numéro 1

PHILOSOPHIE

LES ÉDITIONS DE MINUIT

Daniel C. Dennett

SYSTÈMES INTENTIONNELS *

Je désire examiner le concept d'un système dont on puisse — au moins parfois — expliquer et prédire le comportement en lui attribuant des croyances et des désirs (des espoirs, des craintes, des intentions, des pressentiments, etc.). J'appellerai de tels systèmes, *systèmes intentionnels*, et de telles explications et prédictions, explications et prédictions intentionnelles, en raison du caractère intentionnel des idiomes de la croyance et du désir (de l'espoir, de la crainte, de l'intention, du pressentiment, etc. [1]).

J'avais coutume de mettre une majuscule à « intentionnel » chaque fois que je visais la notion d'intentionalité au sens de Brentano, afin de distinguer ce terme technique de celui, proche parent, qu'on trouve par exemple dans : heurter quelqu'un « intentionnellement ». Mais l'usage technique est maintenant bien plus courant, et puisque tous ceux qui utilisent le terme semblent s'accommoder du risque de confusion, j'ai décidé, non sans quelque émotion, de renoncer à cette excentricité typographique. Que le lecteur non-prévenu en prenne note : « intentionnel » n'est pas ici employé dans son sens ordinaire [2]. Pour moi, comme

* Traduction de Jean Khalfa.

1. Je dois à Peter Woodruff d'importantes améliorations de ce texte avant sa première publication. Depuis lors, j'ai trouvé des anticipations et des développements de thèmes similaires ou les confirmant chez divers auteurs, particulièrement Carl Hempel, « Rational Action », *Proceedings and Addresses of the American Philosophical Association*, XXXV (1962), repris dans N. S. Care and C. Landesman, eds., *Readings in the Theory of Action* (Bloomington, Indiana : University Press, 1968) ; L. Jonathan Cohen, « Teleological Explanation », *Proceedings of the Aristotelian Society* (1950-1951), et « Can there be Artificial Minds ? », *Analysis* (1954-1955) ; B. A. O. Williams, « Deciding to Believe », dans H. E. Kiefer and M. K. Munitz, eds., *Language, Belief, and Metaphysics* (Albany : SUNY Press, 1970), et David Lewis, « Radical Interpretation », *Synthese*, III, IV (1974), pp. 331-344.

2. Pour une introduction lumineuse au concept et à son histoire, voir l'article « Intentionality » de Roderick Chisholm dans *The Encyclopedia of Philosophy*, P. Edwards, ed. (New York : MacMillan, 1967).

pour beaucoup d'auteurs récents, l'intentionalité * est avant tout une caractéristique d'entités linguistiques — idiomes, contextes — et un idiome peut être qualifié — du moins ici — d'intentionnel dès que la substitution l'un à l'autre de termes co-référentiels modifie la valeur de vérité ou lorsque les « objets » de l'idiome en question ne peuvent être capturés de la façon habituelle, par les quantificateurs. J'ai étudié cela plus en détail dans *Content and Consciousness* [3].

I

La première chose à dire sur les systèmes intentionnels [4] tels que je viens de les définir est qu'une chose particulière n'est un système intentionnel que par rapport aux stratégies de celui qui essaye d'en expliquer et prédire le comportement. Un exemple permettra de le bien dégager. Prenez le cas d'un ordinateur joueur d'échecs, et des différentes stratégies ou points de vue qu'un adversaire peut adopter pour essayer de prédire ses coups. Il y a trois différents points de vue qui nous intéressent. Tout d'abord, le *point de vue du plan* (*design stance*). Si l'on sait exactement quel est le plan qui régit l'ordinateur (y compris la partie éphémère de son plan : son programme), on peut prédire la réponse assignée à chaque coup, en suivant les instructions du programme. La prédiction sera vérifiée à la seule condition que l'ordinateur fonctionne comme le prévoit son plan c'est-à-dire sans panne. On peut distinguer des variétés de prédictions basées sur le point de vue du plan, mais toutes se ressemblent en ce qu'elles s'appuient sur la notion de *fonction*, notion relative à celle de but et téléologique : le plan d'un système le décompose en parties fonctionnelles plus ou moins grandes, et les prédictions

* Nous écrivons « intentionalité » et non « intentionnalité », contrairement à la tradition adoptée par les traducteurs de Husserl, pour marquer que l'auteur utilise ce terme en un sens très différent. (*N. D. T.*)

3. *Content and Consciouness* (Londres : Routledge & Kegan Paul, 1969).

4. L'expression « système intentionnel » apparaît dans le livre de Charles Taylor *The Explanation of Behaviour* (Londres : Routledge & Kegan Paul, 1964), page 62, où l'utilisation qui en est faite suggère sa coextensivité avec l'expression telle que je l'utilise. Mais Taylor ne développe pas la notion en profondeur. Cf. cependant les pages 58 et suivantes. Pour une introduction au concept de système intentionnel avec moins de présupposés philosophiques, cf. les chapitres XII et XIV de *Brainstorms*.

faites du point de vue du plan dérivent de l'hypothèse que chaque partie fonctionnelle va fonctionner correctement. Ainsi les schémas de câblage de l'ingénieur-radio contiennent-ils des symboles pour chaque résistance, accumulateur, transistor, etc. — *Chacun avec la tâche qu'il doit accomplir* — et l'ingénieur peut prédire le comportement d'un circuit du point de vue du plan en supposant que chaque élément remplit sa tâche. On peut de même prédire, du point de vue du plan, la réponse de l'ordinateur à plusieurs niveaux différents d'abstraction, selon que l'on envisage comme plus petits éléments fonctionnels des générateurs-de-stratégie et des vérificateurs-de-conséquences, des multiplicateurs et des diviseurs, ou des transistors et des interrupteurs. (Remarquons que tous les diagrammes ou images ne sont pas des plans dans ce sens, car un diagramme peut ne comporter aucune information relative aux fonctions — voulues ou observées — des éléments qu'il représente).

Nous adoptons généralement le point de vue de plan lorsque nous prédisons le comportement d'objets mécaniques, par exemple : « Lorsque le chariot de la machine à écrire atteindra la marge, une sonnette tintera si la machine est en bon état de marche) », et, plus simplement : « Frottez l'allumette et elle s'enflammera. » Nous adoptons souvent aussi ce point de vue dans des prédictions relatives à des objets naturels : « Un abondant élagage donnera un feuillage plus dense et des branches plus fortes. » Le trait essentiel du point de vue du plan est que nous faisons des prédictions uniquement à partir d'un savoir ou d'hypothèses portant sur le plan fonctionnel du système et indépendamment de la constitution physique ou de l'état des éléments internes de l'objet en question.

Deuxièmement, il y a ce que nous pouvons appeler le *point de vue physique*. Selon ce point de vue, nos prédictions se fondent sur l'état physique effectif de l'objet particulier, et sont élaborées à partir de notre connaissance des lois de la nature. C'est seulement de ce point de vue que nous pouvons prévoir les dysfonctionnements d'un système (à moins que, comme il arrive parfois de nos jours, un système soit *conçu* (*designed*) pour ne plus fonctionner au bout d'un certain temps, auquel cas le dysfonctionnellement devient, en un sens, une partie de son fonctionnement propre). Les exemples de prédictions faites du point de vue physique ne manquent pas : « Si vous tournez le bouton, vous recevrez un mauvais choc », « Lorsqu'il neigera, cette branche cassera net ». Si l'on adopte rarement le point de vue physique lorsqu'on s'occupe d'un ordinateur, c'est que le nombre de varia-

bles critiques dans la constitution physique d'un ordinateur submergerait le calculateur le plus prodigieux. Il est significatif que le point de vue physique soit généralement réservé pour les cas de panne où la cause, empêchant un fonctionnement normal, est généralisée et aisément localisable, comme par exemple : « Il ne se passera rien lorsque vous taperez vos questions, parce que ce n'est pas branché », ou : « Ça ne marchera pas parce qu'il y a plein d'eau à l'intérieur. » Essayer de donner une description ou une prédiction physique de l'ordinateur joueur d'échecs serait une tâche vaine et herculéenne, mais en principe réalisable. On pourrait lors d'une partie d'échecs, prédire sa réponse, en suivant les effets de l'apport initial d'énergie à travers l'ordinateur jusqu'au point où un nouveau caractère frappe le papier, et où une réponse est imprimée. (Du fait du caractère numérique du mode de fonctionnement des ordinateurs, les indéterminations au niveau quantique, s'il y en a, s'annuleront au lieu de s'accumuler. A moins, bien sûr, qu'un *randomizer* (producteur de hasard) ou tout autre amplificateur d'effets quantiques ne fasse partie de l'ordinateur.

De nos jours, les meilleurs ordinateurs joueurs d'échecs sont pratiquement imprédictibles tant du point de vue du plan que du point de vue physique : ils sont devenus trop complexes, au regard même de ceux qui les ont conçus pour pouvoir être considérés du point de vue de leur plan. Pour un homme, le meilleur espoir de gagner une partie d'échecs contre une telle machine est d'en prévoir les réponses en supputant de son mieux, étant donnés les buts et règles des échecs, le meilleur coup ou le plus rationnel. C'est dire que l'on suppose non seulement 1) que la machine fonctionnera comme prévu par son plan, mais aussi 2) que ce plan est optimum, que l'ordinateur « choisira » le coup le plus rationnel. Les prédictions basées sur ces suppositions peuvent échouer si l'une ou l'autre supposition se révèle infondée dans le cas en question ; pourtant ce *mode* de prédiction peut encore nous intéresser, comme le plus fructueux à adopter, lorsqu'on a affaire à tel système particulier. En d'autres termes, quand il n'y a plus d'espoir de battre la machine en utilisant notre savoir physique ou informatique afin d'en anticiper les réponses, on peut encore éviter une défaite en traitant la machine plutôt comme un adversaire humain et intelligent.

Examinons de plus près cette stratégie. Une prédiction reposant sur l'hypothèse de la rationalité du système dépend d'un certain nombre de choses. Tout d'abord, rationalité ne signifie rien de plus, jusqu'ici, que le plan optimum par rapport à un but,

ou la hiérarchie de buts la mieux pondérée (dans le cas des échecs, faire échec et mat, gagner des pièces, maintenir une défense, etc.), et un ensemble de contraintes (les règles et la position de départ). De plus, la prédiction elle-même dépend de la nature et de l'étendue de l'information, concernant son domaine d'activité, que possède le système à un moment donné. Formuler ce type de prédiction revient à poser la question suivante : qu'est-ce que l'ordinateur peut faire de plus rationnel, étant donnés les buts *x, y, z,...*, les contraintes *a, b, c,...*, et l'information dont il dispose (y compris les fausses informations) sur l'état de choses *p, q, r,...* ? Lorsque je prédis la réponse de l'ordinateur à l'un de mes coups, ma supputation du coup le plus rationnel pour l'ordinateur peut dépendre par exemple non seulement de la supposition que l'ordinateur est informé de la position présente de toutes les pièces, mais aussi du fait que je crois (ou non) que l'ordinateur est informé de mon impuissance à prévoir au-delà de quatre coups, des pouvoirs relatifs des cavaliers et des fous, et de mon penchant pour les échanges cavalier-fou. Il se peut enfin que je sois incapable de formuler une très bonne prédiction si je ne puis déterminer avec précision l'information et les buts de l'ordinateur, ou si l'information et les buts que je lui suppose ne déterminent pas un meilleur coup unique, voire simplement si je suis moins capable que l'ordinateur de déduire un coup optimum à partir de ces données. De telles prédictions sont alors très précaires ; non seulement elles dépendent d'un ensemble de postulats sur les buts, les contraintes, l'information, et de la détermination de la réponse optimum dans des situations où nous n'avons aucun critère clair de ce qui précisément est optimum mais de plus, elles sont susceptibles d'être falsifiées par les courts-circuits qui sont imprévisibles de ce point de vue. De même que les prédictions faites du point de vue du plan sont vulnérables en cas de mauvais fonctionnement (car elles reposent sur l'hypothèse d'un bon fonctionnement), de même ces prédictions sont vulnérables en cas de faiblesses ou de défauts du plan (car elles reposent sur l'hypothèse d'un plan optimum). Que ces prédictions précaires se vérifient avec assez de régularité pour rendre la méthode utile, c'est là la mesure du succès de ceux qui conçoivent aujourd'hui ces programmes.

Le dénouement de la longue analyse de cet exemple doit être maintenant évident : le troisième point de vue, avec son hypothèse de rationalité, est le *point de vue intentionnel* : les prédictions qu'on en tire sont des prédictions intentionnelles, on considère l'ordinateur comme un système intentionnel. Dans ce cas,

on prédit un comportement en attribuant au système *la possession d'une certaine information* et en le supposant *dirigé par certains buts*, puis en calculant d'après ces attributions et suppositions l'action la plus raisonnable ou appropriée. De là à appeler les informations dont dispose l'ordinateur ses *croyances*, et ses buts principaux et subordonnés, ses *désirs*, il n'y a qu'un pas. Ce que je veux dire en affirmant qu'il n'y a qu'un pas, c'est que la notion de possession d'information est tout aussi intentionnelle que celle de croyance. La « possession » en question n'est certainement pas, comme on pourrait le croire, la notion facile et innocente de stockage ; c'est, et ce doit être une « possession épistémique », un analogue de la croyance. Considérez ceci : on peut dire qu'un Français qui possède l'*Encyclopedia Britannica* sans connaître l'anglais « possède » l'information qu'elle contient ; mais, si ce sens du mot possession est possible, il n'est pas assez fort pour désigner le type de possession dont l'ordinateur est supposé jouir (*enjoy*), si l'on considère l'information qu'il *utilise* lorsqu'il « choisit » un coup aux échecs. De la même façon, les buts que poursuit un ordinateur doivent être spécifiés intentionnellement, tout comme les désirs.

Les doutes qui pourraient subsister quant à la question de savoir si l'ordinateur joueur l'échecs a *réellement* des croyances et des désirs sont hors de propos ; car la définition des systèmes intentionnels que j'ai donnée ne dit pas que les systèmes intentionnels possèdent *réellement* croyances et désirs, mais qu'on peut expliquer et prédire leur comportement en leur *attribuant* des croyances et désirs. Qu'on nomme ce qui est attribué à l'ordinateur croyances, analogues de croyance, complexes d'information ou tout ce qu'on voudra d'intentionnel ne change rien à la nature du calcul fait sur la base de ces attributions. On parviendra aux mêmes prédictions, que l'on pense carrément en termes de croyances et de désirs de l'ordinateur ou en termes de ses stocks d'information et de ses spécifications-de-buts. Ce qu'il y a ici d'incontournable et d'intéressant, c'est que, pour les meilleurs ordinateurs joueurs d'échecs d'aujourd'hui, l'explication et la prédiction intentionnelles de leur comportement est non seulement courante mais même efficace quand les autres modes de prévision sont inopérants. C'est avec un certain succès que nous traitons ces ordinateurs comme des systèmes intentionnels, et nous le faisons indépendamment de toute considération sur la substance dont ils sont composés, leur origine, leur situation ou absence de situation dans la communauté des agents moraux, leur conscience ou leur conscience de soi, le caractère déterminé ou

non de leurs opérations. La décision d'adopter cette stratégie est pragmatique et non, en elle-même, vraie ou fausse. On peut toujours refuser d'adopter le point de vue intentionnel vis-à-vis de l'ordinateur, et accepter ses mats. On peut changer de point de vue à volonté sans néanmoins s'engager dans une voie incohérente ou inhumaine, adopter le point de vue intentionnel en tant qu'adversaire, le point de vue du plan lorsqu'on le révise, et le point de vue physique en tant que réparateur.

Cette célébration de notre ordinateur joueur d'échecs ne vise pas à en faire un modèle ou une simulation parfaitement adéquate de l'esprit ou de l'activité intelligente humaine ou animale ; et je ne dis pas non plus que l'attitude que nous adoptons envers cet ordinateur est précisément la même que celle que nous adoptons envers une créature que nous jugeons consciente ou rationnelle. Tout ce que j'ai affirmé est, qu'à l'occasion, un système purement physique peut être si complexe, et cependant si organisé que nous trouvons commode, éclairant et pragmatiquement nécessaire à une prédiction de le traiter comme s'il avait des croyances et des désirs, et comme s'il était rationnel. L'ordinateur joueur d'échecs est seulement cela : une machine qui joue aux échecs, ce que n'est aucun homme ni aucun animal ; d'où sa « rationalité » étriquée et artificielle.

Peut-être devrions-nous développer directement l'exemple de l'ordinateur joueur d'échecs en un modèle plus fidèle de la rationalité humaine — ou peut-être pas. Je préfère suivre d'abord une ligne de recherche plus fondamentale.

Quand devons-nous nous attendre à ce que la tactique du point de vue intentionnel soit payante ? — Lorsque nous avons une raison de supposer que l'hypothèse du plan optimum est fondée et que nous doutons de la commodité d'une prédiction faite du point de vue physique ou du point de vue du plan. Supposez que nous nous rendions sur une planète éloignée et que nous la trouvions habitée par des choses qui parcourent sa surface, se multiplient, dépérissent, réagissant apparemment aux événements de leur environnement, mais, pour le reste, aussi différentes des hommes que vous voudrez. Pouvons-nous donner des prédictions et explications intentionnelles de leur comportement ? Si nous avons des raisons de supposer qu'un processus de sélection naturelle a joué, alors nous pouvons être sûrs que les populations que nous observons ont été sélectionnées en vertu des plans qui les organisent. Elles réagiront au moins à certains des types d'événement les plus communs de cet environnement et selon les modes les plus normalement appropriés — c'est-à-

dire contribuant à la propagation des espèces [5]. Dès que nous aurons identifié, à titre expérimental, les dangers et les secours qu'offre l'environnement (par rapport à la constitution des habitants, et non à la nôtre), nous pourrons évaluer les buts et des hiérarchies de buts les mieux adaptées aux *besoins* de ces créatures (pour leur survie et leur multiplication), nous pourrons évaluer quelles sortes d'information sur l'environnement seront *utiles* pour guider une activité finalisée et quelles activités seront appropriées compte tenu des circonstances et de l'environnement. Ayant deviné ces conditions (toujours susceptibles de révisions), nous pouvons continuer et attribuer aussitôt des croyances et des désirs à ces créatures. Leur comportement « manifestera » leurs croyances dans la mesure où on le percevra comme une série d'actions qui, étant donnés les désirs de ces créatures, répondent à des croyances appropriées à la stimulation de l'environnement. Les désirs à leur tour seront « manifestés » par le comportement comme ces désirs appropriés (étant donnés les besoins de la créature) auxquels répondraient les actions, en fonction des croyances de cette créature. La circularité de cet emboîtement de spécifications n'est pas accidentelle. Les attributions de croyances et de désirs doivent être interdépendantes, et les seuls points d'ancrage sont les nécessités vitales démontrables, les régularités du comportement, et l'hypothèse, fondée sur la foi en la sélection naturelle, d'un plan optimum. Toutefois, dès lors qu'on a attribué des croyances et des désirs, on peut aussitôt, en se fondant sur eux, prédire le comportement, et si l'évolution a fait son travail — comme elle le doit à long terme —, nos prédictions seront assez dignes de foi pour être utiles.

On peut penser de prime abord que cette tactique impose, sans justification possible, des catégories et des attributs humains (croyances, désirs, etc.) à ces entités étrangères. Et, certes, c'est une sorte d'anthropomorphisme, mais un anthropomorphisme conceptuellement innocent. Il n'est pas nécessaire de supposer que ces créatures partagent aucune de nos inclinations, attitudes, espoirs, faiblesses, plaisirs ou visions humains ; courir, sauter, se cacher, manger, dormir, écouter ou copuler ne seront peut-être pas au nombre de leurs actions. Nous transférons de notre monde vers le leur, les seules catégories de rationalité, de perception

5. Notez que ce qui est *directement* sélectionné, le gène, est un diagramme et non un plan (*design*) ; toutefois, s'il est sélectionné, c'est qu'il confère à son porteur un certain plan (fonctionnel). Cela m'a été signalé par Woodruff.

(entrée d'information selon une ou des modalités « sensorielles », un radar ou une radiation cosmique peut-être), et d'action. La question de savoir si nous pouvons nous attendre à ce qu'elles partagent quelqu'une de nos croyances ou de nos désirs est délicate, mais l'on peut dès maintenant parvenir à quelques conclusions. Ainsi, on peut supposer qu'en vertu de leur rationalité elles partagent notre croyance dans les vérités logiques [6], et nous ne pouvons supposer qu'elles désirent normalement leur propre destruction.

II

Lorsque, confronté à un système — qu'il s'agisse d'un homme, d'une machine, ou d'une créature étrangère —, on explique et prédit son comportement en invoquant ses croyances et ses désirs, on possède ce qu'on pourrait appeler une « théorie du comportement » pour ce système. Voyons maintenant quel rapport ces théories intentionnelles du comportement entretiennent avec les autres théories du comportement.

S'il est un fait évident au point de passer inaperçu, c'est que le « sens commun » explique et prédit le comportement tant des hommes que des animaux de façon intentionnelle. Nous n'*escomptons* pas des nouvelles connaissances (*acquaintances*) qu'elles réagissent de manière irrationnelle à des sujets de conversation ou des événements particuliers, mais, lorsqu'elles le font, nous apprenons à ajuster nos stratégies en conséquence, de même que lorsqu'on joue aux échecs avec un ordinateur : au départ, on place très haut son niveau de rationalité et l'on révise en baisse cette estimation toutes les fois que son jeu révèle des défauts. La présomption de rationalité est si profondément ancrée dans nos habitudes inférentielles que, lorsque nos prédictions se montrent fausses, nous cherchons, avant de mettre en doute la rationalité du système pris globalement, des justifications dans les conditions de la détention d'information (il n'a pas dû entendre, il ne doit pas savoir l'anglais, il n'a pas dû voir x, avoir conscience de ce

6. Voir l'argument de Quine sur la nécessité de « découvrir » nos connecteurs logiques dans toutes les langues que nous pouvons traduire. *Word and Object* (Cambridge, Mass. : MIT, 1960), section 13. [Trad. fr. Le *mot et la chose*, 1979, Flammarion]. Nous ajouterons des arguments en faveur de ce point plus loin.

que *y*, etc.) ou dans la hiérarchie des buts. Dans des cas extrêmes, des personnalités peuvent se montrer si imprévisibles du point de vue intentionnel que nous en abandonnons ce point de vue, et, si nous avons accumulé entre-temps beaucoup d'informations sur la nature des types de réaction propres à ce sujet, il se peut que nous considérions utile d'adopter une des espèces du point de vue du plan. Telle est l'attitude, fondamentalement différente, que nous adoptons parfois vis-à-vis du fou. Voir un gardien d'asile manipuler un patient obsessivement contrariant, c'est voir quelque chose de radicalement différent de relations interpersonnelles normales.

Notre prédiction du comportement animal selon le « sens commun » est, elle aussi, intentionnelle. Que les âmes sentimentales dépassent ou non la mesure lorsqu'elles parlent à leurs chiens, ou bourrent les têtes de leurs chats de schémas et de préoccupations, les plus endurcis d'entre nous prédisent le comportement animal de façon intentionnelle. Lorsque nous observons une souris dans une situation où elle peut voir un chat qui l'attend à l'une des sorties de son repaire et du fromage à l'autre, nous savons de quel côté elle ira pourvu qu'on ne la dérange pas ; notre prédiction ne se base ni sur une familiarité avec des expériences de labyrinthe ni sur des hypothèses concernant l'entraînement spécial qu'aurait suivi cette souris. Nous supposons que la souris peut voir le chat et le fromage, qu'elle a donc des croyances (des analogues-de-croyance, ou tout ce qu'on voudra d'intentionnel) selon lesquelles il y a un chat à gauche et du fromage à droite, et nous lui attribuons le désir de manger le fromage et celui d'éviter le chat (désirs rangés à assez juste titre sous les désirs plus généraux de manger et d'éviter le danger). Ainsi prédisons-nous que la souris fera ce qui répond à de telles croyances et désirs : aller à droite pour prendre le fromage et éviter le chat. Quelles que puissent être nos obédiences académiques ou nos prédilections théoriques, nous serions étonnés si, en règle générale, le comportement des souris et des autres animaux falsifiait de telles prédictions intentionnelles. En vérité, et quelle que soit leur école, les psychologues expérimentalistes auraient bien du mal à concevoir les situations expérimentales destinées à confirmer leurs diverses théories sans l'aide de prévisions intentionnelles concernant les réactions aux circonstances des animaux testés.

J'ai soutenu plus haut que même les créatures d'une autre planète devraient partager nos croyances aux vérités logiques ; on peut éclairer cette affirmation en se demandant si les souris

et les autres animaux, en tant que systèmes intentionnels, croient aussi aux lois de la logique. Il y a bien quelque chose de bizarre dans l'image d'un chien ou d'une souris cogitant une liste de tautologies, mais nous pouvons faire l'économie de cette image. Supposer qu'une chose est un système intentionnel, c'est supposer qu'elle est rationnelle ; autrement dit, on n'aboutit à rien à partir de la supposition que l'entité *x* a les croyances *p, q, r...* si l'on ne suppose pas que *x* croit ce qui découle de *p, q, r...*, faute de quoi il n'y a aucun moyen d'écarter la prédiction que *x* agira, en dépit de ses croyances *p, q, r...*, de façon complètement stupide. Et, si nous ne pouvons écarter *cette* prédiction, nous n'aurons acquis absolument aucun pouvoir de prédiction. Aussi, qu'on dise ou non que l'animal *croit* aux *vérités* de la logique, il faut supposer qu'il en *suit* les *règles*. A coup sûr, notre souris suit ou croit aux règles du *modus ponens*, car nous lui attribuons les croyances suivantes : (a) *il y a un chat à gauche*, et (b) *s'il y a un chat à gauche, je ferais mieux de ne pas aller à gauche*, et notre prédiction s'appuie sur la capacité de la souris à atteindre la conclusion. En général, il y a interchangeabilité entre les règles et les vérités ; nous pouvons supposer que *x* possède une règle d'inférence menant de *A* à *B*, ou bien nous pouvons attribuer à *x* la croyance au « théorème » : *si A, alors B*. Dans nos prédictions, nous sommes libres d'attribuer à la souris soit quelques règles d'inférence et la croyance en de nombreuses propositions logiques, soit de nombreuses règles d'inférence et peu (ou pas) de croyances logiques [7]. Nous pouvons même prendre une croyance manifestement non-logique telle que (b) et la reformuler comme une règle d'inférence menant de (a) à la conclusion désirée.

Toutes les vérités logiques figureront-elles parmi les croyances de tout système intentionnel ? Si le système était idéalement ou parfaitement rationnel, toutes les vérités logiques y figureraient, mais tout système intentionnel réel sera imparfait, et l'on ne doit donc pas ranger toutes les vérités logiques au nombre des croyances d'un système. Qui plus est, toutes les règles d'inférence d'un système intentionnel réel ne peuvent être valides et toutes ses croyances autorisant des inférences ne peuvent être des vérités logiques. L'expérience peut indiquer où, dans un système

7. Nous acceptons l'argument de Lewis Carroll, dans « What the Tortoise Said to Achilles », *Mind* (1895), repris dans I. M. Copi and J. A. Gould, *Readings on Logic* (New York : MacMillan, 1964), et ne pouvons donc pas laisser remplacer toutes les règles gouvernant un système par des croyances, car cela engendrerait un emboîtement infini et improductif de croyances distinctes sur ce qui peut être inféré, et à partir de quoi.

particulier, résident les failles. Si nous trouvions une créature imparfaitement rationnelle dont l'obéissance au *modus ponens*, par exemple, variait avec le sujet de l'inférence, nous caractériserions ce fait en excluant le *modus ponens* comme règle et en le remplaçant par un ensemble de règles d'inférence non-logiques recoupant l'étape du *modus ponens* pour chacun des domaines où la règle était suivie. Et il n'est pas surprenant que plus nous découvrons d'imperfections (plus nous éliminons les vérités logiques des croyances de la créature), plus nos tentatives de prédiction intentionnelle deviennent maladroites et indécidables, car nous ne pouvons plus alors escompter que les croyances, désirs et actions qui *doivent* s'accorder le feront bien. A la fin, si nous poursuivons ce processus, nous en viendrons à adopter le point de vue du plan, autrement dit nous finirons par abandonner l'hypothèse de rationalité [8].

Ce déplacement des explications et prédictions intentionnelles du sens commun vers celles, plus solides, obéissant au point de vue du plan, déplacement qui s'impose à nous lorsque nous découvrons que nos sujets sont imparfaitement rationnels, est, indépendamment de toute découverte de ce type, le chemin que doivent suivre, autant que faire se peut, les constructeurs de théories. En dernière instance, nous voulons pouvoir expliquer l'intelligence humaine ou animale dans les termes de son plan, et celui-ci dans les termes de la sélection naturelle de ce plan. Aussi, chaque fois que nous nous arrêtons dans nos explications au niveau intentionnel, nous remettons à plus tard un cas inexpliqué d'intelligence ou de rationalité. Cela apparaîtra très clairement si nous examinons la construction des théories au point de vue privilégié d'un économiste.

Chaque fois qu'un constructeur de théorie propose d'appeler un événement quelconque, état, structure, etc., dans un système quelconque (par exemple, le cerveau d'un organisme), *signal* ou *message* ou *commande,* ou le dote de quelque autre façon d'un contenu, il *emprunte une somme d'intelligence.* Il postule implicitement, outre ses signaux, messages ou commandes, quelque chose qui puisse *lire* des signaux, *comprendre* des messages, ou *commander*, sans quoi ses signaux ne serviraient à rien, dépéri-

8. Ce paragraphe doit beaucoup à une discussion avec John Vickers, dont l'article « Judgment and Belief », dans K. Lambert, *The Logical Way of Doing Things* (New Haven, Conn. : Yale, 1969) va au-delà des remarques présentées ici, et considère la force ou le poids relatifs des croyances et des désirs.

raient sans être reçus ni compris. On doit en fin de compte rembourser cet emprunt en trouvant et en réduisant par analyse ces instances de lecture ou de compréhension ; car, sans cela, la théorie comptera parmi ses éléments des analogues d'humains, inanalysés, assez intelligents pour lire ces signaux, etc. La théorie *différera* ainsi sa réponse à la question majeure : en quoi consiste l'intelligence ? L'intentionalité de tout ce discours de signaux et commandes nous rappelle que la rationalité est prise comme allant de soi, et, par là, nous montre le point où une théorie est incomplète. C'est ce trait qui, à mon avis, donne de la valeur à la tâche encore inachevée de construire une définition rigoureuse de l'intentionalité. Car, si nous pouvons revendiquer la possession d'un critère purement formel du discours intentionnel, nous aurons l'équivalent d'une monnaie d'échange permettant d'évaluer les théories du comportement. L'intentionalité *fait abstraction* des détails inessentiels des diverses formes que peuvent prendre les prêts d'intelligence (par exemple : lecteurs de signaux, émetteurs de volitions, bibliothécaires des couloirs de la mémoire, moi et surmoi) et sert à localiser exactement le point où une théorie est *déficitaire* relativement à la tâche d'explication de l'intelligence. Partout où une théorie repose sur une formulation portant les marques logiques de l'intentionalité se cache un homoncule.

L'insuffisance de l'explication intentionnelle en psychologie a été aussi largement ressentie que mal comprise. Les plus influents soupçons, exprimés dans le behaviorisme de Skinner et de Quine, peuvent être brièvement caractérisés dans les termes de notre métaphore économique. Le rejet sans appel par Skinner et Quine des idiomes intentionnels, à tous les niveaux de théorie, ressemble au conservatisme exacerbé de la Nouvelle-Angleterre : défense de se livrer à des dépenses déficitaires lorsqu'on construit une théorie ! Dans le cas de Quine, son horreur des prêts est surtout due à la crainte qu'ils ne puissent être remboursés, tandis que Skinner insiste plutôt sur le fait que ce qu'on emprunte ne vaut de toute façon rien. Skinner soupçonne les affirmations de type intentionnel d'être empiriquement vides, parce que bien trop faciles à accorder aux données, tout comme cette *virtus dormitiva* que les médecins de Molière attribuaient au somnifère (cf. le chapitre IV de *Brainstorms* pour une discussion plus détaillée de ces points). Cependant, on peut faire, à titre temporaire, des pétitions de principe, qui autorisent un mode de prédiction et d'explication qui n'est pas absolument vide. Soit la prédiction intentionnelle suivante : si j'avais à demander à mille mathématiciens amé-

ricains combien font sept fois cinq, plus de neuf cents répondraient que cela fait trente-cinq. (J'ai tenu compte de ceux qui auraient mal entendu ma question, de ceux qui ont mauvais caractère, de ceux dont la langue a fourché.) Si vous doutez de cette prédiction, mettez-la à l'épreuve : je parierai gros là-dessus. Cette prédiction semble posséder un contenu empirique, car elle peut, en un sens, être vérifiée et pourtant, comme prédiction tirée d'une théorie psychologique empirique, elle est peu satisfaisante. Cela marche, bien sûr, à cause du fait contingent, empirique — mais garanti par l'évolution — qu'en général les hommes sont suffisamment bien conçus pour, à la fois, trouver la bonne réponse et vouloir la trouver. On rencontrerait aussi peu d'exceptions chez un groupe de Martiens avec qui nous pourrions converser, car ce n'est pas seulement une prédiction de psychologie *humaine*, mais de la « psychologie » des systèmes intentionnels en général.

Le fait de décider, sur la base d'évidences empiriques disponibles que quelque chose est un morceau de cuivre ou du lichen permet de faire des prédictions basées sur les théories empiriques relatives au cuivre et aux lichens ; mais décider sur la base d'évidences disponibles que quelque chose est (peut être traité comme) un système intentionnel permet des prévisions dont la base est normative ou logique plutôt qu'empirique, et c'est pourquoi le succès d'une prédiction intentionnelle (qui ne se fonde sur aucune représentation particulière du plan du système) ne peut être utilisé pour confirmer ou infirmer des représentations particulières du plan du système.

Skinner a réagi à cela en essayant de ne formuler que des prédictions en langage non-intentionnel, en prédisant des réponses corporelles à des stimuli physiques, mais cela ne lui a pas fourni à ce jour le mode de prédiction et d'explication différent qu'il recherchait, comme un examen très superficiel permet de le voir. Pour mettre en scène une prédiction non-intentionnelle du comportement, il a inventé la boîte de Skinner dans laquelle le comportement qui sera récompensé (l'occupant étant un rat, par exemple) est un mouvement corporel extrêmement restreint et stéréotypé — ordinairement, abaisser un levier avec les pattes antérieures.

On prétend alors découvrir que, lorsque l'animal a été entraîné, une relation d'apparence légale lie des événements caractérisés de façon non-intentionnelle : les stimuli de contrôle et les réponses de pression sur le levier. Certes, on découvre bien une régularité, mais le fait qu'elle lie des événements définis de façon non-intentionnelle est dû à une propriété de la boîte de Skinner et non de

l'occupant. En effet, transformons notre prédiction sur les mathématiciens en une prédiction skinnérienne : attachez un mathématicien dans une boîte de Skinner de sorte qu'il ne puisse remuer que la tête ; placez sous ses yeux une carte où est écrit : « Combien font sept fois cinq ? », mettez à portée de sa tête deux boutons, dont l'un est marqué « 35 » et l'autre « 34 » ; mettez-lui des électrodes sur la plante des pieds et envoyez-lui vivement quelques chocs. Le stimulus de contrôle doit être le son : « Maintenant, répondez ! » Je prévois que dans un nombre statistiquement significatif de cas, avant même d'avoir fait des essais d'entraînement en vue de conditionner cet homme à appuyer sur le bouton « 35 » avec son front, il le fera lorsqu'on lui présentera le stimulus de contrôle. Est-ce, du seul fait qu'elle évite le vocabulaire intentionnel, une prédiction scientifique satisfaisante ? Non, c'est une prédiction intentionnelle travestie par une telle restriction de l'environnement que le sujet ne dispose que d'un seul mouvement corporel pour accomplir l'*action* intentionnelle que quiconque indiquerait comme adaptée aux circonstances de perception, de croyance et de désir. Qu'il s'agisse, dans ces prédictions, d'action et non seulement de pur mouvement, on peut le remarquer à propos de sujets moins intelligents que des mathématiciens. Supposons qu'on entraîne une souris dans une boîte de Skinner à faire exactement quatre pas et à presser un levier avec son museau (en la récompensant avec de la nourriture) ; si les lois de Skinner liaient vraiment des stimuli et des réponses définies en termes de mouvements corporels, Skinner serait obligé de prédire qu'au cas où nous reculerions le levier d'un pouce, de sorte que quatre pas ne permettent plus de l'atteindre, la souris donnerait des coups de museau dans le vide plutôt que de faire un cinquième pas.

Une variante de la théorie skinnérienne, conçue pour prévenir cette objection, admet que la réponse prévue comme résultat de l'entraînement n'est pas réellement saisie par une description du seul mouvement du squelette, mais plutôt par une description de l'effet produit sur l'environnement : abaissement du levier, pression exercée sur le bouton « 35 ». Cela n'ira pas mieux. Supposons que l'on puisse en fait exercer un homme ou un animal à produire un effet sur l'environnement, comme cette théorie le propose. Supposons par exemple que nous exercions un homme à pousser un bouton sous la plus longue de deux images présentées, des dessins ou de simples schémas : nous le récompenserons lorsqu'il pousse le bouton situé sous l'image du plus long crayon, cigare, etc. Si elle était correcte, cette théorie aurait pour consé-

quence miraculeuse que si, après avoir entraîné le sujet sur des images simples, nous devions lui présenter l'illusion optique de la tête de flèche dans l'expérience de Müller-Lyer, il ne succomberait pas à l'illusion ; car, *ex hypothesi,* il a été entraîné à produire un effet *réel* sur l'environnement (choisir l'image qui *est* la plus longue) et non pas un effet *perçu* ou *cru* (choisir l'image qui *semble* la plus longue). La prédiction digne de foi est à nouveau la prédiction intentionnelle [9].

Le plan expérimental de Skinner est supposé éliminer l'intentionnel, mais il ne fait que le masquer. Si les prédictions non-intentionnelles de Skinner fonctionnent dans une certaine mesure, ce n'est pas parce qu'il aurait effectivement trouvé des lois non-intentionnelles du comportement, mais parce que les prédictions intentionnelles vraiment dignes de confiance sous-tendant ses situations expérimentales (le rat désire de la nourriture et croit qu'il en obtiendra en appuyant sur le levier — fait dont on lui a largement fourni la preuve —, donc il appuie sur le levier), sont travesties en ne laissant virtuellement de place dans l'environnement que pour un seul mouvement corporel susceptible d'accomplir l'action appropriée et en ne laissant virtuellement aucune place dans le même environnement pour permettre une discordance entre les croyances du sujet et la réalité.

Où devons-nous alors rechercher une théorie satisfaisante du comportement ? En tant que psychologie, la théorie intentionnelle est vide, car elle présuppose la rationalité et l'intelligence au lieu de les expliquer. Les succès apparents du behaviorisme de Skinner dépendent toutefois de prédictions intentionnelles cachées. Skinner a raison de penser que l'intentionalité ne peut être un *fondement* pour la psychologie, ainsi que de chercher des régularités purement mécaniques dans les activités de ses sujets, mais il y a peu de raisons de supposer qu'elles se trouvent à la surface, dans le comportement global — à moins, comme nous l'avons vu, de passer à une régularité intentionnelle une camisole de force artificielle. C'est plutôt dans le fonctionnement des systèmes internes dont le plan approche de l'optimum (relativement à une fin) que nous trouverons des régularités mécaniques. Notre meilleure tactique pour chercher à connaître un plan interne, c'est de contracter des emprunts d'intelligence, de doter

9. R. L. Gregory, *Eye and Brain* (Londres, *World University Library*, 1966), p. 137, rapporte que des pigeons et des poissons que l'on a entraînés exactement de cette façon sont sujets à l'illusion optique de la grandeur, ce qui n'est pas pour nous étonner.

les événements périphériques et internes d'un contenu, et ensuite de chercher des mécanismes dont le fonctionnement corresponde à de tels « messages », afin de pouvoir rembourser les emprunts. Cette tactique est souvent utilisée. La recherche en intelligence artificielle, qui a produit, entre autres choses, l'ordinateur joueur d'échecs, procède en passant d'un problème caractérisé intentionnellement (comment faire pour que l'ordinateur envisage les informations correctes, prenne les décisions correctes ?) à une solution du point de vue du plan — une approximation du plan optimum. Les psychophysiciens et les neurophysiologues qui décrivent couramment les événements en termes de transmission d'information à l'intérieur du système nerveux font, de même, l'emprunt d'un capital intentionnel — même s'ils sont souvent enclins à ignorer ou nier leurs dettes.

Finalement on ne devrait pas supposer que, du seul fait que la théorie intentionnelle est vide de sens comme psychologie en raison de son hypothèse de rationalité, elle est à tout point de vue vide de sens. La Théorie des jeux, par exemple, est inéluctablement intentionnelle [10], mais, en tant que théorie normative formelle, et non pas comme psychologie, elle n'y perd rien. La justesse des prédictions de la Théorie des jeux relatives à des sujets humains repose sur la garantie, fournie par l'évolution, que l'homme est bien adapté au jeu, qui est un cas spécial de rationalité. De même, l'économie, qui est aujourd'hui la science sociale au plus grand pouvoir de prédiction, n'est pas une théorie psychologique et présuppose ce que la psychologie doit expliquer. L'explication et la prédiction économiques sont intentionnelles (bien que de façon partiellement masquée) et le succès qu'elles obtiennent vient de ce que les individus humains sont en général de bonnes approximations de l'opérateur optimum sur le marché.

III

Le concept de système intentionnel est une notion relativement exempte de confusion et non-métaphysique, car elle fait abstraction des questions de composition, constitution, conscience, moralité, ou divinité des entités qu'elle subsume. Aussi est-il plus facile, par exemple, de décider si une machine peut être un

10. Hintikka note en passant dans *Knowledge and Belief* (Ithaca, N. Y. : Cornell, 1962), p. 38, que la Théorie des jeux est semblable à sa logique épistémique en ce qu'elle présuppose une rationalité.

système intentionnel que de décider si une machine peut ***réellement*** penser, être consciente, ou moralement responsable. Cette simplicité en fait une source idéale de classification et d'organisation dans l'analyse philosophique des concepts « mentaux ». Quoi que puisse être une personne — esprit incarné ou âme, agent moral conscient de soi, forme « émergente » d'intelligence —, c'est un système intentionnel, et tout ce qui dérive du seul fait d'être un système intentionnel, est vrai d'une personne. Il est intéressant de voir exactement dans quelle mesure ce que nous pensons s'appliquer aux personnes et à leurs esprits découle directement du fait qu'elles sont des systèmes intentionnels. Pour revenir un instant à la métaphore économique, la question qui, comme guide ou défi, définit le travail en philosophie de l'esprit est : existe-t-il des trésors mentaux qui ne puissent être achetés avec de la monnaie intentionnelle ? Une réponse négative permet d'entrevoir les grandes lignes d'une large unification de la science. Pour une telle enquête, la sous-classe des systèmes intentionnels qui ont un langage, aptes à communiquer, est d'un intérêt particulier, car ils fournissent le cadre d'une théorie de la conscience. Dans la deuxième partie de *Content and Consciousness* et dans les deuxième et quatrième parties de *Brainstorms*, j'ai essayé d'élaborer une telle théorie ; je voudrais en considérer ici les implications pour l'analyse du concept de croyance. Quelles vérités relatives aux êtres humains qui croient quelque chose découleront-elles de leur seule nature de systèmes intentionnels doués de la capacité de communication ?

Tous les systèmes intentionnels que nous connaissons ne savent pas voler ou nager, tous les systèmes intentionnels ne parlent pas, mais ceux qui le font soulèvent des problèmes et des occasions de réflexion spécifiques, dès lors que nous sommes amenés à leur attribuer des croyances et des désirs. Et, en disant cela, nous sommes bien en deçà de la réalité ; sans les systèmes intentionnels parlants, il n'y aurait, bien sûr, ni attribution de croyance, ni construction de théories, ni hypothèse de rationalité, ni prédiction. La capacité linguistique est sans aucun doute l'accomplissement suprême de l'évolution, un accomplissement qui se nourrit de lui-même pour produire des systèmes rationnels toujours plus adaptables et subtils. On peut cependant considérer cette évolution comme une adaptation soumise aux mêmes conditions d'utilité par rapport à l'environnement que n'importe quel autre talent comportemental. Plusieurs faits remarquables apparaissent lorsqu'on adopte ce point de vue. Un des traits les plus généraux des histoires évolutives est l'interdépendance d'organes

et de capacités distinctes à l'intérieur d'une espèce. Des yeux évolués et autres récepteurs à distance ne servent à un organisme qu'à condition qu'il développe des moyens de locomotion évolués ; des capacités de prédateur ne sauraient profiter à une espèce qui ne développerait pas le système digestif d'un carnivore. Les capacités de croire et de communiquer ont aussi leurs propres conditions préalables. Nous avons déjà vu qu'il n'y a aucun intérêt à attribuer des croyances à un système si les croyances attribuées ne sont pas généralement adaptées à l'environnement et si le système ne réagit pas de façon adéquate à ces croyances. Avec quelque excentricité, on pourrait dire : la capacité de croire n'aurait aucune valeur de survie si elle n'était pas une capacité de croire des vérités. Ce qu'il y a là d'excentrique et de potentiellement trompeur, c'est l'image suggérée d'une espèce qui « essayerait » une faculté engendrant des croyances le plus souvent fausses, et qui, après en avoir éprouvé l'inutilité, l'abandonnerait. Une espèce pourrait « expérimenter » par mutation un grand nombre de systèmes inefficaces mais aucun de ces systèmes ne mériterait le nom de système de croyance, précisément à cause de leurs défauts, de leur irrationalité : c'est pourquoi un système de croyance fausse est une impossibilité conceptuelle. Pour emprunter un exemple à une nouvelle de MacDonald Harris, un poisson soluble est une impossibilité du point de vue de l'évolution, mais un système de croyances fausses ne peut même pas être décrit de manière cohérente. Le même penchant de l'évolution en faveur de la vérité élague la capacité de communiquer au fur et à mesure qu'elle se développe ; une capacité de communication fausse ne serait pas du tout une capacité de communication, mais seulement une tendance à l'émission, sans aucune valeur systématique pour l'espèce. La faculté de communiquer ne trouverait aucune place au cours de l'évolution si elle n'était pas pour l'essentiel la faculté de transmettre des croyances vraies, ce qui signifie seulement la faculté de modifier les autres membres de l'espèce dans le sens d'une structure plus proche de l'optimum.

Une base est ainsi fournie pour l'explication d'un caractère de la croyance dont les philosophes ont eu récemment quelque peine à rendre compte [11]. Le concept de croyance semble avoir

11. Je pense spécialement à l'analyse pénétrante de A. Phillips Griffiths : « On Belief », *Proceedings of the Aristotelian Society*, LXIII (1962-1963), pp. 167-186 ; et au texte de Bernard Mayo « Belief and Constraint », *ibid.*, LXIV (1964), pp. 139-156, tous deux repris dans Phillips Griffiths, ed., *Knowledge and Belief* (New York : Oxford, 1967).

quelque chose de normatif très difficile à saisir. On pourrait l'exprimer ainsi : un aveu tel que « je crois que *p* » paraît impliquer d'une certaine façon : « on devrait croire que *p* ». Ce moyen de présenter les choses n'est pas sans défauts, car nous devons alors rendre compte du fait que « je crois que *p* » semble avoir une force normative que « il croit que *p* », dit à mon sujet, n'a pas. Qui plus est, dire qu'on devrait croire ceci ou cela suggère que la croyance est un acte volontaire, point de vue qui offre des difficultés notoires [12]. Tant qu'on essaye de saisir l'élément normatif en l'exprimant sous la forme d'injonctions morales ou pragmatiques à destination de ceux qui croient, comme : « on devrait croire la vérité » et « on devrait agir en accord avec ce que l'on croit », des dilemmes apparaissent. Comment, par exemple, va-t-on suivre le conseil de croire ce qui est vrai ? Peut-on abandonner l'habitude de croire par négligence des choses fausses ? Si l'on entend ce conseil comme : croyez seulement ce dont vous avez une preuve convaincante, alors ce n'est rien d'autre que ce conseil vide : ne croyez que ce que vous croyez être vrai. Si par ailleurs, on l'entend comme : ne croyez que ce qui est en fait la vérité, alors c'est une injonction à laquelle nous n'avons pas le pouvoir d'obéir.

L'élément normatif de la croyance est chez lui non pas dans de telles injonctions, mais dans les préconditions à l'attribution de la croyance, ce que Phillips Griffiths appelle « les conditions générales de possibilité d'application du concept ». Pour que le concept de croyance puisse trouver une application, nous avons vu que deux conditions doivent être remplies :

(1) En général, normalement, dans la majorité des cas, si *x* croit *p*, *p* est vrai.

(2) En général, normalement, dans la majorité des cas, si *x* avoue que *p*, il croit *p* (et, par (1), *p* est vrai).

Si ces conditions n'étaient pas remplies, il n'y aurait pas de système rationnel communiquant ; nous n'aurions pas de gens qui croient ou qui avouent leur croyance. La norme de la croyance est le bien-fondé dans la preuve (norme qui garantit la vérité à long terme), et la norme de l'aveu de la croyance est la justesse (qui inclut la sincérité). Ces deux normes déterminent des implications pragmatiques de nos énonciations. Si j'affirme que *p* (ou que je crois que *p*, c'est la même chose), j'assume l'obligation

12. Cf., par exemple, H. H. Price, « Belief and Will », *Proceedings of the Aristotelian Society*, suppl. vol. XXVIII (1954), repris dans S. Hampshire, ed., *Philosophy of Mind* (New York : Harper & Row, 1966).

de défendre mon affirmation sur deux fronts : on peut me demander la preuve de la vérité de *p*, et l'on peut me demander la preuve comportementale que je crois effectivement *p* [13]. Je n'ai pas besoin d'examiner mon propre comportement pour pouvoir avouer ma croyance que *p*, mais si l'on met en doute ma sincérité, ou la connaissance que j'ai de moi-même, c'est dans cette direction que je dois me tourner pour défendre mon affirmation. Cependant, encore une fois, les doutes portant sur l'un ou l'autre point doivent être l'exception et non la règle, si l'on veut que la croyance ait une place parmi nos concepts.

Une autre façon d'envisager l'importance de cette prédominance du normal est de considérer le cercle bien connu des implications entre croyances et désirs (ou intentions), cercle qui interdit les définitions comportementales non-intentionnelles des termes intentionnels. Le fait qu'un homme se tienne sous un arbre est un indicateur comportemental de sa croyance qu'il peut, à supposer seulement qu'il désire rester au sec, et, si nous cherchons ensuite la preuve qu'il veut rester au sec, le fait qu'il reste sous l'arbre fera l'affaire, à supposer seulement qu'il croit que l'arbre l'abritera. Lui demandons-nous s'il croit que l'arbre l'abritera, sa réponse positive n'aura valeur de preuve et de confirmation qu'à faire l'hypothèse qu'il désire nous dire la vérité, et ainsi de suite, *ad infinitum*. C'est ce cercle apparemment vicieux qui a rendu Quine hostile à l'intentionnel (et a détruit les efforts de Tolman * en vue d'une définition opérationnelle des termes intentionnels), mais, s'il est vrai que dans n'importe quel cas particulier le fait qu'un homme dise que *p* n'est qu'une preuve conditionnelle de sa croyance, nous pouvons être sûrs qu'à la longue et en général le cercle est brisé. Les affirmations d'un homme sont inconditionnellement révélatrices de ses croyances, tout comme le sont en général ses actions. C'est en reconnaissant qu'en général les croyances et les désirs de quelqu'un doivent être ceux qu'il « devrait avoir » étant données les circonstances que nous contournons le caractère « privé » des croyances et des désirs.

Ces deux normes interdépendantes de la croyance, l'une privi-

13. Cf. A. W. Collins, « Unconscious Belief », *Journal of Philosophy*, LXVI, 20 (oct. 16, 1969), pp. 667-680.

* Cf. Quine : *Le mot et la chose*, ch. VI, « La fuite loin des intentions ». E. C. Tolman : *Purposive Behavior in Animals and Man* (1932), « Operational Behaviorism and Current Trends in Psychology », 1936, repris dans *Behavior and Psychological Man* (1951). [*N. D. T.*]

légiant la vérité et la rationalité de la croyance, l'autre, la justesse de l'aveu, sont normalement complémentaires, mais peuvent, à l'occasion, engendrer des conflits. C'est là le « problème de l'incorrigibilité ». Si la rationalité est la mère de l'intention, nous devons encore sevrer les systèmes intentionnels des critères qui leur donnent vie, et les rendre indépendants. En termes moins imagés, pour pouvoir appliquer le concept de système intentionnel à des cas particuliers, il nous faut, à un moment ou à un autre, cesser de *tester* l'hypothèse de la rationalité du système, adopter le point de vue intentionnel, et admettre sans plus que le système est susceptible d'avoir des croyances et des désirs. Vis-à-vis des animaux muets — et des ordinateurs joueurs d'échecs —, cela se manifeste par une tolérance envers des performances inférieures à l'optimum. Nous continuons à attribuer des croyances à la souris, et à expliquer ses actions en termes de croyance, après l'avoir amenée par ruse à une croyance stupide. Cette tolérance a bien sûr ses limites, et moins le comportement est heureux — surtout, moins il se montre adaptable —, plus nos attributions sont hasardeuses. Ainsi sommes-nous enclins à dire du caneton qui se « fixe » sur la première chose mobile qu'il aperçoit au sortir de sa coquille, qu'il « croit » qu'il s'agit de sa mère, et la suit partout, mais nous soulignons les guillemets entre lesquels est placé « croit ». Pour les systèmes intentionnels qui peuvent communiquer — les personnes, par exemple —, la tolérance prend la forme de la convention que l'homme est incorrigible, qu'il est une autorité spéciale quant à ses propres croyances. Cette convention est « justifiée » par le fait que l'évolution garantit le respect de notre seconde norme. Pourrait-il y avoir une meilleure source sur les croyances d'un système que ses aveux . Toutefois, un conflit apparaît lorsqu'une personne n'atteint pas la rationalité parfaite, et avoue des croyances qui sont ou bien nettement infirmées par les données empiriques disponibles, ou bien contradictoires en elles-mêmes ou par rapport à d'autres aveux de la même personne. Si nous nous appuyons sur le mythe de l'homme comme être parfaitement rationnel, il nous faudra concéder à ces aveux une bien faible autorité : « Vous *ne pouvez pas* vouloir dire — comprendre — ce que vous dites ! » ; si nous nous appuyons sur son « droit » de système intentionnel parlant à ce que sa parole soit acceptée, nous lui accordons un ensemble irrationnel de croyances. Aucune des deux positions ne fournit de point d'arrêt stable ; car, comme nous l'avons vu, l'explication et la prédiction intentionnelle ne peuvent tenir compte ni d'une panne, ni d'un plan inférieur à l'optimum ; c'est pourquoi il

n'y a pas de description intentionnelle cohérente d'une telle impasse [14].

Dans un cas de ce genre, ne pourrait-on invoquer d'autres considérations, pour justifier telle attribution de croyances plutôt que telle autre ? Où rechercher de telles considérations ? Le Phénoménologue sera enclin à supposer que l'introspection individuelle nous livrera des données d'une espèce inaccessible à l'observateur extérieur adoptant le point de vue intentionnel ; mais comment utiliser de telles données ? L'introspecteur peut bien amasser autant d'informations intimes qu'il vous plaira ; il doit ensuite nous les communiquer, mais qu'allons-nous faire de ses communications ? Nous pouvons les supposer incorrigibles (mis à part les erreurs verbales corrigibles, les lapsus, etc.), mais nous n'avons pas besoin de la Phénoménologie pour pouvoir prendre cette option, puisqu'elle revient à s'appuyer sur la norme de justesse de l'aveu aux dépens de la norme de rationalité. Si, d'un autre côté, nous exigeons des énonciations de quelqu'un, certains degrés de consistance et de rationalité avant de leur reconnaître une autorité, quels degrés choisirons-nous ? Exiger une rationalité parfaite, c'est simplement se rabattre sur l'autre norme aux dépens de la norme de justesse de l'aveu. Si nous essayons de fixer des degrés minima inférieurs à la perfection, qu'est-ce qui guidera notre choix ? Ni des données phénoménologiques, car c'est le choix que nous ferons qui déterminera ce que l'on peut considérer comme données phénoménologiques ; ni, non plus, des données neurophysiologiques, car considérer qu'un morceau de structure neuronale est doté d'un contenu de croyance particulier dépend précisément du fait d'avoir admis que le dit système neurologique satisfait aux normes de rationalité d'un système intentionnel, supposition compromise par l'impasse d'où nous essayons de sortir. On pourrait faire la théorie de la neurologie d'un individu, théorie qui permettrait de « lire » ou de prédire les propositions auxquelles il donnerait son assentiment. Mais, quant à savoir si cette théorie a découvert ses

14. Hintikka prend ce taureau par les cornes. On admet que sa logique épistémique n'est valide que pour un être croyant idéalement rationnel ; si nous devions appliquer cette logique à des personnes du monde réel autrement que de façon normative, conférant ainsi aux implications qu'elle permet une *autorité* quant aux croyances réelles, l'autorité des personnes passerait par-dessus bord. Ainsi sa règle A.CBB * (*Knowledge and Belief*, pp. 24-26), qui dit en gros que, si l'on croit *p*, on croit qu'on croit *p*, ne peut être comprise, ainsi qu'on serait tenté de le faire, comme une version de la thèse d'incorrigibilité.

croyances, ou simplement un ensemble d'inducteurs-d'assentiments, cela dépend du degré de consistance, de rationalité, de vérité, que nous attribuons à cet ensemble de propositions.

John Vickers m'a suggéré une manière d'envisager cette question. Considérons un ensemble T de transformations menant de croyances à croyances. Le problème est de déterminer l'ensemble T_s pour chaque système intentionnel S, de sorte que, si nous savons que S croit que p, nous puissions déterminer d'autres croyances de S en voyant quelles sont les transformations de p pour T_s. Si S était idéalement rationnel, toute transformation valide se trouverait dans T_s ; S croirait chaque conséquence logique de chacune de ses croyances (et, idéalement, S n'aurait pas de croyances fausses). Or nous savons qu'aucun système intentionnel ne sera idéalement rationnel ; aussi devons-nous supposer que le T de tout système réel sera moins plein. Mais nous savons aussi que, pour être un système intentionnel, S doit avoir un T relativement complet : T ne peut être vide. Cependant, de quel principe disposons-nous pour fixer un ensemble entre les extrêmes et l'appeler *l'*ensemble nécessaire à la croyance (pour S, pour les habitants de la Terre, ou pour les petites filles de dix ans) ? C'est une autre façon de se demander si l'on peut remplacer la théorie normative de la croyance d'Hintikka par une théorie empirique de la croyance, et, si oui, sur quelles données s'appuyer. « En fait », est-on tenté de dire, « les gens croient occasionnellement à des contradictions, comme le montrent leurs paroles ; aussi toute logique ou analyse conceptuelle adéquate de la croyance doit tenir compte de ce fait ». Mais toute tentative pour *légitimer* dans une théorie de la croyance, la faillibilité humaine, en fixant un seuil de tolérance à l'erreur, ressemblerait à l'ajout d'une règle au jeu d'échecs : une Règle Officielle de Tolérance qui dirait que toute partie d'échecs ne contenant pas plus de k coups illégaux par rapport aux autres règles du jeu est une partie d'échecs légale. Supposez que nous découvrions que, dans telle population importante de mauvais joueurs d'échecs, chaque partie contient en moyenne trois coups illégaux inaperçus par chaque protagoniste. Prétendrions-nous que ces gens jouent réellement à un jeu différent du nôtre, un jeu comprenant une Règle Officielle de Tolérance pour k valant 3 ? Ce serait confondre la norme qu'ils suivent avec ce qui se passe dans leur monde. Dans le même esprit, nous pourrions prétendre que les gens croient *réellement,* par exemple, toutes les conséquences synonymes ou intentionnellement isomorphes de leurs croyances, et non toutes les conséquences logiques. Mais,

bien sûr, les occasions où un homme refuse son assentiment à une conséquence logique d'un de ses aveux sont des cas instables ; il s'expose à une critique et ne peut en appeler pour sa défense à aucun canon qui l'absoudrait de croire des conséquences non-synonymes. Veut-on s'écarter des normes, prédire et expliquer le comportement « réel, empirique », des mauvais joueurs d'échecs, alors il faut cesser de parler de leurs coups aux échecs et parler plutôt de leurs propensions à bouger des pièces de bois ou d'ivoire sur des plateaux à damier ; veut-on prédire et expliquer le comportement « réel, empirique » de ceux qui croient quelque chose, alors il faut, pour s'en rendre compte cesser de parler de croyance et redescendre au point de vue du plan, ou au point de vue physique.

Le concept de système intentionnel exposé dans ces pages est destiné à supporter une lourde charge. On s'en est servi ici pour jeter un pont entre le domaine intentionnel (qui comprend notre monde du « sens commun », celui des personnes et des actions, la Théorie des jeux, et les « signaux neuronaux » du biologiste) et le domaine non-intentionnel des sciences physiques. C'est beaucoup attendre d'un seul concept, mais rien de moins que ce que Brentano lui-même en attendait, quand, en un temps de moindre fragmentation de la science, il proposait l'intentionalité comme la marque qui partage l'univers de la manière la plus fondamentale : séparant le mental du physique.

POST-SCRIPTUM, 1983

Quoique je sois pour l'essentiel satisfait de cette exposition du point de vue intentionnel (la première que j'aie donnée), elle contient néanmoins une suggestion très trompeuse et que j'ai voulu par la suite écarter. En soulignant, comme je l'ai fait, le côté optionnel de l'adoption du point de vue intentionnel, ainsi que les raisons pragmatiques de s'y rallier, j'ai donné l'impression que la question de savoir si quelque chose est un système intentionnel ou non (et donc, finalement si quelque chose a des croyances et des désirs) n'est pas une question de fait mais de politique à suivre, au gré des différents interprètes.

Il n'en est évidemment rien. Une chose est — ou n'est pas — un système intentionnel, qu'on choisisse ou non de la traiter comme un système intentionnel. Mais ce fait, qui est donc indépendant de toute interprétation effective, n'est *caractérisable* que

dans les termes du succès (ou de l'échec) qu'une telle tentative d'interprétation du point de vue intentionnel *rencontrerait.* De la même façon, le fait, que tel morceau de quincaillerie (*hardware*) (assemblé pour une raison quelconque, ou bien sans aucune raison) réalise quelque machine de Turing est indépendant de la tentative de l'interpréter *réellement* comme cette machine de Turing, mais la seule façon de dire ce que c'est que d'être une telle réalisation est celle-ci : *si l'on devait interpréter cet objet comme machine de Turing, cette interprétation aurait un très solide pouvoir de prédiction.*

Durant la douzaine d'années qui se sont écoulées depuis la publication de « Systèmes intentionnels » j'ai clarifié et développé les idées de base, tout d'abord dans mon livre *Brainstorms* (Bradford Book/MIT Press et Harvester, 1978), et plus récemment dans une série d'articles :

— « True Believers : the Intentional Strategy and Why it Works » in A. Heath, ed., *Scientific Explanation* (the 1979 Herbert Spencer Lectures at Oxford), Oxford University Press, 1981.

— « Three Kinds of Intentional Psychology » in R. Healey, ed., *Reductionism, Time and Reality,* Cambridge University Press, 1981.

— « Making Sense of Ourselves » (réponse à S. Stich, « Dennett on Intentional Systems »), tous deux dans *Philosophical Topics,* vol. XII, 1981 (le numéro de cette revue est publié sous le titre *Mind, Brain and Function,* J. I. Biro and R. Shahan, eds., Oklahoma University Press, et Harvester Press, 1982).

— « Intentional Systems in Cognitive Ethology : the "Panglossian Paradigm" Defended » (avec des commentaires et une réponse) dans *Behavioral and Brain Sciences,* août 1983.

Nous remercions le Journal of Philosophy *et l'auteur de nous avoir autorisés à traduire cet article.*