LE FONCTIONNALISME FACE AU PROBLÈME DES *QUALIA*[1]

1.0. *Fonctionnalisme, béhaviorisme et physicalisme*

L'approche fonctionnaliste de l'esprit fait aujourd'hui l'objet d'un large consensus[2]. Comme le béhaviorisme et le physicalisme, le fonctionnalisme cherche à répondre à la question : « Que sont les états mentaux ? » Je ne m'occuperai ici que des versions du fonctionnalisme qui prennent la forme d'une théorie de l'identité. Elles posent, par exemple, que la douleur est un état fonctionnel, de même que, en tant qu'il se présente comme une théorie de l'identité, le physicalisme pose que la douleur est un état physique.

Je commencerai par fournir une présentation du fonctionnalisme et un rapide aperçu de sa critique du béhaviorisme et du physicalisme. Puis je tenterai de montrer que le fonctionnalisme succombe en fait aux mêmes reproches qu'il adresse à ces deux doctrines.

Voici une façon de caractériser le fonctionnalisme suffisamment vague pour être recevable par la plupart de ceux qui s'en réclament : chaque type d'état mental est un état qui, étant donné certains inputs sensoriels et certains états mentaux, consiste dans une disposition à agir d'une certaine façon *et à avoir certains états mentaux*. Ainsi caractérisé, le fonctionnalisme peut apparaître comme une nouvelle incarnation du béhaviorisme. Le béhaviorisme identifie les états mentaux à des dispositions à

1. « Le fonctionnalisme face au problème des *qualia* » est la traduction d'une version écourtée de « Troubles with functionalism » (1978) parue en 1990 sous le titre *Qualia-based objections to Functionalism* dans *Mind and Cognition*, William Lycan (ed.), Basil Blackwell. Reproduit avec l'aimable autorisation de l'auteur.

2. Voir Fodor, 1965 ; Lewis, 1972 ; Putnam, 1966, 1967, 1970, 1975*a* ; Armstrong, 1968 ; Locke, 1968 ; peut-être Sellars, 1968 ; peut-être Dennett, 1969, 1978*b* ; Nelson, 1969, 1975 (voir cependant Nelson, 1976) ; Pitcher, 1971 ; Smart, 1971 ; Block et Fodor, 1972 ; Harman, 1973 ; Grice, 1975 ; Shoemaker, 1975 ; Wiggins, 1975.

Les Etudes philosophiques, n° 3/1992

agir d'une certaine façon quand certains inputs sont donnés. Mais comme l'ont fait valoir plusieurs de ses critiques (Chisolm, 1957 ; Geach, 1957 ; Putnam, 1963), le désir d'atteindre le but B ne peut être identifié à la disposition, par exemple, à faire A dans les circonstances où les inputs font que A conduit à B parce que, après tout, il se peut fort bien que l'agent ne *sache* pas que A conduit à B, et donc ne soit pas disposé à faire A. Le fonctionnalisme remplace les « inputs sensoriels » du béhaviorisme par « des inputs sensoriels et des états mentaux » ; il remplace en outre les « dispositions à agir » du béhaviorisme par des « dispositions à agir et à avoir certains états mentaux ». Les fonctionnalistes veulent individuer les états mentaux causalement, et puisque les états mentaux ont des causes et des effets mentaux aussi bien que des causes sensorielles et des effets comportementaux, les fonctionnalistes individuent les états mentaux en partie en termes de relations causales avec d'autres états mentaux. L'une des conséquences de cette différence entre le fonctionnalisme et le béhaviorisme est qu'il peut y avoir des organismes qui pour le second sont dotés d'états mentaux, tandis qu'ils en sont dépourvus pour le premier.

Les conditions nécessaires pour que quelque chose ait des propriétés mentales qui sont postulées par le fonctionnalisme sont donc d'une certaine façon plus contraignantes que celles qui sont postulées par le béhaviorisme. Selon le béhaviorisme, il est nécessaire et suffisant pour désirer B qu'un système soit caractérisé par un certain ensemble (peut-être infini) de relations input-output ; c'est-à-dire que, selon le béhaviorisme, un système ne désire B que dans le cas où un certain ensemble de conditionnels de la forme « si I, il émettra O » sont vrais de lui. Mais selon le fonctionnalisme, un système pourrait avoir ces relations input-output et cependant ne pas désirer B ; car selon lui, le fait qu'un système désire B ou non dépend du fait qu'il ait ou non des états internes ayant certaines relations causales avec d'autres états internes (ainsi qu'avec les inputs et les outputs). Puisque le béhaviorisme n'exige pas d'états internes, il peut y avoir des systèmes dont le béhaviorisme affirme qu'ils ont des états mentaux, tandis que le fonctionnalisme le nie[1]. En d'autres termes, aux yeux du fonctionnalisme, le béhaviorisme est coupable de *libéralisme* — c'est-à-dire d'assigner des propriétés mentales à des choses qui en sont en fait dépourvues.

Malgré cette divergence, l'esprit du fonctionnalisme n'est pas nécessairement radicalement différent de celui du béhaviorisme[2]. Shoemaker (1975) écrit par exemple que « selon l'une de ses interprétations, le fonctionnalisme est, en philosophie de l'esprit, la doctrine que les termes

1. L'inverse est également vrai.
2. En fait, si l'on définit le béhaviorisme comme la théorie que les états mentaux peuvent être définis en termes non mentaux, le fonctionnalisme *est* une variante du béhaviorisme.

mentaux et psychologiques sont, en principe, éliminables d'une certaine façon » (p. 306-307). Les fonctionnalistes ont en effet généralement accordé, dans leur caractérisation fonctionnelle d'un état mental, un traitement différent aux termes qui désignent des états mentaux et à ceux qui désignent les inputs et les outputs. Ainsi, dans la plus simple version que reçoit la théorie en termes de machine de Turing (Putnam, 1967 ; Block et Fodor, 1972), les états mentaux sont identifiés avec l'ensemble des états d'une machine de Turing, qui eux-mêmes sont *implicitement* définis par une table de machine qui mentionne *explicitement* des inputs et des outputs décrits d'une façon non mentaliste.

Dans le fonctionnalisme de Lewis, les termes désignant des états mentaux sont définis au moyen d'une version de la méthode de Ramsey qui permet d'éliminer le recours essentiel à toute terminologie mentale dans les définitions, mais non pas la terminologie input-output. Ainsi, « douleur » est considéré comme synonyme d'une description définie qui contient des termes d'inputs et d'outputs mais aucune terminologie mentale (cf. Lewis, 1972).

De plus, le fonctionnalisme, tant dans sa version machinique que non machinique, a toujours insisté sur le fait que la caractérisation des états mentaux doit contenir des descriptions d'inputs et d'outputs formulées en langage *physique*. Armstrong (1968) écrit par exemple que

> Nous devons distinguer entre le « comportement physique », qui désigne n'importe quelle action ou passion du corps, et le « comportement proprement dit », qui implique une relation avec l'esprit... Or, si dans notre formule « l'état d'une personne apte à adopter une certaine forme de comportement » « comportement » est considéré comme l'équivalent de « comportement proprement dit », alors nous expliquons les concepts mentaux au moyen d'un concept qui fait déjà appel au mental, ce qui est circulaire. Il est donc clair que dans notre formule, « comportement » doit vouloir dire « comportement physique » (p. 84).

Par conséquent, le fonctionnalisme peut être considéré comme « n'épinglant » les états mentaux qu'à la périphérie — c'est-à-dire au moyen d'une caractérisation physique, ou à tout le moins non mentale, des inputs et des outputs. L'une des thèses principales de cet article est que le fonctionnalisme ne parvient pas, pour cette raison, à éviter le problème au nom duquel il condamne à juste titre le béhaviorisme. Le fonctionnalisme est lui aussi coupable de libéralisme, pour les mêmes raisons que le béhaviorisme. A la différence du béhaviorisme toutefois, le fonctionnalisme peut être modifié sans artifice de façon à éviter le libéralisme — mais sans parvenir alors à éviter un autre écueil non moins grave que le premier.

Cet écueil est précisément celui dont le fonctionnalisme montre qu'il fait échouer le *physicalisme*. Par « physicalisme », j'entends la doctrine

selon laquelle la douleur, par exemple, est identique à un état physique (ou physiologique)[1]. Comme l'ont soutenu de nombreux philosophes (en particulier Fodor, 1965 ; Putnam, 1966 ; voir aussi Block et Fodor, 1972), si le fonctionnalisme est vrai, alors le physicalisme est probablement faux. C'est particulièrement clair en ce qui concerne les versions du fonctionnalisme formulées en termes de machine de Turing. N'importe quelle machine de Turing abstraite peut être réalisée par une grande variété de mécanismes physiques ; il est plausible en effet que, pour quelque correspondance établie entre un état de machine de Turing et un état physique configurationnel (ou physiologique) que ce soit, il existe une réalisation possible de cette machine de Turing qui constitue un contre-exemple de cette correspondance (voir Kalke, 1969 ; Gendron, 1971 ; et Mucciolo, 1974 pour une démonstration peu convaincante du contraire ; voir aussi Kim, 1972). Par conséquent, si la douleur est un état fonctionnel, elle ne peut, par exemple, être un état du cerveau, parce que les créatures sans cerveau peuvent réaliser la même machine de Turing que des créatures avec cerveau.

Je dois insister sur le fait que l'argument fonctionnaliste contre le physicalisme ne fait pas simplement appel au fait qu'une machine de Turing abstraite peut être réalisée par des systèmes dont la *composition matérielle* est différente (vois, métal, verre...). Car ce serait alors comme arguer que la température ne peut pas être une grandeur microphysique de ce que plusieurs objets avec *des* microstructures *différentes* peuvent avoir la même température (1972). Les objets avec des microstructures différentes, tels que les objets faits en bois, métal, verre, etc., peuvent avoir plusieurs propriétés microphysiques en commun, telle qu'une énergie cinétique moléculaire de même valeur moyenne. L'argument fonctionnaliste contre le physicalisme est bien plutôt qu'il est difficile de voir comment *il pourrait y avoir* une propriété physique de premier ordre (cf. n. 4) qui ne soit pas triviale et qui soit commune à toutes les réalisations physiques d'un état de machine de Turing et à elles seules. Essayez de trouver un candidat même vaguement plausible ! A tout le moins, il

1. Je parle ici de types d'états mentaux et non d'états particuliers. Dans tout cet article, j'entends par « physicalisme » la doctrine qui affirme que chaque type d'état mental est identique à un certain type d'état physique ; par exemple que la douleur (l'universel) est un état physique. Le physicalisme des particuliers, d'autre part, est la doctrine (plus faible) selon laquelle chaque douleur particulière donnée est un état de tel ou tel type physique. Le fonctionnalisme montre que le physicalisme des types est faux, mais non pas que le physicalisme des particuliers l'est.

Quand je parle de « physicalisme », j'entends un physicalisme du premier ordre, c'est-à-dire la doctrine que la propriété d'éprouver de la douleur est par exemple une propriété physique de premier ordre (au sens de Russel et Whitehead). (Une propriété de premier ordre est une propriété dont la définition n'exige pas de quantifier sur les propriétés ; une propriété de second ordre est une propriété dont la définition exige de quantifier sur les propriétés de premier ordre.) La thèse que le fait d'éprouver de la douleur est une propriété physique de second ordre est en fait une forme (physicaliste) de fonctionnalisme, voir Putnam, 1970.

incombe à ceux qui pensent que de telles propriétés physiques sont concevables de montrer comment nous pouvons en concevoir une.

En d'autres termes, selon le fonctionnalisme, le physicalisme est une doctrine chauviniste : il refuse d'accorder des propriétés mentales à des systèmes qui en sont en fait dotées. En disant que les états mentaux sont des états du cerveau, par exemple, les physicalistes excluent injustement les pauvres créatures dépourvues de cerveau et néanmoins dotées d'esprit.

La seconde thèse principale de cet article est que l'argument au nom duquel le fonctionnalisme condamne le physicalisme peut être aussi bien employé pour réfuter le fonctionnalisme lui-même ; en fait, n'importe quelle version du fonctionnalisme qui évite le libéralisme succombe, comme le physicalisme, au chauvinisme.

Cet article a trois parties. La première soutient que le fonctionnalisme est coupable de libéralisme, la seconde que l'une des manières d'éviter le libéralisme est de resserrer ses liens avec la psychologie empirique, et la troisième qu'aucune version du fonctionnalisme ne parvient à éviter à la fois le libéralisme et le chauvinisme.

1.1. *Précisions sur la nature du fonctionnalisme*

Pour tenter d'introduire un peu d'ordre au sein de la variété déconcertante de théories fonctionnalistes, on peut distinguer entre celles qui sont formulées en termes de machine de Turing et celles qui ne le sont pas.

Une table de machine de Turing énumère un ensemble fini d'états, $S_1 \ldots S_n$; d'inputs, $I_1 \ldots I_m$; et d'outputs, $O_1 \ldots O_p$. Elle spécifie en outre un ensemble de conditionnels de la forme : si la machine est dans l'état S_i et reçoit l'input I_j, elle émet l'output O_k et passe dans l'état S_1. En d'autres termes, étant donné un état quelconque et un input, la table spécifie un output et l'état suivant. Tout système doté d'un ensemble d'inputs, d'outputs et d'états reliés de la manière spécifiée par la table est décrit par cette table et est une réalisation de l'automate abstrait qu'elle spécifie.

Pour être capable de calculer n'importe quelle fonction récursive, une machine de Turing doit pouvoir contrôler son input de certaines façons. Dans les formulations habituelles, l'output de la machine de Turing est considéré comme ayant deux composants. La machine imprime un symbole sur un ruban, puis fait bouger le ruban, mettant ainsi un nouveau symbole sous l'œil du détecteur d'inputs. Pour qu'une machine de Turing soit aussi puissante que possible, le ruban doit être sans fin dans l'une au moins de ses deux directions et doit pouvoir être bougé dans les deux sens. Si la machine n'exerce aucun contrôle sur le ruban, c'est un « transducteur fini », une espèce assez limitée de machine de Turing. Il

n'est pas même besoin de considérer que les transducteurs finis sont dotés d'un ruban. Ceux qui croient à la vérité du fonctionnalisme machinique doivent considérer la question de la nature exacte du type d'automates que nous sommes comme une question empirique importante. Si nous sommes des machines de Turing dotées du maximum de puissance possible, l'environnement doit faire partie du ruban.

Dans l'une de ses plus simples versions, le fonctionnalisme machinique (cf. Block et Fodor, 1972) affirme que chaque système doté d'états mentaux peut être décrit par au moins une table de machine de Turing d'une espèce qui peut être spécifiée, et que chaque type d'état mental du système est identique à l'un des états de la table de machine. Soit par exemple la machine de Turing décrite dans la table 1 (cf. Nelson, 1975).

TABLE 1

	S_1	S_2
Input : pièce de 5 cents	N'émettre aucun output Passer en S_2	Emettre un coca-cola Passer en S_1
Input : pièce de 10 cents	Emettre un coca-cola Rester en S_1	Emettre un coca-cola et une pièce de 5 cents Passer en S_1

Il est possible de se faire une idée approximative de cette version simple du fonctionnalisme machinique en considérant que S_1 = le désir de pièce de 5 cents, et S_2 = le désir de pièce de 10 cents. Bien entendu, aucun fonctionnaliste ne prétend qu'un distributeur de coca-cola ne désire quoi que ce soit. La version simple du fonctionnalisme machinique que je viens de mentionner fait appel à une table de machine hypothétique bien plus complexe. Il faut souligner que le fonctionnalisme machinique caractérise les inputs et les outputs de manière explicite, et les états internes de manière implicite. (Putnam, 1967, cf. *supra,* p. 329) écrit : « Les S_i, une fois encore, ne sont spécifiés qu'implicitement par la description, c'est-à-dire qu'ils ne sont caractérisés que par *l'ensemble* des probabilités de transition donné par la table de machine. » Pour être décrit par cette table de machine, un mécanisme doit accepter les pièces de 5 et 10 cents comme inputs et fournir des pièces de 5 cents et des coca-cola comme outputs. Mais les états S_1 et S_2 peuvent avoir à peu près n'importe quelle nature (même non physique), aussi longtemps que cette nature permet de connecter les états les uns avec les autres ainsi qu'avec les inputs et les outputs spécifiés dans la table de machine. Nous ne

connaissons de S_1 et de S_2 que ces relations ; on peut donc dire que le fonctionnalisme réduit le mental à des structures input-output. Cet exemple permet de faire sentir la force de l'argument fonctionnaliste contre le physicalisme. Essayez de trouver une propriété physique de premier ordre (voir n. 4) qui puisse être partagée par toutes les réalisations (et elles seules) de cette table de machine !

On peut aussi prendre pour critère de classification des fonctionnalistes le fait qu'ils considèrent ou non les identités fonctionnelles comme relevant de la psychologie *a priori* ou de la psychologie empirique... Les fonctionnalistes *a priori* (tels que Smart, Armstrong, Lewis, Shoemaker) sont les héritiers des béhavioristes logiques. Il tendent à considérer les analyses fonctionnelles comme des analyses du sens des termes mentaux, tandis que les fonctionnalistes empiriques (tels que Fodor, Putnam, Harman) considèrent les analyses fonctionnelles comme des hypothèses scientifiques. Dans ce qui suit, je désignerai la première de ces deux positions au moyen du terme de Fonctionnalisme, et la seconde au moyen du terme de Psychofonctionnalisme). (J'emploierai celui de fonctionnalisme — avec un *f* minuscule — pour désigner une position neutre sur la question qui divise le Fonctionnalisme et le Psychofonctionnalisme. Quand il sera indispensable de distinguer entre les deux, j'emploierai toujours les majuscules.)

Les notions de Fonctionnalisme et de Psychofonctionnalisme, ainsi que la différence qui les sépare, peuvent être clarifiées au moyen de la notion d'énoncé de Ramsey d'une théorie psychologique. Les termes désignant des états mentaux qui apparaissent dans une théorie psychologique peuvent être définis de différentes façons au moyen de l'énoncé de Ramsey d'une théorie. Toute théorie fonctionnaliste de l'identité peut être considérée comme définissant un ensemble d'états fonctionnels au moyen de l'énoncé de Ramsey de cette théorie psychologique — chaque état mental correspondant à un état fonctionnel. L'état fonctionnel correspondant à la douleur sera appelé le « corrélat fonctionnel de Ramsey » de la douleur par rapport à la théorie psychologique en question. La distinction entre le Fonctionnalisme et le Psychofonctionnalisme peut alors être redéfinie en termes de corrélat fonctionnel de Ramsey par rapport à une théorie : le Fonctionnalisme identifie un état mental S avec le corrélat fonctionnel de Ramsey par rapport à une théorie psychologique *du sens commun* ; le Psychofonctionnalisme identifie S avec le corrélat fonctionnel de Ramsey par rapport à une théorie psychologique *scientifique*.

Cette différence entre le Fonctionnalisme et le Psychofonctionnalisme entraîne une différence au niveau de la spécification des inputs et des outputs. Les Fonctionnalistes adoptent une spécification qui puisse raisonnablement faire partie de la connaissance du sens commun ; les Psychofonctionnalistes ne sont soumis à aucune restriction de ce genre. Quoique les deux groupes insistent sur le caractère physique — ou du moins non mental — de la spécification des inputs et des outputs, les Fonctionnalistes doivent faire appel à des classifications qui soient observables d'un point de

vue externe (par exemple, caractérisation des inputs en termes d'objets présents dans l'entourage de l'organisme, caractérisation des outputs en termes de mouvements des parties du corps). Les Psychofonctionnalistes, d'autre part, ont la possibilité de spécifier les inputs et les outputs en termes de paramètres internes, tels que des signaux dans les neurones d'inputs et dans les neurones d'outputs.

Soit T une théorie psychologique soit du sens commun, soit de la psychologie scientifique. T contient des généralisations de la forme : toute personne qui est dans l'état w et reçoit un input x émet un output y, et passe dans l'état z. Ecrivons T sous la forme suivante :

$$T(S_1 \ldots S_n, I_1 \ldots I_k, O_1 \ldots O_n)$$

où les S sont les états mentaux, les I les inputs, et les O les outputs. Les « S » doivent être compris comme des constantes d'états mentaux, et non comme des variables, et il en va de même pour les « I » et les « O ». Aussi pourrait-on encore écrire T sous la forme suivante : T (douleur..., lumière de 400 nm entrant dans l'œil gauche..., le gros orteil gauche se déplace de 1 cm vers la gauche...).

Pour obtenir l'énoncé de Ramsey de T, il faut remplacer les termes désignant les états mentaux — *mais non pas les termes désignant les inputs et les outputs* — par des variables, et préfixer un quantificateur existentiel pour chaque variable :

$$\exists F_1 \ldots \exists F_n\, T(F_1 \ldots F_n I_1 \ldots I_k, O_1 \ldots O_m).$$

Si « F_{17} » est la variable qui a remplacé le mot « douleur » lors de la formation de l'énoncé de Ramsey, la douleur peut alors être définie comme suit :

$$x \text{ éprouve de la douleur} \Leftrightarrow \exists F_1 \ldots \exists F_n$$
$$T[F_1 \ldots F_n, I_1 \ldots I_n, O_1 \ldots O_m) \text{ et } x \text{ a } F_{17}].$$

Le corrélat fonctionnel de Ramsey de la douleur est la propriété exprimée par le prédicat à la droite de ce biconditionnel. Il faut souligner que ce prédicat contient des constantes d'inputs et d'outputs, mais aucune constante mentale, puisque les constantes mentales ont été remplacées par des variables. Le corrélat fonctionnel de Ramsey de la douleur est défini en termes d'inputs et d'outputs, mais non pas en termes mentaux.

Admettons par exemple que T soit la théorie que la douleur est causée par une blessure de la peau et cause une irritation ainsi que l'émission de « Aïe », et que l'irritation, à son tour, cause un froncement des sourcils. La définition à la Ramsey serait alors la suivante :

> x éprouve de la douleur ⇔ Il y a deux états (propriétés), dont le premier est causé par une blessure de la peau et cause à la fois l'émission de « aïe » et le second état, et le second état cause un froncement des sourcils, et x est dans le premier état.

Le corrélat fonctionnel de Ramsey de la douleur par rapport à cette « théorie » est la propriété d'être dans un état causé par la blessure de la peau et qui cause l'émission de « aïe » et un autre état, qui à son tour cause le froncement des sourcils. (Notez que les mots « douleur » et « souci » ont été remplacés par des variables, mais que les termes désignant les inputs et les outputs demeurent.)

Le corrélat fonctionnel de Ramsey d'un état S est un état qui a beaucoup en commun avec S. A savoir les propriétés structurales spécifiées par la théorie T. Mais, il y a deux raisons pour lesquelles il est naturel de supposer que S et son corrélat fonctionnel sont distincts. En premier lieu, le corrélat fonctionnel de Ramsey par rapport à la théorie T ne peut « inclure » au mieux que les aspects de S dont T rend compte ; tout aspect qui n'est pas saisi par T est laissé de côté. En second lieu, le corrélat fonctionnel de Ramsey peut même laisser de côté certains aspects dont T rend compte, car la définition de Ramsey ne contient pas le vocabulaire « théorique » de T. La théorie mentionnée en guise d'exemple au paragraphe précédent n'est vraie que des organismes qui sentent la douleur — mais simplement en raison de la manière dont elle fait usage du mot « douleur ». Cependant, le prédicat qui exprime le corrélat fonctionnel de Ramsey ne contient pas ce mot (puisqu'il a été remplacé par une variable), et peut donc être vrai de choses qui ne ressentent pas la douleur. Il serait facile de fabriquer une machine simple qui ait une peau artificielle, un sourcil, un « aïe » enregistré, et deux états qui satisfont les relations causales mentionnées, mais qui ne ressente pas la douleur.

L'hypothèse audacieuse du fonctionnalisme est que pour *une* certaine théorie psychologique, cette supposition naturelle qu'un état et son corrélat fonctionnel de Ramsey sont distincts est fausse. Le fonctionnalisme dit qu'il existe une théorie telle que la douleur, par exemple, *est* le corrélat fonctionnel de Ramsey par rapport à cette théorie.

Un dernier point préliminaire : j'ai donné l'impression trompeuse que le fonctionnalisme identifie *tous* les états mentaux avec des états fonctionnels. C'est à l'évidence une exagération. Admettons que X soit une réplique cellule par cellule de vous qui vient d'être créée (qui, bien entendu, est fonctionnellement équivalente à vous). Peut-être vous souvenez-vous de votre Bar-Mitzvah. Mais X ne s'en souvient pas. En effet, quelque chose peut être fonctionnellement équivalent à vous et cependant ne pas réussir à savoir ce que vous savez, ou à (verbe) ce que vous (verbe), pour un grand nombre de verbes signifiant la « réussite ». Pire encore, si Putnam (1975*b*) a raison de dire que « les sens ne sont pas dans la tête », des systèmes fonctionnellement équivalents à vous peuvent, pour des raisons similaires, ne pas réussir à avoir plusieurs de vos attitudes propositionnelles. Supposons que vous croyiez que l'eau est mouillée. Selon les arguments plausibles avancés par Putnam et Kripke, une des conditions qui vous permettent de croire que l'eau est mouillée est une certaine connexion causale entre vous et l'eau. Votre « jumeau » sur

la terre jumelle, qui est lié de manière similaire à du XYZ plutôt qu'à du H_2O, ne croirait pas que l'eau est mouillée.

Le fonctionnalisme ne peut être défendu que si on considère qu'il ne s'applique qu'à une sous-classe des états mentaux, ces états « étroits » qui sont tels que les conditions de vérité de leur application sont en quelque sorte « dans la personne ». Mais même si l'on suppose qu'il est possible de formuler une notion satisfaisante d'étroitesse d'un état psychologique, l'intérêt du fonctionnalisme s'en trouve peut-être diminué. Je ne mentionne ce problème que pour le laisser de côté.

Dans ce qui suit, je considérerai le fonctionnalisme comme une doctrine concernant tous les états mentaux « étroits ».

1.2. *Les robots dirigés par des homoncules*

Dans cette section, je vais décrire une classe de mécanismes qui constituent apparemment une source d'embarras pour toutes les versions du fonctionnalisme que j'ai mentionnées, en ce qu'ils révèlent que le fonctionnalisme est coupable de libéralisme — c'est-à-dire qu'il considère comme dotés de propriétés mentales des systèmes qui en sont dépourvus.

Soit la version simple du fonctionnalisme machinique qui a été décrite plus haut. Elle affirme que chaque système doté d'états mentaux est décrit par au moins une table de machine de Turing, et que chaque état mental de ce système est identique à l'un des états spécifiés par la table de machine. Je supposerai que les inputs et les outputs sont caractérisés en termes d'impulsions neuronales dans les organes sensoriels et dans les neurones moteurs. Ce qui n'implique nullement que les considérations qui vont suivre valent plus pour le Psychofonctionnalisme que pour le Fonctionnalisme. Ainsi que je l'ai déjà indiqué, toute version du fonctionnalisme suppose *une certaine* caractérisation des inputs et des outputs. On pourrait donc aussi bien retenir ici une caractérisation fonctionnaliste.

Imaginez un corps qui soit extérieurement comme un corps humain, par exemple le vôtre, mais intérieurement tout à fait différent. Les neurones des organes sensoriels sont reliés à une rangée de voyants lumineux dans une cavité de la tête. Une série de boutons sont par ailleurs connectés avec les neurones moteurs. A l'intérieur de la cavité siège un groupe de petits hommes. Chacun a une tâche très simple : implémenter un des « carrés » d'une table de machine qui est une description adéquate de vous. Sur un mur est accroché un tableau sur lequel on a affiché une carte des états ; c'est-à-dire une carte sur laquelle figure un symbole désignant l'un des états spécifiés par la table de machine. Voici ce que font les petits hommes : quand sur la carte figure un « G », les petits hommes qui implémentent les carrés G — ils s'appellent eux-mêmes « les hommes G » — sont alertés. Supposons que la lumière représentant l'in-

put I_{17} s'allume. L'un des hommes G a pour seule tâche la mission suivante : quand sur la carte figure « G » et que la lumière I_{17} s'allume, appuyer sur le bouton d'output O_{191} et inscrire un « M » sur la carte des états. Cet homme G n'est que rarement appelé à exercer sa mission. En dépit du faible niveau d'intelligence exigé de chacun de ces petits hommes, l'ensemble du système parvient à simuler votre comportement parce que l'organisation fonctionnelle qu'ils ont été entraînés à réaliser est la vôtre. Une machine de Turing peut être représentée par un ensemble fini de quadruples (ou de quintuples si l'input est divisé en deux parties) : l'état actuel, l'input actuel ; l'état suivant, l'output suivant. Chaque petit homme a une tâche correspondant à un seul quadruple. Grâce à leurs efforts, le système réalise la même table de machine (raisonnablement adéquate) que vous et est donc fonctionnellement équivalent à vous[1].

Je vais maintenant décrire un cas de stimulation par homoncules qui ait plus de chance d'être nomologiquement possible. Combien d'homoncules sont nécessaires pour cela ? Un milliard sera peut-être suffisant.

Supposons que nous parvenions à convertir le gouvernement de Chine au fonctionnalisme et à le convaincre de réaliser un esprit humain pendant une heure. Nous fournissons à chacun des mille millions de Chinois (c'est pour cela que j'ai choisi la Chine) un émetteur-récepteur spécialement conçu à cet effet qui les relie les uns avec les autres et avec le corps artificiel mentionné dans l'exemple précédent. Nous remplaçons donc chacun des petits hommes par un citoyen chinois équipé d'une radio. Les lettres sont disposées non plus sur un tableau mais sur une série de satellites visibles de n'importe quel endroit de la Chine.

Le système formé par ce milliard d'individus communiquant les uns avec les autres et par les satellites joue le rôle d'un « cerveau » externe connecté au corps artificiel par radio. Il n'y a rien d'absurde à imaginer quelqu'un relié à son cerveau par radio. Le jour viendra peut-être où nos cerveaux pourront être périodiquement détachés pour être réparés et nettoyés. Cela pourrait être réalisé par exemple en traitant les neurones qui relient le corps au cerveau au moyen d'une substance chimique qui leur permette d'être aussi flexibles que des élastiques, de façon à ce qu'aucune connexion entre le corps et le cerveau ne soit rompue. Mais des hommes d'affaires intelligents s'apercevraient vite qu'il est plus facile d'attirer le client en remplaçant les neurones élastifiés par des liaisons radio, de sorte que les cerveaux puissent être nettoyés sans que le corps de son possesseur ait l'inconvénient d'être immobilisé.

Il n'est pas du tout évident que le système chinois soit physiquement

1. L'idée qui est à la base de cet exemple a sa source dans l'article de Putnam, 1967. Elle a fait l'objet de maintes conversations avec Harty Field que je tiens ici à remercier. J'examine dans la section 1.3 la façon dont Putnam tente de défendre le fonctionnalisme contre les problèmes posés par ce genre de cas.

impossible. Et il pourrait être fonctionnellement équivalent à vous pendant une courte période de temps, disons une heure.

« Mais, pourriez-vous objecter, comment quelque chose pourrait-il être fonctionnellement équivalent à moi pendant *une heure* ? Mon organisation fonctionnelle ne détermine-t-elle pas la façon dont, par exemple, je réagirai au fait de ne rien faire d'autre pendant une semaine que de lire le *Reader's Digest* ? » Souvenez-vous qu'une table de machine spécifie un ensemble de conditionnels de la forme : si la machine est dans l'état S_i et reçoit un input I_j, elle émet un output O_k et passe dans l'état S_l. Ces conditionnels doivent être entendus en un sens *subjonctif*. Ce qui confère une organisation fonctionnelle à un système à un moment donné n'est pas seulement ce que le système *fait* à ce moment-là, mais ce sont aussi les contrefactuels qui sont vrais de lui à ce moment-là : c'est-à-dire ce qu'il *aurait* fait (et ce que ses transitions d'état à état auraient été) s'il avait reçu un input différent ou s'il avait été dans un état différent. S'il est vrai de ce système au moment *t* qu'il *obéirait* à une certaine table de machine quel que soit son état et quels que soient ses inputs, alors le système est décrit à *t* par cette table de machine (et réalise un automate abstrait spécifié par la table), même s'il n'existe que pendant un instant. Au cours de l'heure pendant laquelle le système chinois est « branché », il *a* effectivement un ensemble d'inputs, d'outputs et d'états dont de tels conditionnels subjonctifs sont vrais. C'est par là que n'importe quel ordinateur réalise l'automate abstrait qu'il réalise.

Il y a bien entendu des signaux auxquels le système répondrait et auxquels vous ne répondriez pas — par exemple une interférence radio massive ou un débordement de la rivière Yangtze. De tels événements pourraient entraîner un dysfonctionnement, et par là faire échouer la simulation, tout comme une bombe placée dans un ordinateur peut l'empêcher de réaliser la table de machine que sa construction doit le faire réaliser. Mais de même que l'ordinateur *sans* la bombe *peut* réaliser la table de machine, de même le système composé du peuple chinois et du corps artificiel peut réaliser la table de machine aussi longtemps que ne survient aucune interférence catastrophique, telle qu'une inondation, etc.

« Mais, pourriez-vous encore objecter, il y a une différence entre mettre une bombe dans un ordinateur et mettre une bombe dans un système chinois ; dans le second cas (mais pas dans le premier), les inputs tels qu'ils sont spécifiés dans la table de machine peuvent être la cause du dysfonctionnement. Une activité neuronale inhabituelle dans les organes sensoriels des résidents de la province de Chungking causée par une bombe ou un débordement du Yangtze peut par exemple détraquer le système. »

Réponse : En spécifiant de quel système on parle, il faut aussi délimiter une certaine catégorie d'inputs et d'outputs. Je ne considère comme inputs et outputs que l'activité neuronale normale du corps artificiel connecté par radio au peuple de Chine. Les signaux neuronaux dans le

peuple du Chungking ne sont pas plus des inputs de ce système qu'un ruban coincé entre les relais à l'intérieur d'un ordinateur ne compte comme un input de cet ordinateur.

Bien entendu, l'objet formé par le peuple de Chine et le corps artificiel peut recevoir *d'autres* descriptions en termes de machine de Turing en vertu desquelles les signaux neuronaux des habitants de Chungking compteraient comme des inputs. Un tel système (c'est-à-dire le système caractérisé par cette nouvelle description en termes de machine de Turing) ne serait pas fonctionnellement équivalent avec vous. Pareillement, n'importe quel ordinateur commercial peut être redécrit de telle façon qu'un ruban coincé à l'intérieur compte comme l'un de ses inputs. En décrivant un objet comme une machine de Turing, on trace une frontière entre l'intérieur et l'extérieur. (Si nous ne comptons que des impulsions neuronales comme inputs et outputs, cette ligne passe à l'intérieur du corps ; si au contraire nous ne comptons que des stimulations périphériques comme inputs, cette ligne coïncide avec la peau.) En décrivant le système chinois comme une machine de Turing, j'ai tracé cette ligne de manière à satisfaire un certain type de description fonctionnelle — un type que vous satisfaites *également* et qui, selon le fonctionnalisme, justifie l'attribution de propriétés mentales. Le fonctionnalisme ne prétend pas qu'il existe pour chaque système mental une table de machine d'une espèce qui justifie des attributions de propriétés mentales par rapport à *n'importe quelle* spécification d'inputs et d'outputs, mais seulement par rapport à *une certaine* spécification.

Objection : Le système chinois fonctionnerait trop lentement. Le type d'événements et de processus avec lesquels nous sommes normalement en contact sont bien trop rapides pour que le système puisse les détecter. Par conséquent nous serions incapables de converser avec lui, de jouer au bridge avec lui, etc.

Réponse : On ne voit pas pourquoi la rapidité du système importerait d'une manière ou d'une autre. Est-il véritablement contradictoire ou insensé de supposer que nous pourrions rencontrer une race d'êtres intelligents avec lesquels nous ne pourrions communiquer qu'au moyen d'un procédé comme l'accéléré ? Quand nous observerions de telles créatures, elles sembleraient presque inanimées. Mais quand nous les verrions sur des films projetés en accéléré, nous les verrions converser les unes avec les autres. En fait, nous nous apercevrions qu'elles disent précisément que nous n'avons du sens pour elles que sur des films regardés au ralenti. Accorder une telle importance à la rapidité du système, c'est semble-t-il faire preuve d'un béhaviorisme élémentaire.

Ce qui fait du système dirigé par des homoncules (les deux exemples ne devant être considérés que comme des variantes d'un seul système) qui vient d'être décrit *en apparence* un contre-exemple du fonctionnalisme (machinique), c'est qu'il y a semble-t-il une raison de douter qu'il ait le moindre état mental — en particulier qu'il ait ce que les philosophes ont

appelé tantôt des « états qualitatifs », tantôt des « sensations brutes », tantôt des « qualités phénoménologiques *immédiates* ». (Si vous me demandez : qu'est-ce que ce que les philosophes ont appelé des états qualitatifs ? Je vous répondrai — à demi sérieusement : Comme disait Louis Armstrong quand on lui demandait ce qu'était le jazz, « si vous le demandez, c'est que vous ne le saurez jamais ».) Dans la terminologie de Nagel (1974), on est en droit de douter qu'être comme un système dirigé par des homoncules ressemble à quoi que ce soit[1].

1.3. *La proposition de Putnam*

L'un des moyens qui s'offrent aux fonctionnalistes pour résoudre le problème posé par le contre-exemple du système dirigé par des homoncules consiste tout simplement à l'écarter par une stipulation *ad hoc*. Un fonctionnaliste pourrait par exemple stipuler que deux systèmes ne peuvent être fonctionnellement équivalents si l'un contient des parties qui ont une organisation fonctionnelle caractéristique de celle des êtres sensibles et l'autre non. Dans l'article où il fait l'hypothèse que la douleur est un état fonctionnel, Putnam stipule « qu'aucun organisme capable de sentir de la douleur ne peut être décomposé en parties qui possèdent chacune séparément des Descriptions » (en termes de machine de Turing pouvant être dans l'état fonctionnel que Putnam identifie avec la douleur). Le but de cette restriction est « d'écarter la possibilité pour des "organismes" (pour autant qu'on puisse les considérer comme tels) tels que les essaims d'abeilles de la classe des individus capables de ressentir de la douleur » (Putnam, 1967, cf. *supra*, p. 330).

La condition posée par Putnam pourrait par exemple être satisfaite de la manière suivante : un organisme qui sent la douleur ne peut être décomposé en parties qui ont *toutes* une organisation fonctionnelle caractéristique de celle des êtres sensibles. Mais cela ne permettrait pas d'écarter l'exemple du système dirigé par des homoncules, vu qu'il possède des parties sensibles et non sensibles, tels que le corps mécanique et les organes sensoriels. Il ne servirait à rien d'adopter la thèse complètement opposée et d'exiger *qu'aucune* partie réelle ne soit sensible. Sans quoi les femmes enceintes et ceux qui ont des parasites sensibles ne pourraient être considérés comme des organismes capables de ressentir la douleur. Ce qui semble important dans la simulation par

1. Shoemaker (1975) soutient (en réponse à Block et Fodor, 1972) que l'absence de *qualia* est logiquement impossible ; c'est-à-dire qu'il est logiquement impossible que deux systèmes soient dans le même état fonctionnel et que cependant seul l'un des deux ait un contenu qualitatif. N.d.T. : L'article auquel il est ici fait allusion — What is it like to be a bat ? — figure dans le recueil *Mortal Questions*, traduction française de C. Engel-Tiercelin et P. Engel sous le titre *Questions mortelles*, PUF, « Philosophie d'aujourd'hui », 1983.

homoncules que j'ai décrite, c'est le fait que des êtres sensibles *jouent un rôle crucial* dans l'organisation fonctionnelle que possède la machine. Ce qui suggère une version de la proposition de Putnam qui exige qu'un organisme capable de ressentir la douleur ait une certaine organisation fonctionnelle, mais n'ait par contre aucune partie 1 / qui possède elle-même la même sorte d'organisation fonctionnelle, et 2 / qui joue un rôle crucial dans le fait que l'ensemble du système ait l'organisation fonctionnelle qu'il a.

Quoique cette proposition fasse appel à la notion vague de « rôle crucial », elle est assez précise pour permettre de voir qu'elle ne marche pas. Supposons qu'il y ait une partie de l'univers qui contienne une matière tout à fait différente de la nôtre, une matière qui est indéfiniment divisible. Dans cette partie de l'univers se trouvent des créatures intelligentes de toutes sortes de tailles, y compris des créatures anthropomorphes beaucoup plus petites que nos particules élémentaires. Au cours d'une expédition intergallaxique, ces êtres découvrent l'existence de notre type de matière. Pour des raisons qui ne sont connues que d'elles seules, elles décident de consacrer une centaine d'années à créer, à partir de *leur* matière, des substances qui aient les caractéristiques chimiques et physiques de *nos* éléments (sauf au niveau des particules infra-élémentaires). Elles construisent des hordes de vaisseaux spatiaux de différentes espèces, à peu près de la taille de nos électrons, protons, et autres particules élémentaires, et les conduisent de façon à imiter le comportement de ces derniers. Les vaisseaux contiennent également des générateurs qui produisent le même type de radiation que nos particules élémentaires. Sur chaque vaisseau se trouve une équipe d'experts sur la nature de ces particules. Tout ceci a pour fin de produire une quantité énorme (selon nos standards) de substances dotées des caractéristiques chimiques et physiques de l'oxygène, du carbone, etc. Peu après l'achèvement de ce projet, vous menez une expédition jusqu'à la partie en question de l'univers et vous découvrez « l'oxygène », « le carbone », etc. Ignorants de la véritable nature de ces éléments, vous décider d'y établir une colonie et vous les utilisez pour faire pousser des plantes capables de nourrir cette colonie, de produire « l'air » qui lui permet de respirer, etc. Puisque les molécules qui composent un individu sont constamment échangées avec celles de son environnement, les membres de la colonie et vous-mêmes finissez par être composés (en quelques années) essentiellement de la « matière » composée des individus minuscules qui sont dans les vaisseaux spatiaux. Seriez-vous moins capables de ressentir de la douleur, de penser, etc., simplement parce que la matière dont vous êtes composé contient (et dépend pour ce qui est de ses caractéristiques) des êtres qui eux-mêmes ont une organisation fonctionnelle caractéristique de celle des êtres sensibles ? Je ne crois pas. Les mécanismes électrochimiques de base au moyen desquels la synapse opère sont mainte-

nant assez bien compris. Autant que je sache, les changements qui n'affectent pas ces mécanismes n'affectent pas les opérations du cerveau, et donc les propriétés mentales. Les mécanismes électrochimiques de vos synapses ne subiraient aucune modification si la nature de votre matière changeait[1].

Il est intéressant de comparer cet exemple avec les précédents. D'où nous vient notre intuition que les simulations dirigées par des homoncules n'ont pas de propriétés mentales ? Il est naturel de supposer que c'est parce que nous pensons qu'ils ont *trop* de structure mentale interne. Les petits hommes peuvent parfois s'ennuyer, parfois s'exciter. Nous pouvons même imaginer qu'ils délibèrent quant à la meilleure façon de réaliser l'organisation fonctionnelle qu'ils réalisent, et qu'ils y apportent des modifications destinées à leur donner plus de loisir. Mais l'exemple des particules élémentaires composées d'homoncules suggère que cette supposition naturelle est en fait fausse. Ce qui semble important c'est la question de savoir *comment* les propriétés mentales des parties contribuent au fonctionnement du tout.

Il y a une différence majeure entre la première et la seconde espèce de cas. Dans la seconde, le changement qui s'opère en vous quand votre matière se transforme progressivement en homoncules n'entraîne de différence ni dans votre traitement psychologique (c'est-à-dire dans votre traitement de l'information) ni dans votre traitement neurologique, mais seulement dans votre traitement microphysique. Aucune technique propre à la psychologie ou à la neurophysiologie humaine ne révélerait une quelconque différence en vous. Cependant, les simulations dirigées par des petits hommes du début de cet article ne sont pas des choses auxquelles les théories neurophysiologiques qui sont vraies de nous s'appliquent, et *si elles sont interprétées comme des simulations Fonctionnelles* (plutôt que Psychofonctionnelles), il n'est pas nécessaire que ce soit des choses auxquelles les théories psychologiques (théories du traitement de l'information) qui sont vraies de nous s'appliquent. Cette différence suggère que nos intuitions sont en partie contrôlées par l'idée, qui n'est pas déraisonnable, que nos états mentaux dépendent du fait que nous avons la psychologie et/ou neurophysiologie que nous avons. De sorte qu'une chose qui diffère nettement de nous de ce double point de vue (il s'agit je le rappelle d'une simulation Fonctionnelle, plutôt que Psychofonctionnelle) ne doit pas être supposée pourvue de propriétés mentales pour la simple raison qu'elle est désignée comme fonctionnellement équivalente avec nous.

1. Puisqu'il y a une différence entre le rôle des petits hommes dans la production de votre organisation fonctionnelle dans le cas que je viens de décrire et dans celui qui précède, la condition posée par Putnam doit vraisemblablement pouvoir être reformulée de manière à éliminer le précédent sans éliminer celui-ci. Mais ce serait là une manœuvre parfaitement *ad hoc*.

1.4. *Le doute n'est-il qu'apparent ?*

L'argument de l'absence des *qualia* fait appel à l'intuition que les simulations par homoncules manquent de propriétés mentales ou du moins de *qualia*. J'ai dit que cette intuition conduisait apparemment à mettre en doute la vérité du fonctionnalisme. Mais des intuitions que ne viennent étayer aucun argument de principe peuvent difficilement être considérées comme fondées. En fait, des intuitions incompatibles avec une théorie bien étayée (telle que l'intuition précopernicienne que la terre ne se meut pas) ne tardent pas, Dieu merci, à s'évanouir. Même dans des domaines comme la linguistique, où les données sont essentiellement faites d'intuitions, on rejette souvent (pour des raisons théoriques) les intuitions que des phrases comme les suivantes ne sont pas grammaticales :

— The horse raced past the barn fell.
— The boy the girl the cat bit scratched died[1].

Ces phrases sont en réalité grammaticales, quoique difficiles à traiter[2].

Les appels à l'intuition, quand il s'agit de juger de la possession de propriétés mentales, sont cependant *particulièrement* sujets à caution. Intuitivement, *aucun* mécanisme physique ne semble être un candidat très plausible au titre de siège des *qualia,* le *cerveau* moins que tout autre. Est-ce qu'un bout de matière grise tremblotante est intuitivement plus adéquat qu'une compagnie de petits hommes ? Mais alors, le doute que les systèmes dirigés par des cerveaux aient des *qualia* n'est-il pas qu'apparent ?

Il existe toutefois une différence importante entre les systèmes dirigés par des cerveaux et les systèmes dirigés par des homoncules. Puisque nous savons que *nous sommes des systèmes dirigés par des cerveaux* et que *nous* avons des *qualia,* nous savons que les systèmes gouvernés par des cerveaux ont des *qualia*. De sorte que, même si nous n'avons aucune théorie des *qualia* qui permette d'expliquer comment cela est *possible,* nous avons pourtant de très fortes raisons de rejeter tous les doutes qui pourraient surgir à l'égard des systèmes dirigés par des cerveaux. Bien entendu, mon argument est par là rendu en partie *empirique* — il dépend de la connaissance de ce qui nous fait fonctionner. Mais puisque c'est là une

1. Littéralement : 1) Le cheval entraîné devant la grange est tombé ; 2) Le garçon, que la fille, que le chat a mordue, a griffé, est mort. *(N.d.T.)*

2. Comparez la première phrase avec « Le poisson mangé à Boston puait ». La raison pour laquelle raced fait difficulté est qu'on l'interprète naturellement plutôt comme une forme active que passive. Voir Fodor *et al.,* 1974, p. 360. Pour une justification de la grammaticalité de la seconde phrase, voir Fodor et Garrett, 1967 ; Bever, 1970 ; et Fodor *et al.,* 1974.

connaissance que nous possédons, cette dépendance ne saurait être considérée comme un défaut[1].

Il existe une autre différence entre ceux qui, comme nous, ont la tête pleine de matière grise et ceux qui ont la tête pleine d'homoncules : ces derniers sont des systèmes conçus pour nous imiter, mais nous ne sommes aucunement conçus pour imiter quelque chose (je m'appuie ici une fois encore sur un fait empirique). Ce qui prévient toute tentative de faire appel à la notion d'inférence à la meilleure explication pour le cas des *qualia* des têtes remplies d'homoncules. La meilleure façon d'expliquer les cris et les grimaces des têtes pleines d'homoncules n'est pas qu'elles ressentent de la douleur, mais qu'elles ont été conçues pour imiter nos cris et nos grimaces.

Certains semblent penser que le comportement subtil et complexe des têtes pleines d'homoncules (comportement tout aussi complexe et subtil — même aussi « sensible » aux caractéristiques de l'environnement, humain et non humain — que votre comportement) constitue une raison suffisante pour écarter le doute que celles-ci ont des *qualia*. Mais ce n'est là rien d'autre que du béhaviorisme élémentaire.

Mon argument contre le Fonctionnalisme dépend du principe suivant : si une doctrine conduit à une conclusion absurde qu'il n'y a aucune raison indépendante d'adopter, et s'il n'y a aucun moyen de dissiper cette absurdité ou de montrer qu'elle est erronée ou sans pertinence, ni aucune bonne raison de croire à la doctrine qui conduit à une telle absurdité en premier lieu, alors cette doctrine doit être récusée. Je prétends qu'il n'y a aucune raison indépendante de croire qu'une tête pleine d'homoncules ait des propriétés mentales, et je ne vois aucun moyen d'écarter la conclusion absurde qu'elle en soit dotée (quoique bien entendu, mon argument perde de sa validité si quelqu'un en trouve le moyen). Le problème est donc de savoir s'il y a une bonne raison de croire au Fonctionnalisme. L'un des arguments en faveur du Fonctionnalisme est qu'il constitue actuellement la meilleure solution du problème des rapports entre l'esprit et le corps. Je pense que c'est un mauvais argument, mais puisque je pense par ailleurs que le Psychofonctionnalisme est préférable au Fonctionnalisme (pour les raisons mentionnées précédemment), je renvoie l'examen de sa validité à la discussion du Psychofonctionnalisme.

Le seul autre argument que je connaisse en faveur du Fonctionnalisme est celui que la vérité des identités Fonctionnelles peut être établie

1. Nous échouons souvent à concevoir comment quelque chose est possible parce que les concepts théoriques adéquats font défaut. Avant la découverte du mécanisme de la duplication génétique, Haldane soutenait par exemple de façon très persuasive qu'aucun mécanisme physique concevable ne pouvait effectuer un telle tâche. Mais au lieu d'encourager les savants à développer les idées qui nous auraient permis d'imaginer un tel mécanisme physique, il en concluait qu'il devait s'agir d'un mécanisme non-physique. (L'exemple m'a été fourni pas R. Boyd.)

par le biais de l'analyse du sens des termes mentaux. De sorte que les identités Fonctionnelles doivent être justifiées de la même façon que peut l'être l'affirmation que l'état de célibataire est identique à l'état d'homme non marié. Il existe aussi un argument similaire qui fait lui appel aux platitudes du sens commun relatives aux états mentaux et non à l'idée de vérité en vertu du sens des termes. Lewis dit que les caractérisations fonctionnelles des états mentaux relèvent de « la psychologie du sens commun — de la science populaire, plutôt que de la science professionnelle » (Lewis, 1972, p. 250) (voir aussi Shoemaker, 1975 et Armstrong, 1968. Armstrong fait une erreur sur la question de l'analyticité, voir *ibid.*, p. 84-85 et 90). Plus encore, il insiste aussi sur le fait que les caractérisations Fonctionnelles ne devraient « inclure que des platitudes qui sont connues de tous — chacun les connaît, chacun sait que chacun les connaît, etc. » (Lewis, 1972, p. 256). Je m'attacherai ici essentiellement à la version « plate » de l'argument. Celle qui fait appel à la notion d'analyticité est sujette aux mêmes objections, outre le fait qu'elle prête le flanc aux doutes qu'a soulevés Quine à propos de l'analyticité.

Je suis prêt à concéder, pour les seuls besoins de l'argumentation, qu'il est possible de définir n'importe quel état mental en termes de platitudes concernant d'autres termes mentaux; des termes d'input et des termes d'output. Mais cela ne m'engage nullement à cette sorte de définition des états mentaux qui élimine toute espèce de terminologie mentale au moyen de la Ramsification ou d'un autre procédé. Il est tout simplement erroné de supposer que si chaque état mental peut être défini en termes d'autres états mentaux (ainsi que d'inputs ou d'outputs), alors chaque état mental peut être défini de façon non mentale. L'exemple précédent permettra d'éclairer ce point. Pour simplifier les choses, je laisserai de côté ici les inputs et les outputs. Définissons la douleur comme la cause du souci, et le souci comme l'effet de la douleur. Même une personne assez naïve pour admettre cela n'a pas besoin d'accepter une définition de la douleur qui fasse de celle-ci *la cause de quelque chose,* ou une définition du souci qui fasse de celui-ci *l'effet de quelque chose*, Lewis prétend que c'est une vérité analytique que la douleur est ce qui détient un certain rôle causal. Même s'il a raison s'agissant d'un rôle causal caractérisé de façon partiellement mentaliste, on ne peut en conclure que c'est une vérité analytique que la douleur est ce qui détient un certain rôle causal quand ce rôle est caractérisé de façon non mentaliste.

Je ne vois aucun argument décent en faveur du Fonctionnalisme qui puisse être fondé sur des platitudes ou sur l'analyticité. De plus, le Fonctionnalisme fondé sur des platitudes conduit à des difficultés dans les cas où de telles platitudes font défaut. Souvenez-vous de l'exemple des cerveaux que l'on extrait pour les nettoyer et les remettre à neuf et où les connexions entre le cerveau et le corps sont maintenues par radio. Ce processus prend quelques jours et une fois qu'il est achevé, le cerveau est réinséré dans le corps. Il peut arriver que le corps d'une personne soit

détruit accidentellement au cours de l'opération. Si on le rattachait à des organes sensoriels d'inputs (mais non pas d'outputs), le cerveau ne manifesterait alors *aucune* des connexions habituelles mentionnées par le sens commun entre le comportement et les groupements d'états mentaux et d'inputs. Si, comme cela est à première vue plausible, ce cerveau pouvait avoir presque tous les mêmes états mentaux (étroits) que nous avons (et puisque cet état de choses pourrait devenir typique), le Fonctionnalisme est faux.

Il est instructif de comparer ce cas avec la manière dont le Psychofonctionnalisme tente de rendre compte des cerveaux enfermés dans des bocaux. Selon le Psychofonctionnalisme, ce qui compte au nombre des inputs et des outputs d'un système est une question empirique. Considérer les impulsions neuronales comme des inputs et des outputs permettrait d'éviter les problèmes auxquels il vient d'être fait allusion, puisque les cerveaux dans les bocaux et les paralytiques peuvent avoir les bonnes impulsions neuronales sans avoir de mouvements corporels.

Objection : Une paralysie pourrait affecter le système nerveux, et donc les impulsions neuronales, de sorte que le problème qui fait obstacle au Fonctionnalisme affecte tout autant le Psychofonctionnalisme. *Réponse :* Les maladies du système nerveux peuvent effectivement modifier *les propriétés mentales* : elles peuvent par exemple rendre leurs victimes incapables de ressentir de la douleur. De sorte qu'il pourrait en fait être vrai qu'une maladie généralisée ayant entraîné une paralysie intermittente du système nerveux rende les gens incapables d'avoir certains états mentaux.

Selon les versions plausibles du Psychofonctionnalisme, le problème de savoir quels processus neuronaux peuvent être considérés comme des inputs ou des outputs revient en partie à se demander *quels dysfonctionnements doivent être considérés comme des modifications de propriétés mentales et quels dysfonctionnements doivent être considérés comme de simples modifications des inputs périphériques et des connexions avec les outputs*. Le Psychofonctionnalisme a plus de ressources que le Fonctionnalisme, puisqu'il nous permet de faire coïncider la ligne de partage entre les deux avec la limite entre l'intérieur et l'extérieur de l'organisme, et d'éviter par là les problèmes qui viennent d'être examinés. L'erreur de toutes les versions du Fonctionnalisme est de tenter de tracer cette ligne sur la seule base de la connaissance du sens commun ; et celle des versions « analytiques » du Fonctionnalisme est plus précisément de tenter de le faire *a priori*.

2.0. *Le Psychofonctionnalisme*

Ma critique du Fonctionnalisme repose sur le principe suivant : si une doctrine conduit à une conclusion absurde qu'il n'y a aucune raison indépendante de croire, et si il n'y a aucun moyen de dissiper cette absurdité ou de montrer qu'elle est erronée ou sans pertinence, ni aucune

bonne raison en premier lieu de croire en la doctrine qui conduit à une telle absurdité, alors cette doctrine doit être récusée. J'ai affirmé qu'il n'y avait aucune raison indépendante de croire que les simulations Fonctionnelles par homoncules ont des propriétés mentales. Cependant, *il y a* une raison indépendante de croire qu'une simulation *Psycho*fonctionnelle *a des états mentaux,* à savoir le fait qu'une simulation Psychofonctionnelle de vous serait Psychofonctionnellement équivalente avec vous, de telle sorte que n'importe quelle théorie psychologique qui serait vraie de cette simulation serait également vraie de cette simulation. Quelle meilleure raison peut-il y avoir de lui attribuer n'importe lequel des états mentaux qui appartiennent au domaine de la psychologie ?

Cet argument montre qu'une simulation psychofonctionnelle quelconque de vous partage vos états mentaux *non qualitatifs.* Je vais toutefois tenter de montrer dans la section suivante qu'il est permis de douter qu'elle partage vos états mentaux qualitatifs.

2.1. *Les* qualia *sont-ils des états Psychofonctionnels ?*

J'ai commencé cet article en décrivant un système dirigé par des homoncules et en affirmant qu'il y a apparemment des raisons de douter qu'un tel système possède des états mentaux, et en particulier des états mentaux tels que la douleur, la démangeaison ou la sensation de rouge. Peut-être est-il possible d'expliquer le doute particulier que soulèvent les *qualia* en considérant non pas le phénomène de l'absence des *qualia,* mais celui de leur *inversion.* Il n'est pas insensé, ou du moins il ne semble pas insensé de supposer que les objets que vous et moi appelons verts m'apparaissent en fait de la manière dont les objets que nous appelons rouges vous apparaissent. Il semble que nous puissions être fonctionnellement équivalents même si la sensation que les extincteurs évoquent en vous est qualitativement la même que la sensation que l'herbe évoque en moi. Imaginez une lentille qui, lorsqu'on la place sur l'œil d'un sujet amène celui-ci à proférer des exclamations du genre : « Les choses rouges m'apparaissent maintenant de la façon dont les choses vertes m'apparaissaient avant, et *vice versa.* » Imaginez de plus deux jumeaux identiques sur l'un desquels de telles lentilles ont été posées à la naissance. Les jumeaux grandissent normalement, et à 21 ans sont fonctionnellement équivalents l'un avec l'autre. Il est assez probable que le spectre de l'un est l'inverse de celui de l'autre (la possibilité d'une inversion de spectre au sein d'une même personne a été défendue de manière assez convaincante par Shoemaker, 1975, n. 17). Par contre, on ne voit pas quel sens il peut y avoir à faire une supposition analogue dans le cas des états non qualitatifs. Imaginer deux personnes dont l'une croit que p est vrai et que q est faux, tandis que l'autre croit que p est faux et que q est vrai. Ces deux personnes

pourraient-elles être fonctionnellement équivalentes ? On ne voit pas comment cela serait possible[1]. Il est en effet extrêmement difficile de concevoir comment deux personnes pourraient n'avoir d'autre différence entre leurs croyances que celle-ci et ne manifester cependant aucune différence de comportement. Les *qualia* semblent être dans une relation de dépendance asymétrique[2] vis-à-vis de l'organisation fonctionnelle dont les croyances sont exemptes.

Il y a une autre raison pour distinguer fermement entre les états qualitatifs et les états non qualitatifs à propos des théories fonctionnalistes : à savoir le fait que le Psychofonctionnalisme évite les problèmes auxquels se heurte le Fonctionnalisme à propos des seconds — par exemple, les attitudes propositionnelles comme les croyances et les désirs. Mais le Psychofonctionnalisme n'est peut-être guère plus capable que le Fonctionnalisme de rendre compte des états qualitatifs. La raison en est qu'il se peut que les *qualia* ne tombent pas dans le domaine de la psychologie.

Essayons en effet d'imaginer à quoi ressemblerait une réalisation de la psychologie humaine par un système dirigé par des homoncules. La recherche psychologique actuelle semble essentiellement tournée vers la

1. Supposez qu'une personne qui a une bonne vision des couleurs utilise par erreur « rouge » pour dénoter le vert et « vert » pour dénoter le rouge. Elle confond simplement les deux mots. Puisque cette confusion est purement verbale, quoiqu'elle dise d'une chose verte qu'elle est rouge, elle ne *croit* pas plus qu'elle l'est qu'un étranger qui confond « aschan » et « sandwich » ne croit que les gens déjeunent d'aschans plutôt que de sandwichs. Disons que la personne qui a ainsi confondu « rouge » et « vert » est victime d'une commutation de mots.

Considérons maintenant un mal différent : des lentilles qui inversent le rouge et le vert ont été placées sur les yeux de quelqu'un à son insu. On dira que cette personne est victime d'une commutation de stimuli. Comme la victime précédente, celle-ci applique le mot « rouge » aux choses vertes et *vice versa*. Mais à la différence de la première, elle *a* des croyances fausses à propos des couleurs. Si vous lui montrez une tache verte, elle dit *et croit* qu'elle est rouge.

Supposons maintenant qu'une personne victime de commutation de stimuli soit aussi victime d'une commutation de mots. (Supposons également qu'il s'agit d'un habitant d'un village perdu de l'Arctique qui n'a pas de croyances du genre « l'herbe est verte », « les extincteurs sont rouges », etc.). Elle parle normalement, appliquant « vert » aux taches vertes, et « rouge » aux taches rouges. En fait, elle est fonctionnellement normale. Mais ses *croyances* sont tout aussi anormales qu'elles l'étaient avant qu'elle ne soit victime d'une commutation de mots. Avant elle confondait les mots « vert » et « rouge », elle appliquait « rouge » à une tache verte, et croyait de façon erronée que la tache était rouge. Maintenant, elle emploie (correctement) « rouge », mais sa croyance reste erronée.

Donc deux individus peuvent être fonctionnellement identiques et avoir cependant des croyances incompatibles. Le problème de l'inversion des *qualia* affecte donc tout autant les croyances que les *qualia* (encore qu'il ne s'agisse, semble-t-il, que des croyances qualitatives). Ce qui devrait inquiéter non seulement ceux qui défendent une théorie qui identifie la croyance à un état fonctionnel, mais aussi ceux qui sont attirés par les explications du sens en termes de rôle fonctionnel à la Harman. Notre double victime constitue un contre-exemple de telles théories. Car son mot « vert » joue normalement son rôle dans ses raisonnements et ses inférences, et pourtant, quand elle dit d'une chose qu'elle est « verte », elle exprime la croyance qu'elle est *rouge*, mais elle emploie « vert » dans un sens anormal. Je suis reconnaissant envers Sylvain Bromberger pour l'aide qu'il m'a apportée sur ce point.

2. Supervenience.

description de relations de flux d'information entre des mécanismes psychologiques. Son but est de décomposer ces mécanismes en d'autres mécanismes psychologiquement plus simples, des « boîtes noires » dont la structure interne relève de la physiologie plutôt que de la psychologie (voir Fodor, 1968 ; Dennett, 1975, et Cummins, 1975 , Nagel, 1969, soulève d'intéressantes objections contre une telle perspective). Un mécanisme qui compare deux à deux les éléments d'un système représentationnel et détermine s'ils sont des occurrences du même type est un exemple de mécanisme quasi primitif. Les mécanismes primitifs peuvent également être similaires à ceux que l'on trouve dans un ordinateur digital — ils peuvent par exemple : *a)* ajouter 1 à un registre donné, et *b)* soustraire 1 d'un registre donné, ou, si ce registre contient 0, passer à la *n*-ième instruction indiquée. (N'importe quelle opération réalisable par un ordinateur digital peut être réalisée par une combinaison de telles opérations, cf. Minsky, 1967, p. 206). Considérez un ordinateur dont le langage machine ne contient que deux instructions correspondant à *a)* et *b)*. Si vous vous demandez comment il peut faire des multiplications ou résoudre des équations différentielles ou établir des listes de salaires, on peut vous répondre en vous montrant un programme écrit au seul moyen de ces deux instructions. Mais si vous vous demandez comment il additionne 1 à un registre donné, la réponse adéquate n'est pas un programme mais un diagramme de câblage. La machine additionne les 1 en vertu de ses circuits. Quand l'instruction correspondant à *a)* apparaît dans un certain registre, les contenus d'un autre registre changent « automatiquement » d'une certaine façon. La structure computationnelle d'un ordinateur est déterminée par un ensemble d'opérations primitives et par la manière dont les opérations non primitives sont construites à partir de ces dernières. Il est donc sans importance pour la structure computationnelle d'un ordinateur que ses mécanismes de base soient réalisés par des circuits de tubes, de transistors ou des relais. Pareillement, il est sans importance pour la psychologie d'un système mental que ses circuits primitifs soient réalisés par tel ou tel mécanisme neurologique. Considérez qu'un système est « une réalisation de la psychologie humaine » si chaque théorie psychologique qui est vraie de ce système est vraie de nous. Soit une telle réalisation où les opérations psychologiques primitives sont accomplies par de petits hommes, comme dans le cas des simulations dirigées par des homoncules. Admettons par exemple qu'un petit homme extraie des éléments d'une liste, un par un, qu'un autre petit homme les compare avec d'autres représentations pour voir s'ils concordent, etc.

Il y a de bonnes raisons de supposer que ce système a des états mentaux. Les attitudes propositionnelles en sont un exemple. Une théorie psychologique identifiera peut-être le fait de se souvenir que P avec le fait d'avoir « stocké » un objet de type phrastique qui exprime la proposition que P (cf. Fodor, 1975). Si donc l'un des petits hommes a stocké

un certain objet de type phrastique, nous avons de bonnes raisons de considérer que le système se souvient que P, mais à moins qu'avoir des *qualia* ne soit rien d'autre qu'une certaine façon de traiter de l'information (ce qui est au mieux une hypothèse contestable), il n'y a aucune raison théorique de considérer qu'un tel système a des *qualia*. En bref, les *qualia* de ce système dirigé par des homoncules sont peut-être aussi douteux que ceux de la simulation Fonctionnelle dirigée par des homoncules.

Mais le système dont il est ici question est *ex hypothesi* quelque chose dont n'importe quelle théorie psychologique qui est vraie est vraie. *Donc toute raison de douter qu'il possède des* qualia *est aussi une raison de douter que les* qualia *appartiennent au domaine de la psychologie.*

A quoi on pourrait faire l'objection suivante : « Le type de psychologie que vous avez en tête est la psychologie *cognitive,* c'est-à-dire la psychologie des processus de pensée ; et il n'y a pas à s'étonner que les *qualia* ne relèvent pas de la psychologie *cognitive* ! » Mais *ce n'est pas* là la psychologie à laquelle je pense, et si je donne cette impression, c'est simplement parce que rien dans ce que nous savons des processus psychologiques sous-jacents à notre vie mentale consciente n'a quoi que ce soit à voir avec les *qualia*. Ce qui passe pour de la « psychologie » de la sensation ou de la douleur, par exemple, est soit *a)* de la physiologie, soit *b)* de la psychophysique (c'est-à-dire l'étude de fonctions mathématiques reliant des variables de stimulus à des variables de sensation ; par exemple la dépendance fonctionnelle entre l'intensité du son et l'intensité des ondes sonores) ou *c)* un pot-pourri d'études descriptives (voir Melzack, 1973, chap. 2). Et bien entendu seule la psychophysique pourrait en fait être conçue comme traitant des *qualia per se*. Il est non moins clair que la psychophysique ne s'intéresse qu'au seul aspect *fonctionnel* de la sensation, et non à son aspect qualitatif. Les expériences psychophysiques que l'on peut faire sur vous auraient les mêmes résultats que celles faites sur un quelconque système psychofonctionnellement équivalent avec vous, même en cas d'absence ou d'inversion de *qualia*. Or si les résultats expérimentaux demeurent inchangés, qu'il y ait ou non absence ou inversion de *qualia,* on peut difficilement s'attendre à ce qu'ils jettent un peu de lumière sur les *qualia*.

En fait, étant donné le type d'appareil conceptuel sur lequel s'appuie la psychologie contemporaine, je ne vois pas comment elle pourrait *expliquer* les *qualia*. Nous ne pouvons aujourd'hui concevoir comment la psychologie aurait la capacité d'expliquer les *qualia,* quoique nous *puissions* concevoir comment elle peut expliquer la croyance, le désir, l'espoir, etc. (cf. Fodor, 1975). Qu'une chose soit pour le moment inconcevable ne constitue en aucune façon une bonne raison de penser que cette chose soit impossible. Des concepts pourraient demain voir le jour qui permettront de rendre concevable ce qui est aujourd'hui inconcevable. Mais il faut pour le moment faire avec ce que l'on a, et étant donné ce

dont nous disposons, il semble que les *qualia* n'appartiennent pas au domaine de la psychologie.

Que les *qualia* soient en fait le paradigme de ce qui appartient au domaine de la psychologie ne constitue aucunement une objection valable contre l'hypothèse que ce ne sont pas des entités psychologiques. Ainsi qu'on l'a fait à maintes reprises remarquer, la question de savoir si quelque chose appartient ou non au domaine de telle ou telle branche de la science est une question en partie empirique. Il s'avère que la liquidité de l'eau n'est pas explicable par la chimie, mais plutôt par la physique subatomique. Les branches de la science font à tout moment face à un ensemble de phénomènes qu'elles cherchent à expliquer. Mais il peut fort bien se faire qu'un phénomène qui semblait central à telle branche de la science relève aujourd'hui de telle autre.

L'argument de l'absence des *qualia* exploite la possibilité que l'état Fonctionnel ou Psychofonctionnel que les Psychofonctionnalistes ou les Psychofonctionnalistes voudraient identifier avec la douleur puisse se produire sans qu'aucun *quale* ne se produise. Il semble également concevable que le second puisse survenir sans le premier. Certains faits vont dans ce sens. Après des lobotomies frontales, les patients rapportent en général qu'ils éprouvent toujours de la douleur, quoique cette douleur ne les gêne plus (Melzack, p. 95). De tels patients manifestent tous les signes « sensoriels » de la douleur (tels que reconnaître les piqûres d'épingles comme désagréables), mais ils ne manifestent le plus souvent que peu ou aucun désir d'éviter les stimuli « douloureux ».

Ces observations suggèrent notamment que chaque douleur est en fait un état composite dont les éléments sont un *quale* et un état fonctionnel ou Psychofonctionnel[1]. Ou, ce qui revient à peu près au même, que chaque douleur est un *quale* qui joue un certain rôle Fonctionnel ou Psychofonctionnel. Si cela est vrai, on peut alors comprendre comment on a pu croire à tant de théories différentes sur la nature de la douleur et d'autres sensations ; c'est qu'on a en fait mis l'accent sur tel ou tel composant aux dépens de l'autre. Les tenants du béhaviorisme et du fonctionnalisme avaient un composant en tête ; les tenants de la définition privée ostensive, l'autre. Les deux approches ont commis l'erreur de donner une explication unilatérale de quelque chose qui a deux composants de nature tout à fait différente.

3.0. *Chauvinisme* versus *libéralisme*

Il est naturel de comprendre les théories psychologiques vers lesquelles se tourne le Psychofonctionnalisme comme des théories de la

1. Ce *quale* peut être identifié à un état physico-chimique. Un tel point de vue rejoint une suggestion faite par Putnam à la fin des années 60 dans un séminaire de philosophie de l'esprit ; voir également Gunderson, 1971, chap. 5.

psychologie *humaine*. Selon le Psychofonctionnalisme ainsi entendu, il est impossible pour un système d'avoir des croyances, des désirs, etc., sauf si les théories psychologiques qui sont vraies de nous sont vraies de ce système. Le Psychofonctionnalisme (ainsi entendu) stipule qu'il doit y avoir une équivalence Psychofonctionnelle avec nous pour qu'on puisse parler de propriétés mentales.

Mais même si l'équivalence Psychofonctionnelle avec nous est une condition de la *reconnaissance des propriétés mentales,* quelle raison y a-t-il de penser que c'est une condition de la possession de telles propriétés ? Ne pourrait-il se faire qu'il existe une grande variété de processus psychologiques qui soient compatibles avec la possession de propriétés mentales, et que nous n'exemplifions qu'un type particulier de ces processus ? Supposons que nous rencontrions des Martiens et que nous nous apercevions qu'ils sont à peu près équivalents à nous Fonctionnellement (mais non pas Psychofonctionnellement). Après avoir appris à mieux les connaître, nous découvrons qu'ils ne sont pas plus différents de nous que les êtres humains que nous connaissons. Nous développons toutes sortes de relations culturelles et commerciales avec eux. Nous étudions leurs journaux scientifiques et philosophiques et eux les nôtres, nous allons voir leurs films et eux les nôtres, nous lisons leurs romans et eux les nôtres, etc. Les psychologues martiens et terrestres comparent ensuite leurs notes et finissent par s'apercevoir que, psychologiquement, martiens et terriens sont en réalité plus différents qu'il n'y paraît. Ils finissent également par convenir que cette différence peut être décrite de la manière suivante. Imaginez que les martiens et les terriens soient le produit d'un projet conscient. En élaborant un tel projet il faut nécessairement opérer un certain nombre de choix. Certaines capacités peuvent être intégrées dans leur nature (être innées), d'autres apprises. Le cerveau peut être conçu de manière à accomplir des tâches en utilisant le maximum de mémoire afin de minimiser l'utilisation de sa capacité de calcul ; ou au contraire de manière à économiser sa mémoire et à faire essentiellement appel à sa capacité de calcul. Les inférences peuvent être accomplies par des systèmes qui utilisent peu d'axiomes et beaucoup de règles d'inférence, ou, au contraire, peu de règles et beaucoup d'axiomes. Imaginez maintenant que les psychologues martiens et terriens, quand ils comparent leurs notes, s'aperçoivent que martiens et terriens diffèrent autant que s'ils étaient le point d'aboutissement de projets aussi différents que possible (mais compatibles avec une équivalence Fonctionnelle approximative au niveau des adultes). Devons-nous pour autant rejeter notre hypothèse que les martiens peuvent apprécier nos films, croire à nos résultats scientifiques, etc. ? Devraient-ils « rejeter » leur « hypothèse » que nous « apprécions » leurs romans, « apprenions » dans leurs livres, etc. ? Peut-être n'ai-je pas fourni suffisamment d'informations pour que nous puissions répondre à ces questions. Après tout, il y a peut-être plusieurs manières de décrire les différences entre martiens et

terriens qui rendraient raisonnable de supposer qu'il n'y a tout simplement pas de différence objective, ou même de supposer que les martiens ne méritent pas que leur soient attribués des états mentaux. Mais il y a certainement aussi plusieurs manières de décrire la différence indiquée ci-dessus qui rendraient parfaitement clair que même si les martiens se comportaient de façon différente de nous dans certaines expériences psychologiques subtiles, ils n'en penseraient, désireraient, apprécieraient pas moins, etc. Supposer qu'il en aille autrement ne serait rien d'autre que du chauvinisme élémentaire. (Je rappelle qu'une théorie est chauviniste quand elle nie à tort que des systèmes ont des propriétés mentales, et libérale quand elle *attribue* à tort des propriétés mentales.)

Pour échapper à cette difficulté, il vient naturellement à l'esprit de tenter d'identifier les états mentaux avec des états Psychofonctionnels, en supposant que *toutes les créatures dotées de propriétés mentales,* y compris les martiens, sont incluses dans le domaine de la psychologie. Ce qui revient à définir « le Psychofonctionnalisme » en termes de psychologie « universelle » ou « intersystèmes » plutôt qu'en termes de psychologie humaine comme précédemment. La psychologie universelle est toutefois une entreprise suspecte. Car comment décider si un système doit être inclus dans le *domaine* de la psychologie universelle ? L'une des façons possibles de décider si un système a des propriétés mentales, et par conséquent s'il relève de la psychologie universelle, consisterait à recourir à une *autre* théorie des propriétés mentales telle que le béhaviorisme ou le Fonctionnalisme. Mais ce genre de procédure a aussi peu de légitimité que la théorie utilisée. De plus, si le Psychofonctionnalisme doit présupposer une autre théorie de l'esprit, autant se contenter de cette autre théorie.

Peut-être la psychologie universelle parviendra-t-elle à éviter ce problème de « domaine » de la même façon que les autres branches de la science. Celles-ci commencent par délimiter approximativement leur domaine en s'appuyant sur des versions intuitives et préscientifiques des concepts qu'elles sont censées expliquer. Elles s'efforcent ensuite de développer des espèces naturelles permettant la formulation de généralisations nomiques qui s'appliquent à toutes ou presque toutes les entités qui figurent dans les domaines préscientifiques. Dans bien des cas — y compris pour les sciences biologiques et sociales telles que la génétique et la linguistique — il s'avère que le domaine scientifique permet l'articulation de généralisations nomiques.

Il se pourrait toutefois que nous soyons un jour capables de développer une psychologie universelle comme nous avons été capables de développer une psychologie terrienne. Nous déciderons sur une base intuitive et préscientifique des créatures qui devront être incluses dans un premier temps dans son domaine, et nous travaillerons à développer les espèces naturelles de la théorie psychologique qui s'appliqueront à elles ou du moins à la plupart d'entre elles. Peut-être l'étude d'une grande variété

d'organismes découverts dans des mondes différents conduira-t-elle un jour à des théories capables de déterminer des conditions de vérité pour l'attribution d'états comme la croyance, le désir, etc., applicables à des systèmes qui, à un niveau préthéorique, sont tout à fait différents de nous. En fait, il est certain qu'une telle psychologie intermondaine exigera toute une catégorie de concepts mentalistes. Peut-être y aura-t-il des familles de concepts correspondant au désir, à la croyance, etc. : c'est-à-dire une famille de concepts du type croyance, une famille de concepts du type désir, etc. La nature de cette psychologie universelle dépendra alors des nouveaux organismes que nous découvrirons en premier lieu. Même si une psychologie universelle est effectivement possible, il y aura certainement de nombreux organismes possibles dont le statut mental restera indéterminé.

D'un autre côté, il se peut que cette psychologie universelle ne soit *pas* possible. Peut-être la vie dans l'univers est-elle ainsi faite que fera défaut toute base raisonnable pour décider quels systèmes appartiennent au domaine de la psychologie et quels systèmes n'y appartiennent pas.

Si une psychologie universelle *est* possible, le problème que j'ai soulevé disparaît. Le Psychofonctionnalisme universel évite le libéralisme du Fonctionnalisme et le chauvinisme du Psychofonctionnalisme humain. Mais la question de savoir si une psychologie universelle est possible est certainement une question que nous n'avons pas les moyens de résoudre pour le moment.

En résumé mon argument a donc été jusqu'ici le suivant :

1 / Le Fonctionnalisme a cette conséquence bizarre qu'une simulation de vous dirigée par des homoncules a des *qualia*. Il échoit donc au Fonctionnaliste de fournir une raison de croire à la théorie qu'il propose. Mais l'argument fourni à cet effet par la littérature Fonctionnaliste n'est pas valable, et le Fonctionnaliste semble donc incapable de se justifier.

2 / Les simulations Psychofonctionnalistes de nous partagent avec nous tous les états mentaux qui figurent dans le domaine de la psychologie ; de sorte que la tête Psychofonctionnelle remplie d'homoncules ne jette pas le doute sur les théories Psychofonctionnelles des états cognitifs, mais seulement sur les théories Psychofonctionnelles des *qualia*, c'est-à-dire sur le fait que les *qualia* puissent appartenir au domaine de la psychologie.

3 / Les théories Psychofonctionnalistes des états mentaux qui figurent dans le domaine de la psychologie sont cependant désespérément chauvinistes.

Une des versions du fonctionnalisme se heurte donc au libéralisme, et l'autre au chauvinisme. Quant aux *qualia,* s'ils appartiennent au domaine de la psychologie, alors le Psychofonctionnalisme est aussi chauviniste à leur égard qu'il l'est à l'égard de la croyance. D'autre part, si les *qualia* ne relè-

vent pas de la psychologie, la tête Psychofonctionnelle remplie d'homoncules peut être utilisée contre le Psychofonctionnalisme. Car la seule chose qui protège le Psychofonctionnalisme de l'argument de la tête remplie d'homoncules par rapport à un état mental S est que si vous avez S, alors n'importe quelle simulation Psychofonctionnelle de vous a S, parce que la théorie de S vraie de S s'applique aussi bien à elle qu'à vous.

3.1. *Le problème des inputs et des outputs*

J'ai supposé tout au long (ainsi que le font souvent les Psychofonctionnalistes — voir Putnam, 1967) que les inputs et les outputs peuvent être caractérisés au moyen de descriptions d'impulsions neuronales. Mais c'est là une supposition chauviniste, puisqu'elle exclut que les organismes dépourvus de neurones (telles que les machines) puissent avoir des descriptions fonctionnelles. Comment éviter le chauvinisme dans la description des inputs et des outputs ? Une première façon d'y parvenir serait de caractériser les inputs et les outputs ***uniquement en tant qu'***inputs et outputs. De sorte que la description fonctionnelle d'une personne pourrait simplement distinguer les outputs les uns des autres par des nombres : output 1, output 2... Un système pourrait alors être fonctionnellement équivalent à vous à condition qu'il ait un ensemble d'états, d'inputs et d'outputs causalement reliés les uns aux autres de la même manière que les vôtres le sont, quelle que soit leur nature. Et de fait, quoiqu'une telle solution ne respecte pas l'exigence posée par certains fonctionnalistes de caractériser les inputs et les outputs en termes physiques, d'autres fonctionnalistes — ceux qui exigent seulement que les inputs et les outputs soient caractérisés de manière ***non mentale*** — ont peut-être quelque chose de ce genre en tête. Cette version du fonctionnalisme n' « épingle » pas les descriptions fonctionnelles à la périphérie au moyen de descriptions relativement précises des inputs et des outputs : en fait, elle traite plutôt les inputs et les outputs exactement de la même façon dont toutes les versions du fonctionnalisme traitent les états internes. C'est-à-dire qu'elle se contente, dans sa spécification des états, des inputs et des outputs, d'exiger que ce soit des états, des inputs et des outputs.

Le problème de cette version du fonctionnalisme est qu'elle est sauvagement libérale. Les systèmes économiques ont des inputs et des outputs, tels que les flux de crédits et de débits. Et les systèmes économiques ont aussi une grande variété d'états internes, tels par exemple que celui d'avoir un taux d'augmentation du PNB égal au double du taux de base. Il n'est pas impossible qu'un cheikh richissisme puisse gagner le contrôle de l'économie d'un petit pays, par exemple la Bolivie, et manipuler son système financier de façon à le rendre équivalent à une personne, par exemple à lui-même. Si cela ne vous semble guère plausible,

souvenez-vous que les états économiques, les inputs et les outputs qui, selon le cheikh, correspondent à ses états mentaux, ses inputs et ses outputs, n'ont pas besoin d'être des grandeurs économiques « naturelles ». Notre cheikh pourrait choisir n'importe quelle grandeur économique — par exemple, la dérivée cinquième de la balance des paiements. La seule contrainte qu'il doive respecter est que les grandeurs qu'il choisit soient des grandeurs économiques, que les valeurs qu'elles ont constituent autant d'inputs, d'outputs et d'états, et qu'elles lui permettent de mettre sur pied une structure financière adaptable au moule formel en question. La mise en relation des grandeurs psychologiques et des grandeurs économiques peut être aussi bizarre qu'il le souhaite.

Cette version du fonctionnalisme est bien trop libérale et doit par conséquent être rejetée. S'il y a quelques éléments de certitude dans le débat autour des rapports entre l'âme et le corps, l'un d'eux est certainement que l'économie de la Bolivie ne peut avoir d'états mentaux, quelle que soit la façon dont de riches amateurs puissent s'y essayer. Il nous faut à l'évidence être beaucoup plus précis dans nos descriptions des inputs et des outputs. Le problème est alors le suivant : y a-t-il une description des inputs et des outputs qui soit assez précise pour éviter le libéralisme et cependant assez générale pour éviter le chauvinisme ? Pour ma part, j'en doute.

Toutes les propositions de descriptions des inputs et des outputs que j'ai pu rencontrer ou auxquelles j'ai pu penser se rendent coupables soit de chauvinisme, soit de libéralisme. Quoique j'ai surtout insisté ici sur le libéralisme, le chauvinisme est le problème le plus difficile à circonscrire. Les Psychofonctionnalistes tendent à caractériser les inputs et les outputs à la manière des béhavioristes : c'est-à-dire à caractériser les seconds en termes de mouvements des bras et des jambes, de sons émis et de choses du même genre, et les premiers en termes de lumière et de sons qui affectent les oreilles et les yeux. De telles descriptions sont à l'évidence spécifiques à une espèce. Les humains ont des bras et des jambes, mais pas les serpents — et que les serpents aient ou non des propriétés mentales, on peut aisément imaginer des créatures du type serpent qui en ont. En fait, il est possible d'imaginer des créatures avec des mécanismes d'input et d'output de toutes sortes, par exemple, des créatures qui communiquent et manœuvrent par émission de champs magnétiques puissants. Bien entendu, il serait possible de formuler des descriptions Fonctionnelles pour chacune de ces espèces, et quelque part au paradis des disjonctions, il existe une description disjonctive permettant de traiter de toutes les espèces qui ont jamais existé dans l'univers (cette description pouvant être infinie). Mais même un appel à des entités aussi douteuses que des disjonctions infinies ne permettra pas de sauver le Fonctionnalisme, puisque nous n'apprenons rien par là sur ce qui est commun à tous les organismes qui ressentent de la douleur et en vertu de quoi ils ressentent tous de la douleur. Et cela ne permet pas non plus

d'attribuer de la douleur à d'hypothétiques (mais non existantes) créatures qui ressentent de la douleur. De plus, c'est précisément ce au nom de quoi les fonctionnalistes rejettent en général d'un ton acerbe les théories disjonctives avancées parfois par des physicalistes désespérés. Si les fonctionnalistes accueillaient soudain à bras ouverts des états sauvagement disjonctifs pour échapper au chauvinisme, ils deviendraient incapables de se défendre contre l'accusation de physicalisme.

Les descriptions Psychofonctionnalistes habituelles (par exemple en termes d'activité neuronale) des inputs et des outputs sont tout aussi spécifiques à notre espèce et par conséquent tout aussi chauvinistes.

Le chauvinisme des descriptions habituelles d'inputs et d'outputs n'est pas difficile à expliquer. Le nombre d'êtres intelligents possibles est énorme. Etant donné une quelconque description d'inputs et d'outputs suffisamment précise, n'importe quel écolier épris de science-fiction sera en mesure de décrire un être capable de connaître et de sentir dont les inputs et les outputs ne pourront satisfaire cette description.

A mon avis, *toute description physique* des inputs et des outputs (souvenez-vous que beaucoup de fonctionnalistes ont insisté sur la nécessité d'avoir des descriptions physiques) conduit à une version du fonctionnalisme qui est inévitablement chauviniste ou libérale. Imaginez que vous soyez si gravement brûlé dans un incendie que votre meilleur moyen de communiquer avec le monde extérieur soit de recourir à une transcription en morse de votre électroencéphalogramme. Avoir une pensée amusante produit en vous une certaine configuration électrique que ceux avec qui vous communiquez décident d'interpréter par un point, et une pensée déprimante est interprétée par un « trait ». Cette fiction n'est pas si loin de la réalité. Selon un article récent (*Boston Globe,* 21 mars 1976), « des scientifiques de UCLA travaillent sur l'utilisation de l'électroencéphalogramme pour contrôler les machines... Un sujet place des électrodes sur son cuir chevelu et pense à un objet à travers un labyrinthe ». Le processus « inverse » est en apparence également possible : des personnes peuvent communiquer avec vous en code Morse en envoyant des décharges électriques qui affectent votre cerveau (par exemple, en causant une image rémanente de courte ou de longue durée). Pareillement, si les cérébroscopes dont les philosophes ont souvent rêvé devenaient une réalité, il serait possible de lire vos pensées directement à partir de votre cerveau. Là encore, le processus inverse semble être possible. Dans tous ces cas, *le cerveau lui-même devient une partie essentielle des mécanismes d'input et d'output*. Cette possibilité a d'embarrassantes conséquences pour les fonctionnalistes. Vous vous rappelez que les fonctionnalistes font valoir que le physicalisme est faux parce qu'un seul acte mental peut être réalisé par un nombre infiniment grand d'états physiques qui n'ont pas de caractérisation physique nécessaire et suffisante. Mais si cet argument fonctionnaliste contre le physicalisme est juste, il s'applique *également aux inputs et aux outputs,* puisque la réalisation physique des états mentaux

peut constituer une partie essentielle des mécanismes d'inputs et d'outputs. En d'autres termes, quelle que soit l'interprétation du terme « physique » qui rend la critique du physicalisme correcte, *il ne peut y avoir de caractérisation physique des inputs et des outputs qui s'applique à eux tous et à eux seuls*. Par conséquent, toute tentative de formuler une description fonctionnelle au moyen d'une caractérisation physique des inputs et des outputs soit exclut inévitablement des systèmes dotés de propriétés mentales, soit inclut des systèmes qui en sont dépourvus. En d'autres termes, *les fonctionnalistes ne peuvent éviter et le chauvinisme et le libéralisme*.

Les descriptions physiques des inputs et des outputs ne font donc pas l'affaire. De plus, on ne peut non plus recourir à des termes mentaux ou à des termes d'action (tels que « frapper du poing la personne qui vous a offensé »), puisque cela reviendrait à abandonner le programme fonctionnaliste de caractérisation du mental en termes non mentaux. D'autre part, comme vous vous en souvenez peut-être, caractériser des inputs et des outputs simplement *comme* des inputs et des outputs est inévitablement libéral. Pour ma part, je ne vois aucun vocabulaire pour décrire les inputs et les outputs qui évite à la fois le libéralisme et le chauvinisme. Je ne prétends pas que ce soit là un argument décisif contre le fonctionnalisme. J'y vois plutôt, comme dans le cas de la critique fonctionnaliste du physicalisme, un argument qui met la balle dans le camp du fonctionnaliste. Le fonctionnaliste dit au physicaliste : « Il est très difficile de voir comment il pourrait y avoir une seule caractérisation physique des états internes de toutes les créatures dotées de propriétés mentales et d'elles seules. » Je dis au fonctionnaliste : « Il est très difficile de voir comment il pourrait y avoir une seule caractérisation physique des inputs et des outputs de toutes les créatures dotées de propriétés mentales et d'elles seules. » Dans les deux cas, il est clair qu'il incombe maintenant à ceux qui croient que de telles caractérisations sont possibles d'indiquer comment elles pourraient l'être »[1].

Ned BLOCK.

BIBLIOGRAPHIE

Armstrong D. (1968), *A materialist theory of Mind,* London, Routledge & Kegan Paul.
Bever T. (1970), The cognitive basis for linguistic structures, *in* J. R. Hayes (ed.), *Cognition and the Development of language,* New York, Wiley.
Block N. et Fodor J. (1972), What psychological states are not, *Philosophical Review,* 81, 159-181.
Chisolm R. (1957), *Perceiving,* Ithaca, Cornell University, Press.

1. Je remercie Sylvain Bromberger, Hartry Field, Davil Hills, Paul Horwitch, Bill Lycan, Georges Rey et David Rosenthal pour leurs commentaires détaillés de l'une ou l'autre des versions préliminaires de cet article. Certaines de ces versions préliminaires ont fait, à partir de l'automne 1975, l'objet d'exposés à Tufts University, Princeton University, The University of North Carolina de Greensboro et The State University of New York de Binghamton.

Cummins R. (1975), Functional Analysis, *Journal of Philosophy,* 72, 741-764.
Dennett D. (1969), *Content and Consciousness,* London, Routledge & Kegan Paul, 1969.
Dennett D. (1975), Why the law of effect won't go away, *Journal of the Theory of Social Behavior,* 5, 169-187.
Dennett D. (1978*b*), *Brainstorms,* Montgomery, Vt, Bradford.
Fodor J. (1965), Explanations in Psychology, *in* M. Black (ed.), *Philosophy in America,* London, Routledge & Kegan Paul.
Fodor J. (1968), The appeal to tacit knowledge in psychological explanation, *Journal of Philosophy,* 65, 627-640.
Fodor J. (1974), Special Sciences, *Synthese,* 28, 97-115.
Fodor J., Bever T., Garrett M. (1974), *The psychology of Language,* New York, McGraw-Hill.
Geach P. (1957), *Mental Acts,* London, Routledge.
Gendron H. (1971), On the relation of neurological and psychological theories : A critique of the hardware thesis, *in* R. C. Buck and R. S. Cohen (eds), *Boston, Studies in the Philosophy of Science VIII,* Dordrecht, Reidel.
Grice H. P. (1975), Method in philosophical psychology (from the banal to the bizarre), *Proceedings and Adresses of the American Philosophical Association.*
Gunderson K. (1971), *Mentality and machines,* Garden City, Doubleday Anchor.
Harman G. (1973), *Thought,* Princeton, Princeton University Press.
Kalke W. (1969), What is wrong with Fodor and Putnam's functionalism ?, *Nous,* 3, 83-93.
Kim J. (1977), Phenomenal properties, psychophysical laws, and the identity theory, *The Monist,* 56 (2), 177-192.
Lewis D. (1972), Psychophysical and theoretical identifications, *Australasian Journal of Philosophy,* 50 (3), 249-258.
Locke D. (1968), *Myself and others,* Oxford, Oxford University Press.
Melzack R. (1973), *The puzzle of pain,* New York, Basic Books.
Minsky M. (1967), *Computation,* Englewood Cliffs NJ, Prentice-Hall.
Mucciolo L. F. (1974), The identity thesis and neuropsychology, *Nous,* 8, 327-342.
Nagel T. (1969), The boundaries of inner space, *Journal of Philosophy,* 66, 452-458.
Nagel T. (1974), What is it to be like a bat ?, *Philosophical Review,* 83, 435-450.
Nelson R. J. (1969), Behaviorism is false, *Journal of Philosophy,* 66, 417-452.
Nelson R. J. (1975), Behaviorism, finite automata and stimulus response theory, *Theory and Decision,* 6, 249-267.
Pitcher G. (1971), *A theory of perception,* Princeton, Princeton University Press.
Putnam H. (1963), Brains and Behavior, *Philosophical Papers,* vol. II, London, Cambridge University Press.
Putnam H. (1966), The mental life of some machines, *ibid.*
Putnam H. (1967), The nature of mental states, *ibid.*
Putnam H. (1970), On properties, *ibid.*, vol. I.
Putnam H. (1975*a*), Philosophy and out mental life, *ibid.,* vol. II.
Putnam H. (1975*b*), The meaning of meaning, *ibid.*
Sellars W. (1968), *Science and Metaphysics,* London, Routledge.
Shoemaker S. (1975), Functionalism and qualia, *Philosophical Studies,* 27, 271-315.
Smart J. J. C. (1971), Reports of immediate experience, *Synthese,* 22, 346-359.
Wiggins D. (1975), Identity, designation, essentialism and physicalism, *Philosophia,* 5, 1-30.