ou fait de « données sensorielles réelles ou possibles » (le vieux point de vue empiriste); ou bien elles nous conduisent à nier l'existence d'un monde au sens propre, qui serait autre chose qu'un tas d'histoires que nous inventons pour diverses raisons (avant tout inconscientes). Dans ce livre, mon but est d'esquisser les idées maîtresses d'une conception qui ne soit pas aliénante.

Ce texte, qui est issu de travaux en cours, recoupe en partie la conférence Herbert Spencer que j'ai prononcée à Oxford en 1979 et aussi mon article « Si Dieu est mort, tout est permis... (Réflexions sur la philosophie du langage) », paru dans *Critique*, n° 399-400, 1980.

Une bourse de la National Science Foundation m'a permis d'entreprendre les recherches pour ce livre de 1978 à 1980. Je tiens à remercier ici cet organisme pour son soutien.

Thomas Kuhn et Ruth Anna Putnam ont lu différentes versions de ce livre et m'ont apporté des critiques pertinentes et de sages conseils. J'ai aussi bénéficié des conseils et des critiques de nombreux amis, parmi lesquels Ned Block, David Helman et Justin Leiber, ainsi que de mes différents groupes d'étudiants à Harvard. Plusieurs chapitres ont été présentés sous forme de conférences à Lima au printemps 1980 (ce voyage fut rendu possible par une bourse de la Commission Fulbright) et le chapitre II y fut entièrement rédigé. Pendant cette période, j'ai bénéficié de discussions avec Leopoldo Chiappo, Alberto Cordero Lecca, Henrique Fernandez, Francisco Miro Quesada et Jorge Secada. J'ai présenté une version complète du livre (sous une forme différente) à l'occasion de conférences données à l'Université de Francfort pendant l'été 1980 et j'ai bénéficié des critiques stimulantes et des encouragements de mes collègues allemands (plus particulièrement Wilhelm Essler et Rainer Trapp), de mes étudiants et de mes autres amis allemands (tout spécialement Dieter Heinrich, Manon Fassbinder et Wolfgang Stegmüller).

Tous mes collègues du Département de Philosophie de Harvard mériteraient d'être remerciés individuellement. Au cours de ces dernières années, Nelson Goodman et moi-même avons détecté une convergence d'opinions et bien que la première version de ce livre ait été rédigée avant que j'aie eu l'occasion de voir son livre *Ways of Worldmaking*, la lecture de son ouvrage et nos discussions sur ces questions m'ont souvent été précieuses.

Enfin, que Jeremy Mynott soit remercié pour ses encouragements et ses conseils en tant que directeur de collection.

CHAPITRE PREMIER

DES CERVEAUX DANS UNE CUVE

Une fourmi marche sur le sable. En se déplaçant, elle dessine une ligne. Elle trace des courbes, revient en arrière, de sorte que par un pur hasard son parcours finit par ressembler nettement à une caricature de Winston Churchill. La fourmi a-t-elle dessiné un portrait de Winston Churchill, une image qui *dépeint* Winston Churchill ?

La plupart des gens, après un instant de réflexion, diraient que non. La fourmi, après tout, n'a jamais vu Churchill, ni même un portrait de Churchill, et elle n'avait nullement l'intention de dépeindre Churchill. La fourmi a simplement tracé une ligne (ce qui était *déjà* involontaire), une ligne que *nous* nous pouvons « voir » comme un portrait de Churchill.

On peut exprimer ceci en disant que la ligne n'est pas « en soi » une représentation[1] d'une chose plutôt que d'une autre. Le fait qu'une chose ressemble aux traits de Winston Churchill n'est pas suffisant pour que cette même chose représente ou désigne Churchill. Et ce n'est pas non plus nécessaire : dans notre culture, la forme imprimée « Winston Churchill », les mots énoncés « Winston Churchill », et bien d'autres choses encore sont utilisées pour représenter Churchill (mais d'une façon non-picturale), tout en ne ressemblant en rien aux traits de Winston Churchill, au sens où une image — même une esquisse — peut lui ressembler. Si la *ressemblance* n'est ni nécessaire, ni suffisante pour que quelque chose représente quelque chose d'autre, comment pourrait-il y avoir quoi que ce soit de

1. Dans ce livre, les termes « représentation » et « référence » désigneront toujours une relation entre un mot (ou tout autre type de symbole, signe, ou présentation) et quelque chose qui existe réellement (c'est-à-dire qui n'est pas qu'un « objet de la pensée »). Il existe un sens de « faire référence » selon lequel on peut faire référence à ce qui n'existe pas. Ce n'est pas ce sens que nous utiliserons ici. Un terme plus ancien pour ce que j'appelle la « représentation » ou la « référence » est la « dénotation ». Enfin, je suivrai l'usage des logiciens modernes et j'utiliserai « existe » comme voulant dire « existe dans le passé, le présent, ou le futur ». Ainsi Winston Churchill « existe », et nous pouvons faire référence à lui ou nous le représenter, même s'il n'est plus en vie.

nécessaire ou de suffisant à cet effet ? Comment diable une chose peut-elle représenter (ou « remplacer », etc.) une chose différente ?

La réponse peut sembler évidente. Supposons que la fourmi ait déjà vu Winston Churchill, et supposons qu'elle ait l'intelligence et l'habileté nécessaires pour en faire le portrait. Supposons qu'elle ait dessiné la caricature *intentionnellement*. Alors, le tracé représenterait bien Churchill.

Supposons maintenant que la ligne ait la forme WINSTON CHURCHILL. Et supposons que ce soit accidentel (ignorons le fait que c'est peu plausible). Alors la « forme écrite » WINSTON CHURCHILL *ne représenterait pas* Churchill, même si cette forme écrite désigne effectivement Churchill quand elle apparaît aujourd'hui dans n'importe quel livre.

On pourrait donc penser que ce qui est nécessaire à la représentation, ou ce qui y est avant tout nécessaire, c'est *l'intention*.

Mais si j'ai l'intention de *représenter* Winston Churchill par quelque chose d'autre, même dans mon langage privé (même si je me dis les mots « Winston Churchill », sans les prononcer), il faut que j'aie déjà pu *penser* à Churchill. Si des lignes sur du sable, des bruits, etc., ne peuvent « en soi » représenter quelque chose, alors comment se fait-il que des formes pensées le peuvent ? Mais le peuvent-elles vraiment ? Comment la pensée peut-elle tendre la main et « saisir » ce qui est à l'extérieur ?

Dans le passé, certains philosophes ont hâtivement conclu à partir de ce type de considérations qu'ils tenaient une démonstration de la nature essentiellement non-physique de l'esprit. L'argument est simple : ce que nous avons dit du tracé de la fourmi s'applique à tous les objets physiques. Aucun objet physique ne peut, en soi, désigner une chose plutôt qu'une autre ; pourtant, les *pensées* réussissent de manière évidente à désigner une chose plutôt qu'une autre. Donc, la nature des pensées — et par là-même celle de l'esprit — est essentiellement différente de celle des objets physiques. Les pensées sont caractéristiquement « intentionnelles » — elles peuvent désigner une autre chose ; aucun objet physique ne possède d'intentionnalité, sauf lorsque celle-ci dérive de l'utilisation de cet objet physique par un esprit. C'est du moins ce que l'on prétend et c'est un peu rapide : on ne résout rien en postulant des pouvoirs mystérieux de l'esprit. Mais le problème est bien réel. Comment l'intentionnalité, la référence, sont-elles possibles ?

Les théories magiques de la référence.

Nous venons de voir que le dessin de la fourmi n'a aucun rapport nécessaire avec Winston Churchill. Le seul fait de « ressembler » à Churchill ne fait pas du « dessin » un vrai dessin, ni une représentation de Churchill. A moins que la fourmi ne soit intelligente (ce qu'elle n'est pas), ou qu'elle sache qui est Churchill (ce qu'elle ne sait pas), son tracé n'est ni un portrait ni une représentation de rien. Certains peuples primitifs croient que certaines représentations (en particulier les *noms*) ont un rapport nécessaire avec leurs propriétaires ; ils croient que le fait de connaître le « nom véritable » de quelque chose ou de quelqu'un donne un pouvoir sur lui. Ce pouvoir proviendrait de la *relation magique* qui est censée exister entre le nom et le porteur du nom ; dès que l'on réalise qu'un nom n'a qu'un rapport contextuel, contingent et conventionnel avec son propriétaire, on voit mal pourquoi la connaissance du nom devrait avoir une portée magique.

Ce qu'il faut comprendre, c'est que ce qui vaut pour les portraits physiques vaut aussi pour les images mentales et pour les représentations en général ; les représentations mentales n'ont pas plus de rapports nécessaires avec ce qu'elles représentent que les représentations physiques. L'hypothèse contraire est une survivance de la pensée magique.

Cette remarque est peut-être particulièrement évidente dans le cas des *images* mentales. (Wittgenstein fut peut-être le premier philosophe à comprendre l'immense portée de cette remarque, même s'il ne fut pas le premier à la faire). Supposons qu'il existe quelque part une planète sur laquelle des êtres humains ont évolué (ou ont été déposés par des extraterrestres, ou n'importe quoi d'autre). Supposons que ces êtres humains, qui sont par ailleurs semblables à nous, n'ont jamais vu d'*arbres*. Supposons qu'ils n'ont jamais imaginé d'arbres (les seules formes de vie végétale sur leur planète sont peut-être des champignons). Supposons qu'un jour un dessin d'un arbre est accidentellement déposé sur leur planète par un vaisseau spatial qui n'entre pas en contact avec eux. Imaginons-les en train de s'interroger sur le dessin. Qu'est-ce que ça peut bien être ? Toutes sortes d'idées leur viennent à l'esprit : un bâtiment ? Un baldaquin ? Un animal d'une espèce inconnue ? Mais supposons qu'il ne se doutent jamais de la vérité.

Pour *nous*, le dessin est une représentation d'un arbre. Pour ces humains, le dessin ne représente qu'un objet étrange, nature

et fonction inconnues. Supposons que l'un d'eux, parce qu'il a regardé le dessin, entretienne une image mentale identique à une de mes images mentales d'un arbre. Son image mentale n'est pas une *représentation d'un arbre.* Ce n'est qu'une représentation de l'étrange objet (quel qu'il soit) que représente le mystérieux dessin.

Pourtant, quelqu'un pourrait encore soutenir que l'image mentale est *en fait* une représentation d'un arbre, ne serait-ce que parce que le dessin qui a produit cette image mentale était lui-même, pour commencer, une représentation d'un arbre. Il existe une chaîne causale qui va des arbres réels à l'image mentale, même si c'est une chaîne plutôt bizarre.

Mais on peut même supposer que cette chaîne causale est absente. Supposons que le « dessin d'un arbre » que le vaisseau spatial a déposé ne soit pas en fait un dessin d'un arbre, mais des éclaboussures accidentelles de peinture. Même s'il est en tous points pareil au dessin d'un arbre, ce n'est pas plus un dessin d'un arbre que la « caricature » de la fourmi n'était une caricature de Churchill. On pourrait même supposer que le vaisseau spatial qui a déposé le dessin venait d'une planète sur laquelle il n'y avait pas d'arbres. Ces humains auraient toujours des images mentales qualitativement identiques à mon image d'un arbre, mais il ne s'agirait pas pour autant d'images représentant un arbre plutôt qu'autre chose.

La même remarque s'applique aux *mots.* Un texte sur du papier peut sembler être une description parfaite d'arbres, mais s'il a été produit par des singes tapant au hasard sur des machines à écrire pendant des millions d'années, alors ces mots ne désignent rien. Si une personne mémorisait ces mots et les récitait sans les comprendre, alors, même récités pour soi, ils ne désigneraient rien non plus.

Supposons que la personne qui se récite ces mots ait été hypnotisée. Supposons qu'il s'agisse de mots japonais, et que l'on ait suggéré à la personne qu'elle comprend le japonais. Supposons qu'en se récitant ces mots la personne ait l'impression de « comprendre ». (Bien que si on l'interrompait et si on lui demandait ce que ces mots étaient censés signifier, elle découvrirait ne pas le savoir). L'illusion pourrait être à ce point parfaite que même un télépathe japonais y croirait ! Mais s'il est impossible à la personne d'utiliser les mots dans leur contexte approprié, de répondre à des questions sur ce qu'elle « pense », etc., alors elle ne les comprend pas.

En combinant ces différentes histoires de science-fiction, on

peut inventer un cas où une personne penserait des mots qui seraient en fait une description d'arbres dans une langue donnée, aurait simultanément les images mentales appropriées, mais ne comprendrait pas le sens des mots et ne saurait pas ce qu'est un arbre. On pourrait même supposer que les images mentales ont été produites par des taches de peinture (la personne hypnotisée croirait que ce sont des images de quelque chose d'adéquat à sa pensée, mais si on lui demandait de s'expliquer, elle ne le pourrait pas). Et on pourrait imaginer que le langage dans lequel pense l'individu n'est connu ni de l'hypnotiseur, ni de la personne hypnotisée — c'est peut-être par coïncidence que ces « phrases dénuées de sens », ces phrases que l'hypnotiseur suppose être « dénuées de sens », sont en fait une description d'arbres en japonais. En bref, tout ce qui passerait par l'esprit de cette personne pourrait être qualitativement identique à ce qui passe par l'esprit d'un locuteur du japonais qui pense *réellement* à des arbres — mais rien de tout cela ne désignerait des arbres.

Bien sûr, tout ceci est en fait impossible, tout comme il est impossible que des singes tapent au hasard une copie de *Hamlet.* C'est-à-dire que les probabilités *a contrario* sont si élevées que cela ne se produira jamais (pensons-nous). Mais ce n'est ni logiquement, ni même physiquement impossible. Cela *pourrait* se produire sans violation des lois physiques et cela est peut-être même compatible avec l'état réel de l'univers, s'il existe beaucoup d'êtres intelligents sur d'autres planètes. Et si cela venait à se produire, on aurait là une démonstration frappante d'une importante vérité conceptuelle : même un système de représentations vaste et complexe, verbal et visuel, n'a pas de rapport *intrinsèque,* incorporé, magique, avec ce qu'il représente — un rapport qui serait indépendant de la façon dont il a été produit, et de ce que sont les dispositions du locuteur ou du penseur. Et ceci reste vrai, que le système de représentation (les mots et les images de notre exemple) soit physiquement réalisé — les mots étant écrits ou prononcés, les images étant des images physiques —, ou qu'il soit seulement réalisé dans la pensée. Les images mentales et les mots pensés ne représentent pas *intrinsèquement* ce dont ils sont la représentation.

Le cas des cerveaux dans une cuve.

Voici une histoire de science-fiction discutée par des philosophes : supposons qu'un être humain (vous pouvez supposer qu'il

s'agit de vous-même) a été soumis à une opération par un savant fou. Le cerveau de la personne en question (votre cerveau) a été séparé de son corps et placé dans une cuve contenant une solution nutritive qui le maintient en vie. Les terminaisons nerveuses ont été reliées à un super-ordinateur scientifique qui procure à la personne-cerveau l'illusion que tout est normal. Il semble y avoir des gens, des objets, un ciel, etc. Mais en fait tout ce que la personne (vous-même) perçoit est le résultat d'impulsions électroniques que l'ordinateur envoie aux terminaisons nerveuses. L'ordinateur est si intelligent que si la personne essaye de lever la main, l'ordinateur lui fait « voir » et « sentir » qu'elle lève la main. En plus, en modifiant le programme le savant fou peut faire « percevoir » (halluciner) par la victime toutes les situations qu'il désire. Il peut aussi effacer le souvenir de l'opération, de sorte que la victime aura l'impression de se trouver dans sa situation normale. La victime pourrait justement avoir l'impression d'être assise en train de lire ce paragraphe qui raconte l'histoire amusante mais plutôt absurde d'un savant fou qui sépare les cerveaux des corps et qui les place dans une cuve contenant des éléments nutritifs qui les gardent en vie. Les terminaisons nerveuses sont censées être reliées à un ordinateur scientifique super-puissant qui donne à la personne-cerveau l'illusion que...

Lorsque l'on évoque ce type de possibilité dans un cours sur la théorie de la connaissance, l'idée, bien sûr, est de soulever en des termes modernes le problème classique du scepticisme vis-à-vis du monde extérieur. (*Comment savez-vous que vous ne vous trouvez pas dans cette situation?*). Mais cette histoire fournit aussi un moyen pratique de poser des questions sur les rapports entre l'esprit et le monde.

Au lieu de ne prendre qu'un cerveau dans une cuve, nous pouvons supposer que tous les êtres humains, peut-être tous les êtres pensants, sont des cerveaux dans une cuve (ou des systèmes nerveux dans une cuve, s'il s'avère que certains êtres au système nerveux minimal sont néanmoins des « êtres pensants »). Évidemment, le savant fou devrait se trouver à l'extérieur — mais, au fait, est-ce nécessaire? Il n'y a peut-être pas de savant fou. C'est certainement absurde, mais peut-être l'univers n'est-il qu'une machine automatique qui s'occupe d'une cuve remplie de cerveaux et de systèmes nerveux.

Supposons à présent que la machine automatique soit programmée pour nous faire ressentir des hallucinations *collectives* plutôt que des hallucinations individuelles sans rapport entre elles. Ainsi, lorsque j'ai l'impression de vous parler, vous avez l'impression d'entendre mes paroles. Bien sûr, mes paroles n'atteignent pas réellement vos oreilles — parce que vous n'avez pas d'oreilles, et que je n'ai pas de bouche ou de langue. En fait, ce qui se passe lorsque je prononce des phrases, c'est que les impulsions efférentes vont de mon cerveau vers l'ordinateur et celui-ci fait que j' « entends » ma propre voix et que je « sens » ma langue bouger, etc., et il fait que vous « entendez » ma voix et que vous me « voyez » parler. Dans ce cas, on peut dire qu'en un sens nous communiquons effectivement. Je ne me trompe pas sur votre existence réelle; je me trompe seulement sur l'existence de votre corps et du « monde extérieur », à l'exclusion des cerveaux. D'une certaine manière, peu importe que le monde entier ne soit qu'une hallucination collective; après tout, vous m'entendez bel et bien parler quand je vous parle, même si le mécanisme n'est pas celui que nous croyons. (Mais dans le cas de deux amants en train de faire l'amour, l'idée qu'ils ne sont que deux cerveaux dans une cuve pourrait être inquiétante.)

Je vais maintenant poser une question qui semblera plutôt idiote et évidente (du moins à certaines personnes, y compris des philosophes sophistiqués), mais qui nous conduira très rapidement à des problèmes philosophiques profonds. Supposons qu'en fait cette histoire soit vraie. Pourrions-nous, si nous étions des cerveaux dans une cuve, *dire* ou *penser* que nous sommes des cerveaux dans une cuve?

Je vais tenter de montrer que la réponse est « non, nous ne le pourrions pas ». En fait, je vais essayer de montrer que l'hypothèse que nous sommes réellement des cerveaux dans une cuve, bien qu'elle ne viole aucune loi physique, et bien qu'elle soit parfaitement cohérente avec toute notre expérience, ne peut en aucun cas être vraie. *Il est impossible qu'elle soit vraie* parce qu'en un sens elle est auto-réfutante.

L'argument que je vais exposer est d'un type inhabituel, et il m'a fallu plusieurs années pour me convaincre qu'il est juste. Mais il l'est vraiment. S'il semble à ce point bizarre, c'est parce qu'il est relié à certains des plus profonds problèmes philosophiques. (J'ai pensé pour la première fois à cet argument en réfléchissant à un théorème de la logique moderne, le « théorème de Skolem-Löwenheim »; soudain, j'ai vu un rapport entre ce théorème et certains arguments de Wittgenstein dans les *Philosophical Investigations.*)

Une « hypothèse auto-réfutante » est une hypothèse dont la vérité implique sa propre fausseté. Par exemple, prenons la thèse selon laquelle *tous les énoncés généraux sont faux*. C'est un énoncé général. Donc, s'il est vrai, il doit être faux. Donc, il est faux. Parfois une thèse est dite « auto-réfutante » si c'est *la simple supposition que la thèse est prise en considération ou énoncée* qui implique sa fausseté. Par exemple, « je n'existe pas » est auto-réfutante si c'est moi qui le pense (quel que soit le « moi »). Si on pense qu'on existe, on peut donc en être certain, comme l'a montré Descartes.

Ce que je vais montrer, c'est que l'hypothèse que nous sommes des cerveaux dans une cuve possède précisément cette propriété. Si nous pouvons considérer la question de sa vérité ou de sa fausseté, alors elle n'est pas vraie. Donc, elle n'est pas vraie.

Avant de développer cet argument, demandons-nous pourquoi le fait qu'on puisse formuler un tel argument semble si bizarre (surtout aux philosophes qui défendent une conception de la vérité en tant que « copie »). Nous avons admis que l'existence d'un monde où tous les êtres humains sont des cerveaux dans une cuve est compatible avec les lois physiques. Comme disent les philosophes, il existe un « monde possible » où tous les êtres pensants sont des cerveaux dans une cuve. (Cette façon de parler de « mondes possibles » donne l'impression qu'il existe un *endroit* où toute supposition absurde est vraie, et c'est pourquoi elle peut être très trompeuse en philosophie). Dans ce monde possible, les humains ont exactement les mêmes expériences que nous. Ils ont les mêmes pensées que nous (ou du moins ce sont les mêmes mots, images, formes de pensée, etc., qui leur passent par la tête). Pourtant, je prétends qu'il existe un argument par lequel on peut montrer que nous ne sommes pas des cerveaux dans une cuve. Comment est-ce possible ? Et pourquoi des gens qui sont vraiment, dans ce monde possible, des cerveaux dans une cuve, ne pourraient-ils pas utiliser cet argument ?

La réponse sera essentiellement la suivante : même si les gens dans ce monde possible peuvent penser et « dire » tout ce que *nous* nous pouvons penser et dire, je prétends qu'ils ne peuvent pas *faire référence* à ce à quoi *nous* nous pouvons faire référence. Plus précisément, ils ne peuvent pas penser ou dire qu'ils sont des cerveaux dans une cuve (*même en pensant la phrase « nous sommes des cerveaux dans une cuve »*).

Le test de Turing.

Supposons que quelqu'un parvienne à inventer un ordinateur qui puisse réellement avoir une conversation intelligente sur la même variété de sujets qu'une personne intelligente. Comment pourrait-on décider si l'ordinateur est « conscient » ?

Le logicien britannique Alan Turing a proposé le test suivant[2] : supposons que quelqu'un ait une conversation avec l'ordinateur et une autre conversation avec une personne qu'il ne connaît pas. S'il ne parvient pas à dire qui est l'ordinateur et qui est l'être humain (et si on a répété le test un nombre suffisant de fois avec des interlocuteurs différents), alors la machine est consciente. En bref, une machine est consciente si elle réussit le « test de Turing ». (Évidemment, les conversations ne peuvent avoir lieu en face à face, puisque l'interlocuteur ne doit pas connaître l'apparence visuelle de ses deux partenaires dans la conversation. On ne pourra pas non plus se servir de la voix, puisque une voix mécanique est différente d'une voix humaine. Disons plutôt que les conversations ont lieu par l'intermédiaire d'un clavier électronique. L'interlocuteur tape ses assertions, ses questions, etc., et ses partenaires, la machine et l'humain, lui répondent de la même manière. La machine doit aussi pouvoir mentir — si on lui demande « Êtes-vous une machine ? », elle doit pouvoir répondre « Non, je suis un chercheur de ce laboratoire »).

L'idée que ce test est un test concluant de la conscience a été critiquée par de nombreux auteurs, dont certains ne sont aucunement hostiles en principe à l'idée qu'une machine puisse être consciente. Mais ce n'est pas ce qui nous intéresse ici pour le moment. Je voudrais utiliser l'idée générale d'un *test dialogique de compétence* dans un but différent, pour explorer la notion de *référence*.

Imaginez une situation où le but ne serait pas de déterminer si le partenaire est vraiment une personne ou une machine, mais plutôt de savoir si le partenaire utilise les mots pour faire référence comme nous. Le test évident est encore une fois d'avoir une conversation, et s'il n'y a pas de problèmes, si le partenaire « réussit » le test, c'est-à-dire si on ne peut pas le distinguer d'un locuteur reconnu de notre langue qui fait référence à nos objets habituels et ainsi de suite, alors on conclura

2. A. M. Turing, « Computing Machinery and Intelligence », *Mind* (1950), reproduit dans A. R. Anderson (éd.), *Minds and Machines*.

que notre partenaire fait référence aux objets comme nous. Quand le but du test de Turing est celui que nous venons de décrire, c'est-à-dire de déterminer l'existence d'une référence partagée, je dirai qu'il s'agit d'un *test de Turing pour la référence.* Et, de même que les philosophes ont débattu la question de savoir si le test de Turing original est un test *concluant* de la conscience, c'est-à-dire la question de savoir si la machine qui « réussit » le test non pas une fois mais systématiquement est nécessairement consciente, je voudrais me demander si le test de Turing pour la référence que je viens de proposer est un test concluant de la référence partagée.

En fait, la réponse sera « non ». Le test de Turing pour la référence n'est pas concluant. En pratique c'est certainement un test excellent, mais il n'est pas logiquement impossible (bien que certainement improbable) que quelqu'un puisse réussir le test de Turing pour la référence sans jamais avoir fait référence à rien. Il s'ensuit, comme nous le verrons, que nous pouvons généraliser notre observation selon laquelle les mots (et aussi les textes et les discours) n'ont pas de rapports nécessaires avec leurs référents. Même si l'on considère non pas les mots en tant que tels, mais les règles qui déterminent quels mots peuvent être adéquatement produits dans tel ou tel contexte — même si l'on considère, pour employer le jargon de l'informatique, *les programmes d'utilisation des mots* — à moins que ces programmes eux-mêmes *ne désignent quelque chose d'extra-linguistique,* les mots ne possèderont toujours pas de référence. Ceci sera une étape cruciale pour parvenir à la conclusion que les cérébellocuviens ne peuvent pas faire référence à quelque chose d'extérieur (et ne peuvent donc pas dire qu'ils sont des cérébellocuviens).

Supposons, par exemple, que je me trouve dans la situation de Turing (dans les termes de Turing, je joue « le jeu de l'imitation ») et que mon partenaire soit en fait une machine. Supposons que cette machine soit capable de gagner le jeu (de « réussir » le test). Imaginons que la machine soit programmée pour répondre en un anglais châtié à des questions, des affirmations, des remarques et ainsi de suite, mais qu'elle ne possède pas d'organes sensoriels, ni d'organes moteurs, excepté le branchement pour le clavier électronique et le clavier lui-même. (Pour autant que je sache, Turing ne suppose pas que la possession d'organes sensoriels ou moteurs soit nécessaire à la conscience ou à l'intelligence.) Supposons non seulement que la machine n'a pas d'yeux ou d'oreilles électroniques, mais que

le programme de la machine, le programme pour jouer le jeu de l'imitation, n'est pas conçu pour intégrer des inputs sensoriels, ou pour contrôler un corps. Que faudrait-il dire d'une telle machine ?

Il me semble évident que nous ne pouvons pas et que nous ne devons pas attribuer la référence à cet engin. Il est vrai que la machine peut discourir agréablement sur le paysage de la Nouvelle-Angleterre, par exemple. Mais si elle s'y trouvait confrontée elle ne saurait pas reconnaître une pomme d'un pommier, une montagne d'une vache ou un champ d'une haie.

Ce que nous avons, c'est un appareil qui produit des phrases en réponse à des phrases. Mais aucune de ces phrases n'est reliée au monde réel. *Si l'on branchait ensemble deux de ces machines, et si elles jouaient entre elles au jeu de l'imitation, elles se berneraient mutuellement pour toujours, même si le reste du monde cessait d'exister*! Il n'y a pas plus de raisons de considérer que les propos de la machine sur les pommes font référence à de vraies pommes que de considérer que le « dessin » de la fourmi fait référence à Winston Churchill.

Ce qui produit l'illusion de la référence, du sens, de l'intelligence et ainsi de suite, c'est qu'il existe une convention de représentation que *nous* nous suivons, en vertu de laquelle le discours de la machine fait référence aux arbres, aux haies, à la Nouvelle-Angleterre et autres. Pour la même raison, nous avons l'impression que la fourmi a caricaturé Winston Churchill. Nos propos sur les pommes et les champs sont intimement liés à nos transactions *non verbales* avec les pommes et les champs. Il existe des « règles d'entrée dans le langage » qui nous mènent de notre expérience des pommes à des énoncés comme « Je vois une pomme », et des « règles de sortie du langage » qui nous mènent des décisions exprimées par des formes linguistiques (« je vais acheter des pommes »), à des actions non-linguistiques. Puisque la machine n'a pas de règles d'entrée ou de sortie, il n'y a aucune raison de considérer ses propos (ou les propos de deux machines qui joueraient entre elles le jeu de l'imitation) comme étant autre chose qu'un jeu syntaxique. C'est un jeu syntaxique qui *ressemble* sans aucun doute à un discours intelligent, mais qui y ressemble comme le dessin de la fourmi ressemble à une caricature grinçante.

Dans le cas de la fourmi, on peut soutenir que celle-ci aurait dessiné la même chose même si Winston Churchill n'avait pas existé. Dans le cas de la machine, on ne peut pas vraiment défendre cette position ; si les pommes, les arbres, les haies et

les champs n'avaient pas existé, alors les programmeurs n'auraient vraisemblablement pas écrit le même programme. Bien que la machine ne puisse pas *percevoir* les pommes, les champs, ou les haies, ses créateurs-constructeurs, eux, le peuvent. Il existe une certaine relation causale entre la machine et les pommes réelles, par le biais de l'expérience perceptuelle de ses créateurs-constructeurs. Mais une relation si ténue peut difficilement suffire à la référence. Non seulement il est logiquement possible, bien que très peu probable, que la même machine ait pu exister même si les pommes, les champs et les haies n'avaient pas existé, mais en plus la machine est complètement insensible à *l'existence ininterrompue* de pommes, de champs et d'autres haies. Même si toutes ces choses *cessaient* d'exister, la machine continuerait à discourir allègrement de la même manière. Voilà pourquoi on peut dire que la machine ne fait référence à rien.

L'important pour notre discussion, c'est que rien dans le test de Turing n'exclut que la machine ne soit programmée que pour *jouer* le jeu de l'imitation, et une machine qui ne peut que *jouer* le jeu de l'imitation ne peut *visiblement pas* faire référence, pas plus qu'un électrophone.

(*Encore*) *des cerveaux dans une cuve.*

Comparons nos hypothétiques « cerveaux dans une cuve » aux machines que nous venons de décrire. Il y a de toute évidence des différences importantes. Les cerveaux n'ont pas d'organes sensoriels, mais ils sont *faits* pour en avoir ; ils ont des terminaisons nerveuses afférentes, ils ont des inputs pour ces terminaisons nerveuses, et ces inputs figurent dans le « programme » des cerveaux dans la cuve, tout comme ils figurent dans le programme de nos propres cerveaux. Les cerveaux dans la cuve sont des *cerveaux* ; en plus, ce sont des cerveaux *en état de marche*, et ils fonctionnent selon les mêmes règles que les cerveaux dans le monde réel. Pour ces raisons, il serait absurde de leur nier la conscience ou l'intelligence. Mais le fait qu'ils soient conscients et intelligents ne veut pas dire que les mots qu'ils emploient désignent ce que nos mots à nous désignent. La question qui nous intéresse est la suivante : est-ce que leurs verbalisations qui contiennent le mot « arbre », par exemple, font effectivement référence aux *arbres* ? De manière générale, peuvent-ils faire référence à des objets *extérieurs*, par opposition aux objets dans l'image que produit l'ordinateur, par exemple ?

Pour simplifier les choses, supposons que l'ordinateur soit le fruit d'une sorte de coïncidence cosmique (ou qu'il existe peut-être depuis toujours). Dans ce monde hypothétique, la machine est censée ne pas avoir de créateur-constructeur. En fait, comme nous l'avons dit au début du chapitre, on peut supposer que tous les êtres pensants (aussi minimale que soit leur pensée) se trouvent dans une cuve.

Cette hypothèse n'est d'aucun secours. Car il n'y a aucun lien entre le *mot* « arbre » tel que l'utilisent ces cerveaux, et les arbres réels. Ils utiliseraient le mot « arbre » de la même manière, ils penseraient de la même manière, ils auraient les mêmes images, même si les arbres n'existaient pas. Leurs images, leurs mots et ainsi de suite, sont qualitativement identiques aux images, aux mots et ainsi de suite qui représentent effectivement les arbres dans notre monde à nous ; mais nous avons déjà vu — encore la fourmi ! — que la ressemblance qualitative d'une chose avec l'objet qu'elle représente, que ce soit Winston Churchill ou un arbre, ne suffit pas en soi à en faire une représentation. En bref, quand un cerveau dans la cuve pense « il y a un arbre devant moi », il ne pense pas aux arbres réels, parce qu'il n'y a rien en vertu de quoi sa pensée d' « arbre » représenterait des arbres réels.

Si ceci vous semble un peu rapide, songez à la chose suivante : nous avons vu que les mots ne désignent pas nécessairement des arbres, même disposés en une séquence identique à un discours qui, dans notre tête, *serait* sans aucun doute *à propos d'arbres* dans le monde réel. Et le « programme » des cerveaux, c'est-à-dire leurs règles, leurs pratiques et leurs dispositions au comportement linguistique, ne fait pas nécessairement référence aux arbres, et ne conduit pas à la référence par le simple jeu des liens qu'il établit entre des mots ou des indices *linguistiques* et des réponses *linguistiques*. Si les cerveaux pensent à des arbres, s'ils y font référence et s'ils se les représentent (des arbres réels, à l'extérieur de la cuve), alors ce doit être à cause de la manière dont leur « programme » relie le système du langage aux inputs et aux outputs *non verbaux*. De tels inputs et outputs non verbaux existent effectivement dans le monde cérébello-cuvien (ce sont encore une fois les terminaisons neveuses afférentes et efférentes !) mais nous avons vu que les « données sensorielles » produites par l'ordinateur ne représentent pas des arbres (ni quoi que de soit d'extérieur) même lorsqu'elles sont identiques aux images que *nous* nous avons des arbres. Tout comme une tache de peinture peut ressembler à une image d'un arbre sans *être*

l'image d'un arbre, une « donnée sensorielle » peut être qualitativement identique à une « image d'un arbre » sans en être une. Dans le cas des cerveaux dans une cuve, puisque le langage est relié par le programme à des inputs sensoriels qui, intrinsèquement ou extrinsèquement, ne représentent ni des arbres ni quoi que ce soit d'autre, comment le système de représentation dans son ensemble, le langage-dans-son-emploi, pourrait-il, lui, *faire référence*, c'est-à-dire représenter des arbres ou n'importe quel autre type d'objet extérieur ?

La réponse est que c'est impossible. Le système de données sensorielles dans son ensemble, les signaux moteurs des terminaisons afférentes, la pensée verbale ou conceptuelle reliée aux données sensorielles ou à autre chose par l'input des « règles d'entrée du langage » et par l'output des « règles de sortie », n'ont pas plus de rapports avec les arbres que le tracé de la fourmi n'en a avec Winston Churchill. Dès que l'on a compris que la *ressemblance qualitative* — qui équivaut, si l'on veut, à l'identité qualitative entre les pensées d'un cerveau dans la cuve et les pensées d'une personne du monde réel — n'implique aucunement l'identité référentielle, on perçoit facilement qu'il n'y a aucune raison de considérer que les cerveaux dans la cuve font référence à des choses extérieures.

Les prémisses de l'argument.

Je viens de donner l'argument qui vise à démontrer que des cerveaux dans une cuve ne peuvent pas penser ou dire qu'ils sont des cerveaux dans une cuve. Il ne me reste plus qu'à l'expliciter et à en analyser la structure.

D'après ce que je viens de dire, quand un des cerveaux dans la cuve (dans le monde où tous les êtres pensants sont depuis toujours et resteront à jamais des cerveaux dans une cuve) pense « Il y a un arbre devant moi », il ne fait pas, par sa pensée, référence aux arbres réels. Dans le cadre de certaines théories que nous envisagerons, il pourrait faire référence aux arbres dans l'image, ou aux impulsions électroniques qui produisent des expériences d'arbre, ou à la partie du programme d'ordinateur qui est responsable de ces impulsions électroniques. Ces théories ne sont pas exclues par ce que nous venons de dire, car il existe un lien causal étroit entre l'emploi du mot « arbre » en français-cuvien, la présence d'arbres dans l'image, la présence d'impulsions électroniques d'un certain type, et la présence de certains

traits dans le programme de l'ordinateur. Selon ces théories, le cerveau a *raison* et non *tort* de penser « Il y a un arbre devant moi ». Étant donné la référence d'*arbre* et de *devant* en français-cuvien, et si l'on suppose que l'une de ces théories est juste, alors les conditions de vérité de la phrase « Il y a un arbre devant moi » dans son occurrence franco-cuvienne sont simplement qu'il y ait un arbre « devant » le « moi » en question dans l'image, ou peut-être que le genre d'impulsion qui produit normalement cette expérience d'arbre soit fournie par l'ordinateur, ou que les caractéristiques du programme de l'ordinateur qui sont censées produire l'expérience de « l'arbre devant soi » soient effectivement en activité. Et ces conditions de vérité sont certainement satisfaites.

Par le même argument, « cuve » en français-cuvien désigne des cuves dans l'image ou quelque chose de relié (que ce soit des impulsions électroniques ou des caractéristiques du programme). Mais « cuve » ne désigne sûrement pas de vraies cuves, puisque l'emploi de « cuve » en français-cuvien n'est pas causalement relié aux cuves réelles. (Bien qu'en un sens il y soit relié, dans la mesure où les cerveaux dans la cuve ne pourraient pas utiliser le mot « cuve » s'il n'existait pas une cuve particulière, à savoir la cuve dans laquelle ils se trouvent. Mais cette relation s'applique à *tous* les mots du français-cuvien ; ce n'est pas une relation spéciale entre l'emploi du mot *spécifique* cuve et les cuves en général.) De même, « liquide nutritif » désigne un liquide dans l'image en français-cuvien, ou quelque chose de relié (des impulsions électroniques ou des caractéristiques du programme). Il s'ensuit que si ce « monde possible » est le monde réel et si nous sommes vraiment des cerveaux dans une cuve, alors ce que nous disons lorsque nous disons « nous sommes des cerveaux dans une cuve », c'est que *nous sommes des cerveaux dans une cuve dans l'image*, ou quelque chose de ce genre (si tant est que nous disons quelque chose !). Mais une partie de l'hypothèse que nous sommes des cerveaux dans une cuve est précisément que nous ne sommes pas des cerveaux dans une cuve dans l'image (quoique nous hallucinions, nous n'hallucinons pas que nous sommes des cerveaux dans une cuve). Donc, si nous sommes des cerveaux dans une cuve, alors la phrase « nous sommes des cerveaux dans une cuve » dit quelque chose de faux (si elle dit quelque chose). En bref, si nous sommes des cerveaux dans une cuve, alors « nous sommes des cerveaux dans une cuve » est fausse. Donc, elle est (nécessairement) fausse.

L'idée que toute cette histoire a un sens résulte de la combinaison de deux erreurs : (1) On prend trop au sérieux la *possibilité physique*; (2) on utilise inconsciemment une théorie magique de la référence, une théorie qui veut que certaines représentations mentales désignent nécessairement certaines choses et certains types de choses extérieures.

Il existe un « monde physiquement possible » où nous sommes des cerveaux dans une cuve — qu'est-ce que cela veut dire, sinon qu'il existe une *description* d'un tel état de choses qui est compatible avec les lois de la physique ? Tout comme il existe une tendance dans notre culture, présente depuis le XVII^e^ siècle, à prendre la *physique* pour la métaphysique et à considérer que les sciences exactes fournissent enfin la description de la « constitution ultime et définitive de l'univers », il existe aussi une tendance à prendre la possibilité physique pour le fondement de ce qui peut réellement, véritablement, être le cas. Selon ce point de vue, la vérité, c'est la vérité physique ; la possibilité, c'est la possibilité physique et la nécessité, c'est la nécessité physique. Mais nous venons de voir, même si nous nous en sommes tenus pour l'instant à un exemple assez peu naturel, que ce point de vue est faux. L'existence d'un monde « physiquement possible », où nous sommes tous (et avons toujours été, et serons toujours) des cerveaux dans une cuve, n'implique pas que nous pourrions en fait, réellement, possiblement, *être* des cerveaux dans une cuve. Ce qui exclut cette possibilité, ce n'est pas la physique mais la *philosophie.*

Certains philosophes, à la fois soucieux d'affirmer et de minimiser la position de leur profession (état d'esprit caractéristique de la philosophie anglo-saxonne au XX^e^ siècle), diront « Bon. Vous avez montré que des choses qui semblent être des possibilités physiques sont en fait des impossibilités *conceptuelles.* Qu'y a-t-il de surprenant à cela ? »

Évidemment, on peut décrire mon argument comme un argument « conceptuel ». Mais dire que l'activité philosophique consiste à rechercher des vérités conceptuelles revient à en faire une *investigation sur le sens des mots.* Et ce n'est pas du tout ce que nous avons fait.

Ce que nous avons fait, c'est envisager les *conditions préalables de la représentation, de la pensée, de la référence, etc.* Nous *n'avons pas* conduit cette investigation en analysant le sens de ces mots et de ces phrases (comme pourrait le faire un linguiste, par exemple) mais par un *raisonnement a priori.* Il ne s'agit pas du vieux sens « absolu » du terme (car nous ne prétendons pas que

les théories magiques de la référence sont fausses *a priori*) ; nous avons plutôt cherché à déterminer ce qui est *raisonnablement* possible en *supposant* certaines prémisses générales, ou en faisant certaines hypothèses théoriques très larges. Cette manière de procéder n'est ni tout à fait « empirique » ni tout à fait « a priori » mais fait appel à des éléments des deux modes d'investigation. Bien que ma procédure soit faillible et qu'elle dépende de suppositions qui peuvent être dites « empiriques » (comme par exemple la supposition que l'esprit n'a pas accès aux choses extérieures ou à d'autres propriétés que celles fournies par les sens), elle possède de nombreuses affinités avec ce que Kant appelait un « argument transcendental » parce qu'il s'agit, je le répète, d'une investigation des *conditions préalables* de la référence et donc de la pensée — conditions qui sont inscrites dans la nature même de nos esprits, même si elles ne sont pas (comme l'espérait Kant) complètement indépendantes de toute supposition empirique.

Une des prémisses de l'argument est évidente : les théories magiques de la référence sont fausses — fausses non seulement pour les représentations physiques, mais aussi pour les représentations mentales. L'autre prémisse est que l'on ne peut pas faire référence à certains types de choses, par exemple à des *arbres*, si l'on n'a aucune interaction causale avec elles[3] ou avec des choses permettant de les décrire. Mais pourquoi faudrait-il accepter ces prémisses ? Puisque celles-ci constituent le cadre global dans lequel je me place, il est temps de les examiner de plus près.

Pourquoi il faut nier l'existence de rapports nécessaires entre les représentations et leurs référents.

J'ai dit plus haut que certains philosophes (le plus célèbre étant Brentano) ont attribué à l'esprit un pouvoir, l'« intentionnalité », qui lui permet précisément de *faire référence.* J'ai rejeté cela en disant que ce n'était pas une solution. Mais

3. Si, *dans le futur,* les cerveaux dans la cuve se trouvent causalement reliés aux arbres, alors ils peuvent peut-être y faire référence *maintenant* en utilisant la description « les choses que je désignerai comme des « arbres » à tel moment du futur ». Mais il faut imaginer une situation où les cerveaux dans la cuve n'ont *jamais* accès à l'extérieur de la cuve, et ne sont donc *jamais* causalement reliés à des arbres.

qu'est-ce qui m'en donne le droit ? Suis-je peut-être allé un peu trop vite ?

Ces philosophes ne prétendaient pas que nous pouvons penser à des objets ou à des propriétés sans utiliser des représentations. Et ils auraient admis que l'argument que j'ai présenté plus haut, qui comparait les données sensorielles visuelles et le « dessin » de la fourmi (l'argument basé sur l'histoire de science-fiction du « dessin » d'un arbre produit par des taches d'encre qui fournissait des données sensorielles qualitativement semblables à nos « images visuelles d'arbres », mais que n'accompagnait aucun *concept* d'arbre) démontre que les *images* ne désignent pas nécessairement quelque chose. S'il existe des représentations mentales qui désignent nécessairement des choses extérieures, elles doivent être de l'ordre du *concept* et non de l'image. Mais qu'est-ce qu'un concept ?

L'introspection ne nous livre pas des « concepts » qui se promènent tant que tels dans notre esprit. Que l'on interrompe le cours de la pensée, n'importe où, n'importe quand, et l'on percevra des mots, des images, des sensations, des sentiments. Si je pense à haute voix, je ne tourne pas trois fois ma langue dans la bouche avant de parler. J'entends mes paroles comme vous. Il est vrai que je sens une différence entre le fait de prononcer des paroles auxquelles je crois et des paroles auxquelles je ne crois pas (mais si je suis nerveux, ou si je suis confronté à un public hostile, il m'arrive d'avoir l'impression de mentir alors que je sais que je dis la vérité). Et ce n'est pas la même chose de prononcer des paroles que je comprends et des paroles que je ne comprends pas. Mais je n'ai pas de mal à imaginer quelqu'un qui penserait précisément ces mots-ci, qui se les réciterait, qui aurait vraiment l'impression de les comprendre, de faire des affirmations, et qui, une minute plus tard, en émergeant d'une transe hypnotique, s'apercevrait qu'il ne comprenait pas du tout les mots qui lui passaient par la tête et qu'il ne connaissait même pas la langue dont ils étaient tirés. Je ne prétends pas que ce soit très plausible ; je veux simplement dire que ce n'est pas inconcevable. Et ceci ne démontre pas que les concepts *sont* des mots, des images, ou des sensations, mais que le fait d'attribuer un « concept » ou une « pensée » à quelqu'un est une chose bien différente de lui attribuer une quelconque « présentation » mentale ou n'importe quel autre événement ou entité introspectible. Les concepts ne sont pas des présentations mentales qui désignent intrinsèquement des objets extérieurs,

pour la simple raison que ce ne sont pas du tout des présentations mentales. Les concepts sont des signes utilisés d'une manière particulière ; les signes peuvent être des entités mentales ou physiques, publiques ou privées, mais même lorsque les signes sont « mentaux » et « privés », le signe lui-même, considéré indépendamment de son emploi, n'est pas le concept. Et les signes eux-mêmes ne désignent pas intrinsèquement quelque chose.

On peut voir cela en imaginant une expérience de pensée très simple. Supposons que comme moi vous ne sachiez pas reconnaître un orme d'un bouleau. Nous dirons quand même que la référence d' « orme » dans mon parler est la même que dans le parler de tout le monde, c'est-à-dire l'ensemble des ormes, et que dans votre parler comme dans le mien, l'extension du mot « bouleau » est l'ensemble de tous les bouleaux (c'est-à-dire l'ensemble des choses dont le mot « bouleau » peut- être prédiqué véridiquement). Est-il vraiment plausible de prétendre que cette différence dans nos référents d' « orme » et de « bouleau » résulte d'une différence dans nos *concepts* ? Mon concept d'orme est identique à mon concept de bouleau, j'ai honte à le dire. (D'ailleurs, ceci montre que la détermination de la référence est sociale et non individuelle ; vous et moi, nous faisons confiance aux experts qui *savent* distinguer un bouleau d'un orme.) Si quelqu'un veut tenter de soutenir avec héroïsme que la différence entre la référence d' « orme » et la référence de « bouleau » dans *mon* parler est due à une différence d'état psychologique, qu'il s'imagine alors une Terre-Jumelle où les deux mots seraient intervertis. Cette Terre-Jumelle serait identique à la Terre, au détail près que « orme » et « bouleau » sont intervertis. Supposons qu'il existe une réplique de moi sur Terre-Jumelle, qui m'est identique à quelques molécules près (au sens où deux cravates sont « identiques »). Si vous êtes dualiste, alors supposez que ma réplique a les mêmes pensées verbalisées que moi, qu'il perçoit les mêmes données sensorielles que moi, qu'il a les mêmes dispositions que moi, et ainsi de suite. Il est absurde de dire que son état psychologique est différent du mien : pourtant, son mot « orme » désigne des *bouleaux*, et mon mot « orme » désigne des *ormes*. (De même, si « eau » sur Terre-Jumelle désigne un autre liquide, disons XYZ, au lieu de H_2O, alors « eau » désigne un liquide différent lorsqu'on l'emploie sur Terre-Jumelle et sur Terre, et ainsi de suite.) Contrairement à une doctrine très répandue depuis le XVII^e siècle, *les significations ne sont tout simplement pas dans la tête.*

Nous avons vu que posséder un concept ce n'est pas simplement avoir des images (que ce soit des images d'arbres, ou même des images « acoustiques » ou « visuelles » de phrases ou de discours) puisque l'on pourrait avoir tous les systèmes d'images que l'on veut et être encore dans l'*impossibilité* d'utiliser les phrases de manière appropriée (en supposant, comme nous l'avons déjà dit, que des facteurs à la fois linguistiques et extra-linguistiques déterminent ce qui est « approprié dans une situation »). Un homme pourrait avoir toutes les images qu'on veut, et ne pas savoir quoi faire si on lui demandait « montrez-moi un arbre », même en présence d'un grand nombre d'arbres. Il pourrait même avoir à sa disposition une image de ce qu'il devrait faire, et ne pas savoir quoi faire. Car une image mentale qui n'est pas accompagnée de la capacité d'agir d'une certaine manière n'est rien qu'une *image,* et le fait de pouvoir agir en s'inspirant d'une image est une capacité que l'on peut avoir ou non. (Notre homme pourrait s'imaginer en train de montrer un arbre du doigt parce que ça lui chante d'envisager quelque chose de logiquement possible : lui-même en train de montrer un arbre après que quelqu'un ait produit la séquence — pour lui incompréhensible — « montrez-moi un arbre ». Il ne comprendrait pas qu'il était censé montrer un arbre, et il ne *comprendrait* toujours pas ce que « montrez-moi un arbre » veut dire.)

J'ai considéré jusqu'ici que la capacité d'utiliser certaines phrases était le critère de possession d'un véritable concept, mais on pourrait facilement étendre les remarques précédentes. On pourrait imaginer, par exemple, un symbolisme composé d'éléments qui ne seraient pas des mots d'une langue naturelle, et on pourrait admettre en tant que phénomènes mentaux des images ou d'autres types de phénomènes internes. Ce qui est essentiel, c'est qu'ils aient la même complexité, les mêmes possibilités de combinaison que les phrases d'une langue naturelle. Car quoiqu'une présentation donnée, disons une lumière bleue, pourrait représenter pour un mathématicien l'expression intérieure de la démonstration du théorème des nombres premiers, si toutefois ce mathématicien n'était pas capable de décomposer sa « lumière bleue » en des étapes distinctes et logiquement reliées, non seulement il n'y aurait pas de raison de l'appeler son « expression intérieure », mais il serait même faux de le faire. Or, quel que soit le type de phénomène que nous admettions à titre d'*expression* possible de la pensée, des arguments identiques aux précédents montrent que la compréhension n'est pas constituée

par ces phénomènes en tant que tels, mais par la capacité du penseur à *utiliser* ces phénomènes, à produire les phénomènes appropriés dans les bonnes circonstances.

Ce qui précède est une version succincte de l'argument de Wittgenstein dans les *Philosophical Investigations.* S'il est juste, alors toutes les tentatives de comprendre la pensée par ce qu'il est convenu d'appeler l'investigation « phénoménologique » sont fondamentalement erronées ; car ce que les phénoménologues ne comprennent pas, c'est qu'ils décrivent l'*expression* intérieure de la pensée, mais que la compréhension de la pensée — la compréhension que l'on a de ses propres pensées — n'est pas une *occurrence* mais une *capacité.* Notre exemple de l'homme qui semblait penser en japonais et qui trompait même un télépathe japonais montrait déjà la futilité des approches phénoménologiques du problème de la *compréhension.* Car même s'il existe une qualité introspectible qui n'est présente que lorsqu'on comprend *vraiment* (ce qui semble faux, d'ailleurs, précisément à la lumière de l'introspection), cette qualité n'est qu'en *corrélation* avec la compréhension et notre homme qui trompait le télépathe japonais aurait pu posséder cette qualité sans toutefois comprendre un mot de japonais.

D'un autre côté, considérons maintenant l'homme parfaitement possible qui n'a aucun « monologue intérieur ». Il parle parfaitement l'anglais, et si on lui demande son avis sur quelque chose, il l'expose longuement. Mais quand il ne parle pas à haute voix il ne pense jamais avec des mots, des images et ainsi de suite, et rien ne lui passe par la tête. Évidemment, il entend sa propre voix, il perçoit les impressions sensorielles ordinaires de son environnement et il a en plus une « sensation globale de compréhension ». (Il a peut-être l'habitude de parler tout seul.) Quand il tape une lettre à la machine, ou quand il va faire ses courses, etc., il n'a pas de « monologue intérieur » mais ses actions sont intelligentes et orientées, et si quelqu'un venait le voir et lui demandait « qu'est-ce que vous faites ? », il donnerait des réponses parfaitement cohérentes.

Cet homme semble tout à fait concevable. Et bien qu'il n'ait pas de pensées conscientes, sauf quand il parle à haute voix, personne n'hésiterait à dire qu'il est conscient, ou qu'il n'aime pas le rock (s'il exprime souvent une aversion pour le rock-and-roll).

Il s'ensuit de tout ceci que (a) aucun ensemble d'événements mentaux, que ce soit des images, des événements ou des propriétés mentales plus abstraites, ne *constitue* la compréhen-

sion et que (b) aucun ensemble d'états mentaux n'est *nécessaire* à la compréhension. En particulier, *les concepts ne peuvent être identiques à aucun type d'objet mental.* Car, si par objet mental nous entendons quelque chose qui est accessible à l'introspection, nous venons de voir que quel que soit cet objet, il peut être absent chez un homme qui comprend le mot approprié et qui possède donc le concept total, et présent chez un homme qui ne possède pas du tout le concept.

Pour revenir à notre critique des théories magiques de la référence (un sujet qui préoccupait aussi Wittgenstein), la conclusion est d'une part que les objets « mentaux » que nous *pouvons* détecter par introspection (les mots, les images, et les sensations) ne désignent pas intrinsèquement quelque chose, pas plus que le dessin de la fourmi ne désigne intrinsèquement quelque chose et pour les mêmes raisons, et d'autre part que les tentatives de postuler des objets mentaux spéciaux, les « concepts », qui eux *auraient* un rapport nécessaire avec leurs référents et que seuls les phénoménologues expérimentés pourraient détecter, sont fondées sur une erreur *logique,* car les concepts sont (du moins en partie) des capacités et non des occurrences. La doctrine qui veut qu'il existe des représentations mentales qui désignent nécessairement des choses extérieures est non seulement de la mauvaise science naturelle mais aussi de la mauvaise phénoménologie, teintée de confusion conceptuelle.

CHAPITRE II

UN PROBLÈME AVEC LA RÉFÉRENCE

Pourquoi est-il surprenant que l'hypothèse des cerveaux dans une cuve soit incohérente ? C'est parce que nous avons tendance à penser que *ce qui se passe dans notre tête* détermine ce que nous voulons dire et ce que désignent nos mots. Mais il est facile de montrer que cela est faux. Des mots indexicaux ordinaires comme *je, ceci, ici* et *maintenant* sont des contre-exemples triviaux. Je me trouve peut-être dans le même état mental qu'Henri lorsque je pense « je suis en retard à mon travail » (supposons, par exemple, qu'Henri et moi nous sommes de vrais jumeaux) et pourtant l'occurrence du mot « je » qui apparaît dans mes pensées me désigne moi et l'occurrence du mot « je » qui apparaît dans les pensées d'Henri désigne *Henri.* Je me trouve peut-être dans le même état mental lorsque mardi je pense « je suis en retard » et lorsque mercredi je pense « je suis en retard » mais dans chaque cas le verbe *être* désigne un temps différent. Les noms d'espèces naturelles sont un exemple plus subtil de la même idée.

Supposons, pour reprendre l'exemple du chapitre précédent, qu'il existe des locuteurs du français sur Terre-Jumelle (par un hasard miraculeux, ils ont évolué comme nous et parlent une langue qui est, sauf pour une différence que je vais mentionner tout de suite, identique au français tel qu'on le parlait il y a environ deux cents ans). Je vais supposer que ces gens n'ont pas encore développé une chimie daltonienne ou post-daltonienne. Donc, en particulier, ils n'ont pas à leur disposition des

1. Je peux me trouver dans le « même état mental » dans la mesure où les paramètres impliqués dans le processus psychologique qui aboutit à cette pensée peuvent avoir la même valeur. Mon état mental *global* est sans doute différent, puisque mardi je crois qu'on est mardi » et que mercredi je ne le crois pas ; mais une théorie qui dirait que la signification des mots change dès que mon état mental *global* change ne permettrait jamais à des mots d'avoir le même sens et reviendrait donc à abandonner l'idée même du sens d'un mot. On pourrait aussi construire une histoire de Terre-Jumelle où moi et ma réplique nous nous trouverions dans le même état mental *global,* mais où la référence de « moi » et de « maintenant » resterait différente (le calendrier de la Terre-Jumelle pourrait ne pas être synchronisé avec le nôtre).