

Les termes théoriques de Carnap à Lewis

Henri Galinon

Nous nous accordons facilement sur le sens des termes « eau », « tigre » ou « rouge », dont le référent est publiquement observable. Mais quel est le sens des termes « électron », « spin », etc., qui figurent dans nos théories physiques ? Dans la perspective épistémologique qui est celle de l'empirisme logique, le problème est d'autant plus pressant que sa résolution est une condition de possibilité de la mise au jour des fondements empiriques de la connaissance scientifique et de la distinction corrélative entre science et non-science.

L'objet de cet article est de présenter le dernier travail de Rudolf Carnap sur ce problème des termes théoriques, tel qu'il apparaît principalement dans les *Fondements philosophiques de la Physique* (1966). La pièce maîtresse de ce nouvel effort théorique est la méthode dite aujourd'hui de Ramsey-Carnap d'analyse du contenu empirique des théories. Nous l'introduisons en détail, en montrant comment elle permet de donner une sémantique *fonctionnaliste* des termes théoriques. Dans une seconde partie, nous prolongeons notre étude en présentant la modification de la méthode de Carnap introduite par Lewis¹ et son application aux termes d'états mentaux.

1. Carnap et la signification des termes théoriques

Pour introduire plus précisément le problème des termes théoriques, tel qu'il se pose chez Carnap, commençons par écarter une première question. La question : « les électrons existent-ils ? » n'est pas, selon Carnap, celle qui

1. Le texte de référence est ici celui de D. K. Lewis, « Theoretical and Psychophysical Identifications », chap.16, in *Papers in metaphysics and epistemology*, Cambridge University Press, 1999.

fait problème. Car de deux choses l'une : soit la question est une question interne que l'on pose dans le langage de la science, et dans ce cas la réponse est triviale : oui, la théorie physique moderne suppose que les électrons existent (le critère quinién d'engagement ontologique : « être, c'est être la valeur d'une variable », s'applique à plein) ; soit la question est comprise comme une question externe, celle de savoir si les électrons existent « absolument » ou « en dehors de la théorie », et alors cette question est dénuée de « contenu cognitif » (pour utiliser un autre terme de Carnap, elle n'est pas une question « théorique ») : en tant que question métaphysique, elle est tout simplement dénuée de sens, en tant que question portant sur le choix du langage de la science (« pourquoi utiliser un langage contenant le concept d'électrons ? »), c'est une affaire de décision pratique². Nous choisissons simplement de parler d'« électrons », ou bien de « propositions », par convenance (indurée en convention), dans la mesure où les théories dans lesquelles nous les introduisons nous permettent de nous rendre intelligible un aspect du « réel » (théories tautologiques et incohérentes sont donc de ce point de vue aussi inutiles l'une que l'autre)³. Mais cette « liberté » dans le choix de nos cadres d'intelligibilité, Carnap n'est disposé à l'accorder que soutenu par la conviction qu'il serait toujours possible de faire une distinction entre ce que nos théories nous disent du monde (*via* les énoncés qui ont un contenu empirique ou « synthétique ») et les effets du langage que nous avons choisis (qui sont le contenu des énoncés « analytiques », vrais ou faux en vertu de la signification des termes qui y figurent), distinction sans laquelle la frontière entre métaphysique et science deviendrait problématique.

Dans le *langage observationnel* de la science (la partie du langage de la science qui ne contient que des termes observationnels et logiques)⁴, faire cette distinction ne pose pas de problème majeur. Nous comprenons parfaitement le sens des mots que nous employons et il ne tient qu'à nous d'inscrire dans notre théorie à titre de « postulat de signification » que, par exemple, « tous les chiens sont des animaux ». Cela est vrai en vertu du sens

2. « Je pense que la question ne devrait pas être discutée sous la forme : 'Est-ce que les entités théoriques sont réelles ?', mais plutôt sous la forme : 'Préférons-nous un langage de la physique (et de la science en général) qui contient des termes théoriques, à un langage qui n'en contient pas ?'. De ce point de vue, la question devient une question de préférence et de décision pratique » (Carnap, *Fondements philosophiques de la physique*, J. M. Luccioni & A. Soulez (trad.), Paris, Armand Colin, 1973, Chap. 26, p. 256). Le point est plus amplement développé dans « Empiricism, Semantics, and Ontology », *Revue Internationale de Philosophie*, 4, 1950, p. 20-40.

3. Mais il est impossible de *décider* si nos règles d'introduction sont ou non cohérentes.

4. Hilary Putnam doute de la pertinence du concept de « langage observationnel » et fait remarquer que des énoncés ne contenant que des termes observables comme « un objet trop petit pour être vu », pourraient très bien être classés dans le langage théorique de la science (Cf. H. Putnam, « Ce que les théories ne sont pas », in E. Nagel, P. Suppes, A. Tarski, *Logic, Methodology and the Philosophy of Science*, Stanford University Press, 1962). David Lewis a étendu ce genre de remarque dans « How to define theoretical terms », *The Journal of Philosophy*, 67 (13), 1970, p. 427-446.

du mot « chien » et ne nous apprend rien sur le monde ; cet énoncé doit donc compter comme « analytique », ainsi que toutes ses conséquences logiques.

Nous en venons alors au problème des termes théoriques. Il est simple : pour distinguer les énoncés analytiques du *langage théorique*, il faut pouvoir déterminer la signification des termes théoriques (de façon équivalente : être en mesure de poser des « postulats de signification » adéquats). Or la signification de ces termes théoriques est problématique puisqu'ils ne réfèrent à aucune entité « observable ».

En première analyse, la seule chose que nous puissions dire de la signification de ces termes est donc qu'elle est entièrement donnée par le contexte théorique dans lequel ils apparaissent. Le problème de la détermination de la signification des termes théoriques doit donc se résoudre dans une analyse des théories qui les introduisent. Or, quel est ce « contexte » d'apparition des termes théoriques ? Selon Carnap, ces derniers sont introduits par un corps de *lois théoriques*, des lois telles que « la température d'un corps est proportionnelle à l'énergie cinétique moyenne de ses molécules » ou bien la loi d'Ohm : « $U = RI$ », où ne figurent, dans ce contexte natif, que des termes théoriques (et des termes mathématiques). Mais d'une part ces lois ne peuvent pas être obtenues par simple généralisation à partir d'énoncés singuliers observationnels ou de lois empiriques, et d'autre part, réciproquement, aucun énoncé observationnel ne peut en dériver : ceci parce que, si nous appelons L_O la partie observationnelle du langage de la science, et L_T sa partie théorique (incluant les termes mathématiques), des considérations de logique⁵ montrent qu'en l'absence de règles de correspondance entre les termes théoriques et observables, un ensemble d'énoncés de l'un des langages ne peut avoir pour conséquence logique un énoncé (non tautologique) de l'autre langage. Par conséquent, si nous suivons Carnap et les arguments précédents, ces lois théoriques, n'ayant aucune contrepartie observationnelle, doivent être dénuées de sens. Plutôt que de théories, il s'agit donc de *formes de théories* mettant en rapport des termes dénués de signification et de référence. La question de la signification des termes théoriques ne peut donc être réglée par le simple examen des postulats théoriques qui les introduisent.

Mais, continue l'explication carnapienne, les théories scientifiques seraient parfaitement inutiles si le langage de la science n'était constitué que de lois observationnelles et de lois théoriques. Or, il est facile de voir qu'en *posant*⁶ maintenant des *règles de correspondance* entre les termes théoriques et les termes observationnels (non nécessairement des règles de

5. Il s'agit du théorème d'Interpolation de Craig. En un mot, ce théorème énonce que si une formule f_1 est conséquence du premier ordre d'une formule f_2 alors il existe une formule f_i , appelée *interpolante*, dont le vocabulaire est commun à f_1 et f_2 , et telle que f_i est conséquence de f_1 et f_2 est conséquence de f_i .

6. Ces règles de correspondance ne peuvent être que posées en effet : si les postulats théoriques ne sont que des formes d'énoncés, en vertu de quel sens pourrions-nous être contraints d'identifier termes théoriques et termes observationnels ?

correspondance pour chaque terme théorique), c'est-à-dire des énoncés contenant, chacun, des termes de L_0 et de L_T , il sera possible en général de dériver de nouvelles lois *empiriques* (des énoncés conditionnels universels de L_0) à partir de nos lois théoriques. Pour Carnap, pour le Carnap de l'*Introduction à la Philosophie de la Physique* (1966), c'est cela, en première analyse, canoniquement, une théorie scientifique : un ensemble de postulats théoriques, T, et un ensemble de règles de correspondance, C. Notons TC leur conjonction.

La situation est donc la suivante : TC est une théorie scientifique ; elle a, *en bloc*, des conséquences observables, et contient un certain nombre de termes théoriques. Il faut donc que ces termes aient à leur tour une signification empirique, cette dernière leur étant conférée indirectement *via* les règles de correspondance dans le contexte *total* de la théorie dans laquelle ils apparaissent. Dans ces conditions, il serait vain d'espérer extraire de notre théorie des définitions individuelles des termes théoriques en termes observationnels : dans chaque règle de correspondance figurent en général plusieurs termes théoriques, qui eux-mêmes apparaissent dans plusieurs règles de correspondance. Un terme qui serait parfaitement isolé des termes théoriques et figurerait dans une règle de correspondance ne serait tout simplement pas un terme théorique : « la température de x est égale à la hauteur, mesurée de telle façon, de la colonne de mercure dans le dispositif tel et tel [suit une description d'un thermomètre en contact avec x]... » n'épuise pas le concept de température, et, réciproquement, si nous définissions ainsi le concept de température, celui-ci n'aurait plus rien d'un concept théorique, et perdrait sa fécondité.

Comment, dès lors, isoler la signification des termes théoriques, dans une théorie qui à la fois les emploie et les définit ? À ce point, la situation semble passablement désespérée : Quine, dans un argument devenu classique⁷, concluait à l'impossibilité de distinguer dans une théorie scientifique une classe d'énoncés théoriques analytiques, et en particulier de séparer ce qui, dans les conséquences de la théorie, est conséquence du sens des termes théoriques employés de ce qui est conséquence « substantielle » de la théorie. En un mot, parce que nos théories affrontent en bloc le tribunal de l'expérience, toute sémantique vérificationniste doit être *holiste*. Dès lors, il ne peut y avoir entre énoncés analytiques et synthétiques aucune différence de nature, mais seulement une différence de degré⁸.

C'est précisément à ce point que le texte de Carnap prend sa valeur spécifique. Certes, les termes théoriques prennent sens en bloc dans le contexte de la théorie *via* les conséquences observables et, certes, il n'est pas possible de définir individuellement ces termes en fonction de termes

7. Quine, « Two dogmas of empiricism », in Quine, *From a logical point of view*, Harvard, 1953. Traduction française S. Laugier : « Deux dogmes de l'empirisme », *Du point de vue logique*, Paris, Vrin, 2004.

8. Pour une évaluation rétrospective générale de la controverse Quine-Carnap sur ce sujet, voir par exemple R. Creath, « Every dogma has its day », *Erkenntnis* 35, 1991.

observationnels. Mais il y a néanmoins un sens précis dans lequel la théorie détermine la signification des termes théoriques. Pour le montrer, Carnap commence par une remarque cruciale due à Frank P. Ramsey⁹ :

Les termes théoriques sont dispensables (mais non les entités !) au sens où, pour chaque théorie TC, il existe une théorie associée ne contenant pas de termes théoriques et ayant *exactement le même contenu observationnel* : c'est l'énoncé de Ramsey de la théorie TC, notons-le ${}^R\text{TC}$. Il est obtenu à partir de TC de la façon suivante :

1. Ecrire TC sous la forme d'un énoncé conjonctif (pour des raisons de commodité) et remplacer les termes théoriques par des noms de classes, relations etc., (écrire par exemple $x \in \text{Mol}$ au lieu de « x est une molécule »). Notons le résultat ; $\text{TC}(\mathbf{t}, \mathbf{o})$, où \mathbf{t} ($= t_1, \dots, t_n$) sont les termes théoriques et \mathbf{o} ($= o_1, \dots, o_m$) les termes observationnels.

2. Remplacer les termes théoriques qui apparaissent dans $\text{TC}(\mathbf{t}, \mathbf{o})$ par des variables appropriées :

$$\text{TC}(\mathbf{x}, \mathbf{o}) \quad \text{avec } \mathbf{x} = x_1, \dots, x_n.$$

3. La clôture existentielle de cet énoncé, $\exists \mathbf{x} \text{TC}(\mathbf{x}, \mathbf{o})$, notée ${}^R\text{TC}$, est l'énoncé de Ramsey de la théorie.

${}^R\text{TC}$ a donc deux propriétés intéressantes : d'une part, elle ne contient aucun terme théorique, et d'autre part, elle a exactement les mêmes L_0 -conséquences que TC, c'est-à-dire le même contenu empirique, le même pouvoir explicatif et prédictif¹⁰. Quel est donc le sens des termes théoriques,

9. Note historique : Carnap retrace l'histoire de sa redécouverte de l'énoncé de Ramsey dans une lettre à Hempel, datée du 12 février 1958 (présentée dans la note éditoriale de Stathis Psillos à l'article inédit de Carnap, « Theoretical Concepts in Science » (1959 - Conférence à Santa Barbara), in *Studies in History and philosophy of sciences*, 31 (1), 2000, p. 151-172) : « [...] Le cas de l'énoncé de Ramsey est un exemple très instructif de la facilité avec laquelle on s'illusionne sur l'originalité des idées. À la conférence de Feigl, ici, en 1955, où Papn Bohert et d'autres étaient présents, j'ai représenté la forme existentialisée d'une théorie comme une idée récente originale venue de moi. Quelque temps après la Conférence, Bohnert a dit se souvenir avoir eu cette idée il y a quelques années et me l'avoir expliquée dans une lettre envoyée à Chicago. Quoique je ne pus pas trouver cette lettre dans les archives, je n'avais aucun doute que Bohnert avait raison, je lui ai donc cédé la priorité. [...] Puis, je crois que c'était l'été dernier, lorsque j'ai lu le reste de votre "Dilemme", j'ai été saisi par votre référence à Ramsey. J'ai cherché le passage à l'endroit auquel vous faisiez référence dans le livre de Ramsey, et il était là, nettement souligné par moi-même. Il n'y a donc pas de doute que je l'avais lu dans le livre de Ramsey. Je pense que c'était au temps de Vienne ou de Prague [...] » (C'est nous qui traduisons).

10. Carnap ne démontre pas ce résultat, mais il est relativement clair. Preuve : La partie intéressante est de montrer que TC ne démontre pas plus de L_0 -énoncés que ${}^R\text{TC}$. Posons $\text{TC} := \psi(\mathbf{t}, \mathbf{o})$, ${}^R\text{TC} := \exists \mathbf{x} \psi(\mathbf{x}, \mathbf{o})$, ϕ un énoncé de L_0 (observationnel). Tout d'abord, toute L_0 -structure M modèle de $\exists \mathbf{x} \psi(\mathbf{x}, \mathbf{o})$ s'étend en une LTC-structure M' modèle de $\psi(\mathbf{t}, \mathbf{o})$: il suffit d'interpréter les constantes (ou les lettres de prédicats) \mathbf{t} par un uplet (d'ensemble) d'éléments du domaine de M que l'énoncé Ramsey nous assure exister (réciproquement un modèle de TC est bien évidemment un modèle de ${}^R\text{TC}$). Soit maintenant ϕ un énoncé de L_0 :

si tout le contenu cognitif (empirique) de la théorie est en fait déjà dans son énoncé de Ramsey ? Le point de Carnap est de faire remarquer à son tour que TC est équivalente à la conjonction des deux énoncés suivants :

- (1) ${}^R\text{TC} (:= \exists \mathbf{x} \text{TC}(\mathbf{x}, \mathbf{o}))$
 (2) ${}^R\text{TC} \rightarrow \text{TC} (:= \exists \mathbf{x} \text{TC}(\mathbf{x}, \mathbf{o}) \rightarrow \text{TC}(\mathbf{t}, \mathbf{o}))$

Or, si nous regardons maintenant notre théorie comme donnée par la conjonction de (1) et (2), le sens des termes théoriques s'éclaire : en effet (1) épuise le contenu empirique de TC (d'après le paragraphe précédent) et ne contient pas de termes théoriques. (2) épuise donc ce que la théorie nous dit du sens des termes théoriques. Mais, de plus, (2) n'a aucun contenu empirique¹¹ : donc (2) est exactement « le postulat de signification » cherché. C'est le premier point.

Or que dit (2) ? Ici il faut faire attention à la lecture que nous faisons de cet énoncé. Si (2) n'a aucun contenu empirique, il n'est pas cet énoncé qui affirme que si $\exists \mathbf{x} \text{TC}(\mathbf{x}, \mathbf{o})$, c'est-à-dire si la théorie, de l'électricité disons, est réalisée, alors les électrons, les champs magnétiques, etc., satisfont $\text{TC}(\mathbf{x}, \mathbf{o})$. Cet énoncé a un contenu factuel si nous regardons les termes « électrons », etc., comme déjà compris. Mais (2) doit au contraire être lu comme conférant leur sens aux termes, comme un énoncé « analytique ». Notons « ÉLECTRONS », « CHAMPS MAGNÉTIQUES », etc., en petites capitales, ces mots nouveaux dont le sens est donné par la théorie, et continuons de désigner en lettres minuscules ces entités mystérieuses dont nous croyons parler, ou, pour le dire autrement, dont nous comprendrions que la théorie nous parle si nous avons eu l'occasion de les observer. Autrement dit, « ÉLECTRON » est ce mot qui a exactement le contenu cognitif ou la signification empirique du mot « électron », d'après l'analyse de Ramsey-Carnap de la théorie qui les introduit. Ceci étant posé, il y a deux façons de comprendre l'énoncé (2). D'un côté, (2) énonce seulement que ${}^R\text{TC}$ est vrai, c'est-à-dire si le monde est tel qu'il y a au moins un uplet de classes d'entités $\langle M_1, \dots, M_n \rangle$, et peut-être $\langle M'_1, \dots, M'_n \rangle$, etc., qui sont tels, chacun, qu'ils satisfont la relation $\text{TC}(\mathbf{x}, \mathbf{o})$ alors le uplet de classes qui interprète les termes théoriques « électrons », etc., est parmi ceux-là. La théorie ne nous dit pas lequel, seulement que c'est un de ceux-là. Si $\text{TC}(\mathbf{x}, \mathbf{o})$ est réalisée, (2) ne donne donc en général, selon le terme de Carnap, qu'une « interprétation partielle » des termes théoriques figurant dans T : la théorie sous-détermine la référence des termes théoriques. Mais il y a bien sûr une seconde lecture : supposons que « ÉLECTRON », le nouveau mot dont tout le sens est conféré par la théorie TC de l'électricité, figure à la première place

supposons que ϕ est une conséquence de TC, mais non de RTC. Alors il existe un LO modèle de RTC qui n'est pas modèle de ϕ . Mais alors on ne peut pas étendre ce modèle en un modèle de TC. Absurde, ϕ est conséquence de RTC et les deux théories ont exactement les mêmes conséquences observables.

11. Son énoncé de Ramsey est une tautologie.

dans l'écriture canonique de TC. La théorie de l'électricité nous dit très précisément ce qu'est un ÉLECTRON :

$$\text{ÉLECTRON}(y) \text{ ssi } \exists \mathbf{x} \text{ TC}(\mathbf{x}, \mathbf{o}) \text{ et } x_1(y)$$

Autrement dit, y est un ÉLECTRON si, et seulement si, si $\langle M_1, \dots, M_n \rangle$ réalise $\text{TC}(\mathbf{x}, \mathbf{o})$ alors $y \in M_1$. Les ÉLECTRONS sont ces entités qui, dans une réalisation de la théorie de l'électricité, jouent le rôle que la théorie décrit être celui des électrons¹².

Quel est le rapport entre les ÉLECTRONS et les électrons ? La sémantique attendue (« naïve ») du terme « électron » est celle d'un terme d'espèce naturelle, comme « tigre » ou « H₂O », qui désignent rigidement leur référent, tandis que la sémantique du terme « ÉLECTRONS » est celle d'un terme fonctionnel, comme « carburateur », défini seulement par un certain rôle. Si la théorie de l'électricité admet une unique réalisation alors x est un ÉLECTRON si, et seulement si, x est un électron. Dans ce cas, le sens théorique et le sens « préthéorique » s'accordent en ceci qu'ils déterminent la même extension dans le monde actuel. Mais si la théorie de l'électricité s'avérait être multi-réalisée¹³ nous ne pourrions distinguer, parmi les différentes entités qui occupent le rôle causal attribué par la théorie aux électrons, celles qui *en fait* sont les électrons, tout au plus pourrions-nous dire que certaines entités « électronisent », se comportent comme des électrons, bref sont des ÉLECTRONS.

Maintenant, pouvons-nous déclarer tout simplement que les électrons ne sont jamais que des ÉLECTRONS ? Cette concession à une vision purement instrumentale de nos théories a quelque chose de contre-intuitif. Les ÉLECTRONS n'ont en commun que le fait de jouer un certain rôle conceptuel défini par la théorie qui les introduit et ne partagent *a priori* aucune autre propriété, de celles qu'une théorie ultérieure pourrait introduire dans le cours de l'approfondissement de notre compréhension du réel. Mais lorsque nous introduisons les électrons dans notre théorie, nous ne changeons pas d'objet, nous continuons à faire de la physique : nous parlons d'entités dont le statut épistémique, s'il n'est pas encore sur un pied d'égalité avec celui d'entités observables, est voué à l'être par la théorie (il n'y a pas deux régimes de

12. Alternativement, dans « Theoretical Concepts in Science » (1959), Carnap utilise l'opérateur de sélection « ε » de Hilbert. La nouvelle constante logique est introduite axiomatiquement par : $Fx \rightarrow F(\varepsilon x Fx)$. La signification intuitive de « $\varepsilon x Fx$ » est « un F quelconque s'il y en a un ». Il s'agit donc d'une sorte de sélecteur arbitraire, ou pour mieux dire, un opérateur de *description indéfinie*. Après avoir noté comment définir les quantificateurs existentiels et universels comme abréviation à partir de « ε » ($\exists x Fx$ ssi $F(\varepsilon x Fx)$ et $\forall x Fx$ ssi $F(\varepsilon x \neg Fx)$), Carnap propose les définitions suivantes :

(a) $t = \varepsilon u \text{TC}(\mathbf{u}, \mathbf{o})$

(b) et pour $i = 1, \dots, n$: $t_i = \pi_i(t)$ où $\pi_i(t)$ est le i ème élément de t .

13. Et après tout, pourquoi pas ? Dans une compréhension « naturelle » ou « naïve » de ce que disent nos théories, comme nous l'avons déjà noté, il semble bien qu'en affirmant TC, nous affirmons seulement que, pour leur part, les électrons, les champs magnétiques, etc., réalisent $\text{TC}(\mathbf{x}, \mathbf{o})$.

la science). Introduire les électrons c'est faire une hypothèse d'existence d'entités réelles, d'objets naturels de la physique, non simplement un moyen commode de systématiser nos observations. Ces entités réelles sont de celles qu'éventuellement nous pourrions être amenés à identifier différemment dans le cours ultérieur du progrès de nos connaissances, par le biais d'une autre théorie par exemple (la théorie des électrons pourrait être réduite par une théorie plus fondamentale), ou par l'observation, sans que ces identifications aient le moins du monde besoin d'être posées artificiellement pour des raisons de « parcimonie » théorique.

Si nous refusons d'identifier les électrons et les ÉLECTRONS, la situation générale a donc quelque chose d'étrange : nous faisons la théorie des ÉLECTRONS avec en vue une réalisation particulière du concept (lesdits électrons passibles d'une étude scientifique, susceptibles d'être observés, etc.), *mais n'importe laquelle de ces réalisations*. Or comment pouvons-nous avoir en vue une réalisation *arbitraire* ? Lewis a proposé une révision de la sémantique des termes théoriques permettant d'éviter cet écueil, nous allons y venir.

Mais disons un mot encore sur Carnap. La différence irréductible qui subsiste entre les ÉLECTRONS et les électrons, ce que l'on pourrait appeler le *reste* de cette analyse « fonctionnelle », Carnap l'appelle le « surplus de signification » des termes théoriques. Or Carnap semble satisfait de l'indétermination qui en résulte dans l'analyse des termes théoriques. Pourquoi ? C'est que cette dernière permettrait sans doute en un sens de rendre compte de la stabilité de la référence à travers les changements de théories (les changements relativement modestes du moins). En modifiant régulièrement notre théorie de l'électricité, nous précisons ce que les ÉLECTRONS nous disent des électrons, nous affinons nos médiations, les instruments que sont nos théories¹⁴. Supposons que nous modifiions notre théorie TC par l'ajout d'une nouvelle règle de correspondance, résultant en TC' : les termes théoriques ont dans TC' un sens différent de celui qu'ils avaient dans TC. Cependant, à suivre l'analyse de Carnap en termes de « signification incomplète », il est tout à fait possible de comprendre TC'

14. Sur ces questions, le passage suivant donne seulement quelques indications : « Les changements de conception sur la structure des électrons, des gènes et autres, ne signifient pas qu'il n'y pas quelque chose « là », derrière chaque phénomène observable ; cela indique seulement que nous apprenons de plus en plus sur la structure de telles entités. Les partisans de la conception « descriptiviste » nous rappellent que les entités inobservables ont pour habitude de passer dans le royaume observable lorsque des instruments plus puissants sont développés [...] » (R. Carnap, *Philosophical foundations of physics*, Basic Books, 1966, p. 256. C'est nous qui traduisons). Dans « Theoretical concepts in science », Carnap écrit : « Cette définition donne autant de spécifications que nous pouvons en donner, pas davantage. Nous ne voulons pas en donner davantage, parce que la signification doit être laissée indéterminée à certains égards, sans quoi le physicien ne pourrait pas – comme il le veut – ajouter demain de nouveaux postulats, et même de nouvelles règles de correspondance, et par là spécifier davantage qu'elle ne l'est aujourd'hui la signification de ces termes » (S. Psillos, « Rudolf Carnap's "Theoretical Concepts in Science" », *Studies in History and philosophy of sciences*, 31 (1), 2000, p.171).

comme parlant toujours des mêmes entités, les électrons par exemple, et de ne reconnaître dans le passage de TC à TC' qu'un *renfort de l'interprétation* des termes théoriques par l'exclusion de certaines réalisations de TC(**x,o**) des réalisations attendues, autrement dit l'élimination de certaines entités comme dénotation possible de nos termes théoriques. Nous ferions la théorie des électrons *via* les ÉLECTRONS. Cette façon de poser les choses, néanmoins, a quelque chose d'étrange, dans la mesure où elle fait intervenir un concept réaliste d'électrons, électrons qui, en ce sens, n'ont chez Carnap aucun statut théorique : ils ne sont que la projection d'un point de vue « descriptiviste » sur nos théories. Mais c'est que l'enjeu ici est extra-théorique : c'est parce que nos théories sont des instruments, instruments que le théoricien est amené à modifier, étendre, spécifier, que les termes doivent avoir une signification incomplète ; que nous interprétions ces modifications comme des approximations successives n'est jamais qu'une façon de nous représenter le processus qui guide ces modifications.

Avec les conséquences que nous venons d'indiquer, conséquences que Carnap est prêt à embrasser, il apparaît donc que le sens exact des termes théoriques tel qu'il est conféré par la théorie peut être isolé par les postulats de signification. Carnap n'a plus alors qu'à définir la classe des énoncés analytiques comme la réunion de l'ensemble des énoncés analytiques empiriques et de l'ensemble des conséquences logiques de ${}^R\text{TC} \rightarrow \text{TC}$, pour obtenir la distinction rigoureuse recherchée entre énoncé synthétique et énoncé analytique. Dans le langage de Carnap, un énoncé est A-vrai (i.e. analytique dans le langage total de la théorie) s'il est L-impliqué (i.e. logiquement déductible) par les A-postulats (les postulats de signification) ; un énoncé est A-faux s'il est la négation d'un énoncé A-vrai ; enfin, si un énoncé n'est ni A-vrai, ni A-faux, il est synthétique, et P-vrai si, et seulement si, il est vrai et L-impliqué par la théorie totale $\text{TC} + \text{A}_0$ ¹⁵.

2. Lewis et les termes théoriques

Nous avons noté que l'analyse carnapienne des termes théoriques ne rendait pas compte de l'intuition selon laquelle nos théories nous parlent d'entités, hypothétiques certes, mais néanmoins *naturelles*. D'un autre côté, cette analyse semblait sourdre de la conjonction de deux faits : tout d'abord le fait que le sens des termes théoriques est entièrement déterminé par la théorie qui les introduit et ensuite, le fait que le contenu empirique de la théorie est donné par son énoncé de Ramsey. Le travail de Lewis permet de surmonter cette difficulté. Lewis conteste l'idée que nous emploierions le terme « électrons » pour désigner *une* réalisation *arbitraire* du concept d'ÉLECTRONS, aussi bien que l'idée selon laquelle les électrons ne seraient

15. Cf. le chapitre 28 de R. Carnap, *Les fondements philosophiques de la Physique*, J.-M. Luccioni & A. Soulez (trad.), Paris, Armand Colin, 1973.

que des ÉLECTRONS, une concession trop lourde à l'antiréalisme et à une vision purement instrumentale de nos théories scientifiques¹⁶. Mais puisque le sens des termes théoriques est totalement déterminé par le contexte théorique dans lequel ils sont introduits, il s'agit à présent d'expliquer comment nos théories pourraient *définir* (complètement) le sens des termes théoriques. Lewis montre que c'est possible, si du moins nous acceptons l'idée que, lorsque nous formulons une théorie scientifique, son statut épistémologique est tel qu'elle est fautive si elle est multi-réalisée : une théorie introduisant des termes théoriques doit être comprise comme une *définition* de ces termes.

L'idée de Lewis est donc la suivante : lorsque nous introduisons de nouveaux termes par le moyen d'une théorie (le terme « électron » par exemple), nous en définissons en fait complètement le sens en désignant exactement par là même *les* « occupants », quels qu'ils soient, de certains rôles causalement définis. Les termes dans leur contexte théorique fonctionnent à la manière de *descriptions définies*. Mais il faut noter que l'intégration du « surplus de signification » des termes théoriques que Carnap écartait produit ici un « surplus de signification », inattendu celui-là, dans le contenu cognitif de la théorie. Ce dernier ne peut plus être équivalent à l'énoncé de Ramsey de la théorie : si tel était le cas, il n'y aurait rien de tel *a priori* que les uniques occupants des rôles décrits par la théorie. Mais le fait que le contenu cognitif réel de la théorie puisse excéder celui de son énoncé de Ramsey et être donné plutôt par son énoncé de Ramsey *modifié*, qui lui ajoute une contrainte d'unicité de réalisation, ce fait peut être justifié, selon Lewis, par le statut de la théorie : cette contrainte d'unicité, qui ne dérive pas logiquement de la théorie prise *stricto sensu*, est *implicite* parce que la théorie est comprise comme devant conférer un *sens déterminé* aux termes théoriques qui y apparaissent. C'est une différence d'analyse de l'acte de « théorisation » qui permet à Lewis de justifier la modification qu'il propose.

Plus formellement, à partir de l'idée que les termes théoriques réfèrent aux entités, quelles qu'elles soient, qui réalisent la théorie de façon unique, ou alors sont impropres (non-dénotants) dans les cas où la théorie soit n'est pas réalisée soit est multi-réalisée, Lewis conduit sa modification de l'analyse de Carnap de la façon suivante. Soit donc T notre théorie introduisant nos termes théoriques. Commençons par réécrire les termes sous forme de constantes : x est une molécule s'écrit $x \in \text{Mol}$. En vertu du rôle définitionnel de la théorie, celle-ci est vraie si, et seulement si, son énoncé de Ramsey *modifié* est vrai :

Énoncé de Ramsey modifié : $\exists! \mathbf{x} T(\mathbf{x})$

16. D. K. Lewis, « How to define theoretical terms », in *The Journal of Philosophy*, 67 (13), p. 427-446.

Le postulat de signification pour les termes théoriques proposé par Carnap doit à présent à son tour être revu de façon à intégrer le nouvel attendu :

Énoncé de Carnap modifié : $\exists! \mathbf{x} T(\mathbf{x}) \rightarrow T(\mathbf{t}) \wedge \neg \exists! \mathbf{x} T(\mathbf{x}) \rightarrow \mathbf{t} = \bullet$

(notations : $\mathbf{t} = \bullet$ ssi $(t_1, \dots, t_n) = (\bullet, \dots, \bullet)$, \bullet étant un terme nécessairement non-dénotant : par exemple, « le plus grand nombre premier »). Le deuxième membre de la conjonction affirme seulement que si T est multi-réalisée ou non réalisée, alors les termes théoriques qui y figurent sont non-dénotants. D'où l'on tire finalement la définition explicite fonctionnelle suivante¹⁷ :

$$\mathbf{t} = \mathbf{t}x T(\mathbf{x})$$

Il résulte de ce traitement que les identifications théoriques sont nécessaires et n'ont pas besoin d'être posées. Pour illustrer le fait, supposons que $T(\mathbf{t})$ soit notre théorie de l'électricité (celle qui introduit, disons, « électrons », « champs magnétiques », etc.) et que nous constatons qu'un uplet, \mathbf{r} , d'entités connues par ailleurs, disons les causatrons, les champs photomécaniques distants, etc., satisfait $T(\mathbf{x})$ (on a donc : $T(\mathbf{r})$). Il faut dire alors, selon l'analyse précédente, que les causatrons (etc.) réalisent le concept d'électron (etc.) au sens cette fois où nous n'avons pas d'autre choix que d'*identifier* les causatrons aux électrons : les causatrons sont des électrons, ceci parce que le terme « électron » dans la théorie T est une description définie visant les occupants d'un certain rôle causal, occupants qui s'avèrent, dans notre scénario imaginaire, être les causatrons, lesquels sont donc les référents du terme « électrons » ! (Par définition nous avons $\mathbf{t} = \mathbf{t}x T(\mathbf{x})$, et nous constatons que $T(\mathbf{r})$. Donc $\mathbf{t} = \mathbf{r}$.) Sur le même modèle, il est nécessaire à présent que $\text{H}_2\text{O} = \text{eau}$. L'identification est impliquée par la *signification* même des termes théoriques.

3. Termes d'états mentaux

La détermination du sens des termes d'états mentaux pose une difficulté analogue à celle des termes théoriques. Car si les états mentaux sont directement accessibles au sujet qui en est le siège (pour certains du moins), c'est leur caractère irrémédiablement privé, cette fois, qui rend la sémantique des termes d'états mentaux problématique. Les langages et les significations sont publics : comment des agents peuvent-ils jamais s'entendre sur le sens du mot « douleur » si nul n'a accès à *ce que cela fait*, à *l'autre*, d'avoir mal ?

17. Carnap parlait aussi de « définition explicite », mais c'était en un sens plus faible : ses définitions ne fixaient pas complètement la signification des termes.

Lewis propose l'hypothèse de travail suivante : le problème de l'accès aux états mentaux étant analogue à celui des entités théoriques, pourquoi ne pas traiter la sémantique des termes d'état mentaux sur le modèle de la sémantique des termes théoriques ? Un obstacle évident à ce traitement est que les termes d'états mentaux ne sont précisément pas des « termes théoriques » : quelle théorie scientifique les a jamais introduits ? La réponse de Lewis est que les termes d'état mentaux de la psychologie populaire, « avoir mal », « avoir peur », « croire que la neige est blanche », peuvent être compris *comme s'ils* étaient introduits par une théorie, non une théorie scientifique cette fois, mais la « théorie populaire » des états mentaux, l'ensemble des « platitudes » qu'en dit la psychologie commune, en sorte qu'il serait analytique que la douleur, la faim, etc., satisfassent ces platitudes.

Cette hypothèse, affirme Lewis, est une bonne hypothèse (en fait un bon mythe) pour au moins deux raisons : 1) tout d'abord elle explique « l'odeur d'analyticité » des platitudes communes de la psychologie ; 2) elle explique ensuite « la plausibilité initiale du béhaviorisme », autrement dit le projet, sinon difficilement compréhensible, de vouloir rendre compte des états mentaux purement en termes de comportements observables¹⁸.

Mais maintenant la conclusion s'impose : supposons couchée sur le papier cette « théorie » totale, T, de nos états mentaux (M_1, \dots, M_n) mettant en rapport chaque état mental avec des stimuli, des réponses motrices et d'autres états mentaux. Si l'analyse précédente en termes de rôle causal est juste, alors : $\langle M_1, \dots, M_n \rangle = \mathbf{ix} T(\mathbf{x}, \mathbf{o})$. Et supposons encore que soit couchée un jour une théorie de nos états neurophysiologiques, (N_1, \dots, N_n) mettant en rapport chaque état neural avec des stimuli, des réponses motrices et d'autres états neuro-cérébraux, en sorte que soit établi : $\langle N_1, \dots, N_n \rangle = \mathbf{ix} T(\mathbf{x}, \mathbf{o})$. Il s'ensuit nécessairement que $\langle M_1, \dots, M_n \rangle = \langle N_1, \dots, N_n \rangle$ (« par transitivité de = »). Les états mentaux seraient donc identiques à des états neurophysiologiques. Mais ici une précision est nécessaire. Il nous faut prêter attention à une remarque¹⁹ de Lewis selon laquelle la théorie psychologique contient des relativisations implicites au type d'organisme auquel elle s'applique (il faut aussi sans doute comprendre une relativisation implicite au temps, selon une suggestion de H. Field)²⁰. Il est clair que notre théorie des états mentaux n'est pas uniquement réalisée : des organismes distincts la réalisent différemment, nous-mêmes la réalisons sans doute différemment à différents moments. Dire que Jean a faim à midi, c'est dire que Jean à midi réalise T (notre théorie des états mentaux) uniquement. Le terme « faim » désigne ici l'occupant du rôle causal associé dans cette

18. Pensons aussi avec quelle facilité nous tendons à attribuer des états mentaux à des machines ou à des animaux : l'ordinateur « réfléchit », « croit que », le chien et le robot (peut-être conçu pour imiter le chien) « ont soif », etc. Peut-il y a-t-il là une économie du « sens figuré » à faire.

19. Cf. la note 12, p. 257, du recueil : *Papers in metaphysics and epistemology*, Oxford University Press, 1999.

20. Cf. « Mental Representation », in H. Field, *Truth and the Absence of Fact*, Oxford U.P., 2001, p. 46-47, en particulier n. 19.

réalisation, par exemple la sécrétion d'un fluide A dans tel canal spécifié. Il est possible qu'en parlant de la faim de Jean à 14 heures, nous référions à la sécrétion d'un fluide B, et qu'en parlant de la faim chez le caniche nain de douze ans, nous référions à l'excitation d'un neurone (sans parler de la faim d'un robot ou d'un ange). Les identifications psycho-physiques que suggère Lewis sont des identifications de propriété à propriété en un sens restreint, du fait de cette relativisation implicite des théories à des types d'organismes ou de dispositifs : si deux dispositifs de types distincts réalisent la théorie des états mentaux différemment, il s'ensuivra que dans le dispositif de type A, les états mentaux m_1, \dots, m_n sont les états neurophysiologiques n_1, \dots, n_n , mais que dans le dispositif de type B, les états mentaux m_1, \dots, m_n sont des états neurophysiologiques différents n'_1, \dots, n'_n . La soif du chien ou du robot n'est pas la soif d'un *homo sapiens*²¹.

Lewis remarque que son analyse permet de rendre compte de la façon dont on peut parvenir à expliquer le comportement de quelqu'un à partir de prémisses qui sont des énoncés singuliers concernant son état mental. Soient en effet $C_1(t)$, $C_2(t)$... de telles prémisses et E le comportement qu'elles expliquent. Un tel raisonnement n'a pas le format d'une explication si nous ne supposons un appel implicite à des lois causales reliant états mentaux et comportements. Comment des états mentaux pourraient-ils expliquer des comportements ? Les règles de correspondance manquantes n'ont jamais été posées. Mais si nous admettons que les termes d'états mentaux désignent *par définition* les occupants des rôles causaux spécifiés par des lois mettant en rapport états mentaux, stimuli et réponses motrices, l'explication précédente devient possible : si $L_1(t)$, $L_2(t)$... sont les lois causales de la psychologie commune mettant en rapport stimuli, réponses et états mentaux (la « théorie » causale de ces états mentaux), l'explication précédente du comportement E s'écrit de façon équivalente, par définition des états mentaux :

$$\exists! t (L_1(t) \& L_2(t) \& \dots \& C_1(t) \& C_2(t) \dots) \text{ donc } E.$$

Les lois L_i , qui contiennent à la fois du vocabulaire psychologique et du vocabulaire comportemental, permettent à présent, en conjonction avec des prémisses affirmant des faits particuliers et contenant du vocabulaire mental, d'expliquer des faits comportementaux particuliers.

Nous avons vu que l'identification des états mentaux aux états neurophysiologiques, obtenue par Lewis, ne permettait pas de dire en général qu'un état mental x est un état neurophysiologique y. Mais elle permet en revanche de dire que pour un organisme donné à un moment donné, être dans l'état mental x c'est être dans l'état neurophysiologique y.

21. Dans « Mad pain et martian pain », Lewis envisage la possibilité que les propriétés du type « avoir mal », ou « avoir faim » soient, elles, des propriétés fonctionnelles. Tandis que la douleur serait l'occupant d'un rôle causal (une propriété physique), la propriété d'avoir mal désigne le rôle causal lui-même : x a mal s'il existe une propriété physique qui réalise chez lui le rôle de la douleur.

Ceci signifie aussi que, si l'homme et le robot peuvent tous deux avoir mal (ils réalisent la théorie convenable), ils n'en sont pas pour autant dans le *même* état mental. Par contraste, et en vertu de l'analogie entre termes théoriques et termes d'états mentaux, on pourrait vouloir appliquer l'analyse de Carnap-Ramsey, à son tour, à la définition des états mentaux. Supposant ces derniers introduits par une théorie, comme précédemment, il semble que l'analyse de Carnap-Ramsey ne permette pas de distinguer la « douleur » du robot de la « douleur » d'un être humain. La « douleur » n'étant définie dans ce cadre que comme un certain rôle fonctionnel (et non comme l'occupant de ce rôle), la différence des dispositifs physiques²² qui chez le robot ou chez l'être humain réalisent la fonction n'a aucune pertinence pour l'attribution de l'état mental considéré.

Conclusion

De ces deux versions de la sémantique des termes d'états mentaux, laquelle rend compte le plus adéquatement de la façon dont *nous* attribuons les états mentaux ? Nos attributions d'états mentaux ont une portée relativement lâche : elles semblent parfois viser un état subjectif qualitatif, comme dans : « La douleur de Pierre semblait indicible » ; parfois un état purement fonctionnel comme, caricaturalement, dans : « Cette côte est trop forte, le moteur souffre » ; parfois un état neurophysiologique particulier, lorsque par exemple une douleur localisée guide un médecin dans son diagnostic (« La douleur au bras droit est probablement causée par une insuffisance cardiaque »). Notons simplement que seule la version de Lewis paraît pouvoir rendre compte de ces cas d'attribution où nous voulons parler plus précisément de ce qui, chez nous, et chez ceux qui nous ressemblent, réalise le concept de douleur (par exemple) de cette façon singulière dont chacun a fait l'*épreuve* : un état physique bien déterminé²³.

22. Pas nécessairement physiques d'ailleurs : « les différences entre les réalisations » suffisent en principe, sans préjuger de la nature de ces réalisations.

23. Je remercie Marion Vorms et deux lecteurs anonymes du comité scientifique de la revue de leurs remarques et suggestions.

Références :

- Carnap, R., « Empiricism, Semantics, and Ontology », *Revue Internationale de Philosophie*, 4, (1950), p. 20-40. Traduction française Ph. de Rouilhan & Fr. Rivenc, « Empirisme, sémantique et ontologie », in *Signification et nécessité*, Paris, Gallimard, 1997.
- Carnap, R., *Philosophical foundations of physics*, Basic Books, 1966. Traduction française : *Les fondements philosophiques de la physique*, J.M. Luccioni & A. Soulez (trad.), Paris, Armand Colin, 1973.
- Creath, R., « Every dogma has its day », *Erkenntnis*, 35, (1991).
- Dubucs, J., « Les états mentaux sur la place publique », *Intellectica*, 21, (1995), p.115.
- Field, H., « Mental representation », in Field H., *Truth and the Absence of Fact*, Oxford U.P., 2001.
- Field, H., *Truth and the Absence of Fact*, Oxford U.P., 2001.
- Hempel, C.G., « The theoretician's dilemma », in Feigl H., Scriven M. and Maxwell G., *Concepts, theories and the Mind-Body Problem. Minnesota Studies in the Philosophie of Sciences*, University of Minnesota Press, vol.2, (1958).
- Lewis, D.K., « How to define theoretical terms », *The Journal of Philosophy*, 67 (13), (1970), p. 427-446.
- Lewis, D.K., « Theoretical and psychophysical identifications », chap.16 in *Papers in metaphysics and epistemology*, Cambridge University Press, 1999.
- Lewis, D.K., « Mad Pain and Martian Pain », *Readings in the Philosophy of Psychology*, vol. I, Ned Block, éd. Cambridge, Mass. : Harvard University Press, 1980, p. 216-222.
- Psillos, S., « Rudolph Carnap's "Theoretical Concepts in Science" », *Studies in History and philosophy of sciences*, 31 (1), (2000), p. 151-172.
- Putnam, H., « Ce que les théories ne sont pas » in Nagel E., Suppes P., Tarski A., *Logic, Methodology and the Philosophy of Science*, Stanford University Press, 1962. Traduit dans *De Vienne à Cambridge*, Paris, P. Jacob éd., 1980.
- Quine, W.O., « Deux dogmes de l'empirisme », in Quine, *Du point de vue Logique*, Paris, Vrin, 2004.