L'animal intentionnel

(2000), Terrains, 34, 23-36.

Joëlle Proust

CREA - Ecole Polytechnique
(Paris)

Résumé

Les animaux forment des représentations mentales dès qu'ils ont la capacité d'extraire de l'information sur les corrélations environnementales, de la fixer dans certains états internes mémorisés, et d'identifier des objets et des événements indépendants de la perception qu'ils en ont. Un dispositif de calibration entre les modalités sensorielles est indispensable pour que les représentations soient appliquées à des objets perçus comme extérieurs. Les animaux qui disposent de ce type de représentation dite objective sont susceptibles de former des concepts. Toutefois les animaux sociaux non humains extraient l'information sociale non sur la base du registre psychologique (croyances et désirs) mais sur la base d'indices comportementaux. Ils ont ainsi une théorie sociale rudimentaire, mais non une théorie de l'esprit.

Mots-clés: théorie de l'esprit, représentation, intentionnalité,

attention conjointe.

L'animal intentionnel

Joëlle Proust

CREA - Ecole Polytechnique
(Paris)

L'ami des bêtes trouve parfaitement naturel d'interpréter le comportement des animaux familiers en leur prêtant des désirs et des croyances analogues aux siens propres. Par exemple, le possesseur d'un chien attribue quotidiennement à son animal, et sans la moindre hésitation, la croyance que c'est le moment de la promenade, ou l'envie de mordre le facteur. Il va même jusqu'à supposer que l'animal est capable de former des croyances et des désirs sur ce que les autres pensent ou désirent. Il lui suffit pour cela d'avoir feuilleté la littérature "animalière" où abondent les anecdotes rapportant qu'un animal a rusé avec les attentes de ses congénères, leur a tendu des pièges, a intentionnellement transmis son savoir à ses rejetons, ou a tenté de travestir la réalité pour tromper ses rivaux. C'est évidemment en primatologie que les exemples sont les plus nombreux. Par exemple, Franz de Waal rapporte que "les orangs-outangs se mettent des végétaux sur la tête

pour observer l'effet produit". Certaines femelles chimpanzés "poussent le raffinement jusqu'à s'accrocher des plantes grimpantes au cou pour s'embellir (de Waal 1997 : 91). Un chimpanzé peut cacher de sa main son érection en présence d'un mâle dominant. Une femelle peut faire un geste de réconciliation pour mieux mordre sa victime (de Waal 1997 : 100). De nombreux auteurs, parmi lesquels un expérimentaliste aussi exigeant que David Premack (Premack & Premack 1984), ont dit observer des comportements de tromperie sélectivement dirigés contre des individus non-coopératifs.

Ces observations soulèvent plusieurs questions : les animaux ont-ils des états mentaux comme les nôtres, des désirs, des croyances, des intentions, qui auraient des contenus déterminés ? Ceux d'entre eux, s'il y en a, qui ont des états mentaux ont-ils en outre la capacité de se représenter les états mentaux d'autrui (et les leurs) *en tant qu'états mentaux* ?

Que veut dire penser ?

Les éthologistes et les philosophes tendent à estimer que la plupart des oiseaux et des mammifères sont capables de former des représentations et de les utiliser pour contrôler leur comportement. En revanche, personne n'est tenté de dire qu'une moule ou qu'une huître pensent. La propriété qui distingue ces deux types d'organismes, et institue la coupure entre les vivants "cognitifs" et les autres, est nommée par les philosophes "intentionnalité". Ce qu'ils entendent par là, c'est l'aptitude de certains états internes à *porter sur* des propriétés du monde extérieur. En d'autres termes, l'intentionnalité est la

capacité d'utiliser l'information sur le monde extérieur et de la stocker dans des représentations pour l'appliquer à des situations nouvelles et ajuster le comportement au cours des choses.

En quoi consiste cette capacité représentationnelle, et comment se fait l'extraction d'information permettant de former ou d'activer une représentation ? On pourrait supposer que la capacité de représenter implique la capacité d'établir une corrélation entre des états internes et des états externes. Mais cette définition nous conduirait à inclure les organismes élémentaires, comme les huîtres ou les escargots, parmi les êtres qui pensent, parce qu'ils sont effectivement capables, avec un nombre limité de neurones, d'établir de telles corrélations et de réagir de manière appropriée à leur environnement. L'aplysie par exemple, un mollusque bien étudié dans sa structure neuronale (Hawkins & Kandel 1984), possède des organes respiratoires qui comprennent des branchies situées dans un manteau recouvert par une membrane de protection dont l'extrêmité constitue le *siphon*. Quand le siphon est stimulé tactilement, tous les organes respiratoires se rétractent en vertu d'un réflexe de défense. Le fait important, du point de vue de l'extraction des corrélations, est que le réflexe de l'aplysie peut être modifié par apprentissage. L'aplysie apprend à ignorer un stimulus tactile quand il est souvent répété, et à produire sa réponse en présence de stimuli du même type, mais de moindre intensité. Les aplysies peuvent aussi être conditionnées à rétracter leur siphon lorsqu'elles reçoivent des décharges électriques sur leur pied. Tout cela indique qu'elles sont capables de mettre en

corrélation leurs dynamiques neuronales avec les changements du monde ; qui plus est, l'apprentissage dont elles sont capables suit un pattern temporel comparable à celui du conditionnement des vertébrés. Y-at-il un sens à dire que l'état des neurones sensoriels "représente le monde" pour l'aplysie ? Si représentation il y a, s'agit-il d'une représentation mentale ?

Contrairement à une opinion solidement établie, on ne peut considérer qu'une corrélation établie par conditionnement permette *ipso facto* de conclure que la représentation formée n'est pas mentale. Si l'on suit la définition de la représentation mentale de Fred Dretske (Dretske 1986), est mental un état qui remplit simultanément les trois conditions suivantes :

1) Il existe une covariation régulière entre cet état et une situation externe donnée. Dretske appelle "indication" ce type de relation, en vertu de laquelle l'état interne *porte une information* sur la situation ou la propriété externes (ou les *indique*).

2) L' indicateur interne (c'est-à-dire l'état neuronal porteur d'une certaine information) a *la fonction* d'indiquer la situation externe. On dit dans ce cas que l'indicateur *représente* la situation. En parlant de fonction, Dretske entend souligner qu'un mécanisme de sélection a conduit un certain état neuronal - lié par exemple à la fonction visuelle - à être réactivé chaque fois qu'un certain type de forme visuelle était traité.

3) Les états internes à valeur représentationnelle ainsi déterminés doivent enfin être vrais ou faux, c'est-à-dire "sémantiquement évaluables". En d'autres termes, dès qu'une

représentation mentale est acquise, elle doit pouvoir être appliquée de manière véridique ou erronée.

On comprend bien comment cette troisième condition est liée à la précédente : une fois qu'un indicateur a acquis une fonction, le dysfonctionnement *doit* être possible. La représentation peut en principe être activée en l'absence de ce qu'elle représente. Traduite en termes quotidiens, cette troisième condition implique qu'un animal qui a un esprit doit pouvoir de temps en temps croire (à tort) que telle ou telle circonstance est présente et agir conformément à ce qu'il croit. Le lecteur s'étonnera peut-être que la possibilité de se tromper constitue un pas aussi significatif dans l'évolution. Cette possibilité, ainsi que la capacité de modifier l'état interne une fois que l'erreur est repérée, est pourtant conceptuellement liée à la capacité de former des représentations mentales.

Pour revenir au cas de l'aplysie, qui vaut sans doute d'un grand nombre d'animaux, on voit qu'elle remplit les deux premières conditions : elle a des indicateurs internes qui covarient avec des situations extérieures ; ces indicateurs acquièrent la fonction d'indiquer ces situations (il y a apprentissage). Mais on ne peut pas dire que les états internes forment des croyances, parce qu'il n'y a pas de sens à dire que l'aplysie "se trompe". L'équipement neuronal de l'aplysie ne lui permet pas de faire la différence entre ses états internes et ce que ces états représentent. Mais quelles sont les conditions qui permettent à un animal, serait-ce de manière primitive, de distinguer ses propres états de ceux du monde qui l'entoure ?

Dans un ouvrage récent (Proust 1997), j'ai proposé de préciser les conditions auxquelles il est légitime d'attribuer des

croyances à un animal sans langage. L'idée de base est la suivante : pour pouvoir avoir des croyances vraies ou fausses sur le monde, il faut que l'organisme dispose préalablement d'une organisation perceptive lui permettant de situer l'observateur qu'il est lui-même dans un monde d'objets et d'événements perçus. On l'a vu plus haut, une représentation est mentale si elle permet de représenter (ou de méreprésenter) le monde. Cette étonnante capacité paraît impliquer d'une manière ou d'une autre la capacité "de sortir de soi pour atteindre le monde", c'est-à-dire de répondre à des propriétés du monde, et non pas simplement à un quelconque état proximal des récepteurs de l'organisme.

Cette capacité, les philosophes la nomment "objectivité" (Strawson 1959). Ce qu'il faut comprendre par là, c'est l'aptitude à distinguer ce qui relève de l'*expérience* du sujet, de ce qui relève de l'*objet* de son expérience. Le sens commun aborde le problème de l'objectivité de manière métaphorique, en mêlant des considérations spatiales et fonctionnelles : on oppose simplement ce qui se passe "en soi" et "hors de soi", ou ce qui relève du "système" ou de son "environnement". Mais le sens commun commet une pétition de principe, en s'appuyant sur une représentation particulière - l'opposition "dedans/dehors", ou plus généralement les concepts spatiaux, - pour expliquer les capacités représentationnelles dans les esprits non-humains. Car les concepts spatiaux ne peuvent offrir de solution quand il s'agit d'*expliquer* l'usage des relations spatiales pour distinguer des concepts ! Comment expliquer ce qui permet à l'espace d'être, pour ainsi dire, intériorisé par un organisme, et ce qui fait que la distinction intérieur/extérieur

oriente l'ensemble du processus de formation des représentations mentales ?

Espace et pensée objective

De manière intuitive, on pourrait dire que ce qui distingue la manière dont des organismes comme l'aplysie ou l'escargot de mer, d'un côté, et les oiseaux ou les mammifères, de l'autre, utilisent l'information, réside dans le fait que les premiers ne traitent qu'une information au contact de leurs récepteurs, tandis que les autres traitent une information qui concerne le monde environnant. Appelons "proximale" l'information que traite l'aplysie, et "distale" l'information que traite un mammifère. Comment comprendre la différence entre ces deux types d'information sans présupposer l'espace ?

L'espace tel qu'on le conçoit intuitivement est une sorte de cadre vide pour des contenus perceptifs possibles. Mais on peut aussi le comprendre comme un ensemble de propriétés formelles. Si les entrées perceptives sont classées sous la contrainte imposée par la distribution spatiale de l'information (qui est alors, par hypothèse, l'information distale), elles présentent entre elles des relations qui ne se retrouvent pas dans des systèmes capables exclusivement de traitement de l'information proximale (c'est-à-dire l'information portant sur les événements au niveau des récepteurs sensoriels). La relation qui nous intéresse est celle d'*équilocalité* , c'est-à-dire la classe d'équivalence de toutes les expériences perceptives qui concernent le même emplacement péripersonnel. Cette relation a pour nous l'intérêt d'avoir une réalité psychologique

indéniable (nous sommes très sensibles au fait qu'un événement perçu se produise dans telle ou telle partie de l'espace) et d'être définissable en termes purement logiques, qui ne présupposent ni le concept d'espace, ni celui de concept (Carnap, 1928). L'équilocalité est également psychologiquement essentielle, parce qu'elle est à la base d'une aptitude fondamentale, celle de réidentifier des objets au fil du temps. Identifier les régions de l'espace est indispensable pour reconnaître un objet occupant un certain lieu, d'une fois sur l'autre, ou pour découvrir l'identité d'un objet ayant une trajectoire continue dans l'espace. La relation d'équilocalité va ainsi nous donner la clé de ce qui permet à un système de représentation de porter sur le monde, c'est-à-dire à un animal d'avoir des états mentaux. Quel type de dispositif doit-il être présent pour qu'un animal puisse extraire ce type d'information - sur les invariants spatiaux ?

Revenons encore à l'aplysie. Comme chez beaucoup d'invertébrés, la seule procédure permettant de traiter les entrées sensorielles est un mécanisme additif. Les intensités des événements sensoriels des différentes modalités sont additionnées ; les effets de potentiation et de dépotentiation résultent directement de la sommation des entrées. Dans ce cas, aucune erreur ne peut être détectée au niveau des perceptions. Il ne peut y avoir de véritable conflit entre les différentes modalités, par exemple entre les récepteurs de lumière et les détecteurs de gravité. D'autres animaux, tels que les oiseaux, les reptiles, les mammifères, possèdent en revanche un dispositif permettant de *corriger* la réception des données

sensorielles d'une certaine modalité afin de respecter les contraintes d'équilocalité dans le champ perceptif global. Les travaux de Knudsen (Knudsen 1982) montrent par exemple comment une jeune chouette effraie chez qui on a expérimentalement perturbé (en bouchant une oreille) la correspondance entre représentation d'un stimulus visuel et représentation d'un stimulus sonore émanant du même endroit parvient à restaurer la congruence spatiale des événements perçus par réalignement de la carte auditive sur la carte visuelle. On appelle calibration ce processus, par lequel l'information spatiale provenant d'une modalité (comme l'audition ou le toucher) est corrigée par l'information spatiale issue d'une autre modalité (comme la vision). Ce mécanisme est indispensable pour corriger les entrées sensorielles quand elles deviennent incohérentes entre elles (ce qui se produit inévitablement quand l'organisme se développe ou vieillit) ; c'est grâce à ce mécanisme, constitué par l'existence de neurones multisensoriels spécialisés dans l'information spatiale, que l'animal peut acquérir une connaissance spatiale unifiée et cohérente de son environnement.

Les animaux ont-ils des concepts ?

Ce qui précède permet de dire qu'un grand nombre d'animaux non-humains ont vraisemblablement des représentations mentales, au sens défini par les trois conditions de Dretske évoquées plus haut. Ils perçoivent le monde comme composé d'entités distribuées dans l'espace et dotées d'une

certaine autonomie. Du même coup, les représentations qu'ils forment sur la base de leur perception sont elles aussi objectives, ce qui confère à leurs croyances éventuelles la capacité d'être sémantiquement évaluables, vraies ou fausses.

La question suivante est évidemment de savoir si ces animaux capables de former des sortes de proto-croyances forment des croyances effectives et conceptuellement articulées, ou s'ils sont simplement dotés de capacités perceptives non-conceptuelles, c'est-à-dire s'ils se bornent à catégoriser des formes et des événements perçus. Ce qui distingue, on le verra, les concepts des simples perceptions, c'est que les premiers rendent possibles des inférences, et donnent lieu à une forme de rationalité, alors que le contenu perceptif par lui-même ne permet pas à l'animal de se livrer à un raisonnement.

Certains philosophes, parmi lesquels Donald Davidson (Davidson 1991), ont défendu la thèse que les animaux non pourvus de langage se trouvent de ce fait nécessairement dépourvus de concepts. On sait comment Descartes a rejeté l'idée chère à Montaigne que les animaux puissent raisonner. Ce n'est pas en combinant leurs connaissances et les principes d'une logique naturelle que les animaux résolvent les problèmes qui se posent à eux. Ils sont en quelque sorte prédéterminés par leurs comportements innés à résoudre ces problèmes. En contraste avec les solutions locales auxquelles les animaux ont accès, les humains ont par leur raison un instrument universel de résolution de problème ; c'est le langage qui en offre le moyen, en apportant avec la combinaison des signes *l'universalité* propre à la pensée.

Donald Davidson dénie l'accès des animaux à la pensée selon

une stratégie un peu différente de celle de Descartes. Il part de l'idée incontestable qu'utiliser des concepts, c'est avoir des pensées sémantiquement évaluables. Ainsi, on ne peut pas utiliser le concept de *facteur*, par exemple, sans utiliser un certain nombre de caractéristiques qui permettent de dire *à juste titre* que tel individu est ou non un facteur. Or, pour cela, l'utilisateur de concepts doit nécessairement être capable de se représenter que le concept qu'il applique à un individu dans des circonstances données (en pensant, par exemple, "c'est le facteur") peut être employé à tort, incorrectement. Pour y parvenir, il lui faut pouvoir partager son application de concept avec un autre sujet, et tirer de l'échange une ratification de son emploi du concept, ou au contraire une contestation rationnelle de la légitimité d'un tel emploi.

Cette pratique de la triangulation (sujet-objet-interprète), comme la nomme Davidson, paraît impossible dans le cas de l'animal parce que, faute d'une langage symbolique partagé, il est privé de l'accès à l'intersubjectivité. C'est dans la communication intersubjective que deux locuteurs se mettent d'accord sur un objet et sur ses propriétés; c'est à l'occasion de cet échange rationnel que les concepts gagnent leur aspect normatif (chaque concept a des conditions d'applications précises qui déterminent la vérité du jugement correspondant), ainsi que leur détermination propre. Davidson reprend ainsi, en la précisant, l'idée de Descartes selon laquelle l'exercice de la rationalité, dont dépend la pensée conceptuelle, est inséparable de la maîtrise d'un langage. C'est le langage qui confère à la pensée son universalité, c'est-à-dire son indépendance à l'égard de contextes particuliers.

Descartes et Davidson ont raison de souligner en quoi la possession d'un langage public fonde la rationalité intersubjective, et étend considérablement l'empan de la pensée conceptuelle. On peut toutefois contester que l'animal n'ait pas une forme de pensée conceptuelle plus modeste. Cette forme de pensée n'implique pas la possession du concept de concept, mais elle inclut les trois types de dispositions suivantes :

1) quand un organisme possède le concept X, il est disposé à décider si quelque chose est ou non un X et à agir sur cette base.

2) Un concept acquis peut être appliqué à des cas nouveaux et en conjonction avec les autres concepts déjà maîtrisés (ce qu'on appelle "généralisation").

3) Les concepts forment une structure inférentielle (une théorie) et peuvent être modifiés par l'apprentissage.

Il n'est pas extravagant de supposer que certains animaux non humains puissent former des concepts leur rendant intelligibles les aspects de l'environnement les plus importants pour eux. Il est possible que les animaux sociaux (chiens, primates, etc.) disposent de théories sur l'organisation des relations sociales incluant les concepts de dominants, de subordonnés, d'immatures, d'ennemis, d'alliés, ainsi que les liens inférentiels et associatifs entre ces catégories pour le partage de la nourriture, la protection contre un prédateur, la recherche d'un partenaire, etc.

Que devient, dans cette version "modeste", l'idée de *norme*, dont dépend la correction de l'utilisation d'un concept ? On peut défendre l'idée que la norme provient de la sanction du comportement guidé par un certain concept. Si par exemple, le

chien mord le facteur en l'ayant identifié à un prédateur (seul concept disponible dans son "lexique mental" pour un animal étranger revenant à la même heure tous les jours sur le territoire de la meute), il aura confondu le concept de facteur et celui de prédateur. Il sera puni par son maître (qui lui attribue généralement une intention déterminée de mordre par "haine du facteur"), ou soumis à un dressage tentant de corriger son application de concept. La norme est bien inhérente au résultat de l'usage du concept, aux effets (attendus ou non) qui suivront son application.

Admettons donc que les animaux aient une forme de pensée conceptuelle. Qu'en est-il de leurs capacités propres de comprendre autrui, c'est-à-dire d'user pour leur propre compte de cette psychologie ordinaire dont les humains se servent eux-mêmes pour réguler leurs interactions ?

L'animal non-langagier et l'attribution intentionnelle

Etant admis qu'une espèce donnée est capable d'avoir des croyances et des désirs, c'est-à-dire des états internes sémantiquement évaluables, la question se pose de savoir si l'espèce considérée peut se représenter des propriétés et des événements qui sont non pas physiques, mais mentaux, et s'en servir pour agir sur les représentations d'autrui. Depuis le travail de pionnier de deux primatologues, Premack et Woodruff (Premack et Woodruff 1978), on appelle "théorie de l'esprit" le type de connaissance que possède un organisme quand il utilise des concepts mentaux équivalents à ceux de *croyance* et de *désir*, et se sert des régularités qui régissent les

relations entre ces états mentaux pour comprendre et anticiper le comportement des autres agents.

Quand on parle de théorie de l'esprit, on doit donc veiller à la distinguer d'une compétence strictement sociale, et non psychologique, en vertu de laquelle les animaux sociaux tentent d'influencer le comportement de leurs congénères. Comme le note David Premack (Premack 1988), un animal peut avoir l'intention d'agir de manière à affecter soit ce que l'autre individu fait, soit ce qu'il pense. Ce n'est bien entendu que dans le second cas que des connaissances psychologiques doivent être mises en oeuvre.

La distinction rigoureuse entre ces deux types de cas n'est pas toujours facile à faire. Car dans le premier cas - chercher à affecter ce que l'autre fait -, les seules stratégies efficaces, qu'elles soient apprises par leurs conséquences ou bien sélectionnées par l'évolution, sont des comportements de tromperie ou de mimétisme. Ces comportements peuvent être mis en oeuvre sans que les agents aient les concepts mentaux correspondants. Mais dans le second cas, celui où l'animal chercherait à affecter ce que l'autre pense, il est clair que le but ultime demeure de modifier les comportements de l'autre. La différence entre les deux cas tient ainsi non aux buts poursuivis, mais aux moyens utilisés. Si un chimpanzé mâle non dominant est capable de se retenir de vocaliser pendant qu'il copule avec une femelle, cela peut être dû soit au fait qu'il pense que les autres mâles *seraient avertis* de la présence d'une femelle réceptive s'il produisait le cri sexuel, et *voudraient* eux-aussi en profiter. Mais le même comportement pourrait aussi bien être produit parce que l'animal a remarqué la *corrélation* entre la

production de vocalisations et l'approche de rivaux. Cette corrélation peut être observée en l'absence de toute théorie psychologique d'arrière-plan. La version dite "rabat-joie" de l'anecdote (Dennett 1987) explique le comportement en faisant complètement l'économie de l'hypothèse psychologique. L'animal non-langagier peut agir de manière socialement adéquate en utilisant uniquement la corrélation entre des indices physiques ou purement comportementaux.

En dépit de cette évidente possibilité explicative, les éthologistes de terrain ont massivement adopté la stratégie intentionnelle recommandée par Dennett. Adopter la stratégie intentionnelle (Dennett 1987) consiste à présupposer d'emblée la rationalité de l'animal à interpréter, c'est-à-dire à appliquer à l'animal, connaissant la situation où il se trouve et ce qu'il peut en tirer d'avantages ou d'inconvénients, les croyances et les désirs qu'il devrait avoir. Puis on teste les prédictions que l'hypothèse psychologique permet de former sur le comportement de l'animal (en observant évidemment un principe méthodologique de parcimonie). Dennett propose une hiérarchie d'hypothèses d'ordre croissant, selon la profondeur réflexive attribuée à l'animal. Le *premier ordre* se contente d'attribuer des désirs et des croyances : le sujet interprété désire que P ou croit que Q. Le *second ordre* attribue des croyances et des désirs qui ont pour objet des croyances ou des désirs : "le sujet *veut* que le mâle dominant *ignore* qu'il copule avec une femelle". "Le sujet *pense* que, si les mâles *savent* qu'il y a une femelle réceptive, ils *désireront* en profiter". Le *troisième ordre* attribue des croyances et des désirs qui portent sur des croyances ou des désirs de second ordre : "le sujet *veut* que

l'autre *sache* que lui-même *connaît* déjà l'information qui lui est donnée". Etc.

L'attribution de premier ordre peut sembler relativement anodine : il ne s'agit après tout que de déterminer quel est l'objet auquel l'animal pense, l'événement qu'il remarque, anticipe ou désire. Mais elle est en fait très étroitement liée à la saillance qui, du point de vue de l'interprète, appelle l'attribution de second ordre. Si nous reprenons l'exemple de la femelle chimpanzé Mara qui "se pare", on voit que l'intention qui est attribuée, d'être "belle" est inséparable de l'attribution de second ordre "Mara a l'intention de produire chez ses congénères une impression d'admiration". Rien ne nous dit que Mara avait l'intention de produire un effet visuel de ce genre. Peut-être a-t-elle disposé les végétaux sur son corps par jeu ou par plaisir individuel, sans effet extérieur recherché. Le deuxième ordre fait sens parce qu'il est humainement plus rationnel d'organiser son action en fonction non seulement des besoins immédiats, mais de ce que le groupe perçoit du sujet. Le deuxième ordre intervient dès que l'on prête à l'animal le désir explicite de modifier les états mentaux des membres de son groupe : enseigner, tromper, cacher, font ainsi référence à des comportements dont le résultat est un état mental. Le troisième ordre paraît déjà d'une grande complexité. En fait les théoriciens de la communication démontrent qu'il faut pouvoir former des attributions de troisième ordre pour pouvoir interpréter correctement les formules remarquablement économiques d'une conversation ordinaire (Grice 1969, Sperber & Wilson 1989).

L'analyse que nous avons proposée plus haut s'appuie sur la

recherche de critères intrinsèques de cognition animale, tandis que l'approche interprétative de Dennett se satisfait de critères extrinsèques ; l'attribution d'un état de second ordre est une hypothèse que l'on peut selon lui directement confronter aux comportements. Or cette approche donne libre cours à la tendance prononcée des humains à projeter ses concepts psychologiques sur des processus qui en sont manifestement dépourvus : automobiles, chiens et chats donnent à leur propriétaire l'occasion d'exercer ses concepts mentaux. La suggestion de Dennett risque d'avoir pour effet de gonfler, au moins provisoirement, le domaine d'application des explications mentalistes. Nous allons voir toutefois que, sans un contrôle méthodologique très strict, les comportements ne parlent pas d'eux-mêmes.

Les méthodes de l'attribution intentionnelle

Les méthodes utilisées pour tester les attributions intentionnelles se répartissent en deux groupes : les observations des éthologistes de terrain consistent à relever des comportements individuels ainsi que leur fréquence, et dans la mesure du possible, leur contexte d'apparition. Si les conditions de l'observation ne peuvent par définition pas être contrôlées, elles ont le mérite de ne pas ou (très peu) interférer avec la vie normale de l'animal. Ainsi les éthologistes de terrain se flattent-ils d'obtenir des données qui ne peuvent pas être

recueillies en laboratoire. Mais du point de vue des expérimentalistes, ces observations ne livrent que des anecdotes sans valeur. "Aucune anecdote ne peut fournir de preuve convaincante que le comportement d'un animal a été influencé non par l'apparence ou le comportement d'un interactant social, mais par une attribution d'état mental formée à partir de ces stimuli", écrit par exemple l'expérimentaliste Caecilia Heyes (Heyes 1993). Même une pléthore d'anecdotes ne pourrait remédier à ce déficit capital de l'observation de terrain, qui réside dans l'absence de contrôle des indices comportementaux qui ont pu être utilisés par l'animal en lieu et place des concepts mentaux qu'on lui attribue.

La méthode d'*apprentissage discriminant* a été utilisée originellement par Premack et Woodruff pour tester l'existence présumée d'une théorie de l'esprit chez les chimpanzés. Elle consiste à tester les capacités mentales des animaux dans des situations qui semblent requérir l'utilisation de concepts mentaux, tels que "informer X de manière correcte", "mentir à Y". Le paradigme expérimental utilisé consiste à distinguer un moniteur compétitif ou coopératif, et à agir de manière appropriée. Le chimpanzé peut voir où se trouve la nourriture, mais ne peut pas l'atteindre. Le sujet humain ne peut pas voir où elle se trouve, mais il peut être guidé par le chimpanzé qui pointera vers la boîte. Quand un moniteur trouve la nourriture, il peut, dans une première situation expérimentale, la garder pour lui (moniteur compétitif) ou la donner à l'animal (moniteur coopératif). Premack et Woodruff ont montré que les chimpanzés réussissent, pour les plus doués,

à montrer la boîte vide au moniteur compétitif. Ils ont cru pouvoir en conclure que l'animal pouvait se représenter l'état de croyance d'un agent et ses conséquences pour l'action. Mais évidemment, il ne s'ensuit rien de tel. Il est méthodologiquement plus parcimonieux de conclure que l'animal a extrait une régularité plus simple "ne montrer la boîte pleine de nourriture qu'au moniteur coopératif": pas de concept mental, mais une simple application des principes sociaux de la coopération, que les animaux sociaux adultes maîtrisent tous à des degrés divers. Ce type d'expérimentation, remarque Heyes, n'est pas un meilleur test des capacités de mentalisation que ne l'est l'observation de terrain. Une femelle babouin toilette-t-elle un mâle qui a capturé une antilope *parce qu'elle veut le distraire d'un butin* ou le fait-elle simplement *à la première occasion qui se présente* ? Là encore, l'explication parcimonieuse est recommandée.

Une seconde méthode expérimentale critiquée par Heyes consiste dans la méthode dite de "*piégeage*" : il s'agit d'un test expérimental qui ne requiert aucun apprentissage antérieur, et paraît donc éviter les inconvénients de l'apprentissage d'indices comportementaux de type social. Cette méthode a été utilisée en particulier par Cheney et Seyfarth dans leur étude des macaques japonais (Cheney & Seyfarth 1990 : 220). Dans cette étude, on cherchait à mesurer la fréquence relative des cris émis par les femelles concernant la présence de nourriture ou de prédateur, visibles ou non de leurs rejetons dans une arène où ces derniers seuls doivent pénétrer. Le résultat est que la fréquence des cris ne change pas selon les situations, ce qui tendrait à faire penser que les macaques ne maîtrisent pas le

concept général de connaissance acquise par autrui. Heyes considère que ce résultat négatif ne prouve rien, en avançant que le résultat positif aurait pu être interprété soit par la présence d'un apprentissage associatif, soit par une attribution mentale. Toutefois, le résultat négatif paraît montrer que les femelles n'ont pas discriminé une situation dans laquelle les jeunes ignorent l'existence d'un danger ou d'un bénéfice d'une autre où ils en sont avertis ; cela permet au moins de conclure que ni un apprentissage associatif discriminant, ni une attribution mentale n'ont en l'occurrence influencé le comportement des macaques. Mais la méthode de piégeage paraît souvent piégée..

Une troisième méthode, dite de *triangulation*, consiste à articuler deux phases expérimentales. Un apprentissage à la discrimination est effectué, puis transféré sur de nouveaux objets. Quand elle est correctement appliquée, cette méthode permet selon Heyes de distinguer plus nettement les indices observables des indices "mentaux". Les entraîneurs humains peuvent savoir ou seulement deviner si un type de nourriture (convoité par les chimpanzés) se trouve dans une boîte A ou dans une boîte B ; Aucun d'entre eux n'a placé lui-même la nourriture dans la boîte ; le lieu où ils se tiennent est soigneusement contrôlé, et ils intervertissent les rôles d'un essai à l'autre. Ainsi les animaux ne disposent apparemment pas d'indices purement visuels pour reconnaître l'entraîneur qui sait où se trouve la nourriture. Malgré ces précautions, les animaux sont capables de transférer leur discrimination entre le moniteur qui indique la bonne boîte et celui qui en est incapable. Ainsi, dans un premier temps, les

chercheurs ont-ils conclu que les chimpanzés pouvaient "se représenter la perspective visuelle d'autrui" (Povinelli & col. 1992).

Une étude ultérieure les a conduits à changer radicalement d'avis (Povinelli 1996). Elle a montré que de jeunes chimpanzés ne faisaient pas la différence entre deux moniteurs, l'un capable de les voir, l'autre avec la vision obstruée, en particulier dans des situations non familières (par exemple, un entraîneur avec un seau sur la tête, l'autre sur l'épaule ; un entraîneur tournant le dos et regardant devant lui, l'autre le dos tourné mais regardant l'animal par-dessus son épaule). L'apprentissage antérieur suffit à expliquer pourquoi les chimpanzés préfèrent certaines postures pour communiquer ; il ne semble pas qu'une variable intermédiaire psychologique - la compréhension de la vision comme source de connaissance - joue un rôle dans leur comportement. En d'autres termes, la manière dont les chimpanzés comprennent le "voir comme faire attention" n'implique qu'une association motrice (regarder là où l'autre regarde) et non pas un processus d'acquisition de connaissance, qui serait l'amorce d'une théorie de l'esprit.

Les animaux non langagiers ont-ils une théorie de l'esprit ?

Il reste donc, considérations méthodologiques à l'appui, à savoir ce qu'il en est de la capacité des animaux de former des hypothèses relevant d'une théorie de l'esprit. Quels sont les comportements spontanés qui suggèrent que les animaux non-langagiers pourraient posséder une telle capacité ? Notons que le problème ne se pose véritablement que pour les grands

primates, et les mammifères marins tels que dauphins et baleines, encore assez mal connus. Seuls ces animaux sociaux ont également des capacités cognitives justifiant une hypothèse mentaliste à leur propos. Chats et chiens manifestent des comportements sociaux; ils ont un répertoire communicatif relativement étendu, et une aptitude développée par la domestication à interagir avec notre espèce ; mais ils sont loin de manifester les capacités imaginatives et la virtuosité du raisonnement d'un chimpanzé.

L'attention conjointe est un comportement qui paraît impliquer la reconnaissance qu'autrui a découvert perceptivement un objet d'intérêt. Le chimpanzé suit la direction du regard de ses congénères. Mais, comme on l'a vu plus haut, il n'en tire pas l'idée que l'autre *ait vu* un objet ou *connaisse* un état de chose ; il se borne à utiliser la direction du regard de manière pratique et non réfléchie, comme l'indice d'un élément à consommer ou à fuir. La compréhension d'autrui comme sujet d'expériences suppose davantage que le simple mécanisme de détection du regard. Il est remarquable à cet égard que les primates non-humains, à la différence des enfants humains, ne produisent pas spontanément de geste ostensif. Quoiqu'on puisse leur apprendre à montrer du doigt, ils utilisent le geste pour obtenir d'autrui un objet convoité, et non pour montrer un objet à autrui à seule fin de le contempler.

Nous revenons ici à la différence fondamentale entre l'exploitation d'une source d'information et la représentation de cette information comme ayant une source mentale. Même si le chimpanzé tire parti du dispositif inné par lequel il suit le regard de ses congénères, il ne sait pas que ses congénères ont

des états mentaux propres, liés à leur expérience acquise. Quoiqu'il puisse efficacement coordonner son action avec les membres de son groupe, il ne dispose pas de la capacité d'attention conjointe à laquelle les humains ont accès : celle-ci suppose que l'initiateur de la communication ait l'*intention de faire reconnaître* par autrui les propriétés d'un objet ou d'un événement qui sont au centre de son expérience, ainsi que *l'intention de poursuivre l'expérience de manière conjointe* en vertu de la première intention. Cette articulation entre contenus intentionnels joue un rôle capital dans la communication de type humain, mais fait entièrement défaut aux primates non humains.

Dans la mesure où les primates ne se représentent pas l'état des connaissances des autres individus, on estime généralement qu'il est impropre de parler d'*enseignement* à propos de comportements coopératifs. Par exemple, une mère chimpanzé retire des mains de son enfant une plante empoisonnée. S'agit-il d'enseignement ? de prévention ? Certes les primatologues de terrain observent l'existence de cultures spécifiques à une communauté d'animaux (par "culture", on entend la possession de savoir-faire locaux relatifs à la préparation de la nourriture). Par exemple, les macaques japonais de Koshima ont appris en quatre ans à laver leurs patates douces pour les débarrasser de la terre après qu'un seul individu ait inventé cette technique. Mais ce cas relève-t-il de l'enseignement, pratique intentionnelle dirigée vers autrui ? Rien dans le comportement des macaques ayant acquis la technique de nettoyage n'indique le moindre souci pédagogique. La lenteur de l'acquisition tend

à montrer que les animaux n'acquièrent l'innovation ni par enseignement délibéré, ni même par imitation. Là où l'observateur non averti croit voir le fruit d'un enseignement, l'éthologue repère un phénomène d'"intensification du stimulus" : le simple voisinage spatial entre un membre du groupe et l'objet-cible élève l'intérêt des congénères pour ce type d'objet et suscite chez eux des tentatives d'utilisation.

Ces explications "rabat-joie" s'appliquent à d'autres types de comportements de régulation sociale, tels que les gestes d'apaisement ou de réconciliation qui sont indubitablement une condition de possibilité de la coordination sociale telle qu'elle existe dans les groupes humains. Ces gestes ont pu être sélectionnés pour leur valeur régulative, indépendamment de la représentation mentale explicite de la manière dont le geste est compris par le congénère. Les humains font eux-mêmes de nombreux gestes sans savoir pourquoi ils les font, et souvent sans même remarquer qu'ils les font, tant ils sont liés à une situation typique (l'explosion de colère, ou la réaction à la colère d'autrui, etc. s'accompagnent de gestes précis, chez l'homme et chez le primate).

La tromperie tactique, dont nous avons parlé plus haut, recouvre l'ensemble des cas où il semble qu'un animal mente intentionnellement à ses congénères pour des buts privés, en modifiant les signaux qu'il donne de manière à tirer parti d'une situation qu'il est seul à connaître. Les cas de tromperie tactique abondent dans les groupes de primates. Un enfant chimpanzé peut, par ses cris, prétendre être attaqué par un adulte de manière à ce que sa mère chasse ce dernier et laisse l'enfant profiter seul d'un butin. Une femelle peut réprimer ses

cris de jouissance lorsqu'elle s'accouple dans les fourrés avec un mâle non-dominant. Un chimpanzé peut faire semblant de n'avoir pas remarqué un buisson couvert de baies pour y revenir quand le groupe sera passé. Après avoir, pendant un temps, conclu que les primates non-humains peuvent mentir, c'est-à-dire se comporter délibérément de manière à produire chez autrui des croyances fausses, les chercheurs admettent aujourd'hui que ces comportements peuvent s'expliquer par apprentissage des comportements efficaces dans une situation donnée. Il n'est pas nécessaire, pour produire les résultats attendus, de savoir que ses congénères agissent sur la base de leurs croyances. Il suffit simplement de découvrir le type d'actions à faire ou à ne pas faire dans telle ou telle circonstance. Quand un pluvier "prétend" avoir l'aile brisée pour éloigner le prédateur de sa nichée, il ne se représente pas non plus l'état d'esprit de l'attaquant. Il fait ce que la situation exige, en vertu d'une manifestation motrice innée déclenchée par la représentation de cette situation. L'hypothèse de l'existence d'une théorie de l'esprit chez l'animal non-humain ne se soutient plus aujourd'hui que d'anecdotes rassemblées par les propriétaires de grands primates domestiques.

Concluons. Les animaux non-humains sont capables de catégoriser le monde en fonction d'objets et d'événements pertinents pour leur survie, et de stocker en mémoire un grand nombre d'épisodes de leur expérience passée. Ils peuvent ainsi prédire dans une certaine mesure les états du monde, et construire une carte mentale de leur territoire pour orienter

leur recherche de nourriture de manière informée. Les mammifères, les serpents et les oiseaux peuvent former des représentations sur la base de la perception d'objets et d'événements distincts. Les grands primates se représentent le monde physique d'une manière voisine de celle d'un sujet humain qui n'aurait pas reçu de formation scientifique ni n'aurait hérité de son groupe une théorie dite naïve. Mais les pongidés eux-mêmes n'ont pas du tout la même façon de se représenter l'information mentale que les humains, faute de pouvoir former une théorie des états mentaux, ou de pouvoir conjoindre dans la même opération de pensée une représentation factuelle à la représentation d'une situation simplement possible ou entièrement imaginaire. Ils ne représentent pas l'information mentale comme mentale, mais comme une information comportementale.

Ce qui vaut de la compréhension d'autrui vaut aussi de la représentation de soi-même. Il est avéré que les grands singes sont capables de faire référence à eux-mêmes lorsqu'ils ont été entraînés à l'usage de symboles. En outre, à la différence des petits singes, la plupart des chimpanzés et des orang-outans adultes (mais non les gorilles) réussissent à se reconnaître dans un miroir. Mais on ne peut pas en conclure que les primates non-humains ont un sens de l'identité personnelle approchant du nôtre. Dépourvus de concepts mentaux, incapables de faire des inférences utilisant les dispositions mentales, et de surcroît libérés de la tâche de construire et de raconter une biographie socialement appréciée - à la mode humaine - les primates non-humains ont sans aucun doute une manière bien à eux de catégoriser leurs congénères, sur la base de

leurs caractéristiques physiques et comportementales et de leurs dispositions (rang dans le groupe, agressivité, etc.). Mais il faut rappeler que cette capacité de distinguer individuellement ses congénères s'applique à de nombreuses espèces sociales, y compris les poules.

Bibliographie

Allen, C., & Hauser, M.C., (1991), Concept attribution in nonhuman animals: theoretical and methodological problems in ascribing complex mental processes, *Philosophy of Science,* 58, 221-240; reproduced in Bekoff, M. & Jamieson, D., (eds.), *Readings in animal cognition,* Cambridge, MIT Press, (1996), 47-62.

Byrne, R., (1995), *The Thinking Ape, Evolutionary Origins of Intelligence,* Oxford, Oxford University Press.

Byrne, R. & Whiten, A., (eds.), (1988), *Machiavellian Intelligence : Social expertise and the evolution of intellect in monkeys, apes and humans,* Oxford, Clarendon Press.

Carnap, R. (1928-[1967]), *The Logical Structure of the World,* trad. par R. George, Berkeley, University of California Press.

Cheney, D.L. & Seyfarth, R.M., (1990), *How monkeys see the world : inside the mind of another species,* Chicago, University of Chicago Press.

Davidson D., (1982), Rational animals, in *Dialectica*, 36, 318-327, trad. P. Engel in *Paradoxes de l'irrationalité,* Combas, L'Eclat, (1991), 63-75.

De Waal, F., (1997) *Le bon singe. Les bases naturelles de la morale,* Paris, Bayard.
Dennett, D.C., (1971), "Intentional Systems", *Journal of Philosophy, 8,* : 87-106 ; reproduit dans *Brainstorms*, (Montgomery: Bradford Books, 1978).

Dennett, D. C., (1987), *The Intentional Stance,* Cambridge, MIT Press; trad. P. Engel, *La Stratégie de l'intentionnalité,* Paris, Gallimard (1990).

Drestke, F., (1988), *Explaining Behavior, Reasons in a World of Causes,* Cambridge: MIT Press.

Grice, P., (1969), Utterer's Meaning and Intentions, *Philosophical Review,* 78 : 147-77.

Hawkins, R.D. & Kandel, E.R., (1984), Is there a cell biological alphabet for simple forms of learning ?, *Psychological Review*, 91, 375-391.

Heyes, C., (1993), Anecdotes, Training, Trapping and triangulating : do Animals attribute mental states ? *Animal Behaviour,* 46, 177-88.

Heyes, C., (1994), Social learning in animals, categories and mechanisms, *Biological review,* 69 207-31.

Povinelli, D., (1994), Comparative studies of animal mental state attribution: a reply to Heyes, *Animal Behaviour,* 48, 239-241.

Povinelli, D., (1996), Chimpanzee theory of mind ?: the long road to strong inference, in P. Carruthers & P.K. Smith, *Theories of theories of mind,* Cambridge, Cambridge University Press, 293-329.

Premack, D., (1988), "Does the chimpanzee have a theory of mind" revisited, in R. Byrne & A. Whiten (eds.), *Machiavellian Intelligence,* Oxford, Clarendon Press :160-179,

Premack, D., & Premack, A. J., (1984), *L'Esprit de Sarah,* trad. Y. Baudry, Paris, Fayard.

Premack, D., & Woodruff, G., (1978), Does the chimpanzee have a theory of mind ? *The Behavioral and Brain Sciences,* 4, 515-526.

Proust, J., (1997), *Comment l'esprit vient aux bêtes, Essai sur la représentation,* Paris, Gallimard.

Russon, A.R., Bard, K.A. & Parker, S.T., (eds.), (1996), *Reaching into thought : the minds of the great apes,* New York, Cambridge University Press.

Sperber, D., & Wilson, D., (1989), *La Pertinence, Communication et Cognition*, Paris, Editions de Minuit.

Strawson, P.F., (1959), *Individuals,* London, Methuen and Co. ; trad. fr. par A. Shalom et Paul Drong, Paris, Le Seuil, (1973).

Whiten, A. & Byrne, R.W., (1988), Tactical deception in primates, *Behavior and Brain Sciences,* 11, 233-273.