Intellectica, 1995/2, 21, pp. 115-134

Jacques DUBUCS*

Les états mentaux sur la place publique

Notre perplexité est à ce point profonde que nous ne sommes même pas capables de dire ce que nous attendons au juste d'une théorie. (J. Fodor, *Representations*)

Mental states in sight of everybody

Abstract : The meaning of the psychological predicates has to be communicable, thus it cannot be defined by reference to private mental experiences. But these predicates should be only ascribed to creatures who are able to ascribe them to themselves on the ground of their own qualitative experience. The paper deals with the ways of resolving this contradiction.

1. Être ou ne pas être soi-même dans un certain état mental est à la rigueur une différence que l'introspection devrait suffire à décider dans tous les cas. Mais s'accorder avec les autres sur l'utilisation correcte du vocabulaire des états psychiques est une tout autre affaire, qui demande que l'on se rapporte exclusivement à des processus qui se déroulent sous les yeux de tous : en psychologie comme ailleurs, la totalité de ce qui détermine la signification des mots doit être *visible*, et non pas dépendre de ce qui n'est en principe accessible qu'à un seul. Bien que les états mentaux soient subjectifs et personnels, et que l'on ne puisse, en toute rigueur, éprouver la soif ou la douleur d'un autre, la définition même des mots comme "soif" ou "douleur" ne doit donc faire aucune référence essentielle à des données irrémédiablement privées et incommunicables, et ne saurait mettre en jeu que le comportement public, verbal ou non verbal, par lequel s'exprime l'état considéré. Naturellement, ce comportement n'est pas toujours actuellement réalisé : on peut être in-

* Jacques Dubucs, Institut d'Histoire et de Philosophie des Sciences et des Techniques, CNRS-URA 1079, 13 rue du Four, 75006 Paris. Tel (33) 1.43.54.60.36. Télécopie (33) 1.44.07.16.49.

quiet sans le montrer, et croire quelque chose sans le dire. Aussi doit-on plutôt définir l'état mental par la pure et simple *disposition* à se comporter de telle et telle manière, c'est-à-dire par la propension à adopter effectivement tel ou tel comportement lorsque telle ou telle condition est réalisée : la soif est l'état dans lequel vous vous trouvez si vous buvez un verre d'eau lorsqu'on vous en présente un, si vous répondez "Oui" lorsqu'on vous demande "Avez-vous soif ?", si vous vous détournez lorsqu'on vous propose de l'anchoïade, etc. Telle est la façon dont le behaviorisme entend mettre les états mentaux sur la place publique.

Une difficulté se présente ici : à strictement parler, les stimuli sensoriels et les réponses comportementales en termes desquels l'état mental pourrait être ainsi défini doivent être pris dans une acception tellement large que leur observabilité et leur caractère public en deviennent extrêmement douteux. Car ce qui est invoqué dans une définition de ce genre n'est pas la réalité physique de l'objet dont l'image frappe la rétine, ni celle des gestes effectués, ni celle des sons entendus et émis, mais l'objet *en tant qu'*il est appréhendé comme représentant d'une certaine catégorie, les gestes *en tant qu'*ils sont destinés à accomplir une certaine intention, et les sons *en tant que* leur est reconnue une certaine signification : vous avez soif si, en présence d'une substance que vous *prenez* pour un liquide potable, vous faites le geste que vous *pensez* le plus approprié pour vous en saisir et l'absorber, et si vous émettez un son que vous *voulez* d'assentiment lorsque vous en entendez un que vous *interprétez* comme une question relative à votre envie de boire. En somme, chaque définition d'un état mental en termes de stimuli et de réponses observables introduit inévitablement un vaste ensemble d'autres états mentaux, au premier rang desquels des croyances relatives à la nature des stimuli et à l'adéquation des réponses qu'ils suscitent. L'idée d'une définition directe qui associerait à chaque terme du vocabulaire psychologique un ensemble particulier d'observables conduit donc à une impasse.

La difficulté que l'on vient de décrire peut être sommairement formulée de la manière suivante. Nous cherchons à définir les prédicats psychologiques $\psi_1, \ldots, \psi_n$ dans un langage Ω compris de tous, lequel ne permet de se référer à coup sûr qu'à des processus publiquement observables. Nous sommes, par hypothèse, capables de nous accorder sur une "théorie psychologique", c'est-à-dire sur un ensemble $T[\psi_1, \ldots, \psi_n]$ de phrases qui énoncent les relations entre ces prédicats psychologiques et les situations descriptibles dans le langage Ω. Mais chaque prédicat psychologique y intervient dans plus d'une phrase, et surtout chaque phrase

contient plus d'un prédicat de ce genre, en sorte que cette théorie ne contient ni ne permet de déduire aucune équivalence "pure", du type

$\forall x\,(\psi(x) \equiv \omega(x))$ (où ω est un prédicat définissable dans Ω),

que nous pourrions prendre pour définition explicite de ψ : bien qu'elle contienne tout ce que l'on pourrait désirer savoir sur l'expression publique de l'état ψ, nous ne devons donc pas attendre de cette théorie qu'elle nous livre une condition nécessaire et suffisante, exprimable à l'aide du seul vocabulaire observationnel, que devrait satisfaire un individu quelconque pour pouvoir être réputé dans l'état ψ.

À vrai dire, cette difficulté n'est nullement spécifique au vocabulaire des états psychiques, et on la rencontre chaque fois que l'on cherche à définir un ensemble de prédicats "théoriques" à partir d'une collection de phrases considérées comme correctes contenant les prédicats en question. La définition des concepts fondamentaux de la physique fournit un assez bon exemple. En effet, il n'existe pas non plus de condition nécessaire et suffisante, formulable en termes a-théoriques et purement observationnels, pour l'applicabilité d'un prédicat comme "être traversé par un courant d'intensité I" : à l'instar de la loi d'Ohm qui met en relation le potentiel, l'intensité et la résistance, chaque équation de la théorie de l'électricité se réfère simultanément à plusieurs notions théoriques, en sorte que la définition de chacune de ces notions, loin de pouvoir être donnée isolément, par une référence directe à des éléments observables, doit être extraite, en même temps que celle de toutes les autres, à partir de la théorie tout entière.

Les caractéristiques logiques des définitions de ce type apparaissent clairement lorsque le vocabulaire observationnel est réduit à zéro, c'est-à-dire lorsque le langage Ω supposé compris de tous est "purement logique", les seules constantes dotées d'une interprétation fixe étant alors les constantes logiques (connecteurs et quantificateurs). Les concepts à définir sont ici des notions mathématiques, et leurs définitions des définitions "implicites". Soit par exemple T[P,D,S] la théorie qui axiomatise la géométrie plane d'incidence, dont les notions primitives sont celles de point et de droite, ainsi que la relation d'incidence *être sur* entre points et droites. "Etre un point" est un prédicat qui ne peut évidemment pas faire l'objet d'une définition explicite isolée, par le biais d'une équivalence du type

$\forall x\,(P(x) \equiv \omega(x))$ (où ω est un prédicat dont l'écriture ne contiendrait ni P, ni D, ni S).

Ainsi que Frege [1] — qui pensait par ailleurs que l'on ne devrait pas procéder de la sorte pour définir les notions fondamentales de la géométrie — fut le premier à le remarquer, la raison essentielle de cette impossibilité tient à ce que les axiomes d'une telle théorie ne sont tout simplement pas des énoncés authentiques, capables de vérité ou de fausseté, mais de simples *formes d'énoncés*, dans laquelle les symboles P, D et S doivent être compris comme des *variables de prédicats* et non comme des constantes : T[P,D,S], identifiée par commodité à la conjonction de ses axiomes, est en fait une *relation de second ordre*, susceptible d'être ou non satisfaite par des prédicats ? et ? et par une relation binaire ? (dans cet ordre) définis sur un domaine quelconque. Si la théorie détermine une condition nécessaire et suffisante, ce n'est donc certainement pas celle que devrait remplir un objet quelconque pour être considéré comme un point, mais plutôt celle que devrait remplir une multi-relation < ? , ?, ? > constituée de prédicats d'arité appropriée définis sur un domaine quelconque pour que ses constituants méritent d'être respectivement nommés *être un point*, *être une droite* et *être sur*. En quoi consiste alors, si l'on y tient, le fait, pour un objet, d'être un point ? La réponse, à rigoureusement parler, ne saurait être que la suivante : dans l'existence d'une certaine structure **< D**, ? , ? , ? > telle que la multi-relation < ? , ?, ? > satisfasse la relation T (ou, autrement formulé, telle que T(? , ? , ?) soit vraie), et, dans ces conditions, dans la propriété qu'a l'objet en question d'être justement l'un des éléments de **D** qui satisfont le prédicat ? . En bref, être un point, c'est être, parmi les éléments (du domaine) d'une structure qui satisfait (on dit encore : qui *réalise*) les axiomes de la géométrie T, l'un de ceux qui font partie du sous-ensemble particulier par lequel y est interprété le prédicat *être un point* :

$$\forall x \, [\, P(x) \equiv \exists ? \, \exists ? \exists ? \; T(? \, , ?, ?) \; \& \; ? \; (x) \,]$$

L'analyse précédente se généralise sans difficulté au cas où le langage de la théorie comporte également des prédicats dont la signification est supposée "déjà connue", c'est-à-dire des prédicats (notamment "observationnels") dont l'interprétation est censée ne pas plus varier que celle des constantes logiques. Si $\psi_1, \ldots, \psi_n$ sont les nouveaux prédicats introduits par la théorie T, on caractérisera le fait pour un objet de satisfaire le prédicat ψ_i en disant

1) que T possède une réalisation (cette première partie de la caractérisation est "l'énoncé de Ramsey de la théorie")

---

[1] Dans sa lettre du 6 Janvier 1900 à Hilbert (Frege-Hilbert (1899-1900), pp. 232 sq).

2) que l'objet en question appartient à cette partie du domaine de la réalisation considérée qui réalise pour sa part le prédicat ψ_i :

(I) $\forall x [\psi_i(x) \equiv \exists\xi_1 ... \exists\xi_n T(\xi_1, ..., \xi_n) \& \xi_i(x)]$

En particulier, les prédicats qui figurent dans les théories psychologiques sur lesquelles nous sommes susceptibles de nous accorder doivent être définis de cette manière. Telle est en tout cas la conception défendue par le *fonctionnalisme*.

2. Si l'on croit qu'il va pleuvoir, alors on a tendance à prendre son parapluie pour sortir, mais seulement si l'on a le désir de rester au sec. On le prendra également si l'on déteste être mouillé, mais à la condition toutefois de croire que le temps est à la pluie. En caractérisant la croyance à la pluie imminente par la seule disposition à se mettre à couvert, le behaviorisme pèche donc par omission. Et par circularité s'il y ajoute, comme il le devrait, l'aversion pour l'humidité, puisque sa définition de ce dernier état devrait à son tour invoquer, outre la propension à s'abriter, la croyance elle-même à la pluie. Le fonctionnalisme est exempt de ces travers. Il ne réduit pas les états mentaux à des comportements, fussent-ils dispositionnels. Et la définition qu'il en propose n'entraîne non plus aucune aucune circularité. Car si la croyance à la pluie figure à la i-ème place dans une énumération quelconque des états mentaux, la définition de cette croyance résultera simplement de l'équivalence (I) ci-dessus, où $T[\psi_1, ... , \psi_n]$ est la théorie qui énonce les relations des états mentaux entre eux, ainsi que leurs relations aux données publiquement observables pertinentes (d'un côté, les événements susceptibles d'être perçus par le sujet, les *entrées* sensorielles si l'on veut, et de l'autre, son comportement manifeste, ses *sorties* motrices). Et cette équivalence ne mentionne expressément aucun état mental déterminé, si ce n'est la croyance ψ_i elle-même : la quantification existentielle a expurgé le *definiens* de tout le vocabulaire mentaliste. Au total, le fonctionnalisme parvient à faire un sort aux relations mutuelles des états mentaux sans rien supposer de connu, ni même de spécifique, dans le domaine qu'ils circonscrivent. Dire, par exemple, qu'Albert est en train de souffrir, c'est dire

1) qu'il existe une certaine structure, à laquelle participent à la fois les états neuro-physiologiques possibles ϕ_1, ... , ϕ_n d'Albert et certains événements publics qui mettent en jeu son comportement manifeste (chacun peut constater qu'il est exposé à certains types de stimuli, qu'il gémit, que ses gestes ont pour effet de le soustraire à la réexposition à ces stimuli, etc), et que cette structure *réalise* la théorie T, c'est-à-dire que les états neuro-physiologiques en question sont reliés entre eux et aux

événements publics pertinents de telle sorte que l'énoncé $T(\phi_1, \ldots, \phi_n)$ soit vrai)

2) qu'Albert est précisément dans l'état ϕ_i.

La douleur reste ici caractérisée par son rôle médiateur entre les stimulations sensorielles et les réponses motrices, mais la définition ménage la possibilité d'impliquer d'autres états mentaux dans cette médiation : l'émission de plaintes, par exemple, peut être également subordonnée à l'absence de volonté de taire la souffrance, ou bien intervenir là où les stimuli réputés douloureux n'affectent pas réellement le sujet, dès lors qu'entre en jeu la volonté de tromper. On a affaire à une forme raffinée de behaviorisme, capable d'exprimer sans circularité que c'est l'*interaction* des états mentaux entre eux et, de proche en proche, toute l'architecture psychologique, qui est responsable du comportement observable.

Chez un individu donné, un épisode douloureux est donc un état particulier ϕ_i du système nerveux central, qui peut être caractérisé de deux façons distinctes : d'une part d'après sa nature neurologique (ϕ_i consiste, par exemple, dans tel ou tel état d'excitation des D-neurones), et d'autre part par la fonction de médiation qu'il exerce, concurremment avec les autres états neuro-physiologiques de l'individu, entre stimuli et réponses (en bref, il est la i-ième composante d'une architecture qui réalise T).

Pour qu'un individu soit dans un certain état mental défini (disons, la douleur), il n'est ni suffisant ni nécessaire qu'il soit dans un certain état physiologique bien déterminé ϕ_i (disons, par exemple, que les neurones du groupe G soient, chez lui, excités).

1) Ce n'est pas suffisant, puisqu'il faut encore que la totalité de ses états physiologiques soit exactement organisée comme l'énonce la théorie T. Pour qu'il soit dans l'état mental E en question, le prédicat qu'il doit satisfaire n'est pas ϕ_i mais, à strictement parler, $T[\phi_1, \ldots, \phi_n] \,\&\, \phi_i$, c'est-à-dire le prédicat qui est satisfait, *parmi les organismes organisés selon T*, par ceux qui vérifient ϕ_i. Un autre individu peut être dans le même état ϕ_i — c'est-à-dire, supposément, avoir les neurones du groupe G dans tel et tel état d'excitation — *sans* être dans l'état mental E : c'est le cas si dans sa propre organisation cérébrale, l'excitation des G-neurones n'a pas les mêmes relations causales caractéristiques aux autres états neuro-physiologiques. En bref, bien que l'excitation des G-neurones puisse être considérée comme *constitutive* du fait qu'Albert est dans l'état mental E — et, en particulier, que cette excitation ne puisse aucunement être considérée comme une simple *cause* de l'état mental en question —, elle

ne joue ce rôle que dans l'hypothèse et dans la mesure où le cerveau d'Albert est ainsi constitué que l'excitation des G-neurones y joue le rôle causal que définit l'état E.

2) Ce n'est pas, non plus, nécessaire, puisque pour un individu dont l'ensemble des états physiologiques $\phi'_1, \ldots, \phi'_n$ est une autre réalisation (totale) de la théorie psychologique de base (c'est-à-dire dans un cas où c'est $T[\phi'_1, \ldots, \phi'_n]$ qui est vrai), être dans l'état mental E (supposé figurer à la i-ème place dans une énumération de référence des prédicats psychologiques) consiste évidemment à être dans l'état ϕ'_i et non dans l'état ϕ_i.

Au total, le fait d'être dans l'état ϕ_i, pas plus, en fin de compte, que le fait de se trouver dans aucune situation neuro-physiologique *localement* définissable, ne peuvent être considérés comme des conditions nécessaires ou suffisantes pour la satisfaction du prédicat mental E. A strictement parler, la relation de réalisation n'est donc pas *locale*, associant inconditionnellement un état cérébral à un état mental, mais *totale*, associant une architecture cérébrale à une architecture psychologique.

Une théorie psychologique rassemblant toutes les données relatives à l'expression publique des états mentaux $\psi_1, \ldots, \psi_n$ peut être comparée à un système d'équations AX = B à n inconnues, dans lequel les mots du vocabulaire logique et observationnel jouent le rôle de coefficients, et dont une solution est n'importe quel vecteur d'états physiques ou physiologiques $\vec{\Phi} = (\phi_1, \ldots, \phi_n)$ vérifiant $A\vec{\Phi} = B$. La douleur, supposée être la i-ème inconnue du système, est définie par les équations qui disent ses relations aux stimulations reçues et aux réponses retournées, équations dans lesquelles — et c'est là la différence entre fonctionnalisme et behaviorisme — figurent généralement d'autres prédicats psychologiques. Si une architecture matérielle $\vec{\Phi}$, disons l'organisation cérébrale humaine, est solution de l'équation, alors sa i-ème composante est, *ipso facto*, de la douleur. Mais elle n'est pas pour autant *la* douleur, à moins que l'architecture en question ne soit l'*unique* solution de l'équation. La douleur, définie par son rôle causal, ne saurait être identifiée à un état physiologique qui joue ce rôle que si ce rôle est ou peut être seulement joué par cet état. De façon plus générale, être dans un état mental donné consiste toujours à être dans *un* état physique ou physiologique (rien de mental n'est immatériel), mais pas toujours à être dans le *même* état (les prédicats mentaux ne sont donc pas réductibles à des prédicats physiques ou physiologiques). Bien que les propriétés

mentales soient toujours incarnées, c'est-à-dire qu'elles ne puisse être instanciées sans support corporel, il n'existe de substrat physique unique pour aucune d'entre elles.

3. Lorsque nous attribuons à une même entité, disons à une *personne*, des prédicats physiques ou physiologiques *et* des prédicats psychologiques, nous ne procédons pas doublement, attribuant les premiers à son corps, et les seconds à une substance séparée, qui en elle ne serait pas un corps. Nous attribuons les premiers à son corps *simpliciter*, et les seconds à son corps en tant qu'il possède une organisation interne dont la configuration actuelle le dispose à répondre de telle et telle façon lorsqu'il est exposé à telle et telle stimulation. En bref, c'est la même chose qui pèse 80 kilos et qui attend la venue du printemps, et cela, le fonctionnalisme l'explique très bien.

Il est moins sûr, en revanche, que le fonctionnalisme soit en mesure de répondre à une autre demande, selon laquelle les prédicats mentaux devraient conserver le *même* sens, qu'on se les attribue à soi-même ou qu'on les attribue à quelqu'un d'autre. L'adoption d'une définition fonctionnelle des états mentaux, qui découle imparablement du principe selon lequel nous ne pouvons nous accorder sur l'emploi du vocabulaire psychique qu'à la condition de nous rapporter à la manière dont ces états s'expriment publiquement, nous met certes en mesure de garantir que nous nous comprenons bien, vous et moi, lorsque nous attribuons à un tiers, ou bien même nous attribuons l'un à l'autre, le même état mental. Mais pour que nous nous entendions vraiment, il faudrait que la signification des prédicats psychologiques demeure inchangée quelle que soit la personne à laquelle ces prédicats sont attribués *et* quel que soit l'auteur de l'attribution, et notamment que nous les utilisions dans le même sens à la fois lorsque chacun parle de soi-même et lorsque nous parlons tous deux de l'un d'entre nous. Le premier cas se réduit au second, puisque si ce que vous inférez à mon sujet en m'écoutant parler de moi-même et, plus généralement, en observant mon comportement, coïncide bien avec ce que j'éprouve de mon côté, alors je dois être capable d'en faire autant, et nous pouvons donc nous communiquer nos états mentaux sans malentendu. Ce qui doit être établi est donc la coïncidence entre ce que je peux *observer* de vous, ou plus exactement ce que je suis conduit à supposer, compte-tenu de la théorie psychologique, au titre de mécanisme interme expliquant ce qu'il m'est donné d'observer de vous, et ce que je puis *éprouver* pour moi-même. Pour s'en assurer, il convient au moins de vérifier que les états mentaux que l'on s'attribue à soi-même sont de même nature que ceux que l'on attribue à autrui. Même si l'analyse de la signification du vocabulaire des états psychiques doit commencer à la

troisième personne, c'est-à-dire se fonder sur la pratique de l'attribution à autrui, elle ne peut totalement passer sous silence les ressorts de l'auto-attribution, et cela en vertu d'un simple principe de réciprocité : je dois admettre qu'autrui est pour lui-même ce que je suis pour moi-même, et donc qu'il doit être capable, si je lui attribue à bon droit quelque état mental sur la base de ce que son comportement révèle, de s'attribuer lui-même cet état, et cette fois en vertu des mêmes raisons pour lesquelles je m'auto-attribue quelque chose, c'est-à-dire tout simplement parce qu'il le ressent ainsi. Au minimum, nous ne devrions donc attribuer des états mentaux qu'aux seules créatures susceptibles de les *ressentir*.

À première vue, le principe de publicité de la signification ne devrait nullement obvier à la satisfaction de cette exigence, puisque ce principe, d'ordre *linguistique* et non *métaphysique*, n'affirme pas que les sensations privées ne sont rien de réel, mais seulement qu'elles ne peuvent être désignées dans le langage qu'au travers du comportement public de celui qui les éprouve. Une phrase comme 'Je suis seul à éprouver mon mal de dents' est ambiguë, pouvant posséder deux significations contre la confusion desquelles ce principe est précisément destiné à nous prémunir. En un premier sens, vrai mais trivial, la phrase rappelle que c'est moi seul qui éprouve ce que j'éprouve, c'est-à-dire le *token* de sensation douloureuse qui m'échoit : lorsque j'ai mal aux dents (et peu importe ici en quoi cela consiste au juste), ce n'est pas vous qui souffrez. En un second sens, substantiel mais douteux, la phrase affirme que je suis (peut-être) le seul à éprouver une sensation douloureuse de ce *type* : lorsque vous avez (ce que vous appelez) 'mal aux dents', vous n'éprouvez peut-être pas la même sensation que moi. La transition de la première acception à la seconde ne se justifie que si 'douleur' est définie au travers d'une ostension privée, qui conduit chacun à se dire, à l'occasion d'un épisode pénible qui lui arrive en propre : 'Voilà, c'est cela, la souffrance'. Puisque nul autre que l'auteur — et le destinataire — de l'ostension ne saurait traverser l'épisode privé qui lui sert de référence (acception 1), il demeure toujours incertain que le mot 'souffrance' ainsi défini par l'un soit applicable aux états des autres ou coïncide avec ce qu'ils entendent par ce même mot (acception 2). Si la douleur doit être nommée dans un langage commun, il faut au contraire qu'elle soit définie à partir des manifestations publiquement observables auxquelles donne régulièrement lieu son occurrence chez n'importe lequel d'entre nous, c'est-à-dire comme cette chose X, *quelle qu'elle soit*, qui nous arrive à tous lorsque nous sommes ébouillantés, qui nous inspire alors des cris irrépressibles, qui ... , et qui ... : des modalités singulières selon lesquelles chacun reconnaît par ailleurs cette chose lorsqu'elle lui arrive à lui-même, c'est-à-dire de la qualité particulière de ce qu'il éprouve en pareil cas, rien ne doit être retenu qui

resterait en-deçà de ce qu'il est possible de saisir publiquement. Cependant l'existence même de la "chose" reste non seulement hors de contestation, mais strictement impliquée par la définition même de la douleur. Comme l'écrit A. Donagan en commentant les idées de Wittgenstein sur ce point [2], "vous et moi ne pourrions avoir un mot commun pour la douleur si notre comportement naturel de douleur n'était accompagné de quelque chose d'affreux ; quant à savoir si cet accompagnement est le même ou pas pour chacun d'entre nous, ou même s'il change ou pas (à condition que nous ne nous en apercevions pas), ceci est dénué de pertinence". En somme, le principe de publicité demande que l'on renonce, dans la définition d'un état mental, à tenir compte de la nature exacte des sensations qui sont éprouvées par ceux qui sont dans cet état, mais il est supposé compatible avec la décision de n'attribuer des états mentaux qu'aux créatures capables de les éprouver. L'expérience qualitative associée à un état n'est pas niée, mais seulement "publicisée", c'est-à-dire réduite à ce qui peut expliquer le comportement public adopté par ceux qui l'éprouvent : la pénibilité de la douleur, par exemple, *est* ce que l'on cherche à abréger en se soustrayant à ce qui en est la cause.

Or il est loin d'être clair que la mise entre parenthèses de l'expérience privée puisse se limiter à cela, et l'adoption du principe de publicité pourrait bien nous entraîner à repousser la réalité de l'expérience psychique elle-même. Supposons en effet que l'état mental ψ_i soit défini par l'équivalence $\psi_i(x) \equiv \exists\xi_1 ... \exists\xi_n [T(\xi_1, ..., \xi_n) \& \xi_i(x)]$, et qu'il soit réalisé en nous par l'état neurophysiologique ϕ_i, dont l'occurrence s'accompagne d'une certaine sensation caractéristique. Nous sommes obligés d'admettre que toute entité dont l'architecture matérielle $< \phi'_1, ..., \phi'_n >$ satisfait également T, et qui se trouve par là-même fonctionnellement identique à nous, peut elle aussi se trouver dans l'état mental ψ, même dans le cas où les sensations associées chez elle à cette occasion seraient très différentes des nôtres — ce qui est une dissemblance sur laquelle le fonctionnalisme est prêt à fermer les yeux — , et même dans le cas où *aucune* sensation associée ne serait même envisageable pour l'entité en question — ce qui est, cette fois-ci, une différence considérée comme intolérable. Sauf à *postuler* que ce dernier cas ne peut jamais se produire, c'est-à-dire sauf à supposer, comme le fait par exemple P. Churchland, qu'une expérience qualitative d'un certain type est "inévitablement concomitante" [3] de la situation de tout système dont l'organisation fonctionnelle répond à la définition d'un état mental donné,

---

[2] *Wittgenstein on Sensation* (1966), in Donagan (1994), p. 253.

[3] Churchland (1988), p. 41.

on est donc conduit à écarter la réalité de l'expérience psychique elle-même, ou tout au moins à considérer que son intervention est une composante *facultative* de l'état en question. En bref, le caractère "purement linguistique" du principe de publicité apparaît comme une fiction difficilement tenable, et son adoption correspond en vérité à une pure et simple négation de l'expérience mentale privée.

Sans doute aucune objection de ce genre ne saurait-elle être absolument dirimante. D'une part elle affecte, au plus, la manière dont le fonctionnalisme rend compte de *certains* états mentaux, à savoir ceux, comme la souffrance, dont l'occurrence est accompagnée de certaines sensations caractéristiques, et elle laisse entièrement intacte la question des états mentaux dans la définition desquels figure prioritairement des liaisons aux autres états mentaux (en particulier, cette objection ne s'applique aucunement à la définition des "attitudes propositionnelles", et n'hypothèque en rien la possibilité de considérer qu'un ordinateur "croit" que p, lorsqu'une phrase signifiant que p est stockée dans sa mémoire selon telle ou telle modalité, puisque la croyance n'est accompagnée d'aucune sensation caractéristique). D'autre part, et s'agissant maintenant des états mentaux typiquement "ressentis", la question est de savoir comment nous pourrions refuser à un système fonctionnellement identique à nous la possibilité d'avoir de tels états, et en particulier de les éprouver, alors même que nous sommes, par hypothèse, unanimes à convenir que ce système manifeste absolument tout ce que peut manifester une créature qui les a, et donc qui les éprouve. Si, comme le prétend l'objecteur, un système peut se comporter publiquement en tout point comme s'il souffrait, sans pour autant rien éprouver, c'est-à-dire si la présence ou l'absence d'une expérience qualitative de la douleur ne peut faire aucune différence sur laquelle on pourrait se fonder pour distinguer les cas de douleur réelle de ceux de douleur fictive, sur quoi pourrions-nous fonder notre conviction que tel ou tel système est exempt de "vraie" douleur ? S'il n'y a pas de manifestations publiques suffisamment révélatrices de l'existence de *qualia*, il ne saurait non plus y en avoir qui soit suffisamment révélatrice de leur absence, et l'anti-fonctionnaliste n'est donc pas en position d'indiquer en ce domaine un contre-exemple qui pourrait recueillir l'agrément de tous. Aussi son objection doit-elle être différemment construite, et invoquer plutôt le contraste entre notre propre expérience psychique de la douleur et la situation d'un dispositif *artificiel* qui se trouverait, par hypothèse, dans un état fonctionnellement indiscernable de celui qui est le nôtre lorsque nous souffrons, mais que nous ne serions nullement enclins, compte-tenu justement de sa nature d'artefact, à qualifier lui-même de souffrant. Pour construire, au moins hypothétiquement, un tel dispositif, on suppose donc que les énoncés de la

théorie T qui sert de référence sont tous de la forme suivante : exposé à un stimulus σ, un individu qui est dans l'état mental ψ_i adopte le comportement σ', et il est alors dans l'état mental ψ_j (c'est la mention de l'état ψ_j qui distingue un énoncé de ce genre d'une définition purement behavioriste de ψ_i comme propension à retourner σ' en réponse à σ). La théorie T équivaut alors à la donnée d'une machine de Turing définie par une table d'instructions du type $\sigma\psi\sigma'\psi'$, et les états mentaux qu'elle définit sont purement et simplement les états internes de cette machine : être dans l'état ψ_i, c'est être dans le i-ème état physique d'un dispositif matériel (un "ordinateur") qui réalise cette machine sous une certaine de ses descriptions. Or un ordinateur n'a pas la capacité d'éprouver subjectivement ses états. En adoptant une définition purement relationnelle des états mentaux, on s'expose donc à conférer, de manière absurdement libérale, une vie mentale à des entités qui ne sauraient en posséder.

La portée de l'argument anti-fonctionnaliste ainsi reconstruit dépend de la valeur des deux prémisses supplémentaires sur lesquelles il repose, c'est-à-dire d'une part l'impossibilité d'expériences psychiques dans les artefacts, et d'autre part l'hypothèse selon laquelle les théories psychologiques peuvent être rédigées sous la forme d'une table de Turing. Selon cette dernière hypothèse, disons le "Turing-fonctionnalisme", les états mentaux ne sont pas seulement *analogues* aux états de calcul d'une machine, c'est-à-dire implicitement définis par la théorie psychologique dans laquelle ils figurent de la même façon que les états computationnels sont implicitement définis par la table de la machine à laquelle ils appartiennent, mais ils sont strictement *identiques* à de tels états de calcul : l'esprit est, *littéralement*, un automate abstrait. Il est clair que cette variété de fonctionnalisme constitue une *restriction* du fonctionnalisme pur et simple, puisqu'elle consiste à n'admettre que *deux* états mentaux dans chaque loi psychologique (chacune de ces lois a la forme d'une instruction qui ne réfère qu'à deux états de calcul), là où le behaviorisme n'en admet qu'un seul et le fonctionnalisme ordinaire un nombre quelconque. Or il est, *prima facie*, implausible que les lois qui décrivent, par exemple, le rôle causal de la douleur puissent être rédigées en n'invoquant à chaque fois qu'*un* autre état mental (par exemple : l'absence de volonté d'induire en erreur), si bien que la conclusion la plus raisonnable à tirer de toute cette affaire consiste probablement à reconnaître que le Turing-fonctionnalisme sous-estime *grossièrement* la complexité de nos états mentaux, et la difficulté qu'il y aurait à mettre au point un artefact capable de simuler exactement le comportement que ces états nous inspirent. De façon plus générale, la portée *empirique* de

l'objection est faible, en ce qu'elle sous-estime la très vaste différence qui sépare les ordinateurs, dont les états physiques pertinents sont *élus*, et les cerveaux, dont les états neuro-physiologiques pertinents sont simplement *donnés* [4]. Il n'en reste pas moins que cette objection repose sur une situation qui est logiquement, sinon nomologiquement, possible, et qu'il importe donc de l'affronter, fût-ce à titre d'expérience de pensée.

4. La douleur possède un double aspect, qualitatif en tant qu'elle est éprouvée, fonctionnel en tant qu'elle possède des causes et des effets caractéristiques. Une bonne partie de nos attitudes, notamment éthiques, vis-à-vis de la douleur, se définissent évidemment par référence à son aspect qualitatif : c'est en considérant le genre d'expérience qui consiste à l'endurer que nous pouvons éprouver de la compassion ou nous sentir le devoir de soigner. Or il faut bien reconnaître que, même si les choses ont, par certains côtés, assez évolué au cours de la période récente dans le domaine de l'"intelligence artificielle" pour que nous puissions désormais parler sans incongruité d'informations disponibles et d'objectifs poursuivis à propos d'artefacts comme les machines à jouer aux échecs, nous ne nous sentons en revanche nullement tenus par les obligations morales qui devraient être les nôtres si de telles machines devaient s'avérer capables de souffrir ou de ressentir quelque état que ce soit. Cette réticence à accorder aux artefacts la possibilité de connaître des expériences psychiques est à la fois vivace et difficile à fonder. Bien que la distinction entre ce qui est vivant et ce qui n'est l'est pas soit certainement en jeu dans cette répugnance — puisque nous admettrions sans hésiter, par contre, qu'un clone issu d'une créature capable de souffrance puisse lui-même éprouver de la douleur —, il y entre aussi, inévitablement, le fait que nous sommes crucialement sensibles, dans ce domaine, à la question de la *simulation* : même si un dispositif artificiel se comportait durablement, voire indéfiniment, de manière assez complexe et variée pour que nous soyons irrésistiblement enclins à parler de lui en disant qu'il est ému ou qu'il souffre, nous reconsidèrerions notre position, et tiendrions notre compassion pour abusée, si nous en venions à apprendre que ce dispositif avait été expressément mis au point pour mimer le comportement caractéristique de l'émotion ou de la douleur. Une conception des états mentaux qui serait incapable de distinguer la souffrance réellement endurée de la souffrance simulée est menacée d'une absurdité éthique, puisque la première nous crée des obligations, alors que la seconde ne nous engage à rien. Dans sa version "linguistique",

---

[4] En termes plus précis, cette objection repose sur l'hypothèse contestable de la récursivité de l'univers physique (cf. Dubucs (1992)).

qui ne nie pas la réalité de l'expérience psychique, le behaviorisme est, à cet égard, à la merci d'une créature à la fois insensible et insincère, capable d'accueillir sans rien éprouver les stimulations les plus pénibles aux autres, et de feindre en retour leurs manifestations de douleur les plus véhémentes (c'est du reste la raison pour laquelle Wittgenstein voyait dans le comportement "pathétique" non une condition nécessaire et suffisante, mais un *critère* subconclusif et, selon son mot, "défaisable" de la douleur elle-même [5]). De son côté, le fonctionnalisme est immunisé contre ce risque, puisqu'il peut disqualifier les simagrées du comédien en stipulant un certain nombre de conditions collatérales, comme l'absence du désir de tromper, au titre des relations intra-mentales qui contribuent à définir la douleur. Mais il doit aussi, sous peine de la même absurdité éthique, se prémunir contre la possibilité d'un autre type de simulation, à savoir celle du concepteur qui, non nécessairement mû par le désir d'induire en erreur, mais par le simple effet de la réussite de son travail d'analyse, en viendrait à produire un dispositif isomorphe à celui de la créature souffrante qu'il étudie, c'est-à-dire parviendrait non seulement à associer les mêmes "sorties" aux mêmes "entrées", mais obtiendrait ce résultat de manière essentiellement identique, en vertu même du fait que son dispositif possède des états internes qui correspondent terme à terme à ceux qui gouvernent le comportement de l'organisme qu'il a sous les yeux.

En vérité, cette dernière difficulté semble à peu près inextricable, puisque notre refus d'attribuer à de tels dispositifs la possibilité de traverser des expériences psychiques provient désormais d'une décision éthique qui a trait à leur nature d'artefacts, et non aux imperfections qui pourraient être les leurs dans leur genre propre : c'est parce qu'ils simulent, et non parce qu'ils simulent mal, que nous repoussons pour eux toute prétention à une vie psychique. Compte-tenu du fait que nous entendons bien, par contre, ne pas nier la réalité des expériences psychiques pour ce qui *nous* concerne nous-mêmes, le problème ressemble fâcheusement à celui de discriminer les caractéristiques propres de l'original parmi les traits qui sont communs à l'original et à ce qui en est une copie en tout point fidèle. La différence spécifique recherchée ne peut guère être assignée que par la morne ré-énonciation de la séparation qu'elle est censée opérer : des deux états fonctionnellement indiscernables auxquels on a affaire, la vraie douleur est celle qui n'est pas simulée, et la pseudo-douleur est l'autre. En l'absence de la moindre marque publiquement objectivable de l'authenticité, il demeure impossible de s'accorder sur la différence qui fait la vraie douleur. Si détaillée soit-elle, la description des relations qui unissent cette vraie douleur aux autres états mentaux et à

---

[5] Cf. sur ce point Dubucs (1995).

tout ce qui peut être manifesté à la communauté des observateurs est encore trop grossière, justement parce qu'elle ne peut la caractériser que comme un X contraint par des relations, alors qu'il y a en elle un élément monadique, non relationnel, et que ces relations peuvent toujours être satisfaites par un état parasite, dont l'élément monadique crucial est absent. Pour faire le dernier pas vers la référence attendue, et séparer pour finir la souffrance de son fantôme indolore, il incombe à chacun de se tourner vers lui-même pour saisir cet élément, ce qui ne peut se faire qu'à la première personne : nul autre moyen, pour comprendre *que* la douleur est un état que l'on éprouve, que d'expérimenter soi-même *ce que* c'est que de l'éprouver. Mais si les conditions de vérité de la phrase '*x* souffre' ne peuvent être complètement déterminées sans que soit compris en quoi consiste, *pour x,* le fait de souffrir, alors ces conditions ne peuvent être également reconnues par tous les locuteurs, et la signification de la phrase devient inaccessible à tout autre qu'à *x* lui-même ou à ceux qui partagent son "point de vue" : même si y possède toutes les informations — collectables par tous, quant à elles — relatives à la physiologie et à l'architecture fonctionnelle de *x,* il en est réduit, s'il n'est pas *x* lui-même, à chercher à imaginer ce que *x* peut bien ressentir dans l'état où il le voit, alors qu'il ne peut évidemment imaginer que ce qu'il ressentirait, *lui y*, s'il était à la place de *x*. La situation est donc assez sombre, puisqu'elle semble ne laisser le choix qu'entre deux renoncements malencontreux :

— ou bien abandonner la publicité du vocabulaire des états psychiques, c'est-à-dire l'idée que ce vocabulaire doit pouvoir être utilisé par tous sans équivoque, et admettre alors que la signification du terme 'souffrance' peut varier selon celui qui l'emploie et celui à qui il est appliqué : souffrance_x pour *x*, et souffrance_y pour *y*

— ou bien mettre entre parenthèses non seulement la nature mais l'existence des *qualia*, c'est-à-dire renoncer à n'attribuer la douleur qu'à ceux qui la ressentent, et finalement éliminer ainsi la subjectivité caractéristique du mental (comme le résume Nagel [6], "le glissement vers une objectivité plus grande — c'est-à-dire le moindre attachement à un point de vue spécifique — ne nous porte jamais plus près de la nature réelle du phénomène : il nous emporte loin d'elle").

J'aimerais montrer que ce dilemme — en gros : les apories du langage privé, et sinon la négation de l'expérience interne — n'est pas aussi tranché qu'il le paraît, et qu'il reste, à ce que l'on pourrait nommer, par analogie avec la doctrine attribuée à Wittgenstein par Donagan, le fonctionnalisme "linguistique", une marge de manoeuvre conséquente pour

---

[6] Nagel (1974), p. 436.

éviter les deux dangers qui le guettent : un fonctionnaliste qui se laisserait enfermer dans l'alternative en question ne tirerait pas assez parti de tous les avantages que peut procurer la décision de ne retenir que l'*existence* de l'expérience qualitative, et rien d'autre, en elle, de ce qui peut aller au-delà de toute caractérisation fonctionnelle de l'état mental auquel elle est associée.

4.1. Admettons que les difficultés qui proviennent de la simulabilité du rôle fonctionnel ne peuvent être écartées ni par l'éliminativisme (nous n'aurions pas plus de vie mentale que l'artefact qui nous simule) ni par le panpsychisme (l'artefact en question en aurait autant que nous). Ces difficultés proviennent évidemment de la tension entre deux desiderata :

— celui d'établir une connexion essentielle entre les états mentaux et ce qui peut être manifeste à tous, et donc de définir leurs conditions d'attribution de telle sorte qu'il n'y ait personne qui ne puisse en principe se rendre compte qu'elles sont remplies, lorsqu'elles le sont ;

— celui de rendre justice à l'introspectibilité des états mentaux, c'est-à-dire au fait que nul ne peut les avoir sans savoir qu'il les a, et que celui qui les a en est informé immédiatement, sans être lui-même aucunement astreint aux vérifications que doivent effectuer les autres pour savoir si cet état lui est attribuable.

Les paradoxes de la simulation montrent qu'il y a une manière d'obéir à la première exigence qui peut conduire à rendre la seconde insatisfaisable : un usage non critique de la perspective en troisième personne peut conduire à attribuer un état mental à des entités dont il est à présumer qu'elles ne s'auto-attribueraient pas le même (paradoxe du comédien) ou qu'elles seraient incapables de s'en attribuer un seul (paradoxe de l'artefact). Ceci montre certainement que les conditions de l'attribution à autrui d'un état mental doivent considérablement différer des conditions d'attribution d'un état physique : l'imputation d'attitudes propositionnelles repose largement sur des principes de rationalité dont on ne trouve aucune contrepartie dans la théorie physique, et l'ascription d'un état mental à un individu à l'instant t peut exiger que l'observation de l'individu en question soit poursuivie au-delà de t, particularité qui n'a, elle non plus, guère d'écho en physique. Il n'en découle nullement, en revanche, que la psychologie diffère de la physique au point que la perspective en première personne doive y prévaloir. Contrairement à ce qu'écrit Searle [7], pour qui la difficulté de séparer entre les créatures qui éprouvent leurs états et les artefacts qui les simulent repose tout entière sur la thèse métaphysique

[7] Searle (1992), p. 16.

"évidemment fausse" selon laquelle "tout ce qui est réel doit être également accessible à tous les observateurs compétents", l'adoption du point de vue "subjectif" n'est en rien capable de résoudre la question. En effet cette perspective, qui sépare l'ego du reste indifférencié du monde, nous laisse dans un doute *uniforme* au sujet de de la vie psychique des autres hommes et de celle des machines, et ne peut donc contribuer à les distinguer comme il faudrait justement le faire. À cet égard, rien ne résume mieux les limites de la psychologie en première personne, et son incapacité de faire à autrui le sort *particulier* qui lui revient parmi tous les candidats à la vie mentale, que la délicieuse absurdité du dialogue suivant, attribué à Chuang-Tse et Heitse (300 avant notre ère) :

— Regarde le petit poisson, comme il file ! Le voilà bien, le bonheur du poisson !

— Mais tu n'es pas toi-même un poisson ! Que peux-tu donc savoir de son bonheur ?

— Et toi, tu n'es pas moi ! Alors comment peux-tu savoir que je l'ignore ?

Il n'y a aucune espèce d'avantage à attendre du remplacement de la perspective "objective" par la perspective en première personne, puisque cette dernière souffre, quoique pour des raisons opposées, d'un handicap exactement semblable à celui qui affecte l'objectivisme : elle met elle aussi humains et non-humains à la même distance que celle qui sépare les humains entre eux. Le recours à la perspective du "sujet lui-même" repose ici sur un malentendu bien compréhensible. La définition fonctionnelle d'un état comme la douleur est *excessivement libérale*, puisqu'elle conduit à considérer comme capables de souffrance des entités qui se contentent, si l'on ose dire, d'exécuter sans rien endurer les instructions de la théorie de la douleur. De son côté, la perspective de l'ego est *excessivement avaricieuse*, portée qu'elle est à ne considérer comme certains que les états qu'elle s'auto-attribue. On voudrait voir ces deux exagérations symétriques se neutraliser réciproquement, par une sorte d'amalgame des deux points de vue, qui consisterait pour l'essentiel à décrire en troisième personne le processus par lequel le sujet en vient à s'auto-attribuer un état mental. Mais il y a de très sérieuses raisons de douter qu'une telle entreprise ait aucune chance d'aboutir, au premier rang desquelles l'impossibilité de convertir sans perte de signification un énoncé d'auto-attribution comme 'j'ai mal' en énoncé "neutre", où le 'je' serait remplacé par une description objective de la personne qui dit 'je'. Nul doute en effet que ce pronom ne figure dans 'j'ai mal' comme ce que Perry [8] appelle un indexical "essentiel" : la douleur que je m'auto-attribue en l'éprouvant est essentiellement liée à la façon singulière et

[8] Perry (1979), p. 3.

"inobjective" que j'ai de m'apparaître à moi-même, et ne saurait donc être suspendue à ma capacité à me reconnaître dans une quelconque description objective de ma propre personne. Au reste, même là où cette reconnaissance a lieu, les choses ne se passent tout simplement pas comme elles devraient se passer pour que cette "désindexicalisation" soit concevable : il faudrait pour cela que, me reconnaissant dans une description, c'est-à-dire par exemple me voyant dans un miroir comme me voient les autres, et apercevant mon visage ensanglanté, j'en vienne à réaliser brusquement "Tiens, mais cet individu, je le connais, c'est JD !", et qu'*alors* je m'exclame "Aïe". Comme le montre l'absurdité de cette situation, le fait que *j*'éprouve indiscutablement, *quant à moi*, quelque chose de pénible lorsque j'ai mal est un fait qui ne saurait en lui-même servir de base pour amender la définition des conditions de vérité de la phrase 'x souffre', car l'information cruciale que *je* suis précisément cet x ne saurait être véhiculée par aucune phrase ne comportant pas le mot 'je' [9]. En résumé, l'auto-attribution ne pourrait être décrite à la troisième personne, comme il le faudrait, que si le sujet s'attribuait lui-même ses états à la manière dont un tiers les lui attribue, ce qui n'est évidemment pas le cas.

4.2. Une fois reconnue l'impossibilité d'une égologie qui décrirait l'expérience mentale du sujet en des termes qui indiqueraient que c'est lui, et non un autre, qui a cette expérience, il ne reste plus qu'à simplement dire *que* cette expérience a lieu, c'est-à-dire *que* le sujet s'auto-attribue "directement" ses états psychiques. Or cela semble pouvoir se dire dans le langage usuel, sans recourir à un hypothétique mode de désignation canonique qui présenterait à chacun le sujet tel qu'il est pour lui-même. Bien plus, s'il résulte de la nature des états qualitatifs que l'on ne puisse guère décrire l'expérience de ceux qui les éprouvent qu'en utilisant le nom de ces états ("j'ai mal, vous dis-je"), on voit mal comment ceci pourrait empêcher de dire en quoi consiste, pour un état, le fait d'être qualitatif. Un tel état est, justement, un état dans lequel on ne peut être engagé sans savoir qu'on l'est, et cela en vertu même du fait qu'on l'est. A cet égard, la croyance que l'on est dans l'état ψ est indiscutablement un effet de l'état ψ lui-même. Le fonctionnaliste, qui fait droit aux relations causales intramentales, peut saisir cette caractéristique comme une différence discriminante entre les entités capables d'éprouver leurs états et les autres. Il semblerait en effet absurde qu'une entité réalise la définition fonctionnelle d'un état qualitatif — y compris, donc, le fait que cet état est capable de causer la croyance qu'il est bel et bien réalisé — *sans que*

---

[9] Cf. pourtant Peacocke (1979), pp. 172 sq.

cette entité soit du tout dans un état qualitatif. La pseudo-douleur de l'artefact, qui n'est aucunement éprouvée, ne peut donc tout simplement pas être *fonctionnellement* identique à la douleur tout court : même si le "comportement d'entrée-sortie" associé aux deux états est identique, leur rôle fonctionnel ne peut être le même, en vertu du fait que la douleur réellement éprouvée est la seule à ébranler les capacités introspectives du sujet. À y regarder de plus près, cet argument, dû pour l'essentiel à Shoemaker [10] , n'est cependant pas convaincant, car une machine — en vérité, un programmeur — peut être assez retorse pour simuler l'introspection elle-même. En disant que les états qualitatifs sont *causes* de leur introspectibilité, on s'exprime en effet de manière intolérablement confuse, l'énoncé correct étant simplement que pour chaque état ψ de ce type, le conditionnel $\forall x\,(\psi(x) \varnothing \mathrm{BEL}(\psi)(x))$ doit figurer dans une théorie psychologique correcte. Le *quale* associé à ψ, quant à lui, n'est évidemment pas un état mental *sui generis*, c'est-à-dire distinct de ψ : éprouver ce *quale*, *c'est* éprouver ψ, ni plus ni moins, faute de quoi éprouver ψ demanderait que l'on éprouve aussi le *quale* associé au fait d'éprouver le *quale* associé à ψ, etc. En conséquence, ce *quale* n'a pas à figurer dans la théorie, il est seulement le canal par lequel, dans la réalisation particulière (le logicien dirait : dans l'interprétation *attendue*) de la théorie que nous avons en vue, transite la relation de causalité qui relie l'état physique qui réalise ψ à l'état physique qui réalise BEL(ψ). Mais cette spécification d'un chemin causal appartient aux *marginalia* de la théorie, laquelle peut donc fort bien être réalisée par un dispositif qui n'emprunte pas ce chemin. Si ψ et BEL(ψ) sont respectivement réalisés par les états physiques ϕ et ϕ' d'une certaine machine, la transition causale de ϕ à ϕ' peut y être assurée, par exemple, par la saturation du transistor n° 305, situation dans laquelle personne ne serait vraiment disposé à reconnaître l'épreuve pathétique de la vraie douleur.

4.3. La déroute de la défense proposée par Shoemaker montre que le fonctionnalisme est vulnérable à une difficulté dont le type le plus général est le suivant : les propriétés sur la base desquelles les états mentaux sont attribuables à autrui (resp. à soi-même) sont relationnelles (resp. non relationnelles), alors que les théories qui dépeignent les propriétés relationnelles de ces états peuvent toujours être satisfaites par des états physiques *non standard*, qui n'en satisfont pas les propriétés non relationnelles.

---

[10] Shoemaker (1984), pp. 189 sq.

Loin de provenir d'un "préjugé objectiviste" auquel il serait aisé de renoncer, cette difficulté résulte d'une propriété générale du langage : deux interlocuteurs peuvent *toujours* diverger sur la référence d'un ensemble de termes, quel que soit le nombre de phrases contenant ces termes qu'ils s'accordent à considérer comme vraies. Autrement dit, si X découvre, après s'être mis d'accord avec Y sur n phrases contenant les mots ψ_1, ... , ψ_p, que Y n'entend pas ces mots comme il le voudrait ou comme il le faudrait, et qu'il énonce une nouvelle phrase destinée à écarter l'interprétation déviante de Y, le nouvel ensemble de n+1 phrases sera, à nouveau, susceptible d'interprétations divergentes. L'exemple paradigmatique de cette situation est évidemment donné par l'élucidation des mots 'zéro', 'successeur' et 'nombre' au moyen des axiomes de l'arithmétique, lesquels ne contraignent jamais l'interlocuteur qu'à adopter une interprétation *du type* $< u_0, u_1, ... >$, suite infinie à premier terme et sans répétition dont chaque membre n'a qu'un nombre fini de prédécesseurs. Y peut, *aussi bien*, penser à < 0, 1, ... > qu'à < 100, 101, ... > ou à < 2, 4, ... >. Il est de l'essence du langage que la coïncidence des interprétations des locuteurs ne puisse être obtenue qu'à un isomorphisme près au mieux, et l'on peut légitimement qualifier de publics les concepts à propos desquels ce type de concordance a lieu. D'un autre côté, les difficultés qui concernent les ressorts qualitatifs de l'auto-attribution des prédicats psychologiques montrent qu'il est impossible de *totalement* laisser de côté certaines différences intrinsèques, non relationnelles, entre interprétations isomorphes : tout se passe comme si, *mutatis mutandis*, *l*'interprétation correcte était, en ce domaine, < 0, 1, ... > et non < 100, 101, ... >, en vertu du fait, par exemple, qu'il s'agit là de la seule interprétation dont le terme initial peut être écrit sans lever la plume. En somme, une théorie psychologique correcte devrait adjoindre à la liste des conditions que doivent satisfaire les entités qui la réalisent une condition supplémentaire, relative à la manière particulière dont elles devraient la réaliser : les états mentaux sont, parmi les états qui réalisent la théorie psychologique, ceux qui la réalisent comme le font nos propres états neurophysiologiques, nous, les humains, qui éprouvons quelque chose à cette occasion.

Peut-on atteindre cet objectif de façon moins naïve, c'est-à-dire en donnant un statut théorique non dérogatoire, différent du régime *off record* auquel elle semble vouée, à la clause relative au mode de réalisation particulier qui est visé ? Contrairement à une théorie mathématique, une théorie psychologique contient des termes faisant référence à des observables, et c'est là une circonstance favorable dont il convient assurément de tirer parti. Chaque nouvelle clause comportant de tels termes observationnels restreint un peu plus les interprétations

possibles des prédicats psychologiques, et l'on peut concevoir qu'il existe un point limite au-delà duquel les contraintes interprétatives deviennent redondantes, c'est-à-dire ne ferment plus aucune possibilité. Dans ces conditions, la décision la plus raisonnable consiste à n'accepter de définir les termes fondamentaux de la psychologie que sur la base d'une théorie T parvenue à ce point de détermination, quitte pour cela à *forcer les choses* en ajoutant à l'énoncé de Ramsey une clause d'unicité : on n'attribuera un état mental à un individu que si son architecture matérielle est l'*unique* réalisation de T [11] .

La difficulté de cette solution ne consiste certainement pas à trouver les termes observationnels supplémentaires destinés à saturer la théorie. Nous savons bien où les trouver, il suffit de puiser dans le vivier de notre propre répertoire neuro-physiologique, puisqu'il n'est après tout plus question que de dire comment *nous* réalisons nos propres états mentaux, nous-mêmes et les créatures qui nous sont apparentées. Le problème est ailleurs, et il est de comprendre comment cette opération peut préserver l'autonomie de la psychologie. Comme le remarquait en effet Carnap [12] à propos de la physique elle-même, si l'on atteignait un jour un point au-delà duquel on ne pourrait plus renforcer l'interprétation d'un terme théorique, ce terme deviendrait explicitement définissable en termes observationnels, et cesserait du même coup d'être théorique. La réponse est probablement que nous ne ferions en cela que repousser vers l'intérieur de l'organisme les éléments en termes desquels nous formulons nos explications, et que nous n'y trouverions pas de la substance vivante brute, "directement observable", mais, à nouveau, des ingrédients que nous ne savons individuer qu'en termes de leurs relations causales.

Jacques DUBUCS
Institut d'Histoire et de Philosophie des
Sciences et des Techniques - CNRS
13, rue du Four
75006 PARIS

Bibliographie

---

[11] Formellement, la différence tient à un point d'exclamation, puisque la définition d'un état mental est désormais donnée par l'énoncé de Ramsey-Lewis : $\psi_i(x) \equiv \exists! \xi_1 \ldots \exists \xi_n [T(\xi_1, \ldots, \xi_n) \& \xi_i(x)]$.

[12] Carnap (1966), trad. fçse p. 231.

Block N. (1978), *Troubles with Functionalism*, in N. Block (ed.) (1980), pp. 268-305.
Block N. (ed.), *Readings in Philosophy of Psychology* , I, Harvard U.P., 1980.
Carnap R. (1966), *Philosophical Foundations of Physics*, Basic Books; trad. fçse *Les fondements philosophiques de la physique*, A. Colin, 1973.
Churchland P. (1988), *Matter and Consciousness*, Bradford, M.I.T. Press, 2° édition.
Donagan A. (1994), *Philosophical Papers I. Historical Understanding and the History of Philosophy*, Chicago U.P.
Dubucs J. (1992), *Les théories psychologiques sont-elles récursives*, in A. Boyer (éd.), *Méthodologie de la science empirique (2)*, *Cahiers du C.R.E.A.*, XV, pp. 171-181.
Dubucs J. (1995), *Les arguments défaisables*, in *Hermès*, XV-1995, pp. 271-290
Frege G. & Hilbert D., *Correspondance*, in Fr. Rivenc & Ph. de Rouilhan (eds.), *Logique et Fondements des Mathématiques. Anthologie (1850-1914)*, Payot, 1992, pp. 215-235.
Lewis D. (1972) *Psychophysical and Theoretical Identifications*, in N. Block (ed.) (1980), pp. 207-215.
Nagel Th. (1974), *What is it Like to be a Bat ?*, in *Philosophical Review*, LXXXIII, pp. 435-450.
Peacocke (1979), *Holistic Explanation: Action, Space, Interpretation*, Clarendon Press, Oxford.
Perry J. (1979), *The Problem of the Essential Indexical*, in *Noûs*, XIII, pp. 3-21.
Searle J.R. (1992), *The Rediscovery of the Mind*, Bradford, M.I.T. Press.
Shoemaker S. (1984), *Identity, Cause, and Mind*, Cambridge U.P.