d'une conjonction de type « P et Q » constitue une bonne raison d'inférer P (ou Q), et qu'inversement l'acceptation conjointe de P et de Q constitue une bonne raison d'inférer « P et Q ». Pour prendre un exemple très différent, la possession du concept de voiture suppose certainement une certaine capacité cognitive pratique, consistant à pouvoir reconnaître les voitures, à pouvoir les distinguer des

---

que Mary peut accéder à une méthode permettant d'éliminer des mondes possibles qu'elle ne pouvait pas éliminer dans son environnement en noir et blanc. Mais si l'on suppose que les propriétés phénoménales peuvent être identifiées à des propriétés neuronales à l'aide de la méthode de réduction par fonctionnalisation, on ne voit absolument pas quels mondes possibles nouveaux pourraient être éliminés par une telle méthode.

[38] Cette contrainte exprime selon Lewis un lien contingent, et non nécessaire, entre l'expérience vécue et la possession des savoir-faire relevant de la connaissance de «l'effet que cela fait». Lewis souligne en effet que la possession de ces savoir-faire pourrait être causée de façon anormale, par exemple à la suite d'une manipulation neuronale, ou tout simplement par magie. Il n'est donc pas inconcevable qu'un agent sache «ce que cela fait» de voir du rouge, au sens où cet agent possède les savoir-faire correspondants, sans avoir pour autant jamais vécu l'expérience visuelle du rouge. Cf Lewis (1988, p. 264).

motos ou des autobus, à pouvoir imaginer des voitures, se souvenir des voitures, etc.

Pour le dire d'un mot, la question importante à trancher est la suivante : les savoir-faire cognitifs dont Lewis a montré de façon convaincante qu'ils sont liés aux expériences conscientes sont-ils, ou non, constitutifs de la possession de concepts ? En ce qui concerne l'expérience consciente des couleurs, Martine Nida-Rümelin (1996) a développé un argument qui me semble montrer de façon décisive que les capacités identifiées par Lewis sont de nature conceptuelle. Cet argument constitue une variation sur le thème de l'expérience de pensée de Jackson. Imaginons qu'au lieu de Mary, ce soit Marianna qui soit confinée dans une environnement en noir et blanc. Marianna est exactement dans la même situation épistémique que Mary : elle peut donc former des hypothèses sur les expériences visuelles des couleurs et sur leurs effets sur la vie cognitive des humains, mais ces hypothèses sont formulées en utilisant le vocabulaire des sciences de la nature. Marianna a ainsi constaté, en étudiant les faits physiques uniquement, mais sans avoir jamais elle-même eu d'expérience visuelle du bleu, que la plupart de ses semblables appréciaient la couleur bleue. Elle en a déduit que l'expérience de bleu était particulièrement agréable. Supposons par ailleurs que l'on montre à Marianna des échantillons des principales couleurs, mais sans lui indiquer les noms correspondant aux différents échantillons. Ces différentes expériences de couleurs vont lui permettre d'acquérir des savoir-faire cognitifs – par exemple, Marianna va devenir capable de distinguer ce que nous appelons, nous, le bleu, ce ce que nous appelons le rouge, et dont nous supposerons qu'elle les désigne mentalement respectivement comme « cette couleur$_{bl}$ » et « cette couleur$_{ro}$ ». Elle va également devenir capable de reconnaître des présentations du bleu comme des présentations de la même couleur que celle qu'elle a déjà vue auparavant ; de même pour le rouge, le vert, etc. Elle va enfin devenir capable, comme le soulignait Lewis, de se remémorer les différentes expériences associées aux différents échantillons de couleur, d'évoquer ces expériences dans des épisodes imaginatifs, etc.

Ces nouvelles capacités, acquises par Marianna, relèvent-elles de compétences conceptuelles ? A-t-elle acquis, pour le dire autrement, de nouveaux concepts en voyant les échantillons de couleurs de façon répétée, même si elle n'a pas appris à nommer ces échantillons en leur associant les noms de couleurs qu'elle connaît par ailleurs ? Pour tenter de répondre à ces questions, considérons ces deux jugements de Marianna :

> (9) L'expérience que l'on a en regardant le ciel est une expérience du bleu, mais pas une expérience de cette couleur$_{bl}$.
> (10) L'expérience que l'on a en regardant le ciel est une expérience de cette couleur$_{ro}$.

Marianna croit que l'énoncé (9) est vrai en raison de ses connaissances physiques sur le monde. Par ailleurs, elle croit que (10) est vrai car elle a trouvé la vision de la couleur rouge, à laquelle elle réfère comme à « cette couleur$_{ro}$ » particulièrement agréable. Comme elle sait que la plupart de ses semblables trouvent l'expérience du bleu agréable, elle en a inféré à tort que le ciel était de cette couleur$_{ro}$. Les deux hypothèses de Marianna sont raisonnables. Elles sont pourtant contradictoires : comme nous savons que « cette couleur$_{ro}$ » désigne le rouge et non le bleu, nous savons – contrairement à Marianna – que (9) et (10) ne peuvent être tous les deux vrais. Par ailleurs Marianna possède des capacités cognitives qui lui permettent de former des jugements contradictoires sur les même couleurs – ainsi, elle juge, à propos du bleu, que c'est et que ce n'est pas la couleur du ciel. La meilleure explication du fait qu'elle puisse former rationnellement des pensées contradictoires à propos d'une même entité réside dans l'hypothèse suivante : les capacités que Marianna a acquises en percevant les échantillons de couleurs sont des capacités conceptuelles, qui lui permettent de se représenter les couleurs d'une manière différente de la façon dont elle se les représentait dans le cadre de sa connaissance physicaliste du monde. On sait en effet, au moins depuis Frege, qu'un agent rationnel peut former des pensées contradictoires à propos d'une même entité, s'il possède des concepts distincts de cette entité, et s'il ne sait pas que ces concepts ont la même dénotation.

Il semble donc plausible de considérer que Marianna a acquis, à la suite de ses observations répétées de l'échantillon de rouge, un concept recognitionnel de la couleur rouge ainsi qu'un concept dénotant l'expérience de la couleur rouge. Elle ignore pourtant encore l'effet que cela fait de voir du bleu, puisqu'elle croit, à tort, que cette couleur$_{ro}$ = bleu. En sortant de son environnement confiné, elle découvrira que cette couleur$_{ro}$= rouge, et comprendra du coup ce que cela fait que de voir du rouge.

Si l'on accepte cette hypothèse, que l'on appellera « l'hypothèse des concepts phénoménaux », l'on devra formuler une seconde version de la contrainte de l'expérience :

> (Contrainte de l'expérience pour les concepts) Certains concepts dénotant des propriétés phénoménales ne sont effectivement acquis,

> normalement, qu'à condition d'avoir vécu certaines expériences conscientes. Nous les nommerons des concepts phénoménaux.

Dans le cadre de l'hypothèse des concepts phénoménaux, le problème initial posé par Jackson a une solution simple. Les inférences que l'on peut effectuer, à partir d'une base de connaissances factuelles, ne dépendent pas simplement des faits qui sont connus, mais aussi de la façon dont ces faits sont représentés conceptuellement, et donc de la façon dont ils sont décrits. Considérons qu'un cinéphile, Emile, récapitule ses connaissances sur Eric Rohmer. Sa base de connaissance sur ce cinéaste comporte les vérités suivantes : « Eric Rohmer a d'abord publié des romans », « Eric Rohmer a mis en scène les Contes moraux», « Eric Rohmer a été rédacteur en chef des Cahiers du Cinéma ». De ces vérités, il peut inférer par exemple que le metteur en scène des Contes moraux a été rédacteur en chef des Cahiers du Cinéma. Il se trouve par ailleurs qu'Eric Rohmer n'est autre que Maurice Schérer – « Maurice Schérer » est le nom d'état civil de Rohmer, mais Emile l'ignore. A vrai dire, Emile n'a jamais entendu parler de quiconque nommé « Maurice Schérer », et il ne possède donc tout simplement pas le concept associé à ce nom propre. Dans cette situation, il est clair qu'Emile ne peut pas inférer, à partir de sa base de connaissances, les vérités formulées à l'aide du nom « Maurice Schérer ». De ce fait, il ne sait pas, en un sens, que Maurice Schérer a écrit aux Cahiers du Cinéma ; pourtant, on ne peut pas dire en un autre sens que le fait que Maurice Schérer ait écrit aux Cahiers du Cinéma soit inconnu d'Emile, puisqu'il connaît au moins ce fait sous une de ses descriptions. Tout cela est parfaitement trivial. Mais ces trivialités montrent qu'il est tout à fait possible d'inférer un certain fait sous l'une de ses descriptions, sans pouvoir l'inférer sous une autre de ses descriptions. C'est ainsi que Marianna peut inférer que le bleu est la couleur du ciel, mais pas que cette couleur_{bl} est la couleur du ciel.

Or Mary, selon les données de l'expérience de pensée de Jackson, n'a jamais vécu d'expériences visuelles de couleurs. En vertu de la contrainte de l'expérience pour les concepts, elle ne peut donc pas avoir acquis de concepts phénoménaux, du moins pas de façon normale. Il en découle trivialement que certaines descriptions, sous des concepts phénoménaux, des faits qu'elle connaît par ailleurs ne pourront pas être inférées normalement à partir de l'ensemble de ses connaissances physiques. Mais, si l'on suit cette ligne de raisonnement, cela ne permet nullement de réfuter le physicalisme – même pas le physicalisme dans sa version la plus forte, tel qu'il est défendu par Lewis. Cela montre

simplement que comme certaines capacités cognitives de nature pratique dépendent, pour leur acquisition, du fait d'avoir eu des expériences, et que la possession de ces capacités est à son tour une condition nécessaire pour acquérir des concepts phénoménaux, un agent ne peut pas concevoir les vérités phénoménales à l'aide de concepts phénoménaux s'il n'a pas vécu les expériences correspondantes. Il peut cependant parfaitement concevoir ces vérités d'une autre façon, exactement de la même façon qu'Emile peut concevoir qu'Eric Rohmer ait été rédacteur en chef des Cahiers du Cinéma, même s'il ne peut concevoir que ç'ait été le cas de Maurice Schérer.

Faisons le point. Selon Lewis, les expériences conscientes nous permettent d'acquérir certaines capacités cognitives. Savoir l'effet que cela fait de voir du rouge, c'est pouvoir imaginer la couleur rouge, pouvoir la reconnaître, pouvoir la distinguer d'autres couleurs, etc. De la sorte, Lewis, (1995, in Lewis, 1999, p. 327) considère qu'un « concept des qualia – un concept acceptable pour un matérialiste – est le concept des propriétés des expériences susceptibles de causer les capacités (abilities) à reconnaître et à imaginer des expériences d'un même type ». L'hypothèse des savoir-faire n'implique pas, nous l'avons vu, que la connaissance phénoménale ne relève pas de la possession de concepts ; elle implique en revanche que ce qui est acquis par Mary n'est rien d'autre qu'un savoir-faire, fût-il de nature conceptuelle, et non une information spécifiquement phénoménale. Il n'est donc pas exclu que l'on puisse acquérir une connaissance nouvelle sur une expérience, en ayant cette expérience, et en utilisant l'introspection. Mais cette connaissance pourra toujours être acquise par d'autres méthodes que l'introspection, précisément parce que l'information sur laquelle elle porte n'est pas d'une nature non-physique.

C. L'hypothèse du savoir-faire et les concepts démonstratifs

L'hypothèse des savoir-faire a suscité de nombreuses discussions, souvent critiques. Il n'est pas question de discuter ici en détail toutes les objections soulevées dans la littérature philosophique récente[39]. Nombre de ces objections visent à établir que les trois sortes de capacités cognitives mentionnées par Lewis – la capacité à se souvenir, la capacité à imaginer et la capacité à reconnaître – ne sont ni conjointement suffisantes, ni individuellement nécessaires pour pouvoir attribuer à un agent une connaissance de « l'effet que cela fait »[40]. Mais il est facile de voir que de

[39] Pour une revue très complète de la littérature, voir Nida-Rümelin (2010).
[40] Voir par exemple Conee (1994), Alter (1998) et Raymont (1999).

telles objections passent à côté de l'essentiel de la question. Lewis prend certes ces trois capacités comme des exemples importants des savoir-faire associés à la connaissance phénoménale. Rien n'indique cependant qu'on ne puisse pas faire référence à d'autres capacités, et en particulier, comme je l'ai souligné, à des capacités conceptuelles, pour expliquer la connaissance phénoménale.

A cet égard, la discussion de Michael Tye (2000) me paraît particulièrement intéressante, car elle intègre justement cette possibilité. Tye commence par souligner, à juste titre, que dans ses exemples Lewis ne prend pas assez au sérieux la finesse du grain de la connaissance phénoménale. Cette connaissance porte, dans l'exemple central autour duquel tourne l'argument de Jackson, sur l'expérience visuelle du rouge. Il est clair qu'une telle expérience, répétée, produit les capacités mentionnées par Lewis. Mais l'on pourrait s'intéresser non à l'expérience du rouge, mais à l'expérience plus spécifique de la nuance particulière de rouge perçue par Mary – disons, le rouge_{17}. Nous sommes capables de distinguer entre des nuances très fines d'une même couleur lorsque nous les percevons, par exemple de distinguer visuellement entre le rouge_{17}, le rouge_{18} et le rouge_{19}. Ces distinctions peuvent, par ailleurs, être mobilisées dans des inférences. Ainsi, chez un marchant de peinture, nous pourrions faire des hypothèses portant sur l'effet que produirait ces trois nuances de rouge sur le mur d'un salon, et raisonner à partir de ces hypothèses. Nous formulerions ce genre d'hypothèses à l'aide de ce que j'appellerai des concepts démonstratifs expérientiels[41]. Voici un exemple d'utilisation de tels concepts :

> (12) Ce rouge_{17} est un rien trop brillant, mais ce rouge_{19} est un peu foncé.

Il est manifeste que si nous pouvons raisonner ainsi, nous pouvons aussi raisonner sur les caractéristiques des sensations occasionnées par le rouge_{17} ou par le rouge_{19}. Or, comme le souligne Tye, de telles connaissances phénoménales ne s'accompagnent d'aucune des caractéristiques mentionnées par Lewis : une fois le magasin de peinture

[41] Je préfère parler de « démonstratifs expérientiels » plutôt que de « démonstratifs de perception », car il semble possible de former de tels concepts sur la base d'une hallucination, c'est-à-dire sur la base d'une expérience visuelle n'ayant pas le statut d'une perception. Il doit être clair cependant que ces concepts démonstratifs ne dénotent pas des propriétés d'expérience – pour un physicaliste, des propriétés cérébrales, donc – mais des propriétés perceptives, comme le rouge, que l'on peut concevoir comme des propriétés objectives.

quitté, nous ne pourrions plus nous souvenir de la nuance rouge$_{17}$, ni imaginer cette nuance spécifique, ni non plus reconnaître ses occurrences.

Il est difficile, remarque Tye, de trouver des capacités cognitives que l'on pourrait associer à la connaissance de l'effet que cela fait de voir rouge$_{17}$ par opposition à rouge$_{19}$, si ce n'est en faisant précisément référence aux capacités spécifiques associées à la maîtrise du concept démonstratif rouge$_{17}$. D'où la conclusion de Tye :

> « Peut-être devrait-on identifier la connaissance de l'effet que cela fait non à l'ensemble des capacités mentionnées par Lewis – car toutes ces capacités pourraient manquer en présence de la connaissance de l'effet que cela fait –, mais plutôt à la capacité plus fondamentale d'appliquer un concept indexical au caractère phénoménal d'une expérience via l'introspection. » (Tye, 2000, p. 12).

Tye propose donc de modifier l'hypothèse des savoir-faire, ou du moins de la préciser, sous les deux aspects suivants :

(i) les savoir-faire nécessaires à la connaissance phénoménale sont conceptuels ;

(ii) plus précisément encore, il s'agit des savoir-faire que l'on doit posséder pour pouvoir appliquer «un concept indexical au caractère phénoménal d'une expérience via l'introspection».

Selon lui cependant, cette modification ne peut pas « sauver » l'hypothèse des savoir-faire, en raison de l'expérience de pensée suivante. Imaginons que, lorsqu'elle sort de son environnement en noir et blanc, Mary voit clairement une rose rouge d'une certaine nuance, mais qu'elle ait l'esprit occupé par un problème de mathématique, de sorte qu'elle ne prête aucune attention à la couleur rouge de la rose. Selon Tye, Mary a, dans cette situation, une expérience phénoménale du rouge de la rose – ce point pourrait être discuté, en raison complexe du rôle que joue l'attention dans l'expérience consciente, mais nous l'accorderons pour les besoins de la discussion. Elle a également toutes les capacités qui lui permettraient, si c'était nécessaire, d'appliquer un concept démonstratif à son expérience du rouge : « elle a certainement, écrit Tye (2000, p. 14), la capacité de pointer mentalement vers le caractère phénoménal de son expérience dans l'introspection, à l'aide d'un concept indexical ». Et pourtant, selon lui, elle ne sait pas ce que cela fait que de voir du rouge, puisqu'elle ne met pas effectivement en oeuvre cette capacité en appliquant le concept qu'elle pourrait acquérir – mais qu'elle n'a, dans l'expérience de pensée de Tye, pas réellement acquis – à son expérience.

Afin d'évaluer l'objection de Tye, je souhaite faire deux remarques sur son raisonnement. En premier lieu, il n'est pas nécessaire d'introduire des concepts indexicaux pointant « mentalement vers le caractère phénoménal de l'expérience dans l'introspection », comme le fait Tye, pour pouvoir décrire les capacités dont Mary a besoin pour « savoir l'effet que ça fait » de voir du rouge. Tout ce qu'il semble nécessaire d'attribuer à Mary, c'est d'une part une capacité à concevoir démonstrativement la couleur rouge, et d'autre part une maîtrise du concept général d'expérience consciente. Par ailleurs, il semble que Tye confonde deux capacités : (i) la capacité à former un certain concept et (ii) la capacité issue de la formation de ce concept. Dans l'expérience de pensée qu'il propose, Mary n'a en effet pas acquis de concept démonstratif du rouge, puisqu'elle n'a pas focalisé son attention sur cette couleur. Elle a certes la capacité d'acquérir ce concept, mais c'est une chose bien différente de la possession du concept : tant que le concept n'est pas formé, Mary ne possède en effet pas les capacités inférentielles associées, comme la capacité à former des hypothèses sur la couleur rouge.

Il faut donc s'interroger sur le point suivant : supposons que Mary ait bel et bien formé un concept démonstratif de la couleur rouge, qu'elle possède donc toutes les capacités inférentielles associées à la possession de ce concept, qu'elle possède par ailleurs le concept général d'expérience, et qu'elle puisse donc désigner l'aspect qualitatif de son expérience comme « l'effet que cela me fait de faire l'expérience de cette couleur ». Doit-on alors lui attribuer la connaissance de « l'effet que cela fait de voir du rouge » ? Il me semble difficile de voir ce qui manquerait à Mary pour qu'on puisse faire cette attribution : en possession d'un concept démonstratif du rouge et du concept d'expérience, elle peut en effet former des hypothèses sur son expérience de « cette couleur », comparer cette expérience à d'autres expériences, raisonner à propos de son expérience, etc.

La conclusion à laquelle nous parvenons, au terme de notre discussion de l'hypothèse des savoir-faire, est nuancée. David Lewis nous semble avoir raison de soutenir que l'attribution d'une connaissance « de l'effet que cela fait de faire l'expérience E » à un agent repose entièrement sur l'attribution de certaines capacités cognitives à un agent. Il nous est cependant également apparu que ces capacités sont d'ordre conceptuel : comme le montrent les discussions de Nida-Rümelin et de Tye, l'attribution d'une connaissance phénoménale semble être étroitement liée à la possession de certains concepts démonstratifs, au moins en ce qui concerne le domaine des expériences de couleurs. Si c'est bien le cas, l'hypothèse des savoir-faire doit être précisée, de sorte que soit établie la manière dont ces

savoir-faire cognitifs peuvent être mobilisés dans la production de savoir-que propositionnels. Ainsi peut-on admettre en suivant Lewis que le savoir qu'acquiert Mary n'est rien d'autre qu'un savoir-faire : essentiellement, la capacité cognitive issue de l'acquisition du concept démonstratif « cette couleur », désignant la couleur rouge au moment où Mary la voit. Mais il faut ajouter que cette capacité cognitive permettra à Mary de formuler des jugements : par exemple, de comparer l'effet que cela fait de voir cette couleur-ci (le rouge) avec l'effet que cela fait de voir cette couleur-là (le rose). Bien entendu, comme Lewis l'a, me semble-t-il, bien établi, rien dans l'argument de Jackson ne permet de penser que les vérités correspondant à ces jugements, lorsque ceux-ci s'avèrent vrais, ne puissent pas être formulées à l'aide des concepts des sciences naturelles. Ainsi, si nous jugeons que l'expérience du rouge ressemble à l'expérience du rose, rien dans le raisonnement de Jackson ne montre qu'il est impossible de décrire cette vérité dans le vocabulaire des neurosciences.

V. Conclusion

Lewis n'a jamais écrit, à strictement parler, sur les concepts phénoménaux, et on présente parfois l'hypothèse des savoir-faire comme une alternative à la stratégie des concepts phénoménaux. J'ai insisté sur le fait que ces hypothèses pouvaient être conçues comme étant complémentaires plutôt que concurrentes, puisqu'il est plausible que la possession de concepts repose sur celle de certaines capacités cognitives, et donc sur des savoir-faire au sens de Lewis. Plus précisément, notre discussion a montré que ces capacités semblaient liées, au moins dans le cas des expériences de couleur, à la possession de concepts démonstratifs. Plusieurs questions restent ouvertes, que je ne peux que mentionner dans cet article. Quel rôle les concepts démonstratifs jouent-ils dans notre psychologie populaire de l'expérience consciente ? Comment nous permettent-ils d'attribuer des expériences à autrui, ou de nous en auto-attribuer, de façon sans doute spécifique et privilégiée, à nous-mêmes ? Je l'ai souligné : Lewis ne dit rien du fait qu'un agent semble avoir un accès privilégié à ses propres états phénoménaux dans l'introspection. La stratégie de réduction par fonctionnalisation, appliquée au domaine phénoménal, devrait partir de la considération de tous les truismes que nous acceptons, de façon commune, à propos des expériences conscientes – y compris donc de celle des truismes faisant intervenir des occurrences de concepts démonstratifs. Elle devrait intégrer également une prise en compte de notre

psychophysique naïve – un point qui est abordé par Lewis dans (Lewis, 1997) –, ainsi que de notre théorie naïve de la connaissance introspective.

Bibliographie

Alter, T., « A Limited Defense of the Knowledge Argument », Philosophical Studies 90 (1) : 35–56, 1998,.

Alter, T., and Walter, S. (éds.), Phenomenal Concepts and Phenomenal Knowledge. New Essays on Consciousness and Physicalism, Oxford, Oxford University Press, 2007.

Braddon-Mitchell, D., and Nola R. (éds.), Conceptual Analysis and Philosophical Naturalism, Cambridge, Mass., MIT Press, 2009.

Chalmers, D. & Jackson, F., « Conceptual analysis and reductive explanation », Philosophical Review 110 : 315-61, 2001.

Carnap, R., « Empiricism, Semantics and Ontology », Revue Internationale de Philosophie, 4 : 20-40, 1950. Trad. fr. Ph. de Rouilhan et Fr. Rivenc, «Empirisme, sémantique et ontologie», in Signification et nécessité, Paris, Gallimard, 1997.

Carnap, R., Philosophical foundations of physics, New-York: Basic Books, 1966. Trad. fr. J. M. Luccioni et A. Soulez : Les fondements philosophiques de la physique, Paris, Armand Collin, 1973.

Conee, E., « Phenomenal Knowledge », Australasian Journal of Philosophy 72 : 136–150, 1994.

Daly, C., An introduction to philosophical methods, Perterborough, Broadview Press, 2010.

Demopoulos, W and Friedman M., « Critical notice: Bertrand Russell's The Analysis of Matter: its historical context and contemporary interest», Philosophy of Science 52 : 621-639, 1985.

Demopoulos, W., « Carnap on the rational reconstruction of scientific theories », in The Cambridge Companion to Carnap, M. Friedman and R. Creath (éds.), Cambridge, Cambridge University Press, 2007.

Esfeld, M. & Sachse, C., «Theory Reduction by Means of Functional Sub-types», International Studies in the Philosophy of Science, 21 : 1–17, 2007.

Fodor, Jerry, « Special Sciences: Or the Disunity of Science as a Working Hypothesis », Synthese, 28 : 97-115, 1974.

Galinon, H., « Les termes théoriques de Carnap à Lewis », Philonsorbonne 2 : 31-45, 2009.

Hall, N., « David Lewis's Metaphysics », *The Stanford Encyclopedia of Philosophy* (Fall 2012 Edition), Edward N. Zalta (éd.), URL = <http://plato.stanford.edu/archives/fall2012/entries/lewis-metaphysics/>.

Jackson, F., « Epiphenomenal Qualia », Philosophical Quarterly 32 : 127–136, 1982.

Ketland, J., « Empirical adequacy and ramsification, II », in Reduction, abstraction, analysis, A. Hieke and H. Leitgeb (eds.), Lancaster, Gazelle books, 29-46, 2009.

Kim, J., Mind in a physical world, an essay on the mind-body problem and mental causation, Cambridge (Mass.), MIT Press. Trad. fr. F. Athané et E. Guinet, Paris, Syllepse, 1998.

Lewis, D., « An Argument for the Identity Theory, » Journal of Philosophy, 63 : 17–25, 1966.

Lewis, D., « How to Define Theoretical Terms, » Journal of Philosophy, 67 : 427–446, 1970.

Lewis, D., « Psychophysical and Theoretical Identifications, » Australasian Journal of Philosophy, 50 : 249–258, 1972.

Lewis, D., « Mad Pain and Martian Pain, » in Ned Block (éd.), Readings in Philosophy of Psychology, Volume I, Cambridge, Harvard University Press, p. 216–32, 1980.

Lewis, D., « New Work For a Theory of Universals, » Australasian Journal of Philosophy, 61 : 343–377, 1983.

Lewis, D., Philosophical Papers, Volume I, Oxford, Oxford University Press, 1983a.

Lewis, D., « Putnam's Paradox », Australasian Journal of Philosophy, 62: 221–236, 1984.

Lewis, D., On the Plurality of Worlds, Oxford, Blackwell Publishers, 1986

Lewis, D., Philosophical Papers, Volume II, Oxford, Oxford University Press, 1986b.

Lewis, D., « What Experience Teaches, » Proceedings of the Russellian Society, University of Sydney, 13 : 29–57, 1988.

Lewis, D., « Reduction of Mind, » in Samuel Guttenplan (éd.), A Companion to Philosophy of Mind, Malden, Mass., Blackwell Publishers, 1994.

Lewis, D., « Should a Materialist Believe in Qualia?, » Australasian Journal of Philosophy, 73 : 140–144, 1995.

Lewis, D., « Naming the Colours, » Australasian Journal of Philosophy, 75 : 325–342, 1997.

Lewis, D., Papers in Metaphysics and Epistemology, Cambridge, Cambridge University Press, 1999.

Ludlow, P., Nagasawa, Y. & Stoljar, D. (éds.), There is something about Mary: essays on phenomenal consciousness and Frank Jackson's knowledge argument, Cambridge, MA, MIT Press, 2005.

Nemirow, L., « Review of Thomas Nagel, Mortal Questions, » Philosophical Review 89 : 473–477, 1980.

Nemirow, L., « Physicalism and the Cognitive Role of Acquaintance, » in Lycan (éd), 1990, p. 490–499.

Nemirow, L., « So this is what it's like: a defense of the ability hypothesis, » in T. Alter & S. Walter (éds), 2007, p. 32–51.

Nida-Rümelin, M., « What Mary couldn't know », in Thomas Metzinger (éd.), Phenomenal Consciousness, Schoenigh: Paderborn, 1996.

Nida-Rümelin, Martine, 2010,"Qualia: The Knowledge Argument », The Stanford Encyclopedia of Philosophy (Summer 2010 Edition), E. N. Zalta (éd.),<http://plato.stanford.edu/archives/sum2010/entries/qualia-knowledge/>.

Nolan, D., David Lewis, Chesham, Acumen Publishing, 2005.

Nolan, D., « Platitudes and metaphysics », in Braddon-Mitchell, D.& Nola R, (éds.), 2009, p. 267-300.

Peacocke, C., Being known, Oxford, Oxford University Press, 1999.

Polger, T., Natural Minds, Cambridge, MA., MIT Press, 2004.

Putnam, H., « Psychological Predicates. » in W.H. Capitan and D.D. Merrill (eds.), Art, Mind, and Religion. Pittsburgh, University of Pittsburgh Press, 37-48, 1967.

Putnam, H., Reason, Truth and History, Cambridge, Cambridge University Press, 1981.

Ramsey, F. P., « Theories » (1929), dans ses Philosophical papers, D. H. Mellor (éd.), New York, Cambridge University Press, p. 112-139, 1990.

Raymont, P., « The Know-How Response to Jackson's Knowledge Argument », Journal of Philosophical Research 24 : 113–126, 1999.

Sider, T., Writing the Book of the World, Oxford, Oxford Clarendon Press, 2011.

Smart, J. J. C., « Sensations and Brain Processes, » Philosophical Review, 68 : 141–156, 1959.

Tye, M., « Knowing What it is Like: The Ability Hypothesis and the Knowledge Argument, » in M. Tye, Consciousness, Color, and Content, Cambridge, MA, MIT Press, 2000.

Uebel, T., « Carnap's ramseyfications defended», European Journal for the Philosophy of Science, 1 : 71-87, 2011.