DE LA DIFFICULTÉ D'ÊTRE NATURALISTE EN MATIÈRE D'INTENTIONALITÉ *

Ce qu'on appelle les « sciences cognitives » constitue aujourd'hui une nébuleuse de disciplines et de méthodologies différentes, dans laquelle toutefois émergent certaines hypothèses générales communes à la majorité des chercheurs. L'une d'entre elles consiste à postuler que connaître, c'est « traiter » de l'information, c'est-à-dire manipuler des symboles de manière réglée, conformément à des processus qui sont de nature *formelle.* Il est évident que l'un des problèmes fondamentaux que les sciences cognitives doivent résoudre est celui de savoir comment les symboles qui sont censés rendre possibles les processus mentaux peuvent recevoir un *contenu,* c'est-à-dire un sens. Ce problème se pose de façon particulièrement exemplaire dans le cas de l'Intelligence Artificielle : comment un système informatique, soit un ensemble de règles de type *syntaxique,* peut-il prétendre modéliser des processus où l'interprétation du monde joue un rôle essentiel — dépasser la simulation entendue comme bricolage ? De tels systèmes ne sont-ils pas voués à rester purement formels, essentiellement dépourvus de sens en dépit de la tendance de l'utilisateur à projeter du sens sur les formules produites ? Mais le problème se pose de façon tout aussi urgente dans le cas de la linguistique, de la théorie de la vision, ou de la théorie de l'action : si les processus cognitifs de l'esprit — élémentaires ou supérieurs — sont de nature mécanique (c'est-à-dire s'ils sont régis par des règles formelles s'attachant à la seule forme des symboles), *il faut bien* que ces symboles représentent quelque chose, se voient conférer un sens. C'est là le problème dit de l' « intentionalité ».

Notons immédiatement que lorsqu'on parle d' « intentionalité » dans ce contexte, on s'écarte du sens courant en vertu duquel est intentionnel un acte prémédité ou réfléchi, pour retrouver le sens philosophique

* Je remercie Pierre Jacob d'avoir relu cet article et de m'avoir suggéré des éclaircissements ou des précisions sur plusieurs points.

traditionnel, des scolastiques aux phénoménologues : Brentano commente le terme d' « intentionalité » en termes de

> « rapport à un contenu, direction vers un objet (sans qu'il faille entendre par là une réalité), ou objectivité immanente [....] Dans la représentation, c'est quelque chose qui est représenté, dans le jugement quelque chose qui est admis ou rejeté, dans l'amour quelque chose qui est aimé, dans la haine quelque chose qui est haï, dans le désir quelque chose qui est désiré, et ainsi de suite »[1].

L'intentionalité est ainsi le lien sémantique par excellence : dire qu'un symbole *représente*, dire qu'il *renvoie* à quelque chose et, de façon plus générale, dire qu'un état mental *porte sur* un contenu sont autant de façons d'exprimer la propriété d' « avoir un sens », ou de l'intentionalité.

Il n'est donc pas étonnant que du traitement du problème de l'intentionalité dépende pour une bonne part la crédibilité des sciences cognitives. En vertu de quoi un système computationnel — comme l'est selon l'hypothèse cognitiviste l'esprit humain — représente-t-il le monde ? Nous allons voir que pour répondre à cette question, il est inévitable de prendre position sur d'autres problèmes, et en particulier sur l'identification des contenus mentaux (les conditions auxquelles deux contenus mentaux sont des instances d'un seul et même type), c'est-à-dire sur la taxinomie des contenus, et sur la validité de la thèse de la dépendance systématique (*supervenience*) des états mentaux par rapport à des états cérébraux[2]. Les divergences apparaissent dès que l'on pose les conditions auxquelles il est possible de parler d' « intentionalité », de quoi une telle intentionalité doit être faite, ou, en d'autres termes, sur quel terrain il convient de se situer pour en déterminer les composantes. De la grande variété des réponses, émergent deux traditions. La première rassemble de nombreux penseurs qui appartiennent aussi bien à des courants rationalistes qu'empiristes, tels que Descartes, Berkeley, Hume, Kant ou le Carnap de l'*Aufbau* ; elle pose que l'identification des contenus de pensée est méthodologiquement indépendante de leurs causes extérieures. Les *Méditations* de Descartes constituent un exemple canonique de cette démarche, qui consiste à analyser des contenus de conscience accessibles par pure introspection et, de ce fait, points de résistance à

1. Franz BRENTANO, *Psychologie vom empirischen Standpunkt*, 3 vols, Leipzig, Felix Meiner Verlag, 1924-1928, vol. 1, p. 102.

2. Dire qu'un état mental occurrent A d'un sujet donné dépend systématiquement d'un état cérébral B, c'est dire qu'il existe toujours un certain état cérébral B qui sert de substrat à A. Si un état mental A' est différent de A, le substrat B' de A' est en vertu de cette thèse nécessairement distinct du substrat B de A. Pour un examen de la thèse de la dépendance systématique défendue par D. Davidson, cf. ENGEL, 1986.

l'effet dévastateur du doute hyperbolique. Quel que soit l'état du monde, et, en particulier, qu'il existe ou non, je n'en ai pas moins l'impression d'être dans cette chambre, de fumer cette pipe, etc.[3].

L'autre école est naturaliste, et s'inspire des leçons de Dewey et de Quine. La psychologie doit de son point de vue s'efforcer de suivre les leçons de la physique et de la biologie : les événements mentaux ne sont analysables que comme les réponses de l'organisme à des états de l'environnement. On ne peut donc expliquer l'intentionalité du mental ni identifier le contenu des états mentaux sans mettre en relation un état mental ou computationnel avec certains états de choses.

Toutes les analyses récentes de l'intentionalité s'efforcent de déterminer la part qui doit être reconnue respectivement à la thèse computationnelle — qui, comme on va le voir, est étroitement solidaire de la première perspective — et au point de vue naturaliste quant à la détermination des contenus des états mentaux. Le présent travail a pour objectif d'éclairer le développement de la pensée de Jerry Fodor concernant la théorie de l'intentionalité. Une telle mise au point n'est sans doute pas superflue, dans un contexte où les examens critiques du cognitivisme fodorien ont souvent la fâcheuse tendance de combattre une théorie qu'il n'a jamais soutenue, théorie que l'on pourrait résumer en une formule : « le sens d'un symbole se réduit à sa syntaxe »[4].

Jerry Fodor estime en 1981 que, d'un certain point de vue, l'opposition entre solipsisme méthodologique à la Descartes et naturalisme à la Dewey est *résolue* dans le cognitivisme, c'est-à-dire dans la thèse selon laquelle les processus mentaux sont de nature computationnelle. La condition de « formalité » pose, en effet, que tout processus mental est formel, en ce sens qu'il n'a accès qu'aux propriétés formelles des représentations. Peu importe que l'information ainsi « traitée » soit produite par l'effet de l'environnement sur les « capteurs » d'un robot, sur la périphérie sensorielle d'un homme ou par les stimulations électriques qui affecteraient directement un cerveau (version contemporaine de l'hypothèse du « malin génie ») ; dans tous les cas, seules peuvent être déterminantes pour le système les propriétés *formelles* des représentations[5]. La

3. Il faut cependant relativiser cette affirmation dans le contexte contemporain de l'analyse du « contenu étroit ». Comme nous le verrons plus bas, un solipsisme ontologique et non pas seulement méthodologique serait incapable de fournir une notion de contenu, faute de pouvoir rendre compte de la relation causale entre le monde et les représentations.

4. Voir, par ex. PUTNAM, 1988, p. 6 sq., ou SEARLE, 1980.

5. Cette condition de formalité, particulièrement contraignante en Intelligence Artificielle, suggère le rapprochement entre la problématique de chercheurs comme Newell avec l'idéalisme transcendantal. Cf. mon article, « L'Intelligence Artificielle comme philosophie », *Le Débat, 47,* nov. déc. 1987, p. 88-102.

condition de formalité paraît donc interdire aux processus mentaux tout accès aux propriétés proprement *sémantiques* des représentations ou des contenus mentaux : cette condition s'accommode très bien du fait que les premières n'aient aucun référent, ou que les seconds soient faux, c'est-à-dire ne correspondent à aucun état de choses extérieur. Fodor prend l'exemple du programme SHRDLU de Winograd : le micromonde de blocs dans lequel le système s'oriente n'existe pas ; « c'est un simple ordinateur qui rêve qu'il est un robot » (Fodor, 1981, p. 232)**.

Cependant cette condition de formalité est mise à l'épreuve par le type d'interprétation des états mentaux qu'elle suggère. Quelques rappels sur ce qui distingue les contextes opaques des contextes transparents permettront ici de montrer pourquoi.

LA TRANSPARENCE

Dans un système extensionnel (comme l'est généralement celui d'une théorie scientifique), les propositions enchâssées sont telles que 1) leurs constituants sont interchangeables *salva veritate* avec des constituants équivalents ; et 2), qu'on peut toujours effectuer une généralisation existentielle. Dans un tel système, je peux par exemple dériver les propositions (3) et (4) des propositions (1) et (2) :

(1) L'étoile du soir est l'étoile du matin.
(2) Il est vrai que Vénus est l'étoile du soir.
(3) Il est vrai que Vénus est l'étoile du matin.
(4) Il est vrai qu'il existe quelque chose du nom de Vénus.

Or si l'on veut édifier une psychologie des états mentaux destinée à rendre compte du rapport entre état mental et comportement, comme cherche à le faire le psychologue, qu'il soit fonctionnaliste au sens étroit du terme ou naturaliste, on doit bien reconnaître qu'aucune interprétation transparente ne rendra justice aux attitudes propositionnelles qui sont celles des sujets. C'est l'opacité des contenus d'attitudes propositionnelles qui, aux yeux de Fodor (1981), constitue un argument en faveur de l'hypothèse computationnelle.

** Pour plus de précisions concernant les références placées entre parenthèses dans cet article se reporter à la Bibliographie, p. 32.

L'OPACITÉ

Si l'on s'intéresse par exemple à la croyance d'un sujet X relative aux propositions enchâssées 2-4, on peut très bien décrire dans les termes utilisés par X [6] un ensemble de croyances propres à X tel que :

(5) X croit que Vénus est l'étoile du matin.
(6) X croit que Vénus n'est pas l'étoile du soir.

En outre, du simple fait qu'un sujet croie une certaine proposition, dont l'expression *de dicto* est, par exemple,

(7) X croit que les extra-terrestres sont bien disposés à l'égard des humains,

on ne peut dériver

(8) Il existe des extra-terrestres.

Cette impossibilité caractéristique de pratiquer dans le contexte des attributions de croyances *de dicto* n'importe quelle substitution de termes équivalents (c'est-à-dire de même extension) ou de dériver une proposition existentielle non enchâssée vraie — en d'autres termes, cette opacité des attributions de croyance *de dicto* — est ce qui rend le solipsisme méthodologique si séduisant en psychologie :

> « les attributions opaques sont vraies en vertu de la manière dont l'agent se représente *à lui-même* l'objet de ses désirs (intentions, croyances, etc.). Et, par hypothèse, ce sont ces représentations qui ont un rôle dans la causation des comportements de l'agent » (*ibid.*, p. 235).

SEMI-TRANSPARENCE OU OPACITÉ COMPLÈTE ?

Le problème du contenu des états mentaux est cependant compliqué du fait que si la condition de formalité était pleinement adéquate, on ne

6. Il est également possible de décrire les croyances de X *« de re »*, c'est-à-dire dans les termes de celui qui rapporte les croyances de X.

devrait pas rencontrer de cas où l'opacité est en quelque sorte incomplète, c'est-à-dire où l'attribution de croyance fait intervenir des conditions intrinsèquement référentielles. Or ces cas se rencontrent, comme le rappelle Fodor lui-même, tout particulièrement dans les contextes où les attributions de croyances font intervenir des déictiques. On est parfois conduit, par exemple, à considérer comme identiques des contenus formellement distincts mais coréférentiels ; ainsi si Pierre et moi croyons l'un et l'autre que je suis malade, la formule du langage de la pensée qui exprime ma croyance est

(9) je suis malade,

tandis que celle du langage interne de Pierre est

(10) elle est malade.

D'autre part, en raison du fait que les pensées démonstratives supposent l'existence d'un référent, on ne peut décrire à l'aide d'un démonstratif une pensée démonstrative qui serait dépourvue de référent, telle que

(11) Jean croit que c'est un OVNI,

si « c' » n'a aucun contenu. Le contenu strictement individuel d'une pensée est donc ici insuffisant pour rapporter le contenu de croyance[7].

Réciproquement, la pensée que l'on peut exprimer en une certaine occurrence en disant :

(12) Je trouve qu'on est bien ici,

doit pouvoir être la même pensée, survenant dans un autre lieu, que celle que l'on exprime en disant :

(13) Je trouve qu'on est bien là-bas.

Ces remarques conduisent Fodor à découvrir les limites du principe de formalité en tant qu'il devrait conduire à une taxinomie des états mentaux *entièrement opaque*. Une telle taxinomie ne parviendrait pas à recon-

7. L'idée que l'expression d'une pensée dont une composante n'a pas de référent échoue à représenter une pensée déterminée, et donc à avoir un contenu, est défendue par des auteurs tels que G. Evans et J. McDowell (cf. EVANS, 1982). D'autres auteurs, comme PERRY, 1977 et 1979, défendent en revanche l'idée que des énoncés de ce genre ont un contenu.

naître l'identité de type entre des contenus comme (9) et (10), ou (12) et (13). Il ne faut pourtant pas abandonner l'opacité, qui s'impose pour les raisons évoquées plus haut. On doit simplement reconnaître l'existence d'une tension entre l'approche fonctionnelle de l'attribution des croyances et l'exigence sémantique minimale qu'impose l'identification de certains contenus. Tension qui exige seulement, du point de vue qu'exprime Fodor dans cet article, quelques aménagements ou du moins un *modus vivendi* :

> « si l'on construit une taxinomie de manière *purement* formelle, on a une identité de croyance qui va de pair avec une différence de valeur de vérité. D'un autre côté, si on l'élabore à partir de critères préthéoriques, on n'arrive plus à comprendre que ce sont les croyances et les désirs qui font agir les gens » (*ibid.*, p. 238-239).

Car une fois perdue la condition de formalité, on ne peut plus comprendre l'efficacité causale des processus mentaux, on ne peut plus comprendre pourquoi les états mentaux et les comportements s'enchaînent d'une façon qui est descriptible comme une dérivation.

L'article de 1981 se conclut sur un bilan nuancé. Il y a convergence entre une taxinomie des contenus faisant droit à leur opacité et la condition de formalité — tandis que la taxinomie transparente qu'exigerait une psychologie naturaliste est incompatible avec cette condition. Cependant, il ne faut pas confondre ce résultat avec la liquidation de toute ambition naturaliste ; si l'on a montré la plausibilité de l'idée que les *opérations mentales* ne sont sensibles qu'à la forme des symboles manipulés, ou, en d'autres termes, n'ont pas d'accès à la *sémantique* de ces symboles, *il ne faut pas en conclure que les représentations n'ont pas de propriétés sémantiques.*

Tout en étant convaincu de l'intérêt que représenterait une théorie naturaliste du sens, Fodor est convaincu qu'elle est hors d'atteinte ; elle exigerait, en effet, que soit disponible une science universelle achevée, nous permettant de donner *de façon nomologique* les relations causales entre les référents et les représentations mentales telles qu'elles sont établies par la connaissance des référents qu'atteint chaque science particulière : « on ne peut pas faire de psychologie naturaliste de la référence sans avoir une façon de dire ce qu'*est* le sel ; laquelle de ses propriétés détermine ses relations causales » (*ibid.*, p. 250). C'est pourquoi « une psychologie naturaliste reste une sorte d'idéal de la raison pure » (*ibid.*, p. 252).

Les années 1980 ont vu s'amplifier le débat sur le « site » de l'intentionalité, des auteurs tels que Tyler Burge montrant l'incapacité de la thèse solipsiste méthodologique soutenue par Fodor — laquelle identifie les

contenus à des états internes particuliers de traitement de l'information — à rendre compte de certains types d'attributions de croyance assez proches du cas de l'emploi du mot « eau » sur la Terre Jumelle inventé par Putnam[8].

Ce que montrait Putnam dans son célèbre article, c'est que l'extension des mots de la langue naturelle, tels que le terme d'« eau », n'est pas fonction de l'état psychologique du locuteur. Par conséquent, si l'on admet que connaître le sens d'un mot consiste à être dans un certain état psychologique, on ne peut soutenir que le sens d'un terme détermine son extension, comme on le dit habituellement dans une interprétation libre, psychologisée, du célèbre texte de Frege (1971)[9].

Imaginons, en effet, qu'il existe une autre terre entièrement semblable à la nôtre dans ses moindres détails, y compris en particulier dans le fait que tout terrien a une contrepartie, un « Doppelgänger » entièrement identique à lui quant à ses états cérébraux, sur terre jumelle, à l'exception d'un seul fait : la substance nommée « eau » sur terre jumelle, et qui a toutes les caractéristiques phénoménologiques de ce que l'on appelle « eau » sur terre, a en réalité sur terre jumelle une composition chimique différente, XYZ. Supposons, en outre, que deux habitants jumeaux, nommés respectivement par nous Oscar 1 et Oscar 2, pensent tous deux que « l'eau du bain est trop chaude », et qu'ils aient cette pensée à une époque où l'avancement de la chimie est tel que ni l'un ni l'autre ne disposent encore des moyens de distinguer la composition moléculaire des deux types de liquides. Dirons-nous qu'ils ont la même pensée ?

Lors même que l'argument de Putnam concernait le sens des mots de la langue naturelle, il est clair qu'il s'applique aussi aux expressions du langage de la pensée : les deux pensées exprimées par deux formules du « mentalais » de l'un et de l'autre ne peuvent pas être identiques puisque les conditions de satisfaction des concepts exprimés par le mot d' « eau » ne sont pas les mêmes sur terre et sur terre jumelle, lors même que les deux sujets sont dans un état cérébral identique.

L'erreur pour Putnam remonte à une certaine interprétation psychologiste de Frege, erreur qui constitue une tentation permanente pour le fonctionnaliste : elle réside dans le fait de penser que connaître le sens d'un terme consiste à être dans un certain état psychologique, *et* que le sens d'un terme détermine son extension. Mais la leçon de l'expérience de pensée de Putnam va plus loin que la thèse selon laquelle le sens des mots de la langue naturelle « ne sont pas dans la tête » : elle remet

8. Cf. PUTNAM, 1975.

9. Sur la psychologisation des réflexions frégéennes sur le sens et sa portée dans la philosophie contemporaine du langage, cf. PROUST, 1981.

principalement en cause le bien-fondé du solipsisme méthodologique, c'est-à-dire la doctrine selon laquelle, d'après les termes de Putnam, « aucun état psychologique proprement dit ne présuppose l'existence d'autre individu que celui auquel l'état psychologique est attribué » (Putnam, 1975, p. 220), en particulier si cette doctrine a l'ambition de livrer une théorie du contenu. On peut donc tirer de cette expérience de pensée la conclusion qu'aucune théorie des états mentaux à base individualiste ne pourra livrer une théorie de l'intentionalité de ces états.

Burge tire une morale sensiblement différente de l'article de Putnam [10]. Ce qui permet de spécifier le contenu mental de quelqu'un, c'est le sens des occurrences enchâssées, « obliques », dans des phrases telles que « Jean croit que l'eau du bain est chaude ». Si le contenu mental d'Oscar 1 et d'Oscar 2 diffèrent, c'est simplement parce que la pensée de l'un concerne de l'eau, tandis que celle de l'autre ne renvoie pas en fait à de l'eau. Il est donc impossible, du point de vue de Burge, d'invoquer comme le fait Putnam la déicticité cachée de termes de substance, dont l'extension changerait en fonction du contexte. Burge montre que l'on ne peut attribuer au concept EAU exprimé par le mot d'« eau » le changement d'extension du mot « ici » (dans la reformulation proposée par Putnam : dire que *x* est de l'eau, c'est dire que *x* est identique au liquide qu'on appelle « eau » ici) parce que dans ce cas, XYZ serait de l'eau, ce qui est faux.

Renonçons donc à chercher une solution en termes de l'indexicalité cachée des termes de substance : la différence entre les deux emplois du mot « eau » sur Terre et sur Terre Jumelle n'est pas telle qu'elle affecterait les extensions d'un sens linguistique constant ; *car les deux mots d'« eau » ne partagent même pas leur sens linguistique acontextuel.*

Il est, en outre, du point de vue de Burge sinon faux, du moins équivoque de dire que le sens « étroit » d'un terme comme « eau » ne fixe pas l'extension du mot « eau » comme on est tenté de le dire par suite de l'argument de Putnam (par « sens étroit », on entend le rôle fonctionnel d'une représentation ou d'une phrase du langage de la pensée, le « sens large » désignant ses conditions de vérité). On ne peut évidemment conclure directement d'un contenu de croyance pris *de dicto* — sans parler ici de l'attribution à autrui d'une croyance — aux extensions des termes impliqués dans le rapport de croyance correspondant. Mais d'un point de vue purement sémantique, c'est une vérité nécessaire que « eau » fasse référence à l'eau et seulement à elle.

Puisqu'on ne peut pas spécifier le contenu mental d'un sujet sur une

10. Cf. BURGE, 1982, p. 102 sq.

base individualiste, la stratégie de recherche de Fodor dite du « solipsisme méthodologique » se voit enfermée dans des limites assez étroites : elle ne peut pas prétendre comme le pensait Fodor dans son article de 1981 spécifier les attributions de croyance non transparentes (même les attributions de croyance *de dicto* supposent l'existence d'autres entités), ni de façon générale fournir une théorie de l'intentionalité proprement dite.

Fodor (1987) présente une réponse élaborée aux arguments externalistes. D'un côté, il tire d'une réflexion générale sur le raisonnement causal dans la formation des concepts scientifiques l'idée que l'individualisme méthodologique — à distinguer, comme on va le voir bientôt, du solipsisme méthodologique — est rationnellement justifié dans tous les cas, même quand il s'agit de domaines typiquement relationnels comme l'est en l'occurrence celui de la référence.

De l'autre, il propose son propre diagnostic du problème des terres jumelles, diagnostic qui situe la vraie difficulté non pas dans le rapport entre des intensions « dans la tête » et des extensions qui, précisément, resteraient mentalement indifférenciées, mais dans l'impossibilité où nous sommes par principe de spécifier les contenus étroits (du mentalais) dans la langue naturelle (c'est-à-dire sans « toujours déjà » les ancrer dans un contexte). Revenons sur ces deux types d'arguments.

L'argument méthodologique de fond, auquel Fodor a recours[11], invoque contre l'argument de Putnam-Burge le concept de pertinence causale qui, selon lui, caractérise nécessairement toute taxinomie scientifique. En bref, Fodor distingue de façon très nette deux hypothèses qui étaient restées jusqu'alors inextricablement confondues ; le solipsisme méthodologique et l'individualisme méthodologique. Le premier est une théorie *empirique* sur l'esprit, qui en pose la nature computationnelle. Le second est une règle générale de méthodologie scientifique : les catégories scientifiquement pertinentes sont celles qui permettent des généralisations causales.

Le fait qu'Oscar 1 fasse référence à H_2O tandis qu'Oscar 2 fait référence à XYZ n'est fonctionnellement d'aucune conséquence quant à la suite des attitudes propositionnelles ayant un rapport avec le liquide que l'un et l'autre appellent de l'eau. Négliger cette différence ne constitue pas de ce fait une lacune de la théorie ; c'est une mise en application d'un principe universellement reconnu dans les sciences : n'est théoriquement pertinente qu'une propriété ayant des implications causales. Par exemple,

11. Cf. FODOR, 1987, p. 42-43.

deux prédicats parfaitement bien définis en termes purement physiques tels que

« être une particule P au temps t », et
« être une particule F au temps t »,

qui s'appliquent l'un et l'autre à toute particule physique telle respectivement que, au temps t, ma pièce de 1 F est tombée sur pile (P) ou sur face (F), ne peuvent évidemment permettre aucune généralisation causale, et sont donc pour cela dénués de tout intérêt théorique. Certaines catégories scientifiques pertinentes peuvent en revanche, sans paradoxe, être relationnelles et individualistes : elles sont individualistes dans la mesure où elles sélectionnent seulement celles des propriétés relationnelles qui jouent un rôle causal. La propriété « être une planète », par exemple, est une propriété relationnelle mais qui est déterminée de manière purement causale.

Résumons-nous : si c'est l'explication causale du comportement qui fait l'objet de l'investigation psychologique, il n'y a aucune raison de renoncer à dire que l'état mental qui s'exprime chez les deux Oscars par la formule du langage de la pensée correspondant à « l'eau du bain est trop chaude » est identique, même si l'on s'écarte en cela des intuitions du sens commun. L'identité du contenu étroit de cet état mental correspond au fait que les mêmes types d'inférences seront tirés par l'un et l'autre de la pensée en question, et conditionneront des comportements identiques [12].

La première partie de la réponse de Fodor consiste ainsi à dire que toute « individuation » *dans les sciences* est de type « individualiste », l'individualisme méthodologique soutenant dans le cas de la psychologie que les états mentaux sont individués par leurs capacités causales ; et à renvoyer à Burge l'ascenseur causal : à quoi peut bien servir une théorie psychologique externaliste qui découvre des différences de contenu mental *sans contrepartie* dans les états cérébraux : « comment les différences de contexte pourraient-elles affecter les capacités causales des états mentaux d'un sujet sans affecter l'état de son cerveau ? » (Fodor, 1987, p. 41-42). On ne peut évidemment abandonner la dépendance systématique du mental sur le cérébral — ce qui serait l'une des issues possibles hors du labyrinthe de Putnam-Burge et qu'emprunte précisément Burge — sans en même temps se priver du moyen de rendre compte de la causation mentale.

Venons-en alors au diagnostic que Fodor propose, diagnostic destiné

12. *Ibid.*, p. 34.

à préserver à la fois la dépendance systématique du mental à l'égard du cérébral *et* le lien entre contenus et extensions. « Les exemples de Terre Jumelle ne suppriment pas le lien entre contenu et extension ; ils le relativisent seulement au contexte » (Fodor, 1987, p. 47). En d'autres termes, il suffit de dire que mon jumeau et moi partageons le même contenu étroit de croyance, désir, etc. Mais, dans le contexte de la Terre, l'extension déterminée par ce contenu est H_2O, tandis que dans le contexte de la Terre-Jumelle, c'est XYZ, que la pensée concernée soit celle de mon jumeau ou la mienne. Deux pensées sont donc de même contenu à la seule condition que leurs conditions de vérité coïncident pour un contexte donné.

On reconnaît dans ce diagnostic la solution *de dicto* écartée par Putnam dans l'article princeps selon laquelle « l'eau est ce qui est semblable à ce qu'on appelle "eau" ici », puis rejetée, pour des raisons différentes, par Burge. La difficulté majeure de cette solution est que le contenu étroit paraît à peine mériter le nom de contenu dans la mesure où il n'est *pas encore sémantiquement évaluable,* puisqu'un contexte n'est pas encore donné qui rende possible cette évaluation. Par opposition à un tel contenu étroit, ce qu'on appelle contenu dans la réflexion sémantique traditionnelle comme la pensée frégéenne, la proposition en soi de Bolzano ou la phrase éternelle de Quine ont une valeur de vérité déterminée en ce sens qu'elles *incluent* les déterminants contextuels du sens. Il est difficile de voir dans le contenu étroit autre chose qu'un déterminant du contenu, clairement inspiré du « caractère » de Kaplan [13], déterminant qui ne fournit pas une condition suffisante mais seulement dans le meilleur des cas une condition nécessaire de l'intentionalité.

Cette difficulté se double du fait que le contenu étroit, comme on l'a vu, n'est jamais spécifiable dans la langue. C'est là aux yeux de Fodor un fait empirique : « le contenu que la phrase anglaise exprime est *ipso facto* un contenu *ancré,* par conséquent *ipso facto* un contenu qui *n'est pas* étroit » (Fodor, 1987, p. 50). La difficulté se précise alors de la manière suivante : qu'est-ce qui autorise ici le théoricien à opposer un contenu seulement « potentiel » à un contenu « en acte » ? Comment la relation de référence (d'un mot à ce qu'il dénote) peut-elle s'articuler sur une relation potentielle de sens (d'un mot à ce qu'il signifie) dans une théorie *naturaliste* des contenus (entreprise parfaitement étrangère à Frege et à

13. Cf. KAPLAN, 1989. Rappelons que Kaplan distingue le « caractère », comme fonction fixée par des conventions linguistiques d'un contexte à un contenu, du contenu qui est une fonction des circonstances d'évaluation à une extension appropriée (cf. *ibid.*, p. 500-507). Par exemple, le mot « je » a pour caractère ' « je » fait référence à l'agent dans le contexte considéré ', et a pour contenu dans les présentes circonstances l'auteur de cet article, J. P.

Bolzano et qui a conduit Quine, comme on le sait, à renoncer aux contenus) ? Comment, enfin, est-on assuré que le « contenu étroit » constitue *déjà* un contenu, par opposition à une simple forme ? Ce contenu « en forme de caractère » *(characterlike)* est-il autre chose qu'une combinaison (que la tradition aurait jugée « nominale ») de jetons symboliques pourvus d'une forme et d'une fonction ? N'est-ce pas une pétition de principe que d'y lire les prémices d'un contenu [14] ?

On peut être tenté de donner raison à Fodor lorsqu'il reconnaît que « la théorie qui est ici en train d'émerger est, en un sens, une théorie "sans contenu" du contenu étroit », mais hésiter à reconnaître avec lui qu'il s'agisse encore « d'une théorie pleinement intentionnaliste » (Fodor, 1987, p. 53). Il faut cependant se souvenir que la théorie du contenu comporte désormais deux niveaux, et chercher dans la suite de l'ouvrage la seconde partie de la théorie, celle qui doit permettre d'apporter une réponse aux questions laissées en suspens, telle que celle de l'évaluation sémantique des attitudes propositionnelles ou celle du lien entre contenu « étroit » et contenu « large ». Le premier niveau, dit du « contenu étroit », est solidaire de la version fodorienne de la théorie représentationnelle de l'esprit, en vertu de laquelle les états et les processus mentaux sont computationnels, en ce sens que les représentations obéissent à des règles formelles de combinaison, et reçoivent leurs propriétés causales en partie en vertu de leurs propriétés formelles. Le contenu étroit est ainsi requis, comme on l'a vu, pour garantir le caractère causal des relations entre états mentaux. Reste alors à comprendre en naturaliste ce qui est en jeu dans l' « interprétation d'un symbole primitif » — non logique — « du mentalais dans un contexte donné » (Fodor, 1987, p. 98).

La théorie causale de la référence paraît ici fournir le cadre général de la solution recherchée. De façon générale, une telle théorie explique la relation de référence entre une expression et une entité par l'existence d'une relation converse de causalité entre l'objet ou la propriété dénotés et le nom propre ou le prédicat qui les dénotent ; initialement destinée à rendre compte de la référence des expressions du langage naturel, cette théorie peut aussi bien s'appliquer aux symboles du langage de la pensée.

En fait, une théorie causale est notoirement trop sommaire pour rendre compte du contenu des états mentaux. Comme le montre Dretske (1981), il convient de distinguer entre causalité et régularité nomique pour pouvoir rendre compte du lien informationnel qui existe entre, par exemple, deux séries d'événements. L'existence d'une relation causale entre A et B (une mouche dans le champ de vision d'une grenouille

14. Sur l'opposition traditionnelle entre définition nominale et définition réelle, cf. PROUST, 1986, p. 66 sq.

provoque une excitation neuronale particulière qui déclenche à son tour la réponse « happer au vol ») ne suffit pas à dire qu'un flux d'information se produise de A vers B. La théorie informationnelle du sens que suggère le travail de Dretske pose que l'intentionalité présente dans la transmission et la réception de l'information dépend non d'une relation de causalité, mais de « régularités nomiques », c'est-à-dire de corrélations réglées entre événements qui ne sont pas nécessairement déterministes, mais qui sont contrefactuellement stables.

Dans la mesure où elles éclairent la complémentarité du contenu étroit et du contenu large, les analyses de Dretske apportent de l'eau au moulin de Fodor : le *« narrow content »* renvoie à la structure cognitive d'un concept (c'est-à-dire à ses effets et conséquences fonctionnels), tandis que le *« wide content »* renvoie aux origines informationnelles des concepts. Dretske admet ainsi à la suite de Putnam que l'on peut parfaitement posséder un concept sans connaître les conditions nécessaires et suffisantes de son application. L'apprentissage du sens d'un mot se fait par exposition à des signaux porteurs d'information, laquelle dépend des régularités physiques de l'environnement, et non pas des caractéristiques physiques isolées des expériences d'un sujet. Ce sont ces régularités qui décident du fait que le concept d' « eau », par exemple, ait le contenu informationnel qu'il a. Pour le comprendre, revenons à l'exemple de Putnam. Imaginons maintenant que l'on compare le contenu informationnel de deux « presque-jumeaux », le terrien Oscar 1 et un non-terrien jumeau Oscar 3, qui ont été exposés à deux types de régularités : le premier est entouré d'eau dont la composition moléculaire est H_2O, tandis que le second a appris que l'eau avait deux types de composition moléculaire possibles : H_2O ou XYZ. Imaginons enfin que, par hasard, et à son insu, toutes les expériences de l'eau qui ont été faites par Oscar 3 (le non-terrien) étaient des expériences d'H_2O. Le fait que rien ne permette de distinguer les échantillons d'eau — qu'il s'agisse en fait dans les divers cas rencontrés par nos deux sujets de la substance nommée « eau » sur terre — n'empêche pas qu'ils aient une valeur informationnelle différente ; cette différence vient de la différence entre les types d'information auxquels les sujets ont été exposés pendant leur période d'apprentissage, différence qui détermine le contenu des concepts en question [15].

15. L'une des difficultés de la thèse défendue par DRETSKE, 1981, réside dans le caractère flou et non rigoureux de la notion de « période d'apprentissage ». Comme l'écrit FODOR, 1987, p. 103 : « Il n'y a pas de moment à partir duquel l'usage que l'on fait d'un symbole cesse d'être simplement en voie de formation et commence à se faire, en quelque sorte, pour de bon. »

LE PROBLÈME DE LA ROBUSTESSE DU SENS

Néanmoins, la solution de Dretske, de même que les autres théories causales ou informationnelles du contenu, n'échappent pas à une série de difficultés évoquées dans les publications les plus récentes de Fodor, difficultés qui sont très largement imputables à ce que Fodor appelle la « robustesse » du sens, c'est-à-dire à la propriété en vertu de laquelle les occurrences d'un symbole peuvent être causées par les moyens les plus divers sans pourtant cesser de signifier une seule et même chose.

La première difficulté que Fodor a remarquée, dans *Psychosemantics,* tient à l'incapacité de ces théories de rendre compte de la *méprise représentationnelle (misrepresentation).* Prenons le cas d'une représentation de « vache » qui est causée par une occurrence de vache, mais qui peut également être causée, dans certaines circonstances (obscurité, brouillard...) par autre chose qu'une vache — par exemple un orignal. Faut-il en conclure que tantôt les occurrences de « vache » sont causées par des vaches, tantôt elles sont causées par des orignaux ? Dans ce cas, le contenu du symbole « vache » est la propriété disjonctive « être soit une vache soit un orignal ». Une telle théorie causale ne peut donc rendre compte de l'erreur de représentation. Elle ne parvient pas à distinguer le cas où le concept représenté est disjonctif (comme « être une vache ou être un orignal ») du cas où le concept de vache a par erreur été appliqué à quelque chose qui ne tombait pas sous lui (cf. Fodor, 1987, p. 102). Le défi qui se présente à une théorie *naturaliste* est de résoudre ce problème de l'erreur — qu'on nomme « problème de la disjonction » — sans s'appuyer circulairement sur des concepts intentionnels, en invoquant par exemple des « circonstances normales », ou « idéales », ou « sélectionnées par l'évolution », etc.

Fodor (1987) propose de résoudre le problème de la disjonction en faisant appel à l'asymétrie — laquelle n'est pas selon toute apparence intentionnelle — de la dépendance entre, respectivement, l'occurrence d'une propriété causant « normalement » le symbole mental de « vache » et ce symbole mental, et l'occurrence de la propriété « orignal » causant « accidentellement » ce même symbole et le symbole. C'est parce que les vaches donnent lieu à la représentation de « vache » que les orignaux conduisent parfois à dire, ou à se représenter « vache », tandis qu'il n'est pas vrai symétriquement que ce soit parce que les orignaux produisent le symbole de « vache » que les vaches conduisent à se représenter « une

vache ». Cette théorie permet d'expliquer l'erreur en termes non intentionnels, empiriques, de dépendance asymétrique entre des relations causales. Selon la formule de Fodor (1988b), « les occurrences fausses dépendent métaphysiquement des vraies » (p. 29).

Cependant, le problème de la disjonction, qui dans *Psychosemantics* est associé à la méprise représentationnelle, est d'application beaucoup plus large, comme Fodor le montre dans des textes encore inédits (1988a, 1988b). Il concerne non seulement l'application erronée des symboles, mais plus généralement *tous les cas où les occurrences symboliques ne sont pas causées par les objets ou propriétés entrant dans l'extension du symbole.* Le problème de la disjonction naît, en fait, précisément de la confusion ou de la distinction insuffisante entre « signifier » et « être porteur d'information ». L'information, dont un symbole est le porteur, est solidaire du lien causal qui s'établit entre une entrée perceptive et une représentation symbolique ; mais le sens est très largement indépendant de ce lien causal. En d'autres termes, à cause différente d'occurrences d'un certain type symbolique, information différente ; en revanche, « le sens d'un symbole fait partie des choses que toutes ses occurrences ont en commun, quelle qu'ait été leur histoire causale » (Fodor, 1988b, p. 28). Le sens a une « robustesse » qui fait défaut à l'information dont un symbole ou un objet, ou un événement peuvent être porteurs.

La généralité de la difficulté apparaît si l'on remarque qu'un problème de disjonction se manifeste dans un type de cas qui n'a rien à voir avec l'erreur : celui où un symbole est produit non par la perception d'un objet (soit dans l'usage où il fonctionne comme étiquette — cette forme appauvrie de langage que Wittgenstein évoque en tête de ses *Recherches*), comme dans le cas précédent (« c'est une vache », ou seulement « vache » !), mais par un autre symbole avec lequel il entretient une certaine relation (comme « une vache est un mammifère » ou bien « il y a des vaches en Suisse »), ou bien, les productions mentales « du coq à l'âne » si l'on peut dire, comme lorsqu'une occurrence mentale de taureau évoque une occurrence de vache.

Fodor étend à cette classe de cas la solution qu'il avait appliquée au cas de la méprise représentationnelle. Si un locuteur utilise, par exemple, dans le langage « augustinien » imaginé par Wittgenstein, l'expression « brique ! » non pas pour rapporter qu'il voit une brique, mais pour en demander une, cet usage est asymétriquement dépendant à l'égard d'un usage en quelque sorte fondamental, usage dans lequel existe une certaine relation causale entre une brique et mon expression « brique ». Ce que l'on exprime en termes de langage naturel vaut tout aussi bien du mentalais : on peut faire l'hypothèse qu'il puisse exister, à ce niveau, des

mécanismes qui sous-tendent les relations asymétriques entre divers usages des représentations.

Cette explication, pour ingénieuse qu'elle soit, pose un certain nombre de difficultés que je dois me contenter d'évoquer brièvement. Mais je commencerai par défendre Fodor contre une objection qu'on ne manquera pas de lui faire, et qui est fréquemment élevée contre toutes les tentatives de naturalisation. La théorie de Fodor ne contient-elle pas une circularité vicieuse, en se donnant le droit d'identifier les contenus — des référents de symboles mentaux tels qu'une vache ou un orignal — et en examinant les diverses situations contrefactuelles où ces référents pourraient causer conjointement ou non les symboles mentaux correspondants ? La réponse à cette question n'est pas aisée. Car en un sens, il y a circularité, puisque la théorie ne cherche pas à réduire les faits, les transformer dans un format où ils cesseraient de constituer un certain découpage des phénomènes que la langue naturelle ou scientifique nous transmet. Mais ce n'est pas parce que la théorisation serait restée en quelque sorte insuffisamment poussée qu'elle reste « au niveau des faits ». C'est qu'elle estime que c'est à ce niveau que se forment les conditions ultimes de l'intelligibilité, c'est-à-dire de l'objectivation scientifiquement effectuable.

Ce n'est pas dire pourtant que la solution de Fodor soit, selon ses termes, parfaitement « kasher » du point de vue naturaliste. L'une des exigences du naturalisme est de fournir une explication qui soit de type extensionnel, c'est-à-dire qui puisse être donnée dans un vocabulaire physicaliste. Or la théorie qui identifie les conditions suffisantes du contenu intentionnel dans l'information + l'asymétrie doit postuler le réalisme des propriétés. Ce réalisme permet entre autres de garantir que l'on puisse avoir des relations réglées entre des propriétés qui n'ont pas de porteur, et de leur appliquer de ce fait les manipulations contrefactuelles requises par la condition d'asymétrie. Par exemple, on a le droit de dire que toutes les occurrences de non-licornes ne peuvent causer une occurrence du concept Licorne que si des licornes causaient l'occurrence du concept Licorne au cas où elles existeraient (cf. Fodor, 1988b, p. 37). Mais on peut douter que la solution proposée soit naturaliste si quelques-uns des problèmes essentiels d'une théorie du contenu sont résolus par le truchement d'une ontologie intensionnelle. Par exemple, la question de l'indétermination de la référence, la question de la traduction, la question du vocabulaire logique et mathématique, la question de la nature des termes primitifs sont prérésolues — à bon compte d'un point de vue naturaliste strict — par une telle théorie qui peut identifier les contenus mentaux à des contenus objectifs transindividuels.

Une seconde limitation du naturalisme de cette solution réside dans la

postulation que l'on doit faire de l'existence de mécanismes régissant les divers emplois d'un symbole mentalais. Ces mécanismes sont une hypothèse qui reste pour le moment purement spéculative ; l'idée qu'il existe un contenu en quelque sorte fondamental des concepts en faveur duquel pourrait jouer la dépendance causale paraît difficilement recevable pour un naturaliste. Entre le fait de la robustesse sémantique et l'explication en termes de mécanismes mentaux sous-jacents qui en est finalement proposée, la théorie paraît se borner à proposer comme une solution ce qui ressemble encore à une reformulation du problème, ou, tout au moins, à un programme de recherche.

Fodor évoque lui-même un second ensemble de difficultés, qui tiennent au fait que la condition qui est présentée pour rendre compte du rapport intentionnel est satisfaite probablement par de nombreuses chaînes causales qui constituent elles aussi des cas de dépendance asymétrique (Fodor, 1988b, p. 55). Ne risque-t-on pas dès lors de tomber dans un « pansémantisme » ? Fodor propose de restreindre la classe des cas indésirables en considérant que la robustesse n'est pas caractéristique des rapports causaux non symboliques. Si l'on suppose que

> « les A comme A causent les B comme B, et que les B comme B causent les C comme C, la loi A → C dépend asymétriquement de la loi B → C [...] La dépendance des C à l'égard des A n'est robuste que s'il y a des C non causés par des A. Mais dans la chaîne causale A → B → C, tous les C causés par B sont aussi causés par A »,

ce qui montre que la relation A → C n'est pas robuste.

Un exemple permet pourtant de se demander si la robustesse n'est pas un phénomène courant dans les relations causales. Si un vent de force 8 fait chavirer un voilier, et si le naufrage entraîne la noyade des occupants, le rapport causal entre la force du vent et la noyade est symétriquement dépendant du rapport causal entre naufrage et noyade : si les organismes humains pouvaient s'oxygéner sous l'eau (c'est-à-dire si le lien causal entre B et C était rompu), la force du vent ne pourrait pas entraîner de noyades. La robustesse fait-elle défaut à cette chaîne causale, comme Fodor le dit à propos du cas général ? Ce serait le cas si la noyade des occupants ne pouvait être imputée à d'autres événements qu'à un vent de force 8. Or on peut imaginer qu'un vent de force 9 ou 10, qu'un raz de marée, qu'un tremblement de terre sous-marin, auraient pu avoir le même résultat. Quant au lien causal B → C, on peut imaginer que l'effet aurait pu être causé par une baignade imprudente, un suicide collectif, etc. On peut donc être justifié par la théorie à dire que la noyade signifie la force 8.

Ce que cet exemple montre accessoirement, c'est que la relation

causale elle-même paraît difficilement caractérisable de façon strictement non intentionnelle. Il paraît clair qu'une situation est causale sous une certaine description, soit en vertu de certains traits sémantiquement pertinents. On dira sans doute que l'épistémologie de la causalité ne doit pas être confondue avec l'ontologie physicaliste : la seconde est donnée dans le réel et non dans le discours. Mais ce que l'exemple précédent permet de soupçonner, c'est que l'intuition des contenus, omniprésente, risque de contaminer les critères naturalistes de l'intentionalité. En termes humiens, il paraît difficile en ce point d'abstraire la causalité de l'habitude, et de l'utiliser comme s'il s'agissait d'une relation qui n'était pas déjà construite sur un certain type de pertinence sémantique. Comme on le voit, l'explication naturaliste du contenu risque bien de rester encore longtemps un « idéal de la raison pure ».

Joëlle PROUST,
C.N.R.S.,
C.R.E.A., École polytechnique, Paris.

BIBLIOGRAPHIE

BURGE (Tyler), 1979, « Individualism and the Mental », eds. P. A. FRENCH, T. E. UEHLING, H. K. WETTSTEIN, Minneapolis, University of Minnesota Press, *Midwest Studies in Philosophy*, vol. 4.

BURGE (Tyler), 1982, « Other Bodies », *in* A. WOODFIELD, ed., *Thought and Object*, Oxford, Clarendon Press.

DRETSKE (Fred I.), 1981, *Knowledge and the Flow of Information*, Cambridge, MIT Press.

ENGEL (Pascal), 1986, « L'anomalie du mental », *Critique, 474*.

EVANS (Gareth), 1982, *Varieties of Reference*, ed. J. MCDOWELL, Oxford, Oxford University Press.

FODOR (Jerry), 1981, *Representations. Philosophical Essays on the Foundations of Cognitive Science*, Cambridge, MIT Press.

FODOR (Jerry), 1987, *Psychosemantics. The Problem of Meaning in the Philosophy of Mind*, Cambridge, MIT Press.

FODOR (Jerry), 1988a, « Information and Representation », manuscrit.

FODOR (Jerry), 1988b, « A Theory of Content », manuscrit.

FREGE (Gottlob), 1971, « Sens et dénotation », in *Ecrits logiques et philosophiques*, trad. Cl. IMBERT, Paris, Le Seuil.

JACOB (Pierre), 1989, « Le problème du rapport du corps et de l'esprit aujourd'hui ; essai sur les forces et les faiblesses du fonctionnalisme », manuscrit.

JACOB (Pierre), 1990, « Externalism Revisited : Is There such a Thing as Narrow Content ? », *Philosophical Review*, à paraître.

KAPLAN (David), 1989, « Demonstratives », *in* J. ALMOG, J. PERRY, H. WETTSTEIN, eds, *Themes from Kaplan*, Oxford, Oxford University Press.

PERRY (John), 1977, « Frege on Demonstratives », *Philosophical Review*, LXXXVI, p. 474-497.

PERRY (John), 1979, « The Problem of the Essential Indexical », *Noûs*, XIII, p. 3-21.

PROUST (Joëlle), 1981, « Sens frégéen et compréhension de la langue », in *Meaning and Understanding*, Berlin/New York, De Gruyter, p. 304-323.

PROUST (Joëlle), 1986, *Questions de forme. Logique et proposition analytique de Kant à Carnap*, Paris, Fayard.

PUTNAM (Hilary), 1975, « The Meaning of Meaning », in *Mind, Language and Reality, Philosophical Papers*, vol. 2, Cambridge, Cambridge University Press, p. 215-271.

PUTNAM (Hilary), 1988, *Representation and Reality*, Cambridge, MIT Press.

SEARLE (John R.), 1980, « Minds, Brains and Programs », *The Behavioral and Brain Sciences, 3*, p. 417-457.