1 Fonctionalisme RL

• La seconde forme de fonctionnalisme que nous allons étudier est celle dont on trouve l'exposition canonique dans Lewis 1972. Je l'appelle le fonctionnalisme RL parce qu'il est élaboré par Lewis, à partir de certains outils développés dans l'entre-deux-guerres par le philosophe et logicien britannique F.P. Ramsey. Dans la littérature anglosaxonne, le fonctionnalisme TP a un nom plus ou moins canonique : *"machine functionalism"*. Il n'y a pas d'analogue qui fasse l'unanimité pour le fonctionnalisme RL : pour mémoire, J. Kim parle de *"causal-theoretic functionalism"*, Braddon-Mitchell et Jackson de *"common-sense functionalism"* ou d'*"analytic functionalism"*.

• Deux remarques préliminaires :

1. Le fonctionnalisme RL se présente comme une théorie sémantique : elle propose une analyse de la *signification* des termes mentaux. On a vu que tel n'était pas l'objectif du fonctionnalisme TP.
2. Le fonctionnalisme RL se présente comme une application aux termes mentaux d'une théorie générale de la signification des *termes théoriques.*

1.1 La sémantique fonctionnelle des termes théoriques

• La sémantique fonctionnelle des termes théoriques est une réponse à une question lancinante de la philosophie des sciences du XXème siècle : la question de la signification des termes théoriques. C'est en particulier une question qui était centrale chez les positivistes logiques. Voici, en quelques mots, pourquoi.

1. On a vu que le positivisme logique avait une doctrine vérificationniste de la signification : le sens d'un énoncé est l'ensemble de ses conditions de vérification. Les positivistes ajoutaient que vérifier, c'est observer. D'où l'idée que seuls les concepts observables ont un sens immédiat.
2. Or, les sciences modernes, et en premier lieu la physique, font appel à des concepts qui ne sont pas, selon la typologie positiviste, observables.

> Carnap, *Philosophical Foundations of Physics*, Basic Books, 1966 ; trad.fr. *Les fondements philosophiques de la physique*, A. Colin, 1973

> "Dans mon vocabulaire, les lois empiriques sont celles qui contiennent des termes correspondant à des phénomènes soit directement observables par les sens, soit mesurables par des techniques relativement simples."

> "...les termes d'une loi théorique renvoient non pas à des termes observables, mais à des entités telles que les molécules, les atomes, les électrons, les protons, les champs électromagnétiques et d'autres qui ne peuvent pas être mesurées par des procédés simples et directs."

3. D'où la question : comment les termes théoriques acquièrent-ils une signification ?

• La distinction chère aux positivistes entre termes observationnels et termes théoriques a été extrêmement critiquée. C'est pourquoi Lewis 1972 place le débat à un niveau supérieur d'abstraction : supposons que nous ayons des termes que nous comprenions avant l'introduction d'une certaine théorie T : appelons-les des O-termes (pour *"old terms"*). Pour rendre compte de phénomènes couchés dans le vocabulaire des O-termes, la théorie T introduit des nouveaux termes, les T-termes ou termes théoriques.

Exemple : pour expliquer la mort de M. Corps, le détective Sherlock H. introduit les T-termes t_1, t_2, t_3 qui désignent des individus qui ont complotés à la mort de M. Corps.

(1) t_1 et sa soeur t_2 sont les neveux et nièces de M. Corps. t_1 et t_2 sont ses seuls héritiers. Ils ont embauché le tueur professionnel t_3 pour que le 23 au soir il assassine M. Corps.

• L'idée de base de l'approche fonctionnelle, comme le dit Carnap (1966, p. 241), c'est qu'un terme théorique

> "tire sa signification du contexte de la théorie."

Tout le travail de Lewis (et d'autres avant lui) consiste à élaborer rigoureusement cette idée selon laquelle les termes théoriques tirent leur signification de la théorie.

• Supposons qu'une théorie T soit formulée par un unique énoncé T où figurent les termes théoriques $t_1, ..., t_n$. Supposons également que ces termes

soient des constantes d'individus, et non des symboles de prédicat ou de fonction ; selon Lewis, cette hypothèse ne constitue pas une perte de généralité car on peut traduire dans un tel langage des langages contenants des symboles de prédicats (et de fonctions) théoriques.

L'énoncé T peut alors s'écrire

(2) $T[\mathbf{t}]$

(3) t_1 et sa soeur t_2 sont les neveux et nièces de M. Corps. t_1 et t_2 sont ses seuls héritiers. Ils ont embauché le tueur professionnel t_3 pour que le 23 au soir il assassine M. Corps.

• Construisons maintenant l' énoncé de Ramsey de T (d'après l'article "Theories" de F.P.Ramsey) s'obtient en remplaçant les T-termes par des variables liées par des quantificateurs existentiels :

(4) $\exists \mathbf{x} T[\mathbf{x}]$

(5) $\exists x_1, x_2, x_3$ tels que x_1 et sa soeur x_2 sont les neveux et nièces de M. Corps. Ce sont ses seuls héritiers. Ils ont embauché le tueur professionnel x_3 pour que le 23 au soir il assassine M. Corps.

L'énoncé de Ramsey de T affirme qu'il existe un n-uplet qui réalise $T[\mathbf{x}]$. T implique bien sûr son énoncé de Ramsey, mais le point important est que les deux énoncés ont le même O-contenu : tout énoncé sans T-termes qui suit de l'une suit de l'autre, et réciproquement.

• Une première innovation de Lewis consiste à proposer de renforcer cet énoncé : l'énoncé de Ramsey modifié de T affirme non seulement qu'il existe une réalisation de $T[\mathbf{x}]$, mais que cette réalisation est unique :

(6) $\exists ! \mathbf{x} T[\mathbf{x}]$

(7) il existe un unique triplet d'individus x_1, x_2, x_3 tel que x_1 et sa soeur x_2 sont les neveux et nièces de M. Corps et que x_1 et x_2 sont ses seuls héritiers et que x_1 et x_2 ont embauché le tueur professionnel x_3 pour que le 23 au soir il assassine M. Corps.

• Rappelons que l'objectif de l'analyse est de fournir une explication du sens des T-termes ; or, à lui seul, l'énoncé de Ramsey de T (même dans sa version modifiée) n'aide pas à déterminer la signification des T-termes puisque précisément il les fait disparaître. D'où une seconde idée importante, que l'on doit à Carnap, qui est de *factoriser* la théorie T d'une part en son contenu observationnel, exprimé selon lui par l'énoncé de Ramsey, d'autre part en la spécification du sens des T-termes (Carnap 1966, pp. 261 et sq.).

C'est l'énoncé de Carnap modifié qui remplit cette fonction. L'énoncé de Carnap de T ne dit pas qu'il existe un n-uplet qui réalise T, mais que s'il en existe un, alors les entités qui sont dénotées par les T-termes le réalisent :

(8) $\exists \mathbf{x} T[\mathbf{x}] \rightarrow T[\mathbf{t}]$

L'énoncé de Carnap de T est impliqué par T ; la conjonction de l'énoncé de Carnap et de l'énoncé de Ramsey est logiquement équivalente à T ; il n'y a pas de O-énoncé non tautologique impliqué par l'énoncé de Carnap de T.

• Lewis propose un énoncé de Carnap modifié qui est à l'énoncé de Carnap ce que l'énoncé de Ramsey modifié est à l'énoncé de Ramsey :

(9) $\exists! \mathbf{x} T[\mathbf{x}] \rightarrow T[\mathbf{t}]$

(10) **S'**il existe un unique triplet d'individus x_1, x_2, x_3 tel que x_1 et sa soeur x_2 sont les neveux et nièces de M. Corps et que x_1 et x_2 sont ses seuls héritiers et que x_1 et x_2 ont embauché le tueur professionnel x_3 pour que le 23 au soir il assassine M. Corps, **alors** t_1 et sa soeur t_2 sont les neveux et nièces de M. Corps et t_1 et t_2 sont ses seuls héritiers et t_1 et t_2 ont embauché le tueur professionnel t_3 pour que le 23 au soir il assassine M. Corps,

Lewis ajoute une condition supplémentaire selon laquelle si $T[\mathbf{x}]$ n'est pas uniquement réalisé, les T-termes ne dénotent rien ; si $\star$ est un nom qui ne dénote rien, alors cette condition supplémentaire se formule comme suit :

(11) $\neg \exists! \mathbf{x} T[\mathbf{x}] \rightarrow \mathbf{t} = \star$

(12) **S'**il n'existe pas un unique triplet d'individus x_1, x_2, x_3 tel que x_1 et sa soeur x_2 sont les neveux et nièces de M. Corps et que x_1 et x_2 sont ses seuls héritiers et que x_1 et x_2 ont embauché le tueur professionnel x_3 pour que le 23 au soir il assassine M. Corps, **alors** t_1, t_2 et t_3 ne dénotent aucun individu

La conjonction de ces deux énoncés est logiquement équivalente à un énoncé qui définit les T-termes à la manière d'une description définie :

(13) $\mathbf{t} = \iota T[\mathbf{x}]$

(14) t_1, t_2 et t_3 sont les 3 individus x_1, x_2, x_3 tels que x_1 et sa soeur x_2 sont les neveux et nièces de M. Corps et que x_1 et x_2 sont ses seuls héritiers et que x_1 et x_2 ont embauché le tueur professionnel x_3 pour que le 23 au soir il assassine M. Corps,

C'est le point où Lewis voulait aboutir : (13) fournit ce qu'il appelle une définition fonctionnelle (explicite) des termes théoriques de T, qui pour sa

part les définit implicitement. Dans (13), la signification de ces termes est détérminée par leur rôle ou leur fonction au sein de la théorie.

• Remarque 1 : Lewis ajoute que ces rôles sont des rôles causaux, :

> "...a general hypothesis about the meanings of theoretical terms : that they are definable functionnally, by reference to their causal roles."

> "According to their definitions, the T-terms name the occupants of the causal roles specified by the theory."

• Remarque 2 : Lewis considère donc les termes théoriques comme des *descriptions définies*, ainsi qu'on les appelle en philosophie du langage. Exemple : "le président de la République française". Si la description est fausse, alors pour Lewis la description ne dénote rien. En l'occurence, cela signifie que si la théorie T est fausse, alors ses T-termes ne dénotent rien. A la différence d'une description définie "monadique", il suffit que la théorie T soit fausse à un endroit pour qu'*aucun* des T-termes ne dénote quoi que ce soit.

1.2 La sémantique fonctionnelle des termes mentaux

• Nous pouvons désormais revenir au fonctionnalisme RL. On l'a dit, il se présente comme une pure et simple application de la sémantique fonctionnelle que je viens d'esquisser aux concepts mentaux.

• Dans la sémantique fonctionnelle des termes théoriques, ce qui est crucial dans l'acquisition de la signification par un terme théorique t, c'est la théorie T au sein de laquelle il figure. Dans le cas des concepts mentaux ("ressentir une douleur", "croire que", "avoir l'intention de"), que peut être cette théorie qui confère la signification aux concepts mentaux ?

• La réponse de Lewis est que c'est la **psychologie naïve**. Ce que propose Lewis, c'est donc de *considérer les termes mentaux comme les termes théoriques introduits par la psychologie naïve.*

Dans ce cas, quels seraient les O-termes, c'est-à-dire les termes compris "avant" la théorie ? Ce sont les termes qui renvoient aux stimuli sensoriels d'une part et aux réponses comportementales d'autre part.

• Quelle est exactement cette théorie ? Lewis propose de prendre toutes les *platitudes* qui établissent des relations entre états mentaux, stimuli sensoriels et réponses comportementales :

> "When someone is in so-and-so combination of mental states and receives sensory stimuli of so-and-so kind, he tends with so-and-so probability to be caused thereby to go into so-and-so mental states and produce so-and-so motor responses."

• Lewis restreint les platitudes aux platitudes qui sont de *connaissance commune* : une platitude P est de connaissance commune ssi chacun sait que P, chacun sait que chacun sait que P, chacun sait que chacun sait que chacun sait que P, etc. Pourquoi cette restriction ? Parce que Lewis estime par ailleurs que la signification d'un terme est partagée dans une communauté linguistique si elle est de connaissance commune parmi les individus de cette communauté.

• L'exigence de connaissance commune est-elle trop forte ?

D. Braddon-Mitchell et F. Jackson, *Philosophy of Mind and Cognition*, Blackwell, 1996, p. 55

> "There is, however, a problem with setting the standard as high as Lewis does in the passage quoted. There may not be enough clauses that meet the standard. Consider the difference between pride and vanity. This is a subtle matter. Is it really plausible that the difference is captured by what we can write down that meets the standard of being a platitude that is common knowledge and is known to be common knowledge."

Braddon-Mitchell et Jackson préfèrent concevoir le contenu de la théorie qui détermine la signification des termes mentaux sur le modèle de la connaissance tacite. Les connexions entre stimuli, états mentaux et comportements sont plutôt des régularités que nous connaissons tacitement, comme ce serait le cas pour l'essentiel de la grammaire.

• Pour revenir au projet initial, on voit que si l'on accepte les hypothèses qui précèdent, la sémantique *fonctionnelle* aboutit bien à une conception *fonctionnaliste* des états mentaux : les états mentaux sont bien définis par le rôle causal qu'ils occupent relativement aux stimuli sensoriels, aux autres états mentaux et aux réponses comportementales.

• Exemple : reprenons notre pseudo-théorie de la douleur. Son énoncé de Ramsey est le suivant :

(15) Il existe x_1, x_2 tel que si y est piqué par guêpe, alors il est dans l'état x_1 et s'il est dans l'état x_1 alors il se trouve peu après dans l'état x_2 et crie "Aïe".

L'énoncé de Ramsey ne contient aucun terme psychologique. Il peut donc servir de base à une définition des termes psychologiques. Les termes psychologiques sont *interdéfinis*.

> D. Braddon-Mitchell et F. Jackson, *Philosophy of Mind and Cognition*, Blackwell, 1996, p. 52
>
> "This shows that there is no vicious circularity in common-sense functionalism. Although the most natural way of stating common-sense functionalism *interdefines* the mental, we can recast the story so as to yield an account of when a subject...is in any particular state that makes no explicit reference to other mental states. The essential idea is shorn of the technicalities is that common-sense functionalism defines mental states holistically by their place in a network."

1.3 Fonctionnalisme RL et théorie de l'identité

• Lewis 1972 développe sa sémantique des termes mentaux au sein d'un argument *en faveur* de la théorie de l'identité, et plus précisément de la théorie de l'identité *typique ("type-type")*. Cela a de quoi surprendre : on a vu que la première forme de fonctionnalisme, celle de TP, était taillée pour être compatible avec l'argument de la réalisabilité multiple. Avec le fonctionnalisme RL, on a une conception assez proche des états mentaux, mais on aboutit à une attitude radicalement opposée vis-à-vis de la théorie de l'identité.

• Structure de l'argument lewisien en faveur de l'identité typique :

P1. Etat mental M = l'occupant du rôle causal R (définition)
P2. Etat cérébral C = l'occupant du rôle causal R (neurosciences)

CC. Etat mental M = état cérébral C

Pour revenir à l'exemple du détective. Supposons que l'on sache ceci :

(16) T[(Jean Dupont, Marie Dupont, Paul "le dingue")]

(17) Jean Dupont et sa soeur Marie Dupont sont les neveux et nièces de M. Corps. Ce sont ses seuls héritiers. Ils ont embauché le tueur professionnel Paul "le dingue" pour que le 23 au soir il assassine M. Corps.

Il en découle alors

(18) $\mathbf{t} = \mathbf{r}$

(19) Jean Dupont$= t_1$ et Marie Dupont $= t_2$ et Paul "le dingue" $= t_3$

• Comme Lewis le souligne, son argument pour la théorie de l'identité *ne fait pas appel à des considérations de simplicité* comme c'est le cas, par exemple, chez Smart - la manière la plus simple de rendre compte des corrélations psychocérébrales, c'est d'identifier types psychologiques et types cérébraux.

• Block (1980) propose de clarifier la situation de la manière suivante. Il distingue

1. *functional identity claim* (Putnam, Harman) : un état mental M est un état fonctionnel. Etre dans l'état mental M = être dans un état qui a les propriétés fonctionnelles décrites par T.
2. *functional specification claim* (Lewis, Armstrong) : un état mental M est spécifié de manière fonctionnelle. L'état mental M = l'état qui a les propriétés fonctionnelles décrites par T.

2 Problèmes et objections

2.1 L'argument des qualia absents

• Scénario :

1. un corps artificiel identique en apparence au nôtre, avec des organes sensoriels et des organes moteurs. Ce corps artificiel a des neurones sensoriels qui proviennent des organes sensoriels et des neurones moteurs qui vont vers les organes moteurs
2. chaque chinois dispose d'un émetteur/récepteur
3. un ensemble de satellites est capables de représenter des symboles dans le ciel, symboles que tous les Chinois peuvent voir
4. chaque chinois doit suivre *une* instruction qui est déclenchée par la réception d'un certain input neuronal et d'un certain symbole écrit par satellite ; en retour, il ordonne que les satellites représentent un nouveau symbole et déclenche un certain *output* neuronal

Supposons qu'il soit possible de faire en sorte que le système S formé par le peuple chinois, le corps articiel et les satellites aient globalement la même

organisation fonctionnelle (réalise la même machine de Turing) que Pierre. S est un *double fonctionnel* de Pierre.

- Pourquoi cette possibilité pose-t-elle problème au fonctionnaliste ?

  N. Block, "Troubles with Functionalism", 1978, repris dans N. Block (ed.), *Readings in the Philosophy of Psychology*, 2 vol., Harvard UP, Harvard, 1980

  > "What makes the homonculi-headed system just described a prima facie counterexample to (machine) functionalism is that there is a prima facie doubt whether it has any mental states at all - especially whether it has what philosophers have variously called "qualitative states", "raw feels", or "immediate phenomenological qualities"."

  > "My claim is not that every sort of Functional simulation of you must lack qualia...My point rather is that not every sort of homonculi-headed Functional simulation *need* have qualia. If there is even *one* possible Functional simulation of you that has no qualia, Functionalism is false."

L'idée est donc qu'un double fonctionnel comme le système S ne semble pas avoir de *qualia*. L'argument de Block est destiné à montrer que le fonctionnalisme est *trop libéral* : il accorde la mentalité à des systèmes qui, selon lui, en sont dépourvus.

- Réponse possible : aussi étrange que cela puisse paraître, le système S a des *qualia*. A cause de la différence d'échelle, cela nous paraît étrange. Imaginons une créature qui ferait la taille d'un neurone et qui se promènerait dans notre cerveau. A elle aussi, il pourrait paraître impossible que nous ayons des *qualia*.

2.2 L'argument des qualia inversés

- Les scénarios du type "qualia inversé" remontent au moins à J. Locke dans son *Essay concerning Human Understanding* (1689).

- Scénario

  N. Block, "Troubles with Functionalism", 1978, repris dans N. Block (ed.), *Readings in the Philosophy of Psychology*, 2 vol., Harvard UP, Harvard, 1980

> "It makes sense, or seems to make sense, to suppose that objects we both call green look to me the way objects we both call red look to you. It seems that we could be functionnally equivalent even though the sensation fire hydrants evoke in you is qualitatively the same as the sensation grass evokes in me."

Invert et Nonvert ont le rouge et le vert inversé : l'effet que fait une fraise (mûre) à Invert est le même que celui que fait une herbe (bien grasse) à Nonvert. Invert et Nonvert utilisent le langage de manière indiscernable : quand Invert voit une fraise mûre, il dit "la fraise est rouge", exactement comme Nonvert. Bien plus, il semble que l'état mental dans lequel se trouve Invert et celui dans lequel se trouve Nonvert quand ils voient la fraise mûre puisse être fonctionnellement identique : ce sont les mêmes stimuli qui les causent, ils engendrent les mêmes réponses, les mêmes états mentaux, etc.

- Pourquoi cela pose-t-il problème au fonctionnaliste ? Parce que deux systèmes fonctionnellement équivalent auraient des états mentaux différents par certains aspects. Il y a donc des aspects de la mentalité qui échappent au fonctionnalisme.

> D. Braddon-Mitchell et F. Jackson, *Philosophy of Mind and Cognition*, Blackwell, 1996, p. 124
>
> "Intuitively, phenomenal nature is **intrinsic**. The perceived redness of a sunset is a feature of how the experience is here and now ; you cannot capture its nature fully by talking of its similarities and difference, of what causes it, and of what it causes. It is as intrinsic as squareness. But then perceived redness in particular, and colour experience in general, cannot be captured in functional terms of causal relations."

- Quelques réponses :

1. Quand Invert et Nonvert voient une fraise mûre, ils ont différentes croyances sur la couleur de la fraise - même s'ils disent tous les deux : "la fraise est rouge". (Harman, 1990)
2. ("containment response", Block 1990) : le fonctionnalisme est capable de rendre compte des aspects cognitifs de l'esprit, pas de ses aspects qualitatifs.

3 Références

3.1 Sur le fonctionnalisme RL

D. Braddon-Mitchell et F. Jackson, *Philosophy of Mind and Cognition*, Blackwell, 1996, **chap.3**

J. Kim, *Philosophy of Mind*, Westview Press, Boulder, 1998, **chap.5**

3.2 Sur le fonctionnalisme

N. Block, "What is Functionalism ?", dans N. Block (ed.), *Readings in the Philosophy of Psychology*, 2 vol., Harvard UP, Harvard, 1980

J. Levin, "Functionalism", *The Stanford Encyclopedia of Philosophy*, Fall 2004 Edition, E.N. Zalta (ed.)

3.3 Sur les difficultés du fonctionnalisme

N. Block, "Troubles with Functionalism", 1978, repris dans N. Block (ed.), *Readings in the Philosophy of Psychology*, 2 vol., Harvard UP, Harvard, 1980

N. Block et J. Fodor, "What Psychological Are Not", *The Psychological Review*, vol. 81, n°2, 1972, pp. 159-81

A. Byrne, "Inverted Qualia", *The Stanford Encyclopedia of Philosophy*, Summer 2005 Edition, E.N. Zalta (ed.)

G. Harman, "The Intrinsic Quality of Experience", dans J. Tomberlin (ed.), *Philosophical Perspectives*, vol.4, 1990

M. Tye, "Qualia", *The Stanford Encyclopedia of Philosophy*, Summer 2003 Edition, E.N. Zalta (ed.)