Externalisme, Rationalité et explanandum de la psychologie intentionnelle

Elisabeth Pacherie

Pacherie, E. 1995. Externalisme, rationalité et *explanandum* de la psychologie intentionnelle, *Dialogue*, 34, 2: 237-57.

EXTERNALISME, RATIONALITE ET EXPLANANDUM DE LA PSYCHOLOGIE INTENTIONNELLE[1]

Elisabeth Pacherie
Institut Jean Nicod
CNRS-ENS-EHESS, Paris
pacherie@ehess.fr

A l'occasion de la tenue à Paris des Conférences Jean-Nicod de philosophie cognitive dont il a été le premier invité au printemps 1993, Jerry Fodor a présenté le dernier état de la théorie qu'il défend en philosophie de l'esprit[2]. Cette théorie se fonde sur trois thèses sur la signification et l'esprit qui, prises séparément, semblent chacune dotée d'une certaine plausibilité, mais dont la cohérence mutuelle apparaît sujette à caution. L'objet des conférences données par Fodor était de montrer que, contrairement aux apparences, ces trois doctrines peuvent être conciliées. Toutefois, de l'aveu même de Fodor, une telle conciliation n'est pas sans avoir un coût. Il concède notamment que sa stratégie de conciliation impose une révision du statut de certains phénomènes traditionnellement considérés comme des *explananda* pour une psychologie intentionnelle. Il soutient cependant que le coût de cette opération reste raisonnable puisque les phénomènes écartés sont de toute façon des phénomènes rares et marginaux.

Je voudrais ici contester les modes de calcul de Fodor, montrer qu'il sous-estime gravement le coût de cette conciliation -- qui implique en réalité une restriction drastique du domaine d'*explananda* de la psychologie intentionnelle. Je m'intéresserai plus spécifiquement au traitement que Fodor propose des cas frégéens et m'efforcerai de montrer que la tentative de marginalisation de ces cas opérée par Fodor prend appui sur une conception inadéquate de la

rationalité des actions. Afin de situer mon propos, je présenterai d'abord succinctement les trois thèses soutenues par Fodor et j'exposerai les difficultés que suscite leur conjonction. J'en viendrais ensuite à la solution proposée par Fodor et aux sacrifices demandés par cette solution, au nombre desquels le sacrifice des cas frégéens. J'examinerai le mode de calcul employé par Fodor pour estimer le prix de ce dernier sacrifice et contesterai le principe d'équilibre épistémique qui est à la base de ce calcul et qui, soutiendrai-je, repose lui-même sur une conception erronée de la rationalité des actions. Il sera alors temps de se demander, si étant donné une évaluation plus réaliste des coûts, le jeu en vaut encore la chandelle.

1. Les trois thèses fodoriennes

Les trois principales thèses aujourd'hui défendues par Fodor, telles qu'elles sont énoncés dans les Conférences Jean-Nicod, sont les suivantes :

(1) Hypothèse psychologique : les généralisations explicatives fiables de toute psychologie que nous pouvons aujourd'hui envisager seront intentionnelles de part en part.

(2) Hypothèse métaphysique : le contenu intentionnel se réduit à de l'information.

(3) Hypothèse computationnelle : les lois psychologiques sont implémentées par des processus computationnels.

De la première et de la troisième de ces thèses, Fodor s'est fait de longue date l'avocat. Son adhésion à la seconde marque une évolution plus récente de sa pensée, parallèle à l'évolution qui a eu lieu ces dernières années en philosophie de l'esprit en matière de réflexion sur la nature des contenus. D'autre part, la première de ces thèses témoigne de l'attachement de Fodor aux intuitions de la psychologie ordinaire tandis que la seconde et la troisième témoignent, chacune à sa manière, d'un souci naturaliste.

Dans les explications de la psychologie ordinaire, on considère que les comportements sont causés par des états mentaux intentionnels, désirs et croyances dotés d'un contenu. Fodor estime que le modèle explicatif de la psychologie ordinaire est fondamentalement correct même si dans leur détail les explications psychologiques ordinaires peuvent être erronées. Il considère

par conséquent que la tâche d'une psychologie scientifique est d'énoncer les généralisations intentionnelles (généralisations qui subsument les états mentaux en termes de leur contenus) qui sous-tendent le comportement des individus. Fodor admet que ces généralisations n'ont pas le caractère de lois strictes -- les lois strictes ne se rencontrant que dans les sciences fondamentales --, mais il considère que, à l'instar de ce qui se passe dans les autres sciences spéciales, les généralisations de la psychologie doivent avoir le caractère de lois *ceteris paribus*.

D'autre part, Fodor considère qu'une théorie n'a vraiment sa place dans l'édifice scientifique que pour autant qu'elle répond à des exigences naturalistes. Dans le cas de la psychologie, cela impose, d'une part, que l'on soit susceptible de formuler une théorie naturaliste des contenus, d'autre part, que l'efficacité causale supposée des contenus mentaux soit elle-même explicable en termes naturalistes. On peut considérer que la seconde et la troisième thèse défendues par Fodor s'efforcent chacune de répondre à ces exigences.

Dans l'hypothèse (2), l'hypothèse métaphysique, Fodor se rallie à une conception informationnelle externaliste des contenus mentaux : «Le contenu d'une pensée dépend de ses relations *externes*; de la manière dont la pensée est reliée au monde, *non de la manière dont elle est reliée à d'autres pensées*» (1994, p. 4). Fodor indique que deux types de raisons lui font soutenir cette hypothèse. La première a à voir avec le souci naturaliste qui doit, selon lui, inspirer une psychologie intentionnelle sérieuse. Les psychologues n'ont pas le droit de postuler l'existence de contenus intentionnels, s'ils ne pensent pas que ces contenus sont en principe naturalisables. Or, selon Fodor, de toutes les tentatives de naturalisation des contenus mentaux, seules les théories informationnelles semblent avoir une chance de réussite.

Le deuxième attrait des théories informationnelles est leur approche atomiste des contenus mentaux. Si la notion de contenu est explicable en termes d'information et si celle-ci est à son tour explicable en termes causaux, il doit être possible de déterminer le contenu d'un état mental en examinant la nature de ses connexions causales avec des objets de l'environnement sans avoir à prendre en compte où à supposer l'existence d'autres états mentaux et sans avoir à faire intervenir le relations causales que ces autres états mentaux,

s'ils existent, entretiennent avec l'environnement. On sait Fodor allergique au holisme sémantique[3], qui ne permet pas la formulation de lois intentionnelles générales et ainsi se trouve être incompatible en principe avec la première hypothèse défendue par Fodor. L'atomisme des théories informationnelles constitue donc un point en leur faveur car il garantit que ces théories ne sont pas en principe incompatibles avec une psychologie intentionnelle.

La troisième hypothèse, l'hypothèse computationnelle, manifeste une autre facette du souci naturaliste de Fodor. Fodor tient pour nécessaire qu'à toute loi d'une science spéciale, non fondamentale (*non basic*), doive être associé un mécanisme qui en assure l'implémentation. Toutefois, le point essentiel de l'hypothèse de Fodor n'est pas tant le fait qu'elle suppose nécessaire l'existence de tels mécanismes que l'idée que, dans le cas des lois de la psychologie, ces mécanismes doivent être computationnels. Qu'est ce qui motive cette hypothèse computationnelle? Premièrement, nous savons depuis Turing que les processus computationnels peuvent être réalisés mécaniquement par des systèmes physiques. La notion de mécanisme qui est ici à l'œuvre n'est donc pas métaphorique. Cette seule raison n'est toutefois pas suffisante. Il serait en effet plus simple de dire que les lois intentionnelles sont directement implémentées par des mécanismes biologiques et ainsi de faire l'économie de cette étape computationnelle intermédiaire. Ce sont, pour Fodor, certaines caractéristiques des processus mentaux régis par ces lois qui imposent cette étape intermédiaire. Ils manifestent en effet une cohérence sémantique -- préservation de la vérité --, une systématicité et une productivité dont seuls des mécanismes d'implémentation computationnels sont susceptibles de rendre compte.

Il semble que le rôle de l'hypothèse métaphysique de Fodor et celui de son hypothèse computationnelle soient de permettre la conciliation de sa première hypothèse sur le caractère intentionnel des lois de la psychologie avec ses préoccupations naturalistes, L'hypothèse informationnelle vise à rendre compte en termes naturalistes de la notion de contenu et l'hypothèse computationnelle vise en quelque sorte à naturaliser les lois de la psychologie en expliquant comment elles peuvent être implémentées. Quelle est donc la nature du problème? En première approximation, on pourrait dire que la stratégie préconisée par l'hypothèse métaphysique pour naturaliser le contenu

apparaît difficilement compatible avec la stratégie préconisée par l'hypothèse computationnelle pour implémenter les lois de la psychologie.

Cette incompatibilité apparaît plus nettement lorsque l'on examine de plus près les conditions que semble devoir remplir une théorie de l'implémentation. Une théorie de l'implémentation de la loi selon laquelle les Fs causent les Gs doit expliquer comment les Fs causent les Gs. Pour ce faire elle doit expliquer comment un mécanisme peut répondre aux trois conditions suivantes : (1) l'instantiation de F est une condition suffisante de l'instantiation d'une propriété MF du mécanisme implémentateur; (2) le mécanisme est tel que l'instantiation de MF conduit de manière fiable à l'instantiation de MG; et (3) l'instantiation de MG constitue une condition suffisante de l'instantiation de G.

Pour expliquer comment les conditions du type de (1) et (3) peuvent être remplies, les théories de l'implémentation peuvent avoir recours, dans les cas autres que les cas psychologiques, à deux possibilités : la réduction et, de manière plus controversée, la réalisation multiple. Le problème dans le cas de la psychologie est que la relation entre les propriétés intentionnelles et les propriétés computationnelles ne semble pouvoir être décrite ni dans les termes de l'une ni dans les termes de l'autre. Selon l'hypothèse métaphysique, les propriétés sémantiques d'un état mental dépendent des relations causales entre cet état et l'environnement externe. Selon l'hypothèse computationnelle, les propriétés computationnelles d'un état dépendent de la structure syntaxique interne de cet état. En d'autres termes, les propriétés sémantiques étant par nature extrinsèques et les propriétés computationnelles étant par nature intrinsèques, on ne voit pas comment on pourrait identifier propriétés syntaxiques et propriétés sémantiques, on ne voit pas non plus comment les unes pourraient constituer des réalisations fonctionnelles des autres et, de manière générale, on ne voit pas comment il pourrait y avoir des conditions computationnellement suffisantes pour la satisfaction de propriétés intentionnelles ou des conditions intentionnellement suffisantes pour la satisfaction de propriétés computationnelles

Face à cette difficulté, on serait tenté de penser que les trois thèses dont Fodor se fait l'avocat ne pouvaient être soutenues simultanément et que l'une au moins devait être abandonnée[4]. Bien entendu, ce n'est pas la réaction fodorienne.

2. La solution fodorienne

Fodor fait l'hypothèse que l'on peut sortir de ce trilemme apparent si l'on adopte une nouvelle conception de l'implémentation qui ne fasse appel ni à la notion de réduction ni à celle de réalisation multiple. Il propose de concevoir l'implémentation en termes d'harmonie ou de corrélation fiable. L'idée qu'il développe est que les propriétés intentionnelles et computationnelles, quoique distinctes et non corrélées nomologiquement, peuvent néanmoins être maintenues en phase de manière fiable et explicable. Pour expliquer son idée, il fait appel à l'analogie suivante. Soit les deux propriétés : *être un billet d'un dollar* et *avoir l'apparence d'un billet d'un dollar*. Il existe un certain nombre de généralisations qui font intervenir la propriété d'être un billet d'un dollar. Ces généralisations ne sont pas fondamentales et doivent être implémentées. De manière typique elles le sont par des processus qui mettent en jeu la propriété d'avoir l'apparence d'un billet d'un dollar.

Toutefois, cette relation d'implémentation ne peut être analysée en termes de réduction ou de réalisabilité multiple. La propriété que peut avoir une chose d'être un billet d'un dollar est une propriété extrinsèque qui dépend de son origine causale (avoir été émise par les autorités monétaires compétentes). En revanche, la propriété qu'a une chose d'avoir l'apparence d'un dollar dépend de propriétés internes à cette chose (sa couleur, sa taille, sa texture et ainsi de suite). Comme dans le cas des propriétés intentionnelles et computationnelles, ces deux propriétés sont de types différents, l'une est une propriété relationnelle extrinsèque, l'autre une propriété intrinsèque. Il n'y a pas de nécessité nomologique à ce qu'elles aillent de pair. Toute l'activité des faux-monnayeurs est au contraire fondée sur la possibilité qu'une chose ait l'une de ces propriétés sans avoir l'autre. Par ailleurs, le Département du Trésor américain pourrait décider d'émettre désormais des billets d'un dollar de couleur rouge sans que cela remette en cause leur propriété d'être des billets d'un dollar. Il existe néanmoins une corrélation fiable entre le fait de posséder une de ses propriétés et le fait de posséder l'autre. Cette corrélation est expliquée par l'existence d'un mécanisme chargé de maintenir les deux

propriétés en phase (à savoir un système de répression assuré par la police et chargé de mener la lutte contre les contrefaçons).

La proposition de Fodor consiste à soutenir que, de la même façon, la coinstantiation des propriétés intentionnelles et computationnelles, quoique métaphysiquement contingente, est néanmoins fiable et explicable. Evidemment, cette coinstantiation n'est pas explicable par l'existence de forces de police, mais elle l'est, nous dit Fodor, par l'existence de faits très généraux au sujet de l'organisation du monde réel et des mondes possibles proches où les lois intentionnelles de la psychologie sont implémentées de la même manière.

La stratégie de Fodor va donc consister à essayer de montrer que le monde réel et ses proches voisins sont ainsi faits qu'ils garantissent dans le cas général le maintien d'une correspondance bi-univoque entre propriétés intentionnelles et propriétés computationnelles. Cette harmonie entre propriétés intentionnelles et propriétés computationnelles peut en principe être perturbée de deux façons : soit que plusieurs propriétés intentionnelles correspondent à une seule et même propriété computationnelle; soit, inversement, que plusieurs propriétés computationnelles correspondent à une seule et même propriété intentionnelle. Or la littérature philosophique abonde de cas illustrant l'une ou l'autre de ces perturbations. L'existence de perturbations du premier type a été mise en évidence par les expériences de pensées de Putnam. La célèbre expérience de la Terre-Jumelle met en scène deux jumeaux moléculaires et computationnels placé l'un dans un environnement qui contient du H_2O, l'autre dans un environnement qui contient un liquide qui possède les mêmes qualités phénoménales mais dont la composition est XYZ. Selon Putnam, ces deux jumeaux, quoiqu'étant dans le même état computationnel lorsqu'ils ont des pensées au sujet de ce que chacun appelle «eau», sont néanmoins dans des états mentaux dont les propriétés intentionnelles sont différentes. Dans un cas «eau» fait référence à H_2O, dans l'autre à XYZ. En outre, selon Putnam, il n'est pas besoin de faire appel à la machinerie compliquée des mondes possibles pour faire surgir ce genre de situation. Tout individu qui ne sait pas distinguer les hêtres des ormes ou encore l'aluminium du molybdène, tout en sachant que des experts sont capables d'opérer cette distinction, se trouve dans une situation similaire.

Quand il pense à des hêtres, il est dans le même état computationnel que quand il pense à des ormes, alors même que ses pensées ont des conditions de vérité différentes et donc des contenus différents.

Inversement, les cas frégéens fournissent les exemples paradigmatiques de perturbations de la seconde espèce. Le principe de non-substituabilité des identiques dans les contextes de croyance est censé refléter le fait qu'un individu peut croire que Fa et ne pas croire que Fb alors même que a = b. Par exemple un homme peut croire que l'Etoile du Matin est lointaine et ne pas croire que l'Etoile du Soir est lointaine. Autrement dit, à la croyance que l'Etoile du Matin est lointaine et à celle que l'Etoile du Soir est lointaine correspondent chez cet individu des états mentaux et donc des états computationnels distincts alors même que les conditions de vérité des deux contenus de croyances sont identiques et donc que, selon une sémantique purement externaliste, ces deux contenus sont eux-mêmes identiques. L'existence de cas frégéens et le fait, que ces cas sont censés illustrer, que la référence ne détermine pas le sens, sont au cœur des raisons qui ont motivé la construction de théories du contenu étroit.

Si dans le monde réel les cas putnamiens ou les cas frégéens prolifèrent, la stratégie de Fodor, consistant à défendre une sorte de principe de parité entre propriétés intentionnelles et propriétés computationnelles, n'est plus tenable. D'une part, elle laisse en effet échapper des généralisations psychologiques importantes; d'autre part elle conduit à la formulation de généralisations inadéquates qui se solderont par un échec prédictif. Les cas putnamiens illustrent la première de ces difficultés, les cas frégéens la seconde. Si en présence de jumeaux, par exemple, on maintient que les lois intentionnelles computationnellement implémentées doivent être larges, on ne se donne pas les moyens de rendre compte de la similitude psychologique entre les deux jumeaux computationnellement identiques et l'on perd une généralisation importante. Si, dans un cas frégéen, on maintient que la croyance que Fa est identique à la croyance que Fb et donc que les mêmes généralisations intentionnelles valent dans les deux cas, on court à l'échec prédictif. Rien ne garantit que les sujets qui croient que Fa se comporteront comme ceux qui croient que Fb, puisque les mécanismes computationnels différents implémentent leurs croyances. Si le maintien des trois hypothèses chères à

Fodor a pour prix une théorie psychologique qui souffre à la fois d'un défaut de généralité et d'inadéquation prédictive, on peut se demander si le jeu en vaut encore la chandelle.

Pour maintenir la plausibilité de son entreprise, Fodor doit donc absolument montrer non pas que les cas putnamiens et les cas frégéens ne se rencontrent jamais -- ce qui en effet n'est pas nécessaire puisque les lois de la psychologie ne prétendent pas au statut de lois strictes --, mais simplement qu'ils sont rares et marginaux et que par conséquent ils n'entament que fort peu la généralité et l'adéquation prédictive de la théorie psychologique. L'objectif principal des conférences Jean-Nicod était de fournir une telle démonstration. Je voudrais quant à moi soutenir qu'au moins dans les cas frégéens la démonstration proposée n'est pas concluante.

3. Les cas frégéens

Fodor consacre une partie de sa deuxième conférence ainsi que la totalité de la troisième aux cas frégéens. La troisième conférence est toutefois consacrée à un cas frégéen bien spécifique, lié au puzzle quinien de l'inscrutabilité de la référence : le cas où deux propriétés sont invariablement coinstanciées quoiqu'elles ne soient pas coextensives (par exemple, *lapin* et *partie-non-détachée-de-lapin*). Je m'en tiendrai au traitement des cas frégéens familiers proposé par Fodor dans sa seconde conférence. L'ambition de Fodor est de montrer que l'on peut raisonnablement expliquer «pourquoi peu de cas frégéens peuvent surgir; ou, plus précisément, pourquoi peu d'entre eux peuvent surgir d'une manière qui conduirait à des échecs prédictifs / explicatifs des théories psychologiques du contenu large» (1994, p. 39). En bref, Fodor soutient que, dans le cas général, lorsqu'un agent croit que Fa et lorsque a = b, l'agent croit également que Fb et il se propose de donner une explication de ce phénomène.

On peut donner la reconstruction suivante de l'argument développé de manière plus informelle par Fodor (1994, pp. 40-43) :

(1) Les comportements rationnels sont normalement réussis.

(2) La relation entre rationalité et comportement réussi n'est pas accidentelle.

(3) Une psychologie intentionnelle doit rendre compte du fait que la relation entre rationalité et comportement réussi n'est pas accidentelle.

(4) Une psychologie intentionnelle crédible doit expliquer le comportement d'un agent comme largement déterminé par les interactions causales entre ses croyances et ses utilités.

(5) Le succès d'une action est accidentel à moins que les croyances sur la base desquelles l'agent agit ne soient vraies.

(6) Un agent ne peut pas rationnellement choisir A de préférence à B à moins de croire qu'il choisirait A de préférence à B si tous les faits lui étaient connus.

(C1) Principe d'équilibre informationnel : Les agents (rationnels) sont normalement en équilibre épistémique relativement aux faits sur la base desquels ils agissent. Avoir toute l'information -- avoir toute l'information que Dieu possède -- ne conduirait normalement pas l'agent à agir autrement qu'il n'agit.

(C2) un agent rationnel sait normalement que a = b, lorsque le fait que a = b possède une importance comportementale (*is behaviorally significant*).

(C3) Une psychologie intentionnelle ne peut rendre compte du fait que le relation entre rationalité et comportement réussi n'est pas accidentelle que si elle suppose (C1) et (C2).

Avant d'examiner plus en détail certaines des prémisses offertes, quelques mots sur l'économie générale de l'argument. La structure de l'argument est complexe dans la mesure où deux niveaux d'argumentation s'y trouvent imbriqués. A un premier niveau, l'argument porte sur ce qui est requis pour expliquer sans en faire un accident le fait que les comportements rationnels sont réussis. C'est à ce niveau qu'interviennent les prémisses (1), (2), (5), (6) et les conclusions (C1) et (C2) Ce premier niveau d'argumentation se trouve enchassé dans un second qui porte lui sur l'*explanandum* de la psychologie et sur ce qui est requis pour en rendre compte. En bref, étant donné ce qu'est son *explanandum*, la psychologie intentionnelle doit accepter les conclusions (C1) et (C2).

La prémisse (1) est présentée par Fodor comme une vérité empirique («*as a matter of fact*» (1994, p. 41)), comme un fait général sur l'organisation du monde réel. Le statut de la prémisse (2) n'est pas clairement explicité par Fodor. Il est quelque peu ambigu. La prémisse énonce-t-elle un fait empirique, un cas particulier d'une vérité générale selon laquelle les régularités empiriques ne

sont pas normalement accidentelles, ou bien un cas particulier d'un précepte de méthodologie scientifique qui prescrit de rechercher une explication systématique aux régularités observées ? Le cas n'est pas clairement tranché par Fodor. La prémisse (3) définit, en s'appuyant sur (1) et (2), un *explanandum* pour la psychologie intentionnelle. Elle hérite dans une certaine mesure de l'ambiguïté de la prémisse (2) S'agit-il pour la psychologie intentionnelle de proposer une explication en termes non accidentels du fait que les comportements rationnels sont généralement réussis ou bien s'agit-il d'expliquer le fait non accidentel que les comportements rationnels sont en général réussis? La prémisse (4) constitue une hypothèse sur la forme que peuvent prendre les explications psychologiques. Les prémisses (5) et (6) sont présentées par Fodor comme des truismes. La conclusion (C1) découle de (5) et (6) et (C2) découle de l'application de (C1) aux cas frégéens et en tire que les cas où un agent croit que Fa mais non que Fb alors que a = b constituent l'exception et non la règle. Enfin (C3) est une conséquence de (3), (5) et (6) prises ensembles.

3. 1 Rationalité

Examinons maintenant plus en détail les prémisses correspondant au premier niveau de l'argument et les relations qui les unissent. Fodor présente la première prémisse comme une vérité empirique. Mais son statut de vérité et en particulier de vérité empirique dépendent de ce que l'on entend, d'une part, par agent rationnel, d'autre part, par comportement réussi. La glose proposée par Fodor pour comportement réussi est «comportement tendant à la satisfaction des désirs de l'agent». Or, si un agent rationnel est défini comme un agent dont le comportement tend à la satisfaction de ses désirs ou si la définition qui est donnée de la rationalité est telle qu'elle implique des comportements tendant à la satisfaction des désirs, la prémisse (1) n'apparaît plus comme une vérité empirique mais comme une vérité conceptuelle. En outre, si la prémisse (1) est en fait une vérité conceptuelle, la prémisse (3) tombe. Il n'appartient pas à la psychologie en tant que discipline empirique de rendre compte de vérités conceptuelles.

Fodor nous doit donc une définition de la rationalité qui ne prenne pas la forme d'une pétition de principe, autrement dit, qui soit telle qu'elle n'implique pas logiquement la réussite comportementale. La prémisse (4) nous donne une première indication sur la forme que pourra prendre une telle définition de la rationalité. La rationalité va devoir être définie en termes de relations déterminées entre les croyances et les utilités ou préférences d'un agent[5]. Cette première indication est complétée par l'indication donnée dans la prémisse (6) qui énonce non une définition mais une condition nécessaire, selon Fodor, de la rationalité. Dans la mesure où cette condition porte sur les relations entre croyances et préférences et non sur les relations entre les croyances et le monde -- sur la vérité des croyances --, Fodor peut être lavé de l'accusation de pétition de principe. La réussite comportementale n'est pas une conséquence de la seule rationalité de l'agent (définie par les relations entre ses croyances et ses désirs), elle est une conséquence de la rationalité de l'agent jointe à la vérité de ses croyances. Le rôle de la prémisse (5) est précisement d'introduire cette deuxième condition. Toutefois, si l'on peut considérer avec Fodor que cette prémisse (5) peut être vue comme un truisme, il n'en va pas de même de la prémisse (6) qui non seulement n'a rien d'un truisme mais est encore extrêmement problématique.

Cette prémisse (6) est susceptible d'au moins trois interprétations entre lesquelles Fodor ne se prononce pas explicitement dans les Conférences Jean-Nicod. Mais quelle que soit l'interprétation retenue, elle impose sur l'action rationnelle des contraintes extrêmement fortes et singulièrement peu plausibles. Ces trois interprétations de la condition de rationalité de Fodor peuvent être mises en évidence soit au moyen du formalisme des mondes possibles, soit au moyen du formalisme de la théorie de la décision[6].

Suivant le modèle de formalisation en termes de mondes possibles proposé par Dubucs (1992), on se donne W, un ensemble de mondes possibles, C, un sous-ensemble de W, correspondant à l'ensemble des mondes possibles compatibles avec les croyances de l'agent et D, un sous-ensemble de W, correspondant à l'ensemble des mondes possibles compatibles avec les désirs de l'agent. Une action A est définie comme une application de W dans lui-même qui à chaque monde possible w_i élément de W associe le monde possible

w_j qui résulterait de l'action A dans w_i. On peut alors définir la notion d'optimalité d'une action. Une action A est optimale si et seulement si :

$\forall w$ (w C A(w) D)

Toutefois, pour pouvoir définir dans ce formalisme la notion de préférabilité entre actions, il nous faut prolonger la formalisation proposée par Dubucs et introduire une fonction de désirabilité ou, comme le disent les économistes, d'utilité, ce qui peut se faire en attribuant des utilités soit aux propositions atomiques, soit directement aux mondes possibles. La première option est à la fois plus complexe à mettre en œuvre et plus restrictive car, d'une part, elle nécessite la définition d'une relation de composition entre utilités pour pouvoir attribuer une utilité aux mondes possibles et, d'autre part, elle interdit la possibilité d'utilités conditionnelles[7]. La seconde option demande simplement que l'on définisse une application U, de W dans, par exemple, l'intervalle [0, 1]., telle notamment que $\forall w$ (w D U(w) = 1).

Si l'on s'en tient à la définition couramment utilisée de la préférence rationnelle en termes de maximisation, celle-ci peut alors être définie de la manière suivante. Il est rationnel pour un agent de préférer A à B étant donné ses croyances et ses désirs si et seulement si :

$\mathfrak{R}_{w \quad C}$ U (A (w)) > $\mathfrak{R}_{w \quad C}$ U (B (w)).

Or, la condition de Fodor est plus forte. L'énoncé hypothétique «l'agent croit qu'il choisirait A de préférence à B si tous les faits lui-étaient connus» peut être compris de trois manières :

(1) L'agent croit que tous les faits lui sont connus, autrement dit il croit que C = {w*} et il croit que U (A (w*)) > U (B (w*))

Cette première interprétation revient à dire que l'agent n'agit rationnellement que s'il se croit omniscient, non pas simplement logiquement omniscient -- ce qui est déjà beaucoup --, mais *empiriquement* omniscient!

(2) L'agent admet que ses croyances peuvent ne pas être exhaustives ou, au moins pour certaines d'entre elles, être erronées, mais il croit que A domine B dans tous les mondes possibles, autrement dit que

$\forall w$ W, U (A (w)) > U (B (w))

Selon cette seconde interprétation, il n'est pas nécessaire qu'un agent se croie omniscient pour agir rationnellement, mais il est nécessaire qu'il croie que l'action qu'il préfère est absolument dominante.

(3) Enfin, on peut suggérer une troisième interprétation, en quelque sorte intermédiaire entre les deux précédentes. L'agent ne croit pas que tous les faits lui soient connus mais il croit que tous les faits pertinents lui sont connus et que, eu égard à ces faits, l'action A domine l'action B. En termes plus formels, il croit que w* C et que ∀w C, U (A (w)) > U (B (w)). Cette troisième interprétation conjugue, pourrait-on dire, un principe d'omniscience locale et un principe de dominance locale.

La première interprétation est outrageusement implausible. La condition de Fodor ainsi interprétée apparaît moins comme une condition de rationalité que comme la définition d'une forme aiguë de mégalomanie. Or, les agents rationnels humains sont des êtres finis et, exception faite des mégalomanes patentés, ils en sont généralement conscients. La deuxième interprétation apporte une restriction considérable quant aux situations dans lesquelles un agent peut se comporter rationnellement. Selon cette interprétation, une action ne peut être rationnellement choisie par un agent que s'il la croit dominante quelles que soient les circonstances (*i. e.*, quel que soit le monde possible effectivement réalisé). Mais il s'agit là d'une conception extrêmement frileuse de la rationalité qui limite son exercice aux situations qui sont crues dépourvues de risques, alors que, semble-t-il, l'intérêt principal de l'exercice de la rationalité est de nous permettre d'agir en situation d'incertitude en prenant des risques calculés. Soit, par exemple, la situation suivante : on propose à un agent de choisir entre deux boîtes A et B, dans l'une desquelles on a placé au hasard et hors de sa présence, une boule blanche. Si l'agent choisit la boîte A et que celle-ci contienne la boule blanche, il gagne 1000 F, si elle est vide, il gagne 100 F; s'il choisit la boîte B et qu'elle contienne la boule blanche, il gagne 101 F, sinon 1 F. Supposons que l'agent croie ce qu'on lui dit et désire gagner le plus d'argent possible. Intuitivement, mais aussi selon la théorie de la rationalité économique, lorsque la fonction d'utilité adoptée prend une des formes courantes, il est rationnel de choisir A plutôt que B. Mais, selon la condition de rationalité de Fodor, il n'y a pas de choix rationnel entre A et B, puisque ni A ni B ne dominent absolument.

Prima facie, la troisième interprétation pourrait sembler la plus raisonnable. Mais à y regarder de plus près, tel n'est pas le cas, car elle conjugue les difficultés des deux interprétations précédentes. D'une part, comme la seconde

interprétation, elle restreint l'exercice de la rationalité aux situations de dominance, ou aux situations crues telles. D'autre part, la notion d'omniscience locale est sujette à caution. Un agent serait (croirait être) localement omniscient s'il connaissait (croyait connaître) l'ensemble des faits pertinents pour un problème donné, sans pour autant connaître (croire connaître) la totalité des faits. Mais comment est-il possible de croire que l'on connaît tous les faits pertinents alors même que l'on admet ne pas connaître tous les faits? Comment peut-on croire *a priori* que des faits que l'on ignore ne sont pas pertinents pour le problème qui nous intéresse? De l'avis même de Fodor (1983, 1987), dans la mesure où la situation à laquelle un agent est confronté est une situation empirique et où le résultat de ses actions dépend du système de relations causales qui existe dans le monde, toute restriction dans le choix des données considérées est arbitraire et irrationnelle. Parce que nous ne savons pas à l'avance comment les relations causales sont disposées, nous devons être prêts à changer d'avis sur ce qui est pertinent. Se croire localement omniscient est donc aux yeux de Fodor irrationnel, à moins de supposer que cette omniscience locale n'est que la conséquence d'une omniscience empirique générale. En bref donc, si un agent se croit localement omniscient, sa croyance ne peut être rationnelle que s'il se croit aussi généralement omniscient et dans ce cas, l'interprétation (3) de la condition de rationalité se ramène à l'interprétation (1). S'il se croit localement omniscient sans se croire généralement omniscient, sa croyance est irrationnelle. L'interprétation (3) de la condition de rationalité de Fodor devient alors terriblement suspecte, puisqu'elle revient à définir la rationalité en termes d'irrationalité : un agent ne peut pas choisir rationnellement A de préférence à B à moins de croire, irrationnellement, qu'il est localement omniscient.

L'interprétation donnée de la prémisse (6) rejaillit sur (C1), le principe d'équilibre informationnel, qui est une conséquence de (5) et (6) prises ensembles. Si l'on adopte la première interprétation de la prémisse (6), le principe d'équilibre épistémique peut être reformulé de manière lapidaire : les agents rationnels sont normalement omniscients. (C2) est bien évidemment une conséquence immédiate de (C1) ainsi interprété : un agent omniscient ne peut pas manquer de savoir que Fb si Fa et a = b. Si, en revanche on adopte la seconde interprétation de la prémisse (6), l'exercice du choix rationnel est

soumis au principe de dominance. Le choix d'une action n'est rationnel que pour autant que l'agent croit qu'elle est dominante quels que soient les faits. De la prémisse (5), prise avec la prémisse (6) ainsi interprétée, il s'ensuit qu'un agent agit rationnellement lorsque l'action qu'il choisit est dominante quels que soient les faits. Il est alors trivialement vrai que les agents rationnels sont en équilibre épistémique, autrement dit qu'«avoir toute l'information -- avoir toute l'information que Dieu possède -- ne conduirait normalement pas l'agent à agir autrement qu'il n'agit», pour la simple raison que l'action est préférable quels que soient les faits. Mais si l'on admet cette interprétation triviale car vide de (C1), on est conduit à une interprétation également triviale et tout aussi vide de (C2) : un agent rationnel sait normalement que a = b, lorsque le fait que a = b possède une importance comportementale et, en tout état de cause, dans les situations de dominance où la rationalité peut s'exercer, le fait que a = b est sans importance comportementale.

3. 2. *Explanandum* de la psychologie

Il nous reste maintenant à examiner les conséquences que l'on peut tirer de tout cela pour le second niveau de l'argument. Rappelons pour commencer que le but de Fodor est d'expliquer ce qui maintient en phase propriétés intentionnelles et propriétés computationnelles et, en particulier d'expliquer ce qui empêche la prolifération des cas frégéens. Dans son analogie avec les dollars, Fodor suggérait qu'il fallait, pour atteindre cet objectif, mettre en évidence des faits sur l'organisation du monde réel qui jouent à l'égard des propriétés intentionnelles et computationnelles le rôle joué par les forces de police à l'égard des propriétés d'être un billet d'un dollar et d'avoir l'apparence d'un billet d'un dollar. On s'attendrait donc, pour ce qui est des cas frégéens, à ce que Fodor nous donne une description des mécanismes chargés d'empêcher leur prolifération. Or tel n'est pas l'objet de l'argument qu'il propose. La forme générale de son argument est celle d'une démonstration d'existence : le fait que les comportements rationnels sont normalement réussis ne pourrait apparaître que comme un accident si les cas frégéens proliféraient, or ce fait n'est pas accidentel, donc les cas frégéens ne prolifèrent pas. L'argument, on le voit est

totalement muet quant à la nature des mécanismes qui empêchent cette prolifération.

Le seul fait empirique général sur lequel s'appuie l'argument est que les comportements rationnels sont normalement réussis, or ce fait ne constitue pas une explication, au contraire il a lui-même besoin d'être expliqué. Il constitue un *explanandum* pour la psychologie. Plus exactement, le raisonnement de Fodor -- au second niveau de l'argument -- semble être le suivant : les comportements rationnels constituent l'*explanandum* de la psychologie intentionnelle, ces comportements rationnels ont pour caractéristique essentielle d'être normalement réussis, par conséquent, la psychologie intentionnelle se doit de rendre compte de cette caractéristique. Comme il a été établi au premier niveau de l'argument que cette caractéristique ne s'expliquait qu'à condition de supposer que les agents sont en état d'équilibre épistémique et que, par conséquent, les cas frégéens ne prolifèrent pas, il s'ensuit que la psychologie intentionnelle doit faire cette supposition. Remarquons incidemment que, même si l'on accepte les prémisses sur lesquelles se fonde le raisonnement fodorien, la conclusion tirée semble trop faible. Admettons que la psychologie ait pour objet d'expliquer le succès des comportements rationnels et admettons également que ce succès ne soit explicable que si les agents sont en état d'équilibre épistémique, il s'ensuit, semble-t-il, que la psychologie intentionnelle, pour autant qu'elle se veuille entreprise naturaliste, doive rendre compte du fait que les agents sont en équilibre épistémique et non simplement supposer ce fait.

Toutefois, il n'est peut-être pas nécessaire de s'engager dans cette voie. Dans la mesure où les doutes que peut inspirer l'argument examiné au premier niveau rejaillissent sur le second niveau, on peut se demander si Fodor ne fait pas fausse route dès le départ. Etant donné les deux interprétations possibles de la prémisse (6) et par conséquent du principe d'équilibre informationnel, il semble que nous n'ayons le choix qu'entre Charybde et Scylla. Si l'on choisit la première interprétation, la psychologie intentionnelle doit pour expliquer le succès du comportement rationnel admettre que les agents rationnels sont omniscients. A mon sens, elle ne doit d'ailleurs pas se contenter de l'admettre, elle doit démontrer que l'omniscience est possible pour des agents finis dotés de capacités cognitives limitées. Les chances de succès d'une telle entreprise

semblent pour le moins limitées. Si l'on choisit la seconde interprétation, on échappe certes au Charybde de l'omniscience, encombrant *explanans*, mais c'est pour réduire l'*explanandum* de la psychologie intentionnelle à la portion congrue. Si une action n'est caractérisable comme action rationnelle que pour autant qu'elle est totalement dominante -- que son succès n'est pas lié à des faits qui ne dépendent pas de l'agent -- la psychologie intentionnelle pourra certes expliquer sans trop de peine la réussite d'une telle action. Mais à ainsi modeler la rationalité sur des principes stoïciens, on donne à l'*explanandum* de la psychologie intentionnelle des allures de peau de chagrin. Seule une infime partie des comportements d'un agent peuvent être qualifiés de rationnels et seule cette infime partie des comportements constitue un *explanandum* pour la psychologie intentionnelle. Si le prix à payer pour la conciliation des trois hypothèses fodoriennes consiste soit à admettre que les agents rationnels sont omniscients, soit en un extrême rétrécissement de l'*explanandum* de la psychologie, on peut se demander si le jeu en vaut vraiment la chandelle.

Il est bien sûr possible de penser que les formulations employées par Fodor ont dépassé sa pensée et qu'il se suffirait d'une conception moins exigeante de la rationalité. Mais il semble qu'une reformulation moins exigeante de la prémisse (6) ne permette plus d'aboutir à la conclusion désirée par Fodor. Supposons qu'à la prémisse (6) de Fodor, on substitue la prémisse suivante :

> (6') Un agent ne peut rationnellement choisir A de préférence à B à moins que l'utilité attendue de A ne soit supérieure à l'utilité attendue de B, étant donné ses croyances.

Selon cette formulation, la rationalité n'est plus définie en termes de dominance mais simplement de maximisation. Si l'on reprend l'exemple des boîtes donné plus haut, il devient possible de dire qu'un agent agit rationnellement en choisissant la boîte A de préférence à la boîte B. Maintenant toutefois, même si les croyances sur lesquelles l'agent a fondé son choix sont vraies (la croyance que la situation n'était pas truquée, qu'il y avait une chance sur deux pour que la boule blanche soit dans la boîte A, etc.), le succès de ce choix rationnel n'est pas garanti. Il se pourrait qu'en choisissant la boîte A l'agent gagne un franc de moins que s'il avait choisi la boîte B. Ceci oblige à modifier à son tour la prémisse (1), que l'on peut alors reformuler de la manière suivante :

(1') les comportements rationnels sont normalement plus réussis que les comportements non-rationnels.

Mais pour rendre compte de (1'), il n'est pas nécessaire de supposer que les cas frégéens sont marginaux. Il suffit de supposer que les agents rationnels ne sont pas plus exposés à des cas frégéens que les agents qui agissent irrationnellement. Etant donné que les agents, qu'ils soient rationnels ou qu'ils ne le soient pas, ne sont pas omniscients, il est toujours possible qu'une action entreprise en vue de la satisfaction d'un désir ne rencontre pas le succès escompté parce que l'agent ignore l'identité de a et de b, alors même que cette identité est importante dans le cas considéré. Mais il n'y a pas de raison de penser qu'un agent rationnel soit plus susceptible d'ignorer ces identités importantes qu'un agent irrationnel. Supposer qu'il en aille ainsi, reviendrait à faire l'étrange supposition d'une conspiration du monde contre la rationalité. Si l'on refuse cette supposition, on peut parfaitement expliquer qu'à ignorance égale, l'agent rationnel rencontre plus de succès que l'agent irrationnel (parvienne à une meilleure satisfaction de ses désirs), puisque contrairement à l'agent irrationnel, l'agent rationnel entreprend les actions qui maximisent son utilité.

On pourrait nous reprocher ici d'avoir substitué à une définition trop exigeante de la rationalité une conception purement instrumentale, trop faible, et en particulier de n'avoir pas préservé le lien établi par Fodor entre rationalité et information. Pour éviter ce reproche, supposons qu'à la prémisse (6) de Fodor , soit maintenant substituée la prémisse suivante :

> (6'') Un agent ne peut rationnellement choisir A de préférence à B à moins de croire que ce choix prend en compte toute l'information dont il dispose ou qu'il pourrait acquérir en temps voulu.

Selon cette formulation, la rationalité n'est plus définie en termes de dominance ou d'omniscience. Toutefois, cette définition de la rationalité n'est pas purement instrumentale. Elle ne reprend pas simplement le principe de maximisation mais préserve un caractère essentiel de la prémisse originale de Fodor, à savoir le lien entre rationalité et information. La prémisse modifiée a toutefois une plus grande plausibilité dans la mesure où elle prend en compte les contraintes temporelles et cognitives auxquelles sont soumis les agents rationnels finis. Etant fini, un agent ne peut acquérir et traiter qu'une quantité finie d'information sur la base de laquelle prendre une décision, étant en outre

soumis à des contraintes temporelles -- les décisions doivent être prises quand il en est encore temps --, un agent ne dispose que d'un temps fini pour collecter des informations. Selon la prémisse (6''), ce qui distingue un agent rationnel d'un agent irrationnel n'est donc ni l'omniscience ni la soumission au principe de dominance, mais le fait d'agir de manière réfléchie et non précipitée. Toutefois, ces précautions n'immunisent pas l'agent rationnel contre l'échec. Ainsi, selon cette prémisse (6'') on peut encore dire, si l'on reprend une dernière fois l'exemple des boîtes, qu'un agent agit rationnellement en choisissant la boîte A de préférence à la boîte B. Quelque vraies que soient les croyances de l'agent et quelque rationnelle que soit sa décision, l'échec est encore possible. Et donc le choix de la prémisse (6''), comme celui de la prémisse (6') nous oblige à abandonner la prémisse (1) au profit de la prémisse (1').

Cette fois-ci nous n'expliquerons plus (1') en disant qu'à ignorance égale, l'agent rationnel maximise et l'agent irrationnel ne maximise pas. Nous dirons que les actions de l'agent rationnel soit normalement plus réussies parce que mieux informées. Toutefois, cette réussite est relative et non absolue. Etant donné que les agents, même rationnels, ne sont pas omniscients, il est toujours possible qu'une action entreprise en vue de la satisfaction d'un désir ne rencontre pas le succès escompté parce que l'agent ignore l'identité de a et de b, alors même que cette identité est importante dans le cas considéré. Un agent peut parfaitement être rationnel et néanmoins -- étant donné les contraintes auxquelles il est soumis -- ne pas avoir accès une information importante, il peut même, comme dans le cas des boîtes, faire un choix rationnel, alors qu'il sait précisément quelle information lui manque -- à savoir, dans quelle boîte se trouve la boule blanche.

Si la psychologie intentionnelle a pour objet d'expliquer non pas que les comportements rationnels sont normalement réussis, mais, plus modestement, que les comportements rationnels, quoique susceptibles d'échouer, sont normalement plus réussis que les comportements non-rationnels, il n'est pas nécessaire pour donner une explication de ce phénomène de supposer que les cas frégéens sont marginaux. On peut parfaitement maintenir que les cas frégéens sont relativement fréquents et dire que le propre du comportement

rationnel est d'essayer dans la mesure du possible de prévenir ceux parmi les cas frégéens qui pourraient influer négativement sur le résultat des actions qu'il envisage.

Toutefois, on peut contester que le rôle de la psychologie intentionnelle se limite à expliquer pourquoi les comportements rationnels sont normalement plus réussis que les comportements non-rationnels. Dans la mesure où l'échec de comportements rationnels n'est pas une exception liée à la prévalence de circonstances anormales, mais une chose relativement fréquente, explicable par les contraintes internes et externes auxquelles sont normalement soumis les agents rationnels, il semblerait que le rôle de la psychologie intentionnelle doive être d'expliquer les comportements rationnels en général, qu'ils soient ou non réussis. Mais alors, si l'on veut encore expliquer les comportements rationnels en termes de lois intentionnelles, il faut que les lois intentionnelles soient définies en termes de contenus étroits, car définies en termes de contenu large, elles conduiraient à des échecs prédictifs.

Enfin, on peut aussi se demander si donner pour but à la psychologie intentionnelle de rendre compte des comportements rationnels, quels qu'en soient les résultats, n'est pas encore restreindre à l'excès son *explanandum*. Il semblerait que bon nombre de comportements que nous ne considérerions pas comme rationnels mettent en jeu des états mentaux représentationnels. Pour ne citer que quelques exemples, les travaux de Kahnemann et Tversky sur les biais qui influent sur nos jugements de probabilité, les travaux sur les prototypes en psychologie de la catégorisation ou encore les travaux sur la dissonance cognitive -- qui suggèrent qu'à défaut de pouvoir mettre nos actes en accord avec nos désirs et nos croyances, nous avons tendance à modifier nos croyances et désirs pour les mettre en conformité avec nos actes -- mettent tous en évidence certaines formes de systématicité de nos comportements. Il semblerait naturel d'exprimer ces systématicités en termes de généralisations portant sur les contenus, pris au sens de contenus étroits.

Fodor a été détourné des contenus étroits par la crainte qu'une théorie des contenus étroits ne mène tout droit au holisme sémantique, qui rendrait caduc le projet d'énoncer des généralisations intentionnelles intéressantes. J'ai essayé de montrer dans un autre article (Pacherie, à paraître) que cette crainte était exagérée et qu'il était possible de définir une notion substantielle de contenu

étroit sans tomber dans le holisme sémantique. Si l'on fait usage d'une telle notion, le projet d'une psychologie intentionnelle peut, me semble-t-il, être préservé sans rétrécissement abusif de son *explanandum*. Cela ne signifie pas que la tâche soit aisée. Il faut encore expliquer comment contenu étroit et contenu large, défini en termes informationnels, s'articulent; il faut également expliquer comment les contenus -- larges ou étroits -- sont implémentés. Nous ne sommes donc pas au bout de nos peines, mais, *prima facie* à tout le moins, une telle entreprise n'a pas le caractère auto-destructeur de la démarche fodorienne.

Appendice

Modélisation dans le formalisme de la théorie de la décision.

Soit E, un ensemble fini exhaustif d'éventualités mutuellement incompatibles, soit Pr : E -> [0, 1], qui à chaque e, élément de W, associe un degré de crédibilité, soit U : E -> [0, 1], qui à chaque e associe un degré de désirabilité (une utilité), soit enfin A, une une action définie comme une application de E dans E qui à chaque éventualité e_i élément de W associe l'éventualité e_j qui résulterait de l'action A dans e_i.

L'utilité attendue de A est définie de la manière suivante :

$V(A) = \Re_{e \quad E} Pr(e) . U(A(e))$

A est préférable à B si et seulement si :

$V(A) > V(B)$.

L'avantage de ce formalisme dans sa version standard par rapport au formalisme des mondes possibles est qu'il permet de représenter des degrés de croyance et de désir, son inconvénient en retour est de ne pas permettre de représenter l'ignorance absolue, Pr et U doivent être définies sur tout E.

Selon Dubucs (1992), si l'on exige que Pr et U prennent leurs valeurs dans la paire {0, 1} et non dans l'intervalle réel [0, 1], P et U peuvent être considérées comme des fonctions caractéristiques des ensembles C et D respectivement et l'utilité attendue de A se ramène alors à $\Re_{e \quad E} U(A(e))$ Dans ce cas particulier, la théorie de la décision coïncide, selon Dubucs, avec le formalisme des mondes possibles. Toutefois, dans la version standard de la théorie de la décision, E étant un ensemble d'éventualités mutuellement incompatibles, on exige de la fonction Pr qu'elle vérifie la condition suivante, que l'on peut appeler principe de cohérence :

$\Re_{e \quad E} Pr(e) = 1$.

Si l'on conserve cette condition, on aboutit à ce que C, tel que le définit Dubucs, soit tel que $|C| \leq 1$, ce qui semble trop restrictif. Pour que le formalisme de la théorie de la décision coïncide véritablement avec le formalisme des mondes possibles, on a donc le choix entre exiger de Pr qu'il prenne ses valeurs dans la paire {0, 1} et renoncer au principe de cohérence, afin de pouvoir considérer Pr comme la fonction caractéristique de B ou bien

encore laisser Pr prendre ses valeurs dans l'intervalle [0, 1] et définir C de la façon suivante :

e C <=> pr (e) > α , avec $0 \leq \alpha < 1$

Si l'on choisit $\alpha = 0$,

V_C (A) = $\Re_{e \quad C}$ pr (e). U(A(e))

se ramène à

V (A) = $\Re_{e \quad E}$ Pr (e) . U(A(e))

Dans ce formalisme, les trois interprétations possibles de la condition de rationalité de Fodor (l'agent croit qu'il choisirait A de préférence à B si tous les faits lui étaient connus) peuvent être représentées de la manière suivante :

(1) Omniscience :

L'agent croit que C = {e*} et que U (A (e*) > U (B (e*)).

(2) Principe de dominance :

L'agent croit que $\forall$e E U (A (e)) > U (B (e))

(3) Omniscience et dominance locales :

L'agent croit que e* C et que $\forall$e C U (A (e)) > U (B (e))

Références bibliographiques

Dubucs, J. «Omniscience Logique et Frictions Cognitives», dans D. Andler *et al.* (éds), *Epistémologie et Cognition*, Liège : Mardaga, 1992, pp. 115-131.

Fodor, J. A., *The Modularity of Mind*, Cambridge Mass. : MIT Press, 1983.

Fodor, J. A., «Modules, Frames, Fridgeons, Sleeping Dogs, and the Music of the Spheres», Dans Z. Pylyshyn (éd.) *The Robot's Dilemma - The Frame Problem in Artificial Intelligence*, Norwood, N.-J. : Ablex, 1987, pp. 139-149.

Fodor, J. A. *The Elm and the Expert*, Cambridge, Mass. Bradford Books, 1994.

Pacherie, E., «Holophobia», *Acta Analytica*, 12, pp. 105-112, 1994.

Notes

[1] Une version antérieure de ce texte a fait l'objet d'un exposé au Séminaire d'Epistémologie Comparative d'Aix-en-Provence. Je remercie les membres du Séminaire, notamment François Clementz, Pierre Livet et Elisabeth Schwartz, pour leurs remarques et suggestions. Je remercie également de leurs critiques et éclaircissements Daniel Andler, Frank Döring, André Orléan et Joëlle Proust, ainsi que les deux rapporteurs anonymes de *Dialogue*.

[2] Cf. Fodor,1994, pour une version remaniée de ces conférences.

[3] Cf. Fodor, 1987, notamment chapitre III et Fodor & Lepore, 1992.

[4] Trois types de réactions sont en gros possibles. La première, illustrée par Stich et d'une certaine façon Dennett, consiste à défendre une théorie syntaxique de l'esprit et à renoncer à l'hypothèse que les lois psychologiques sont intentionnelles. La seconde, préconisée par les partisans d'une sémantique fonctionnaliste, consiste à dénoncer l'hypothèse informationnelle et à soutenir l'idée d'une théorie sémantique duale ou internaliste: les lois de la psychologie sont bien intentionnelles mais elles font référence aux contenus étroits des pensées. La troisième, défendue par Searle, Dreyfus, Chomsky et parfois Churchland, consiste à faire l'impasse sur l'hypothèse computationnelle et à soutenir que les lois intentionnelles sont directement implémentées par des mécanismes biologiques.

[5] Quoique Fodor (1994) n'évoque les théories de la décision qu'en termes très généraux et ne se réclame pas explicitement d'une théorie particulière, ses propos dans *The Elm and the Expert*, comme, à ma connaissance, dans ses autres écrits, laissent penser qu'il accepte le modèle standard de la décision rationnelle.

[6] Voir appendice pour une modélisation dans le formalisme de la théorie de la décision.

[7] Si l'on attribue une utilité directement aux propositions atomiques, il devient très difficile de définir une quelconque relation de composition entre utilités, même intuitivement implausible, telle que par exemple:

U (p & q) > U (- p & - q) > U(- p & q) > U (p & - q)

et en tout cas totalement impossible que si U (q) = U (r) on ait U (p & q) ≠ U (p & r). En d'autres termes, on s'interdit de faire dépendre l'utilité d'une proposition donnée de la vérité ou de la fausseté d'une autre proposition donnée.