'Does the chimpanzee have a theory of mind?' revisited
DAVID PREMACK

Over many years, Premack has shown an uncanny skill in using the tools of behaviourism to pose apparently mentalistic questions to captive chimpanzees. For this chapter, we asked him to assess the results of perhaps his most ambitious such attempt: can chimpanzees be said to understand the minds of others?

In 1978, we raised the question 'Does the chimpanzee have a theory of mind?' (Premack and Woodruff 1978), by which we meant, does the ape do what humans do: attribute states of mind to the other one, and use these states to predict and explain the behaviour of the other one? For example, does the ape wonder, while looking quizzically at another individual, What does he really want? What does he believe? What are his intentions? Notice that in raising this question we do not ask, Is the ape an intentional system? Does it have wants, beliefs, hopes, plans, etc.? We take it for granted that the ape is an intentional system and ask, instead, 'Does the ape think that other apes are intentional systems?' In other words, does the ape attempt to account for ape behaviour in the same way that humans attempt to account for human behaviour?

In even so much as raising this question, we already imply that language or speech is not a necessary condition for having a theory of mind. There is a widespread tendency to see language as the source of virtually all the 'interesting' properties of the human mind (e.g. Schwartz 1980; Bickerton 1987). We do not subscribe to this view. Many important competences, such as that of social attribution are, we believe, competences in their own right, not secondary properties derived from a competence in language. This is not to say that language cannot influence these competences. Language can amplify competences in several ways, the most obvious being that of giving the individual a greater awareness of the competence, thereby enabling him to use it in more powerful ways.
breaking the lamp, taking money from his mother's purse, or about lying itself. We would look closely at such a 'person', wondering whether it was child or robot. Deception in the ape and monkey is not so restricted; for this reason alone, the primate's behaviour is of interest with respect to the question of 'theory of mind', while the bird's is not.

One of the frequently reported cases of primate deception concerns the sexual behaviour of the male chimpanzee (Goodall 1986). Although orgasm in the male is ordinarily accompanied by a distinctive cry, when copulating with a favourite female the male sometimes suppresses his sexual cry. In this way he avoids having to share the female with others. For if other males were drawn to the scene, they would discover the receptivity of the female and, given the promiscuity of the chimpanzee, the male would lose his exclusive rights.

In calling this observation an anecdote, I do not mean to question the reliability of the report or even, for that matter, the general accuracy of the account as to the circumstance in which the act occurs. Although, in fact, these are serious problems surrounding the use of anecdotes, for the sake of discussion, let us assume that none of them apply. Let us assume that suppression of sexual cries does not occur except as sexual rivals are near by—the base level of the act is zero—and that the agreement among judges as to the occurrence of the act is respectably high. Given observations of this character, what more could one ask? Where is the need for experiment?

The need for experiments arises with regard to the interpretation we place on the act. In general, an individual can engage in social behaviour for either of two quite different reasons. In one case, he acts so as to affect what the other individual does; in the other case, so as to affect what the other individual believes. Only in the second case is the behaviour relevant to and evidence for a theory of mind; for only in the second case does the actor attribute a state of mind to the other one. How can we distinguish between the two cases?

We can picture the hypothetical situation in which the behaviour itself will tell the difference. For instance, suppose I do not want you to fish in my pond. I could bring this about in either of two quite different ways. I could attempt to influence your belief concerning the presence or absence of fish in the pond. I could assure you that there are no fish there, i.e. I could do so provided you spoke. However, speech or language is not a necessity. If you lacked language, I could substitute perceptual evidence: stir up the water and point out the absence of fish, or have you sit with me for hours, watching while I failed to catch anything. These observations, arranged specifically for your benefit, are likely to influence your belief about the lack of fish in the pond. In an alternative approach, I could ignore your belief and deal with your behaviour directly. Whenever I caught you fishing in the pond, I could punish you. Punishment will reduce the likelihood of your fishing in the pond and, if severe enough, eliminate it altogether. Now if we found two species, one of which specialized in the first approach (always arranged perceptual evidence for the benefit of the 'listener'), the other of which specialized in the alternative approach (did nothing but punish the 'listener'), we could reasonably entertain rather different hypotheses about the mentation of the two species. Only in the former would it seem reasonable to ascribe a belief in belief. Unfortunately, cases of this hypothetical kind—where distinctive behaviours invite distinctive interpretations—are hard to come by; actual social behaviour will not have this helpful character. Most acts that are carried out to affect what another one believes will not look any different from acts that are carried out to affect what another one does.

Consider again the ape who suppresses his sexual cry. Does he do this to affect the beliefs of rival males, or merely to affect what they do? Were human adults to behave in a comparable manner we would not hesitate to assign a rich interpretation. We would say that the actor knows that if he does not suppress his cry, his rivals will hear him; if they do hear him they will know what he is doing; and if they know what he is doing, they will want to do the same. This series of assumptions credits the actor with at least the following: he understands:

1. the distinction between seeing and hearing, recognizing that sound can be effective where vision is not;
2. that belief or knowledge depends upon perception;
3. that desire (want) depends upon knowledge or belief.

Which of the assumptions that we grant in the human case can be maintained for the chimpanzee?

It may be thought that, at this point, we could advance our cause by turning to other anecdotes, using one to disambiguate the other. Regrettably, this is not the case. All anecdotes are about equally lacking in resolving power so far as the present issue is concerned. Each of them can be interpreted frugally (he acts so as to affect what the other one does) or in the more profligate way. To resolve this issue we must turn to experiments.

Consider, for example, our first two assumptions:

1. the actor understands the difference between seeing and hearing;
2. he recognizes that knowledge depends upon perception.

Fortunately, both of these assumptions have been tested with young children. By the age of 2½, children appreciate the difference between seeing and hearing, recognizing the different conditions on which they depend (Flavell 1978). Do they take the next step, and understand that...
knowledge or belief depends upon perception? This is evidently a more demanding task for, though by the age of 3 children already attribute knowing/believing to the other one, they do not understand that knowledge depends upon perception. That is, they use the concepts of know and believe, but not in the adult manner, for they do not know how these states come about. They do not have a causal theory of knowledge.

When children are allowed (or not allowed) to see what is in a box, or are told or not told what is in the box, and then subsequently asked, 'Do you know what’s in the box?' they answer correctly. However, if they are shown the same conditions for another child—either he is or is not permitted to look, either he is or is not told—and asked the comparable question, ‘Does he/she know what’s in the box?’ they answer incorrectly. In his own case, the child apparently answers simply by determining whether or not he can answer the question, ‘What is in the box’, but in using this internal knowledge he does not understand how the knowledge came about—that it depends upon his having either seen or been told what’s in the box. Hence, he cannot answer the question in the case of the other child for, of course, he has no access to that child’s internal knowledge. Not until the child is over 4 years old does he develop a causal theory of knowledge (Wimmer et al. in press). Now what bearing does the child’s performance have on the claims we can make for the chimpanzee?

A considerable bearing, for in 20 years of comparing chimpanzees with children we have only one case in which a chimpanzee passed a test that was failed by 3½-year-old children (Woodruff et al. 1979). Barring the one exception, a good rule of thumb has proved to be: if the child of 3½ years cannot do it neither can the chimpanzee. In this case, moreover, we also have two tests that were done on chimpanzees. The results of these tests tend to agree with those from the children’s tests.

In the first of these tests, my former colleague Guy Woodruff put the juvenile apes into a circumstance in which they needed the assistance of the trainer to obtain food from a locked container. The trainer (carrying the key on a chain around his neck) willingly followed the animals across the ½-acre compound to the locked container provided that the ape did not dash off, but took care to ‘lead’ him, i.e. turned around, glancing back from time to time, waiting for him to catch up if he fell too far behind. Evidently, all four animals found this a natural form of communication, for they were immediately successful, regularly ending up at the container with the trainer only a few steps behind. He promptly unlocked the container, giving the animal the long-awaited banana. Then one day we blindfolded the trainer.

Now the blindfolded trainer did not follow the ‘beckoning’ animal, but being unable to see, remained seated on the ground. Three of the four animals, finding that they had headed into the field without being followed,
The test involved three steps, the first two of which were pretraining. In the first step, the animal was simply familiarized with two opaque containers that were placed before it; food was placed in one of them on each trial and the animal was allowed to choose between them. In the second step, the baiting was obscured from the animal’s view by placing an upright box before the animal and inserting the containers in the box. Although the animal could see the trainer place the containers in the box, take out a piece of fruit, kneel down behind the box, rise up with the fruit no longer in hand, it could not see the actual baiting. When the box was removed, leaving behind the containers, the animal was again allowed to choose; it chose willingly but now, of course, at chance level. In the third and essential step, a screen was placed perpendicular to the animal’s cage (sometimes on the right, sometimes to the left), and on each trial, two trainers took up positions on opposite sides of the containers. One stood on the side without the screen so that she could readily see the containers, whereas the other stood on the far side of the screen so that she could not see the containers. (The position of the screen, trainers, and of the baited container was, of course, varied from trial to trial in a counterbalanced fashion so that each condition occurred equally often on both sides.) The animal could readily see which trainer on a given trial had an unimpeded and which an impeded view of the baiting, and it was given an opportunity to exploit this knowledge. The box was removed revealing the two containers, and the screen was taken away; the two trainers then came forward and stood behind the containers equidistant from them. The animal chose between the two trainers by pulling either of two strings, one of which was attached to each trainer. The trainer chosen stepped forward and tapped one container or the other. The trainer who had witnessed the baiting always tapped the wrong container. Ordinarily, of course, the unwitnessing trainer would have responded at chance level, but we downgraded his advice in this manner to help the animal see the advantage of choosing the witnessing trainer.

All four animals readily adopted the practice of pulling the trainer forward by the string and, to a degree that varied over animals, followed the trainer’s advice, choosing the container that she indicated. Three of the four animals solved the problem in the sense of choosing significantly more often than chance the trainer who had witnessed the baiting. These animals chose correctly essentially from the beginning: if we divide the 24 trials given them into blocks of four, we find that they performed as well on the first block as on the last. However, one of the three animals who chose the correct trainer did not consistently follow her advice; she followed the trainer’s tapping only 15 out of 24 trials ($P > 0.05$) and thus could not be said to have genuinely solved the problem. A fourth animal neither chose the correct trainer above chance level nor followed her advice (a further possibility, that the animal choose at chance level and yet follow the trainer’s advice was not found). Hence, only two of the four animals solved the problem fully, both choosing the trainer who witnessed the baiting and following her advice; two others failed, either fully or in part.

These results suggest that two of the four juveniles understand the conditions on which seeing depends and recognize the advantage that seeing has for knowing, i.e., for obtaining the kind of information that the animal wants. This is a strong interpretation to place on the results of one test, of course, and we need controls to eliminate weaker alternatives. For instance, we must deal with the possibility that the animal simply has a preference for the trainer who sees the baiting — and will always choose this trainer whether or not it needs the information the trainer possesses.

Suppose the animal has observed the baiting and therefore knows which container is baited; nevertheless, when offered a choice between trainers who did and did not witness the baiting, it still chooses the witnessing trainer. In fact, these are the results we obtained with the two animals who were fully successful on the earlier test. When we eliminated the box that enclosed the containers — thus restoring the baiting to the animal’s view — and repeated the experiment, two of the animals continued to choose the trainer who witnessed the baiting. What does this mean?

It need not be taken to mean that the animal acted out of ‘blind’ preference for a ‘seeing’ trainer, and has no comprehension of the relation between seeing and knowing. In a choice between trainers who do and do not know, why not choose the knowing one? (The animal might even realize that it has difficulty avoiding a trainer’s advice, and therefore should not choose the ignorant trainer.) The point is, choosing between the trainers cost the animal nothing. We must arrange an experiment in which this choice is costly, and then determine whether the animal will elect to pay for the right to seek advice only when its view of the baiting is blocked and it does not know which container to choose.

Although, as things stand, we cannot definitively interpret the results of the present tests, for the sake of discussion I shall proceed as though they can be interpreted and will treat the outcome for the two successful animals as defending the strong claim. Then the implications of these results for the interpretation of the sexual anecdote obviously depend on the animal; for the successful animals they are of one kind, for the unsuccessful another. Consider first the implications in the case of the unsuccessful animals.

The rich interpretation we would give in the case of comparable sexual behaviour by a human adult, ‘The individual suppresses his cry because he knows that if he is overheard his rivals will know what he is doing’, etc., is unwarranted. For a creature who does not understand the relationship between perception and knowledge, the most we can say is: ‘The individual suppresses his cry because he knows that if he does not do so his rivals will
rush in and destroy his alliance’, etc. This claim entails a relationship between two pieces of behaviour, suppression of cry by the one individual and interference by others, and has no implications for a ‘theory of mind’. The appreciation of the causal relationship between the two pieces of behaviour does not entail any attribution, neither of want, belief, or any other state of mind. Moreover, we could give a still weaker account, one that did not credit the ape with causal understanding of any kind.

How could an animal, without attributing any state of mind to the other one, learn the correlation between his sex cry and interference by other animals, and thus learn to suppress his cry? The standard answer of the tough-minded learning account is well known. The sex cry is the last act in a sequence that began with courtship (presentation by the female) and culminated in copulation. The arrival of the other males and their taking over of the female is aversive. Aversive events, by definition, suppress the occurrence of preceding acts, especially the act that is last or most temporally proximal to the aversive event.

If this account were correct, then close observation might reveal that the sex cry was not the only act suppressed. Suppose that the ape, while copulating, clung with one hand to an overhead branch (as is sometimes shown in videotapes of apes copulating); interference by rivals could lead to suppression of this act too. That is, if ‘last’ acts are especially prone to suppression by aversive events, grasping-the-branch-with-right-hand might be no less subject to suppression than the sex cry itself. If the data actually bore this out, we would be obliged to take seriously the traditional learning account. On the other hand, if the sex cry is uniquely prone to suppression, the learning account would be in trouble. For learning can explain the suppression of an act only on the grounds of its temporal proximity to the aversive event; all ‘last’ acts must be equally prone to suppression.

A cognitive treatment of the learning may be more compelling. On this view, the sex cry is not like other acts but has special status (after all, clutching a branch will not give one away to other animals). The animal recognizes the unique connection between the sex cry and the likelihood that others will find him out. When he hears someone else’s sex cry—and begins to run towards its source—he runs with a definite expectation of what he will find: animals are copulating (he may even be able to picture which animals depending on how intimately he knows the social group). Knowing that other animals’ sex cries bring him running, he infers that his sex cry will have the same effect on them. Consequently, when he copulates (and does not want to be disturbed by rivals) he suppresses his sex cry (more so than other acts).

Notice, incidentally, that the two accounts are not mutually exclusive. It is often assumed that if cognitive variables apply, conditioning or learning ones do not, and vice versa. This view fails to recognize that we may find,
produced by reasoning of this kind should be quite unemotional. I would think, and therefore readily confinable to just the right occasion. It would not spill over onto occasions when rivals were not present. Moreover, on the assumption that copulation-with-cry is more enjoyable than without cry (suppression presumably would detract from the pleasure) if there were no countermanding force (no rival in the vicinity) we should expect to hear the sex cries. If, however, we find suppression even though rivals are not present, we might consider that the suppression was not the exclusive product of analogical reasoning, but was based at least in part on painful learning. On some occasion the animal must have failed to suppress its cries and paid the price, losing its exclusive prerogatives. We may be advised to adopt this view in any case—even without the hypothetical results of our Gedankenexperiment—for analogical reasoning is not a standard practice of the chimpanzee. We find it only in the language-trained chimpanzee (Premack 1984; Matsuzawa and Premack unpublished data). Hence, even apes who understand the relation between knowledge and perception may nonetheless be unable to anticipate the results of their sex cries. They may have to suffer one or more painful experiences before learning to suppress the cry, and only come to understand the relationship between the two events after the fact.

Consider the analogy we looked at earlier. 'I interfere with him on the occasion of his sex cries, he will do the same to me', etc. The analogy as written is strictly behavioural or non-attributional: it makes no reference to the other one’s mind. It could just as well be written in an attributional form, however. For example, 'When I hear sex cries, I expect to find some one copulating; when others hear my sex cries they will expect to find me copulating', etc. Having given thought to his own expectations, the individual now attributes expectations to the other one—giving the analogy an attributional form. To ask which is correct, the attributional or non-attributional, is, of course, to ask the question we are trying to answer.

Perhaps we could gain some leverage on this question by combining the results of field and laboratory. Our current laboratory finding is that some adolescent apes appear to understand the relationship between knowledge and perception. Suppose, in addition, that unless the animal is at least pre-adolescent—6 or 7 years old—we do not find suppression. Suppression not only of sex cries, of course, which naturally would be confined to the older animal, but suppression of any kind so long as it qualified as deceit or deception. If pre-adolescent animals neither showed deceit nor understood the relationship between perception and knowledge, we might suppose that a causal theory of knowledge was a prerequisite for the practice of deception; but this is almost certainly false. Children show both deceit and
learning; whereas advanced cognition, such as analogical reasoning, is confined to the laboratory.

Belief is special

So far we have proceeded as though all states of mind were of comparable complexity—entailed the same psychological processes—such that if an individual instantiated or attributed one state of mind it would instantiate or attribute all of them. This is almost certainly false. Some states of mind are more complex than others. Belief, for example, is a more complex state than is perception or desire. Though the distinctions we require here are controversial and suffer from too close contact with the vernacular, we can make some progress simply by dividing all states of mind into two groups, simple and not so simple. Simple states are those produced by processes that are hard-wired, automatic or reflex-like, and encapsulated (for enlightening discussion see Pylyshyn 1980; Fodor 1983). In perception, for example, the prototypic simple state, the proximal stimulus leads more or less inexorably to the end state, with little or any input from ‘outside’ sources. While perception is the prototypic simple state, we may add others: first, certain basic motivational states; and secondly, somewhat more controversially, expectancy, a state that is produced by conditioning or simple learning. These three states—seeing, wanting, expecting—have in common a restricted and automatic production process that is independent of language both at the level of input to the system and of internal representation. We might call all such states sensory—even though this entails stretching ‘sensory’ a bit—to underline their simplicity and independence of conceptual processes.

Complex states, of which belief is the prototype, are of course everything that simple states are not. Belief is not automatic, encapsulated, or hard-wired; moreover, it definitely depends on language, most certainly at the level of the internal representation though often also at the level of input to the system. Belief is not a completely unitary entity; we can distinguish at least two forms. In its simplest form, belief concerns the reliability of sensory states. Ordinarily, sensory states lead directly to action, but an individual (or species) may reach a stage of development where it calls these states into question. ‘Do I really see X?’ ‘Do I really want X?’ ‘Are my expectations of X well founded?’ The discovery that one’s sensory states are subject to unreliability has a social parallel: the discovery that the information told one by a second party is subject to unreliability. Both cases lead to the same outcome. The action to which these states would normally lead is deflected, and the states themselves are examined in the light of available evidence. The examination can be quite simple, as in the case of a child who is just discovering the disparity between seem and real (e.g. Flavell et al. forthcoming), or quite elaborate as in the case of a sophisticated adult. However, even when simple the process is not hard-wired, automatic, or encapsulated, and the mental representation on which it depends has a discursive or language-like form. Normally, the process culminates in a decision either to accept or reject the information that one is given—by one’s own sensory states or by another party—and this decision is belief in its simplest form.

Have any non-human species reached a stage of development such that they do not invariably act directly on the information given them, but sometimes ‘call to question’ the information itself? We answer ‘no’ unhesitatingly for at least some species, e.g. we do not believe that bees debate the reliability of the information encoded in the dances of their conspecifics. The bee does not, we think, distinguish between conspecifics whose dance it believes and others whose dance it doubts or rejects. So much for the bee. By the time we reach the chimpanzee we are no longer certain. There are field reports claiming that the adult ape responds differently to the food grunts of mature and immature conspecifics. Given the food grunts of another adult, the ape is said to climb directly up the tree without so much as a glance; but given the food grunt of a juvenile, the animal is said to pause, to peer up into the branches, having a check for itself before possibly wasting its time climbing a tree for which the message is unreliable (Nishida 1987). As in the case of all anecdotes, however, we do not know what interpretation to put on this behaviour. Is it the result of simple learning, a discrimination based on good experience with old conspecifics, mixed experience with young conspecifics? ‘Mixed’ experience can produce exactly the kind of vacillation that is described for the one stimulus. If so, this is no more evidence for the informed investigative process culminating in belief (simple form) than is any other case of vacillation. That is, not every species that vacillates does so because it has discovered the unreliability of sensory states or social messages. We recognize a by now familiar bind. There is no experimental work on this issue and, therefore, there can be no sound judgement.

Perhaps we should fall back on our rule of thumb: capacities that do not appear in the 3½-year-old child will not be found in the ape. Children are about 4 years old when they begin to distinguish between ‘what X looks like and what X really is’, thus offering at least some evidence for the discovery of the unreliability of sensory states. Taken in the light of our informal rule, these data imply that the discovery of sensory unreliability—and thus of belief even in its weak form—lies beyond the chimpanzee, requiring a stage of development the species does not reach. Of course, the conclusion is risky or inadvisable. The children’s data are based inescapably on a test that makes specific demands as to level of
consciousness or degree of knowing. If we relax those requirements we may obtain positive evidence at an earlier age. There is no substitute for testing the animal itself (not at least until our theories of test demands become more than the informal suggestions which they presently are).

Belief has a more advanced form, one that goes beyond questioning the reliability of information. In its advanced form belief is the decision to accept as an explanation (of one phenomenon or another) conditions that do not depend on sensory states. Familiar examples (in our culture) include: the earth is round, the soul is immortal, species evolve, germs cause disease, God is omniscient, the mind is in the brain, and so forth. Although some of these beliefs can be more easily driven back to sensory states than others, none of them really depend on sensory states. The vast majority of people who believe, for example, in the germ theory of disease and in the immortality of the soul have never seen either germ or soul; nor are they waiting to see them, suspending their belief in one or the other until such time as they do. Belief of this kind has two essential prerequisites. First, a high level of language competence not only for the sake of internal representation, but also for the communication of the information that is typically the source of the belief. Secondly, a conscious high level use of explanation based on causal theory, i.e. the belief that one event causes another. Indeed, the belief that events are caused is itself an example of the present form of belief: it differs from more ordinary beliefs only in that it is somewhat more species-specific (the genetic-experiential mix varies from one belief to another) and it is obviously foundational, a belief that makes other beliefs possible. Both these requirements and particularly their combination make it unlikely that we shall find belief of this form in non-human species, though a definitive judgement awaits the experimental analysis. We may find interesting precursors.

A special variant of the advanced form of belief is to be found in the personality traits we assign to one another. Although this case does not introduce any new properties—it is a classical case of belief pretending to depend more on sensory states than it actually does—it has the virtue of enabling us to contrast belief with expectancy, and thus of showing how much stronger the one state is than the other. In humans, a social encounter can produce either belief, expectancy or both. Thus, Bill, having encountered John, may be led to expect that if he does X, John will do Y. Alternatively, Bill’s encounter may lead him to believe that John is a particular kind of person and, therefore, likely to do not only Y in situation X, but a diversity of things in indeterminately many situations, some of which Bill can specify, others of which he can only dimly sense. In the human, these are not mutually exclusive states; he may have expectancies (‘narrow predictions’) based on conditioning, beliefs (‘broad predictions’) based on more cognitive processes. The question is whether the chimpanzee resembles the human in having both, or is unlike the human in being restricted to expectancy. One way to find out is to examine the chimpanzee’s anticipations of another one’s behaviour. Are they broad or narrow? Dependent on specific situations or largely independent of situation?

We tested this, in a sense, inadvertently, when we accidentally allowed the four young apes who participated in the benign/hostile trainer experiment to encounter the hostile trainer outside the experimental room. He was passing through the hall, still wearing the costume identifying him as hostile trainer (he had removed his mask, but not his neck-to-floor white gown) when the animals chanced to be taken out of the test room. Those of us who beheld the encounter were taken aback by the animals’ acceptance of the hostile trainer, the readiness with which they either climbed upon him or took his hand and were led off to the compound. The observation taught us that chimpanzee view of liar (the hostile trainer regularly pointed them to the unbaited container) is not necessarily the human view. Humans are inclined to believe that someone who would intentionally misdirect them, who would lie to them, is a bad person generally and should be avoided or dealt with warily. The young chimpanzees appeared to have a much narrower view; their negative expectations concerning the hostile trainer did not appear to extend outside the test room. Perhaps this is only because of their extreme youth; three-year-old children might show a similar ‘innocence’ (if so, I should be greatly interested in the possible developmental transition from expectancy to belief in the child). Sarah, when tested in a manner comparable to that of the young animals, reacted far more aggressively, hurling objects at the hostile trainer from below the mesh of her cage at a speed so dangerous we terminated the experiment. But this aggression took place in the test room itself during the experiment proper; we have no idea how she might have reacted to him outside the test space. Once again, to resolve the issue we need experiments not serendipitous comparisons or anecdotes.

**The chimpanzee’s theory of mind**

The immediate implication of the preceding section for present purposes is this: if the chimpanzee does have a theory of mind, it will be weaker than the human one. We have seen that the states of mind the chimpanzee is most likely to instantiate are the sensory ones—seeing, wanting, expecting. Belief is more doubtful, the advanced form especially, but even the weaker one concerning the reliability of information. Now a species will not attribute to others states of mind that it does not instantiate itself. Hence, the only states of mind the chimpanzee may attribute—if it attributes any at all—may be the simple ones—seeing, wanting, expecting.
Anecdotes aside, the evidence we have for evaluating theory of mind in the ape (or any animal) is painfully thin. We have two experimental paradigms, one previously reported (that I will summarize briefly in a moment), a new one that I will describe in a later section. Both have their limitations, the former being weak, designed to avoid false negatives (but at the possible expense of false positives), the latter stronger and designed for the opposite reasons.

In the original paradigm, the animal was shown videotapes of a human actor in a cage experiencing ‘problems’ of one kind or another. In one series, he encountered food that was inaccessible in various ways, and in another series, equipment that was deficient in one respect or another. The videotape depicting the actor in the throes of the problem was put on hold, and the animal was given either two or three photographic alternatives, one of which constituted a solution to the problem. For example, the actor was shown struggling to reach bananas overhead, and the alternatives consisted of the actor stepping onto a chair in one case and reaching to the side with a stick in another. The animal’s task in each series was to pick one of the photographs and place it in a designated location.

Sarah performed well above chance, picking the correct alternative on the first trial on nearly all problems (10/12 correct on one series, 12/14 on another). How does behaviour of this kind support the claim that the ape attributes state of mind to the actor? Notice that the selection of solutions presupposes problems, and a problem is not physically instantiated by the videotape. A videotape is merely a sequence of events, e.g. an actor jumping up and down in a cage with bananas overhead. That the actor is not merely jumping up and down, but ‘wants’ the bananas and is ‘trying’ to get them, etc., is an interpretation of the actor’s behaviour, one the animal will pick, I assume, when presented the tape, stopping it exactly where the actor’s next act is hitching up his pants. We offer as alternatives: the actor stepping up onto a chair, reaching out with a stick, and hitching up his pants. What will Sarah choose? The animal will pick, I assume, not next acts, but relevant next acts where relevance is defined by the intention the animal attributes to the actor. To obtain inaccessible bananas, stepping up onto a chair is a relevant act; hitching up one’s pants is not. A sophisticated individual such as human adult or even perhaps an older child could, I assume, be taught to pick specifically next acts no matter how irrelevant, but it would be a difficult idea to teach a chimpanzee.

New paradigm

The new paradigm was designed to avoid the inferential indirection of the old one, to obtain more direct evidence (direct, even though one can’t speak to the animal). Sarah was the subject again, and again she was tested in home cage. We mounted a cabinet on the wall immediately in front of her cage in a position fully visible to Sarah. The cabinet was divided into left and right halves, and painted white and black, respectively. The left side was stocked exclusively with good things, the right with bad. Good consisted of pastries that a favourite trainer, Bonnie, shared with Sarah during a daily tea time. Bad consisted of rubber snakes, putrid rubber rubber and a cup of faeces; items so untoward that Bonnie did not merely feign dislike of them, but signalled her disgust by quite genuine gestures including wearing rubber gloves when handling the material.

Both for the intermittent restocking of the cabinet and the daily tea time, Bonnie needed Sarah’s co-operation. For she could not open the cabinet door alone. The lock on the door was controlled by Sarah; only when she
pushed a button inside her cage was the trainer able to open the door. Sarah's latency on opening the door (pushing the button) once Bonnie entered her cage area was almost immediately stable. After three of the daily visits by Bonnie, Sarah's latency fell to an average of about 7 seconds and so remained for the rest of the test.

On day 18 we introduced the first experimental variation. A 'villain' (masked and concealed in a gown) entered Sarah's room, forced the cabinet door with a crowbar, took out all the material, placing them on the floor before Sarah, and then replaced them, totally reversing their position in the cabinet. Videotapes confirmed the 'villain's' report: Sarah responded hostilely, throwing things out under the mesh of her cage at the 'villain'. Bonnie arrived 15 minutes after the 'villain' had left, at her usual time.

What was Sarah's response? Did she greet Bonnie in an unusual manner? More important, did she hesitate in opening the cabinet door, i.e. in pushing her button? One might have thought she would because she was in a position to know, first, that the location of the goods was now reversed, and secondly, that Bonnie did not know this (because, of course, the 'villain' had made the change in Bonnie's absence). Any one knowing this would see that Bonnie was in jeopardy: she might put her hand where it did not belong! And an individual taking this into account might be expected to show some hesitation in opening the door.

However, Sarah did not show any change, either in general demeanour or in latency on pushing the button (she did only one aberrant thing: after Bonnie had already opened the door, Sarah pushed the button a second time), and she showed the same equanimity following four subsequent intrusions by the 'villain'. Either Sarah did not know the pertinent facts, knew but did not care, or both.

Negative outcomes seldom lend themselves to diagnosis. Sarah could have failed for any of dozens of reasons, but one possibility is more interesting than the others. To react appropriately on this test, i.e. to show evidence for theory of mind, Sarah must have had separate representations of what she knows and of what she knows the trainer knows; this was not required by the first paradigm. What Sarah knew and what the actor knew did not differ. For example, both Sarah and the actor could see that the bananas were out of reach, the cord was not plugged in, the flame was out, etc. There was no disparity in the knowledge of the two parties and, therefore, no need for separate representations, or strictly speaking for a representation of a representation. Sarah could have represented the actor's knowledge or perception simply as being the same as her own; she was not obliged to represent her knowledge on the one hand, and her knowledge of the actor's knowledge on the other. On two earlier occasions we applied the old paradigm to test material different from the original

material in that it did require that Sarah represent separately her knowledge and her knowledge of the other's (different) knowledge; she failed both tests, perhaps not accidentally (Premack and Premack 1983).

There is, on the whole, only suggestive evidence for theory of mind in the chimpanzee (or any non-human). Moreover, even the positive evidence leaves open the possibility of a theory of mind weaker than the human one. First, should the ape attribute any states at all, these are certain to be a small subset of those the human attributes. Secondly, the ape may be incapable of attributing different states of knowledge than its own. We may wish to distinguish three degrees of theory of mind: (a) species that make no attributions of any kind, presumably the case for the vast majority of species; (b) species whose attributions are unlimited in any respect except perhaps for number of embeddings (e.g. John thought that Mary believed that Bill thought that . . .), presumably the case for humans (by the time they are 4 years old) (e.g. Wimmer and Perner 1983); and (c) species that make attributions but attributions that are limited in a number of respects, possibly the case for the chimpanzee.