Theories of theories of Mind
P. Carruthers + P. Smith (eds)
Cambridge Univ Press (1996)

17 When does smart behaviour-reading become mind-reading?

Andrew Whiten

1 Introduction

The question of whether or not an individual is discriminating between others' states of mind is commonly addressed through a contrast with the alternative that it is merely discriminating between others' behaviour patterns. This is a frequent point of debate in the case of pre-verbal infants (e.g. Perner, 1991a, p. 128) and non-verbal animals (e.g. Cheney and Seyfarth, 1990b, p. 235), where it is usually assumed that mind-reading is a more advanced cognitive achievement than behaviour-reading, and that the latter will precede the former in either evolutionary or ontogenetic mental change.

However, mind-reading is not telepathy. So, the recognition of another's state of mind must somehow rest on observation of certain components within the complex of others' *behaviour patterns* together with their *environmental context:* that's all we can see – we can't see their minds in the direct way suggested by the idea of telepathy. This means that the contrast of mind-reading with behaviour-reading is not so straightforward as it may first appear: mind-reading, one might say, must be some sort of 'behaviourism'! At least, it must reflect some special form of behaviour analysis – special because it must differ from what we typically consider to be the mere perception of behaviour patterns (Whiten, 1993; 1994). Thus, rather than ask where mind-reading differs from behaviour reading, I shall tackle the question of when behaviour-reading *becomes* mind-reading, regarding the latter as some sort of sub-category of the former.

The question of what shall count as making this transition is surely a profound one. It underlies the debates which have occupied so much of the present century over whether the academic/scientific version of psychology (as opposed to folk psychology) should be mentalist or behaviourist, and how it can, in the mentalist case, be objective. The question is of fundamental importance in our attempts to specify both the development of human mentalism from its origins in infancy and its evolutionary origins as inferred from comparative studies of other

species. It is also of relevance to a much wider range of disciplines including philosophy of mind, social anthropology, psychiatry, and law, which are concerned with issues like recognising deception, or intent to harm, in an individual's actions.

2 What would non-verbal mentalism look like?

In the case of verbal humans, the mental states which are the 'building bricks' of folk psychology are neatly labelled (beliefs, wants, and so on) and conventionally defined well enough in each linguistic culture that we can tell each other about our mental states, and the mental states we may assign to others. Mental verbs like *think* can be shown to be a special class, distinguishable from action verbs like *hit* in logical aspects of the way they relate to their referents (referential or logical opacity). We need not delve into the technicalities of this here (see Dennett, 1988, for a clear exposition) because, although the exercise helps to delineate a list of mental states humans have found it useful to label, it does not in itself help us to recognise the *non-verbal* mentalist in any direct way.

What the discovery of referential opacity has done, perhaps, is to demonstrate the possibility of a sharp mental/behavioural divide: a point of view which according to Preston (1993) 'appears to be tacitly and nearly universally accepted among philosophers of mind and psychology', and which is summed up in a quotation from Fodor (1968b, p. 55): 'The distinction between mentalism and behaviourism is both exclusive and exhaustive. You must be either a mentalist or a behaviourist.' But Preston goes on to show the senses in which the two are in fact neither mutually exclusive nor conjunctively exhaustive. With respect to exclusivity, she argues that each approach, as applied by psychologists, tends in practice to incorporate some element of the other.

When we turn to the case of potential non-verbal mind-readers, like preverbal infants (Trevarthen, 1977, 1993; Reddy, 1991) and non-verbal apes (Premack and Woodruff, 1978; Whiten, 1993), the blurring of the distinction becomes more acute. We cannot appeal to the analysis of language which in the case of referential opacity appears so elegantly to pick out 'the mental'. In a group of chimpanzees there is no conversation about what Flo thought Fifi wanted, just individuals behaviourally interacting with each other. If they are mentalists, how would their behaviour look, compared to if they are behaviourists? Where we are tempted to say *A responded to the mental state of B* (the state of B *wanting* A's banana, for example), shall we not always be faced with the problem that A responded directly to whatever was the evidential basis for the mental state (probably B's behaviour and/or the situation it faced)? And when, to look at the reverse causal pathway, we

are tempted to say *A attempted to change the mental state of B so it would perform act X*, shall we not have to concede that A was just as likely to be attempting *to get B to do X*? More fundamentally, what makes the mental and non-mental alternatives really different in practice? The same issues arise in the case of pre-verbal infants.

These dilemmas remain in the case of experimental tests. Thus, when Premack and Woodruff (1978) showed that a chimpanzee would choose a photograph depicting the correct behavioural solution to a problem faced by another individual, the criticism was made that perhaps the chimpanzee simply knew, not the individual's *purpose*, but rather just what was the next thing to be *done* by such an individual in those circumstances (Bennett, 1978).

When, then, would it become valid to say that a non-verbal creature was reading behaviour in a way which made it of real interest to say they were mind-reading? Below I shall consider several alternatives. They do not allow us simply to say, *that's mentalism* or *that's not*: rather, there are grades of mind-reading (Whiten, 1994). The point is that some grades are non-trivially, interestingly different from others which are more easily thought of as 'just behaviour-reading'.

3 Implicit mind-reading

In our language of mental terms, we *explicitly* recognise and differentiate states of mind: states have specific labels, like *belief* and *desire*. But perhaps one could investigate whether mental states must be acknowledged as *implicit* in certain patterns of action in others which some non-verbal animals respond to. This point of view has been elegantly expressed by Gómez (1991). Gómez studied a young gorilla faced with the problem of shifting a door-latch which was out of her reach. Initially, she used a human caretaker as an object, pulling them to the door and then climbing on them to reach the latch. Later, a more interesting strategy developed in which she gently led the human towards the door, alternating her gaze between his eyes and the latch, and then taking his hand a little way in the direction of the latch. Humans easily interpret this as a request.

In this second strategy, Gómez remarks, 'she seemed to use eye contact to monitor if the human was *attending* to her request that he *acted*. Thus, she seemed to understand that in subjects, perceiving is causally related to acting. And here is where the mind appears, since the co-ordination between perceptions and actions is carried out by the mind' (p. 201). This phenomenon may be important as a potential precursor of the more sophisticated discriminations which develop in children, between states of ignorance, knowledge, and false belief, for in all of these the

observer translates from observed relationships between an individual's perceptual access in relation to environmental circumstances, to their later, causally linked, actions. From this perspective, then, it is an important insight to recognise the ape as performing a certain grade of (implicit) mind-reading. However, we must acknowledge that what the ape apparently recognises can be economically described in terms of direct observables: i.e. the relationship between the human's gaze behaviour and the latch, and the later, contingent action of opening the latch.

This appears similar to the 'behaviour-reading' interpretation of the experiment of Premack and Woodruff mentioned earlier. It also appears to be compatible with a conception of mind-reading which has been developed by ethologists, who have noted the usefulness to one animal of being able to predict the future behaviour of others with whom it must to do business, whether this be mating, fighting, or a host of other interactions where the outcome bears on survival and reproductive success. Krebs and Dawkins (1984) noted that: 'Animals can, in principle, forecast the behaviour of other animals, because sequences of behaviour follow statistical rules. Ethologists discover the rules systematically by recording long sequences of behaviour and then analysing them statistically, for example by transition matrices, and in the same way an animal can behave as if it is predicting another individual's future behaviour.' And they go on to suggest that:

> we may use the term mind-reading as a catch-word to describe what we are doing when we use statistical laws to predict what an animal will do next. For an animal, the equivalent of the data-collection and statistical analysis is performed either by natural selection acting on the mind-reader's ancestors over a long period, or by some process of learning during its own lifetime (Lorenz 1966). In both cases, 'experience' of the lawfulness of the behaviour of victims becomes internalised in the brain of the mind-reader. In both cases its mind-reading ability enables it to exploit its victim's behaviour by being 'one jump ahead of it'. The mind-reader is able to optimise its own behavioural choices in the light of the probable future responses of its victim. A dog with its teeth bared is statistically more likely to bite than a dog with its teeth covered. This being a fact, natural selection or learning will shape the behaviour of other dogs in such a way as to take advantage of future probabilities, for example by fleeing from rivals with bared teeth . . . mind-reading refers to a role that an individual can assume' (pp. 386–7).

I quote this fully, first, because it is important that all readers interested in mind-reading understand the breadth of meaning of the term as it is used in this literature, and Krebs and Dawkins' review represents an authoritative view in ethology. Second, it well expresses a hypothesis about the functional utility of mind-reading in predicting others' future actions, which is often put forward as the power and raison-d'etre of human mentalism (e.g. Premack and Woodruff, 1978; Wellman, 1991).

However, the quite direct sense in which mind-reading as defined by Krebs and Dawkins is a form of behaviour-reading is clear from the quotation above. Notwithstanding my earlier comment about the positive value of Gómez's interpretation of his gorilla as evidencing an implicit theory of mind, the gorilla's action might also, and as easily, be accounted for as 'using statistical laws to predict what an animal will do next', these laws referring to linkages between the human's gaze and actions on the latch.

At an even more general level, it could be argued that mind is implicit in much, or even (according to how broadly we define mind, as the causal mechanism for animals' behaviour) most animal behaviour: so to respond to another's threat behaviour, for example, is to respond implicitly to their aggressive state of mind. In this sense, implicit mind-reading is common, but this approach to mind-reading would not take us interestingly beyond behaviour-reading. Gómez's handling of the idea of implicit mind-reading in the particular case he discusses is important for the reasons discussed earlier, but we need to look further afield for a more general solution to identifying non-verbal mind-reading.

Before moving on, however, we should note what appears to be a rather different sense of 'implicit mind-reading' which has recently been applied by Clements and Perner (1994). Clements and Perner have shown that children may fail a standard test of false belief attribution (in which they are required to predict where a person will search for an object moved without their knowledge), yet their eye gaze indicated a correct prediction. They suggest the latter shows an 'implicit understanding of belief', contrasting it with a lack of 'explicit' understanding as revealed in verbal responses. This may be a useful measure to apply in experiments with non-verbal subjects like non-human primates. However, it would appear to be equivalent to any other measures one could use with such subjects: since they cannot give an explicit response in the form of speech, non-verbal measures like gaze *must* be used to record the distinctions they make. It is not obvious that gaze is any different to other behaviours which have been used to this effect, such as pointing (Povinelli *et al.*, 1990) or vocalising (Cheney and Seyfarth, 1990b). Thus, Clements and Perner are distinguishing 'implicit' by reference to how the putative *mind-reader* demonstrates behaviourally that they are indeed mind-reading; whereas in the earlier discussion of this section I have been concerned with whether the distinctions the putative mind-reader makes are about *others'* behaviour patterns, or mental states. Whether the subject uses gaze or speech to indicate where she thinks a subject in the situation used by Clements and Perner will search is a different question from what she is discriminating: is she discriminating states of mind, or does she just know

how individuals will behave in these circumstances (e.g. will search where they last saw the object)?

Thus, whichever sense of 'implicit' versus 'explicit' is applied, the behaviour-reading versus mind-reading distinction has not been satisfactorily cracked.

4 Recognising deception

Arguably the more interesting aspect of the Krebs and Dawkins analysis outlined here lies in their reasoning about the nature of communication, which they suggest may explain the *evolution* of mind-reading. In an earlier paper, Dawkins and Krebs (1978) questioned the conventional ethological wisdom which stated that animals' communication signals have evolved to transmit to others good information, particularly about their internal motivational state (motivation to attack, or court, and so on). Dawkins and Krebs argued that this idea is fundamentally flawed because animals are not expected to evolve actions merely for the benefit of others: rather, signals should evolve which most successfully manipulate others to the genetic benefit of the sender (but note that this more general formulation incorporates the possibility that if such benefits *do* arise through helping others, this will be selected for too). The deception and other forms of manipulation which thus evolve, in turn provide the selection pressure for mind-reading: because it now becomes important for animals to read not just the surface behaviour (which may be misleading), but to distinguish from this the underlying 'true' state of mind – the other animal's real intentions, such as the intent to attack while appearing friendly (Krebs and Dawkins, 1984).

This analysis seems to me to offer an important conceptual basis for what it could be for an animal to be 'really mind-reading': viz., the situation in which it identifies the true state of the mind of its protagonist, discounting the surface behaviour which on a particular occasion is at variance with that state. An appreciation of teasing, whether in apes (Kohler, 1927) or human infants (Reddy, 1991) may also fall into this category. Ironically perhaps, this essential contrast which appears to necessitate reference to a mind/behaviour distinction, is lost in Krebs and Dawkins' 'behaviourist' definition of mind-reading cited in the previous section.

However, consider how an animal *could* come to see through deceptive ploys to the true state of the protagonist's mind. One way ('history') would be that it learned by experience that under certain conditions a different interpretation was warranted: for example, a particular individual might come to be recognised as untrustworthy because they so often cry wolf. A second way ('leakage') would be to recognise tell-tale cues in the

protagonist, such as the equivalent of blushing in humans. And a third ('contradiction') would be to recognise other inconsistent cues, such as when A acts as if seeing some interesting object, the non-existence of which B has independent grounds for recognising. Records of primate behaviour exist which appear to correspond to each of these three criteria (Whiten and Byrne, 1988; Byrne and Whiten, 1990; 1991). However, each of these ways of discounting can be re-described as an observational analysis, with the mind-reader succeeding at its task because it discounts certain conventional signals in favour of some other observable criteria which are given more weight.

As in the case of the implicit mind-reading described for Gómez's gorilla, we must acknowledge that such phenomena are of great interest with respect to the emergence of the mature mentalism humans come to apply, because even though they *may* be adequately characterised as sophisticated behaviour-analyses, they require the mind-reader to tap into cues which represent likely important observational foundation stones even for the adult human mentalist. Thus when the latter is dealing with another's intent to deceive them, they must surely do this on the basis of discerning one or more of the three types of 'give-away' sketched above – history, leakage or contradiction.

However, in the human mentalist, we assume there must be a brain state which corresponds specifically to (encodes) the state of mind in the protagonist – 'deceptive intent'. It is not apparent that this is necessary in the case of the animal counter-deceiver described above: they may just know to distrust certain behavioural and contextual signs in favour of others, and we need say only that their brain encodes these signs. In the next section, we consider what is additionally required to diagnose encoding of others' *states of mind* by a non-verbal mind-reader.

5 Mental states as intervening variables

I have explained the concept of mental states as intervening variables quite fully elsewhere (Whiten, 1993; 1994), and so provide a concise description here. My starting point has been the usage of the idea of intervening variables in an earlier phase of animal psychology. It was shown that a number of different aspects of rats' drinking behaviour (the amount they would drink, how hard they would work to get water, and so on) could each be caused by any one of a number of conditions (eating dry food, going long without a drink, and so on). This could be represented as a large number of S-R (stimulus-response) links, joining each of these inputs to each of the behavioural outputs (fig. 17.1a). But alternatively, an intervening variable can be posited which underlies the pattern of

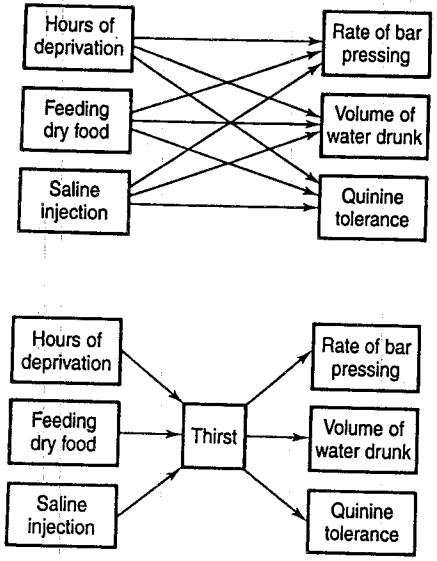

Figure 17.1 A simple example of an intervening variable.
(a) The relationships between three independent and three dependent variables in the case of rats' drinking; (b) Recognising an intervening variable (here called 'thirst') permits a more economic representation of the causal linkages, in this case reduced from the nine shown in (a), to six (after Miller, 1959; Hinde, 1970; Whiten, 1993).

results: this variable is not directly observable, and is thus an 'intervening variable' the value of which can be affected by any or all of the input variables, and having changed, can itself affect each of the outputs (Miller, 1959) (fig. 17.1b). It does not really matter what we call the intervening variable, but in this case 'thirst' would not be unreasonable. We now have a representation of these phenomena which is more *economic of representational resources*, because the multitude of S-R links no longer need to be coded. The observer needs only to know that the 'Ss' can each lead to the state (thirst, in this case) and that for an organism in this state certain 'Rs' can be forecasted. The more different conditions which can affect the intervening variable, and the more outputs it can affect in turn, the more efficiency of representation can be gained by its replacing a profusion of specific S-R links.

The suggestion is that recognising a state of mind in another follows this

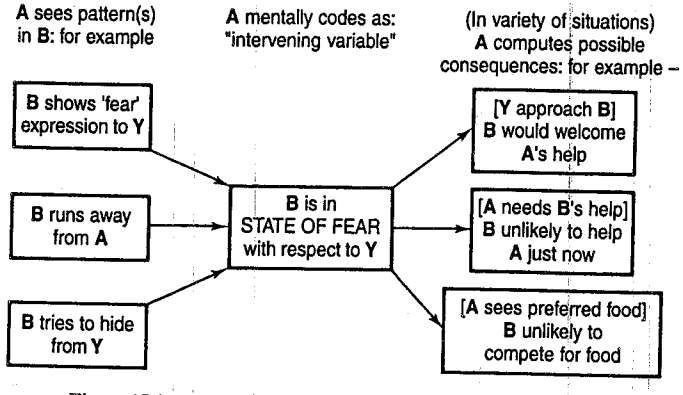

Figure 17.2 The recognition of a mental state as an intervening variable: fear.
Here, a hypothetical primate, A, reads the mental state of fear in an individual B, coding this state as an intervening variable generated on different occasions by a variety of circumstances like those shown on the left, and in turn giving rise to various predictions appropriate to different circumstances such as those shown on the right. Such mentalising gains the same economy of representation expressed in Figure 1(b) contrasted with 1(a) (after Whiten, in press).

general pattern: for the folk mind-reader, attributing a mental state is in important respects the same as recognition of intervening variables by the professional psychologist. Any specific state, such as B *knowing* a certain thing, may be recognised by A on the basis of a number of different observable conditions which can cause this knowledge: and once the knowledge is attributed, that state itself could lead to a multitude of outcomes predictable according to circumstance. Thus recognition of such states in others can be a powerful way of representing and predicting their behaviour patterns, economic of neural resources. Figures 17.2–4 illustrate hypothetical examples for different mental states.

A number of points should be made about the significance of the intervening variables conception:

(1) *Power*. As outlined in the previous paragraph, the intervening variables conception suggests why we or other mentalist creatures might be mentalists, in terms of that same cognitive power or efficiency obtained by crediting the rat with 'thirst'.

(2) *Nature of mentalism*. It also proposes what it can *be* to be a mentalist, as opposed to a behaviourist who instead learned all the specific S-R rules of

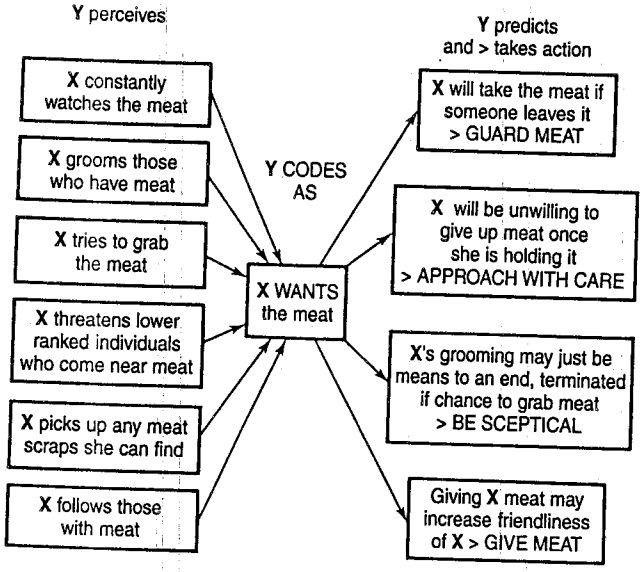

Figure 17.3 Recognising another individual's state of desire as an intervening variable.
This hypothetical example concerns a baboon Y, who recognises a state of wanting in another baboon X, with respect to competition over a desirable food source. The general scheme is as for Figure 2. As the number of eliciting circumstances on the left rises (and so too for the relevant consequences on the right) so the gains in economy of recognising such intervening states rises (after Whiten, 1994).

others' behaviour patterns. In other words, it provides perhaps the most defensible definition of mentalism, compared to other approaches outlined earlier. However, mentalism is still a term which grades into behaviourism, because it is applied on the basis of *how complex* webs such as those shown in Fig. 17.1b and 17.2–4 are. If such a web had only one line of links (for example, if fig. 17.2 only included the link between (on the left, input side) X putting something in a location and (on the right, output side) X's ability to recover it, we would have nothing distinguishable from a behavioural rule).

(3) *Abstraction*. Intervening variables constitute *states which others are classed as being in*: it is not necessary that they are seen as 'internal' to others, even less that they are seen as being 'in the head'. But see point 5, below.

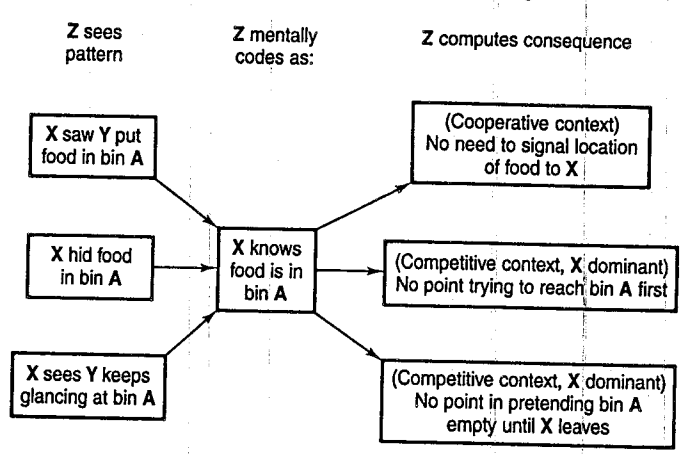

Figure 17.4 Recognising another individual's state of knowledge as an intervening variable.
Here Z reads X's state of knowledge, portrayed in the scheme developed in Figures 17.2–17.3 (after Whiten, 1993).

(4) *Cleverness*. If mentalism conceived in this way owes its existence to cognitive economy, it may appear a paradox if our current working hypothesis that it is refined only in particularly clever species like apes (Whiten, 1993; Povinelli, 1993) is confirmed. The answer may depend on distinguishing the process of *acquisition* of mental state discrimination from the *application* of mentalism on subsequent particular occasions.

Thus, the capacity to recognise in the first place the complex pattern which is covered by an intervening variable (see figs. 17.1–4) may require considerable neural resources, of a level we see only in apes. It is once this recognition has taken place that application in behavioural analysis can become efficient on any one occasion, facilitating fast and sophisticated tactics to be deployed in, for example, what has been described as political manoeuvring in chimpanzees (de Waal, 1982; Byrne and Whiten, 1988).

(5) *Insight*. I have remarked before that the recognition phase may be likened to an insight into the underlying pattern (Whiten, 1993). This would suggest that experience could play an important part in the acquisition of a theory of mind, however much guided by innate preparedness. Although the idea of innateness rests on evolutionary theory, the fact of evolution itself presents a difficulty if innateness is thought of as a sort of pre-formation which solves the problem of recognising certain complex patterns in the world, like language structure, or mental states: would not

some recognition have been necessary before natural selection could operate to shape a special-purpose mechanism to facilitate such recognition (a 'theory of mind' module – Leslie and Thaiss, 1992)? At a more specific level, I suggest it is plausible that mental states must have been recognised in ancestral populations before words (names) came to be designated for them, and thus before children were raised in an environment suffused with mental state names.

(6) *Explicit mentalism*. Continuing this line of argument, there is a sense in which mental states as intervening variables could be said to be *explicitly* recognised by the mind-reader, even if the mind-reader is itself non-verbal. Explicit here means that in the brain of the mind-reader, there is a state which uniquely codes for the state (i.e. the intervening variable) of the mind that they are reading, this latter being an inclusion class which could be identical to the one which you and I name with a word like *knowing* or *wanting* (see Bennett, 1976). If the mind-reader picks out this phenomenon in the same way we do, even though they cannot name it, they are doing something more than the implicit mind-reading discussed earlier, where no such corresponding inclusion class is recognised.

(7) *Methodology*. An implication of this analysis is that mentalism cannot be diagnosed using a one-shot test, like the standard Sally-Anne 'false-belief' task, which could be solved by applying a single behavioural rule, such as that people search for things where they last watched them put. According to the analysis above, mentalism is diagnosed only when the mind-reader classes different situations as leading to the same mental state, and puts this classification to different uses according to context. Few animal experiments approach this (see Whiten, 1993, 1994; Heyes, 1993). Ironically, it *may* be our ability to recognise this complex of evidence in large batches of everyday situations which permits us to assign our children to different levels of mind-reading competence with some confidence, without performing any experiments at all. It is interesting that at this stage of our science, we do not know for sure how we *do* do this!

6 Experience projection

Povinelli and his colleagues have performed experiments on role reversal with both monkeys and apes. Each subject first played one of two roles. In one, their task was to indicate gesturally to an ignorant human which of several containers was baited with food: if successful, they received a share of the food. Other subjects played the role of the ignorant partner, and thus had to learn to make use of the helpful gestures of a human in order to choose the correct container. When performance had reached a high criterion of success, the roles of the non-human and the human were in each case

reversed. Povinelli, Nelson, and Boysen (1992) found that three of four chimpanzees showed evidence of role reversal, whereas Povinelli, Parks, and Novak (1992) found that rhesus monkeys did not, having to learn their new role afresh. The behaviour of the chimpanzees is consonant with a report by Savage-Rumbaugh (1986), that chimpanzees did not begin to use lexigrams to indicate the identity of the specific food in a container of which their partner was ignorant, until they had themselves been in that role of being ignorant, yet needing to know in order to obtain a reward.

It is not obvious how the chimpanzees would succeed in such a task by behaviour analysis. In the condition of Povinelli *et al.* where they are initially the ignorant partner, we might suppose that after reversal, they are showing (delayed) imitation of the behaviour of the human, who earlier pointed: and this sounds like an explanation resting only on behaviour analysis. However, given that chimpanzees do not mimic everything humans do, it is not obvious why the chimpanzees would point unless they recognised the other as being in a state where this could influence their behaviour, and apparently their only basis for doing this would be their own previous experience of ignorance, which had made the gesture of pointing helpful to them. In the other condition, where the chimpanzee initially pointed, an ability after role reversal to benefit from the pointing of the human is more difficult to see as (delayed) imitation of the human's behaviour, since the human had only selected a container, as indicated by their partner's gesture. The alternative explanation is along the lines of the chimpanzee 'understanding what the gesture means'; or (relating this more directly to mentalism) extrapolating to the human their own earlier communicative intent, perhaps of encouraging a correct choice in the partner.

Heyes (1993) has argued that in the experiments of Povinelli *et al.*, the chimpanzees may merely have learned what to do on reversal *more quickly* because they had learned most of what they need to know during the pre-training and previously (see Povinelli, 1994a; Heyes, 1994b). However my point here is one of principle rather than problems in the practice or reporting of one particular experiment. What I am suggesting is that one can envisage situations in which a subject adopts a reversed role which is novel for them and not explicable as a copy merely of the behaviour of their partner before reversal: the alternative explanation remaining is that they are using their own past experience in one of the roles as the basis for constructing appropriate actions towards another individual in that role. That would appear to be one further valid sense in which such a non-verbal subject would be acting as a mentalist rather than behaviour analyst.

This, one might argue, would be a case of the working of the simulation theory, discussed in chapters 2–8. This is probably true, but I have chosen to talk about 'experience projection' here to emphasise a difference in

way in which this could work would be where the desire can always be assumed, as some sort of default; in the scenario of figure 17.1c, for example, the food may be one which X will always want if it can get it (clearly, if sometimes X does not want the food, recognising beliefs but ignoring this motivational variability will be useless for prediction). One might object that a reliance on such default desires means that Z is really a belief-desire reasoner because desire is implicit in all the computations. However, one might equally argue that beliefs are implicit in Y's model of X in Fig.1b, where Y is just a desire-psychologist: in this case it is implicit that X *believes* the object to be a piece of edible meat, that it *believes* approaching it is a good first step to getting it, and so on. What is really at stake is whether the desire element in such reasoning could be handled by a behavioural analysis, and it seems to me that in principle it could. Thus, instead of Y coding X's state of desire as an intervening variable of the sort shown in figure 17.1b, it might utilise a straightforward S-R rule that (for example) if X has once tried to grab the meat, it will try to grab the meat again: and this in conjunction with a diagnosis of X's beliefs about the location of the meat will lead to novel and appropriate predictions of X's actions.

This may appear to be a merely academic issue if the earlier suggestion is true, that recognition of belief will tend to emerge only after recognition of desire is in place. However, it may be of more relevance to what happens in practice, insofar as it raises the possibility that a number of mental states might be recognised, with each used somewhat in isolation from each other, and in conjunction with more obviously behavioural variables and circumstances, to predict and explain actions. A theory of mind, in which multiple mental states are manipulated, may be worth distinguishing as an additional achievement: the analogy with language acquisition alluded to earlier would be that child or chimp might have a significant vocabulary of mental states it recognises before starting to combine them predictively and syntactically *as* mental states, in the 'intervening variables' sense described earlier.

ACKNOWLEDGEMENTS

I am grateful to the following for discussions which have been particularly helpful in thinking through some of the issues discussed in this paper: Jonathan Bennett, Peter Carruthers, Daniel Dennett, Juan-Carlos Gómez, Paul Harris, and Annette Karmiloff-Smith.

18 Chimpanzee theory of mind? the long road to strong inference.

Daniel Povinelli

1 Timing of the evolution of theory of mind

Here is an extreme view of the evolution of theory of mind:[1] prior to about four million years ago no organism ever paused to consider its own mental experiences or the mental experiences of others. This view carries with it the implication that the reproductive payoffs that led to the selection for theory of mind began to be realised only during the course of human evolution. It also implies that for some (as-yet-unknown) reason the complex social groups common to many mammals had not produced the right mixture of social or physical problems sufficient to drive the evolution of neural material capable of representing mental states. In short, this view implies that it was something about the unique history of human evolution that led to our pervasive and unshakeable folk psychology of mind. Of course, there are even more extreme views than this. For example, it has been maintained by some that theory of mind emerged coincident with the evolution of human language or that it is merely an illusion created by linguistic conventions (e.g., Wittgenstein, 1953; Langer, 1942; Lutz, 1992). Still more extreme would be the view espoused by some cultural anthropologists that beliefs about the mind are relative constructs peculiar to the cultures in which they are formed (e.g., Geertz, 1973; Mauss, 1984; La Fontaine, 1984).

The extremity of the views described above are in one direction only. It is possible to construct equally extreme views about the antiquity of theory of mind. For example, one could argue that theory-of-mind-like abilities are innovations that emerged during the evolution of the last common ancestor of the great apes and humans, that they were primitive mammalian innovations, or even that it was an innovation primitive to all vertebrates (for different views on the antiquity of consciousness and theory of mind see Fox, 1982; Gallup, 1982; Griffin, 1976; Rollin, 1989; Harris, this volume). Central to these views is the common denominator that knowledge about the mind is not restricted to the human species.

Some investigators will find some of the possibilities outlined above

Theories of theories of mind

edited by

Peter Carruthers

Professor of Philosophy and Director, Hang Seng Centre for Cognitive Studies, University of Sheffield

and

Peter K. Smith

Professor of Psychology, University of Sheffield

Published in association with the Hang Seng Centre for Cognitive Studies, University of Sheffield

CAMBRIDGE
UNIVERSITY PRESS