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.

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

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

difficult to accept. They will argue that good science should not give space to what does not cohere well with the rest of our beliefs: thus, we should reject the extreme positions regarding the widespread distribution for theory of mind on *a priori* grounds (Peter Carruthers, personal communication). After all, is it not absurd to suppose that birds and snakes have access to some of their own mental states? Conversely, do humans from diverse cultures really possess fundamentally different theories of mind? Unfortunately, as long as we leave it up to our intuitions to decide, almost any position concerning the evolutionary history of theory of mind can be justified. The reason that this state of affairs continues to exist is because views about the minds of other species have largely been driven by what particular theorists view as plausible, not by what is testable. Indeed, some commentators chastised our early empirical attempts to compare chimpanzees and rhesus monkeys' understanding of mental states on the grounds that the results merely supported widely held beliefs 'that chimpanzees are smarter than monkeys in almost all ways' (Mitchell *et al.*, 1994, p. 762)! Ironically, one result of this emphasis on intuition is that the current data set remains far too impoverished to allow an easy rejection of any of these positions.

In contrast to the view described above, this essay starts from the assumption that good science is about strong inference, and that strong inference is best arrived at by using the method of multiple working hypotheses (Chamberlin, 1897). In this fashion, various alternative hypotheses are outlined up front, and each are used to generate a set of predictions. Hypotheses which generate the most useful predictions ascend as the most viable until new alternatives are proposed and evaluated. Accordingly, in this chapter I stake out no *a priori* claims about which species (or cultures) possess (or possessed) a theory of mind. My aim is to provide theoretical justification for the conceptual and methodological sensibility of a research agenda which in the long run can answer a very basic question: do any species other than humans possess some kind of appreciation of the mental world? In short, do other species represent mental states, and if so, which species and which states?

2 Species that might have a theory of mind

I begin by examining some theoretical reasons why at least one group of organisms – the great apes – might be a reasonable place to begin to search for theory of mind in other species. Evolutionary biologists might think that I am about to commit an egregious error by selecting one representative of the great ape/human clade, *Pan troglodytes*, and then using them to make a general claim about the group in question. However, I focus

on chimpanzees for strictly practical reasons, and I fully acknowledge the striking species differences among the living great apes in social organisation, ecology, and perhaps even cognitive abilities. My purpose is to show that there is a set of plausible reasons why chimpanzees (at least) might have a theory of mind. Thus, I set aside worries about other species[2] and merely note that if chimpanzees possess some kind of theory of mind, then depending upon one's assumptions about the ancestry (polarity) of the trait and the exact phylogeny of the great ape/human clade, then there might be reasons for suspecting that other species in this clade – or even most species in the primate order – might as well. However, in examining these logical possibilities I neither take as obvious the claim that chimpanzees have a theory of mind, nor do I view the claim as a straw man to be immediately refuted. Instead, I see the examination of the possibilities as a crucial first step in outlining a theoretical and methodological framework for a long-term research program designed to determine if theory-of-mind skills are a uniquely derived feature of the human lineage, or if other species possess at least some portions of a common epigenetic[3] program governing the construction of representations of the mental states.

2.1 *Control of the timing of cognitive developments*

There are several facts which could lead one to suspect that chimpanzees have at least some understanding of the mind. But in order to appreciate them, it is necessary to consider two constructivist views of the timing of cognitive developments (see fig. 18.1). One reasonable view of cognitive development is that abilities which emerge at later ontogenetic time-points build in some fundamental manner upon earlier ones, and that most new cognitive structures emerge from domain-general shifts in representational abilities (fig. 18.1a). Thus, detectable differences in task performances across a seemingly wide range of domains (spatial, temporal, verbal) could be treated as being linked to underlying shifts in representational abilities. One particular version of this approach was Piaget's monolithic theory of intellectual development. A second variant of this constructivist approach to cognition is to assume that many skills are domain-specific, but within these domains the representational changes build upon earlier ones. The fundamental difference between this view and the first is that the domain-specific view argues that in many cases there may be little relation between shifts in one domain and shifts in others (fig. 18.1b). From an evolutionary point of view, this would imply the presence of different control mechanisms across domains, allowing for evolution to proceed at different rates across domains.

These two views are clearly artificial. For when it comes to specifying

.Figure 18.1 Two views of the relation among cognitive developments in 'separate' domains. In (a) changes across apparently distinct cognitive domains are necessarily yoked to fundamental core shifts in representational abilities or processing abilities, in (b) epigenetic interactions which determine the rate (and perhaps even ordering) of cognitive developments are specific to each domain.

exactly what we mean by a given 'domain' it becomes obvious that these views really just represent extreme ends on a spectrum of possibilities. As long as we set aside extreme cases of modularity, each domain will at best refer to a psychological territory with sketchy borders. This vagueness will arise precisely because although the concept of a domain may have heuristic value, precise neural boundaries within the brain may be as difficult to define as are the boundaries of other heuristically useful, but difficult to define biological constructs such as species (e.g., Mayr, 1957; Burma, 1954; Ghiselin, 1975). Thus, I proceed cautiously using the distinction, but I do not abandon it because depending on which end of the spectrum one falls, the same data set may give rise to very different interpretations.

2.2 *Reasons for suspecting that chimpanzees harbour a theory of mind*

If we were to accept some version of the domain-general argument advanced above, then existing evidence concerning general cognitive development in chimpanzees might provide support for the possibility that they have some ability to represent mental states. For example, suppose that we could demonstrate that in some domains these apes possessed clusters of abilities which young children develop at the same time they are beginning to attribute certain mental states. If so, then a domain-general view of cognitive development might lead one to suspect that chimpanzees also possess the ability to represent mental states. If theory of mind is just another arena in which general transitions in representational abilities manifest themselves, and if chimpanzees display clear evidence of such transitions through their performance on standardised tasks, then we would have a coherent reason for suspecting that they may also possess theory of mind.

In order to move beyond theoretical claims, let us consider two possibilities concerning the onset of theory of mind in human development. First, it is possible that infants as young as 12 months form representations of mental states, albeit simple ones. For example, Baldwin and Moses (1994) have proposed an early understanding of attentional focus through studies which have investigated the capacity of 12- to 13-month-olds to understand that others refer to things in the external world. Baron-Cohen (1994) interprets joint attention behaviours (proto-declarative pointing, gaze-following) as evidence of a similar kind of goal-desire psychology of infants by 12 to 14 months. A second possibility is that the capacity for genuine representations of mental states does not emerge until 18-24 months (Gallup and Suarez, 1986; Barresi and Moore, in press). Indeed, 18-24 months seems to represent an especially striking turning point in human development. As noted in figure 18.2, a wide range of behaviours emerge at this

Figure 18.2 Comparison of selected cognitive developmental pathways in humans and chimpanzees

point: pretend play, self-recognition in mirrors, correct personal pronoun use, an explosion in naming skills, co-operative play, simple acts of altruism, level-1 perspective-taking, mastery smiles, self-descriptive utterances, closure of stage 6 object permanence, directives to adults, along with several others. Thus, in addition to an emphasis on interesting cognitive changes which may be occurring between three and five years of age (Wimmer and Perner, 1983; Perner, 1991a; Wellman, 1990), two additional periods may be of special interest in looking for early theory-of-mind skills: one is between 12 and 16 months and the other is between 18 and 24 months.

With these two periods of human development in mind, we can now ask if chimpanzees display any behaviours akin to those which are emerging in infants during these two time periods. In terms of general development, there is clear evidence that chimpanzees display some *behaviours* which typically emerge in young children in both of these time-frames. For example, great apes (including chimpanzees) and humans are clearly travelling along a similar developmental trajectory (although at different rates) in terms of general development and sensorimotor intelligence up until the complexity of behaviours shown by 18 to 24-month-old human infants (Chevalier-Skolnikov, 1983; Parker and Gibson, 1977; Mathieu and Bergeron, 1981; Vauclair and Bard, 1983; Mignault, 1985; Hallock *et al.*, 1989; Bard *et al.*, 1992; Poti and Spinozzi, 1994). However, by the eighteenth month or so in human development, we reach a point at which the typical human and chimpanzee pattern are beginning to diverge – although perhaps not completely. For example, there is some evidence that chimpanzees reach some of the same landmarks typically achieved by the 18 to 24-month-old child, such as stage 6 aspects of sensorimotor intelligence, spontaneous second-order classification skills, language comprehension skills similar to 12 to 18-month-old human infants, and self-recognition in mirrors, although many may not be achieved until four to eight years of age (Mathieu and Bergeron, 1981; Spinozzi, 1993; Savage-Rumbaugh *et al.*, 1993; Gallup, 1970; Povinelli, Rulf, Landau, and Bierschwale, 1993). The point of this cursory review is to point out that if chimpanzees and humans are displaying similar developmental transitions along these general lines, then there is evidence for substantial commonality in the cognitive-developmental pathways of the two species.

But what does the picture look like if we narrow our search to behaviours which may have a relationship to theory of mind in human development? Here the chimpanzee data set become thinner and less compelling. Although there have been some direct attempts to test for theory of mind in non-human primates, it is not necessary to infer from these data that chimpanzees possess theory-of-mind capacities comparable to three to five-

year-old human children (see 2.3 below). However, chimpanzees do clearly display two behavioural patterns which have been theorised to have a possible relation to human theory-of-mind development: joint visual attention (gaze-following) and mirror self-recognition. To begin, joint visual attention is present in some form by 6 months of age in human infants, and apparently undergoes several developmental changes throughout infancy (Scaife and Bruner, 1975; Butterworth and Cochran, 1980; Butterworth and Jarrett, 1991; Corkum and Moore, 1994). However, one important finding is that although gaze-following is present quite early, it is not until about 18 months that infants will track another's line-of-regard into space outside their immediate visual field (Butterworth and Cochran, 1980; Butterworth and Jarrett, 1991). Baron-Cohen (1994) has proposed that joint visual attention is evidence of a 'shared attention mechanism' which is a precursor to theory of mind development in normal humans. He has recently bolstered this claim by providing preliminary evidence that the absence of gaze-following (together with pretend play and proto-declarative pointing) at 18 months predicts a diagnosis of autism, a syndrome which has been held to be at least partly characterised by theory-of-mind impairments (Baron-Cohen and Swettenham, this volume).

We have recently documented that chimpanzees display the gaze-following response, and have now replicated this finding several times. Our results indicate that chimpanzees will track another's line-of-regard under the following conditions: (a) the subject sees an experimenter orient their head and eyes (or just their eyes alone) to a point above and behind them, and (b) an experimenter is already positioned in an unusual visual orientation before they enter (fig. 18.3; Povinelli and Eddy, in press a; Povinelli and Eddy, in press b). The general phenomena has also been independently experimentally documented by Sanjida O'Connell (personal communication). Indeed, we have extended the finding by demonstrating that chimpanzees even understand how line-of-sight can be impeded by the opaqueness of objects (fig. 18.4; Povinelli and Eddy, in press b). I am especially confident about the replicability of these findings because they require no training, and the results are derived from dependent measures which are not differentially reinforced. However, unlike Baron-Cohen (1994) I am not confident that gaze-following is tapping into a mentalistic appreciation of the attention of others. Indeed, there are good theoretical and empirical reasons for thinking that the early appearance of gaze-following may have nothing to do with a subjective understanding of attention, even though at later ontogenetic time-points it may be imbued with such meaning by organisms with a theory of mind (Povinelli and Eddy, 1994). Nonetheless, it is critical to note that our findings reveal that in terms of the complexity of their gaze-following behaviour, chimpanzees are

probably displaying the level of sophistication shown by 18-month-old human infants (Butterworth and Cochran, 1980).

In addition to gaze-following, I am also quite confident that chimpanzees are also capable of self-recognition in mirrors. This phenomenon emerges by about 18–24 months in human infants (Amsterdam, 1972; Lewis and Brooks-Gunn, 1979). By this age, many young children demonstrate an understanding of the correspondence between the physical appearance of their image in a mirror and their actual appearance. Thus, when confronted with a mirror after having been marked on the nose or forehead they will reach up to touch the mark. Chimpanzees also recognise themselves in mirrors (Gallup, 1970; Povinelli, Rulf, Landau, and Bierschwale 1993; Gallup, *et al.*, in press). Mirror-naive chimpanzees initially respond socially to their mirror images, but the best available evidence now indicates that within several minutes to an hour of continuous mirror exposure many chimpanzees display self-exploratory behaviours which entail orienting to the mirror and using their hands to manipulate parts of the body difficult or impossible to see otherwise (the eyes, nose, ears, teeth, ano-genital region; see fig. 18.5). However, there are debates about the distribution, ontogeny, patterns of emergence, and underlying meaning and cause of this phenomenon (see Lin *et al.*, 1992; Swartz and Evans, 1991; Povinelli, Rulf, Landau, and Bierschwale, 1993; Mitchell, 1993). Nonetheless, the fundamental phenomenon of chimpanzee self-recognition in mirrors is comparable to that displayed by 8- to 24-month-old children, and I do not believe that this fact can be seriously questioned at this point, despite attempts by some to do so (Heyes, 1994a; see reply by Gallup *et al.*, in press).

Gallup (1982; Gallup and Suarez, 1986) has argued for a strong relation between mirror self-recognition and theory of mind both across and within species. Although I cannot do his hypothesis justice here, suffice it to say that he has interpreted self-recognition in mirrors as reflecting an underlying capacity for self-conception, and that such self-knowledge can be used to generate limited inferences about minds of others. Lewis and colleagues have also argued for the relation between the emergence of self-recognition in mirrors and the development of an understanding of certain mental aspects of self and other (Lewis *et al.*, 1989). Some tentative support for this view has come from two directions. First, some researchers have reported significant correlations between self-recognition in human infants and other behaviours which may have a link to early theory of mind (synchronic play: Asendorpf and Baudonniere, 1993; early altruism: Johnson, 1982, Bishof-Kohler, 1988; self-conscious emotions: Lewis *et al.*, 1989). Second, early research suggested a correlation between those species which have shown evidence for self-recognition and successful performance

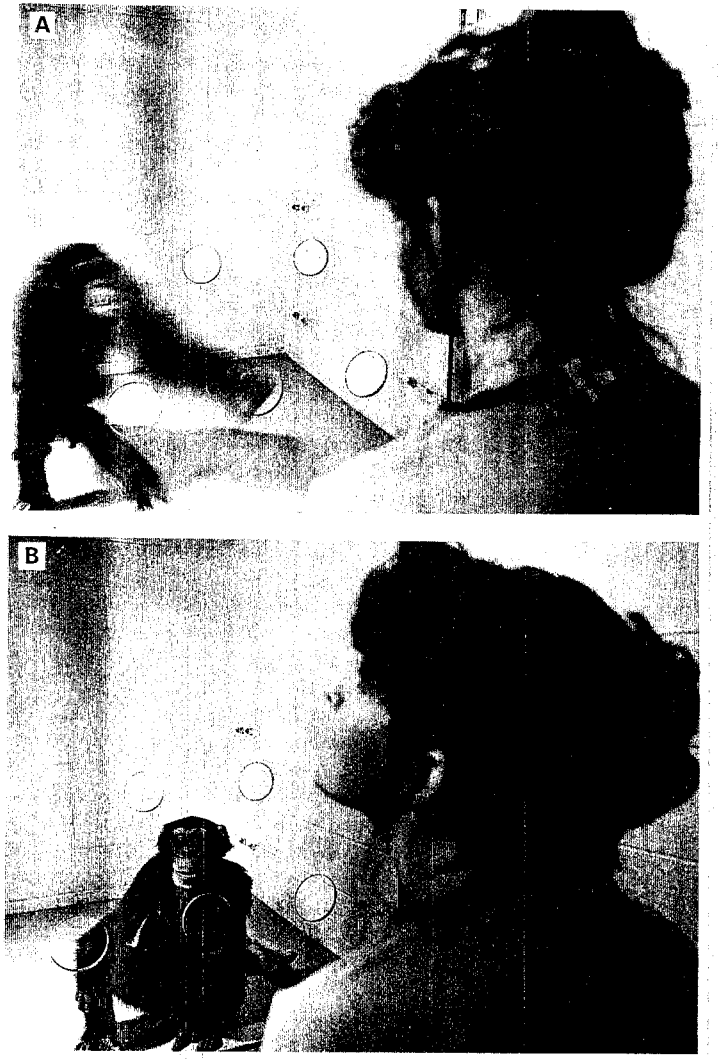


Figure 18.3 Setting and procedure for gaze-following studies in chimpanzees. (a) chimpanzee enters test unit and gestures in front of the experimenter, (b–d) experimenter shifts gaze to predetermined location above and behind the subject and chimpanzee turns head to follow experimenter's gaze. Various control treatments demonstrate that the subject's head-turning is triggered by the experimenter's gaze.

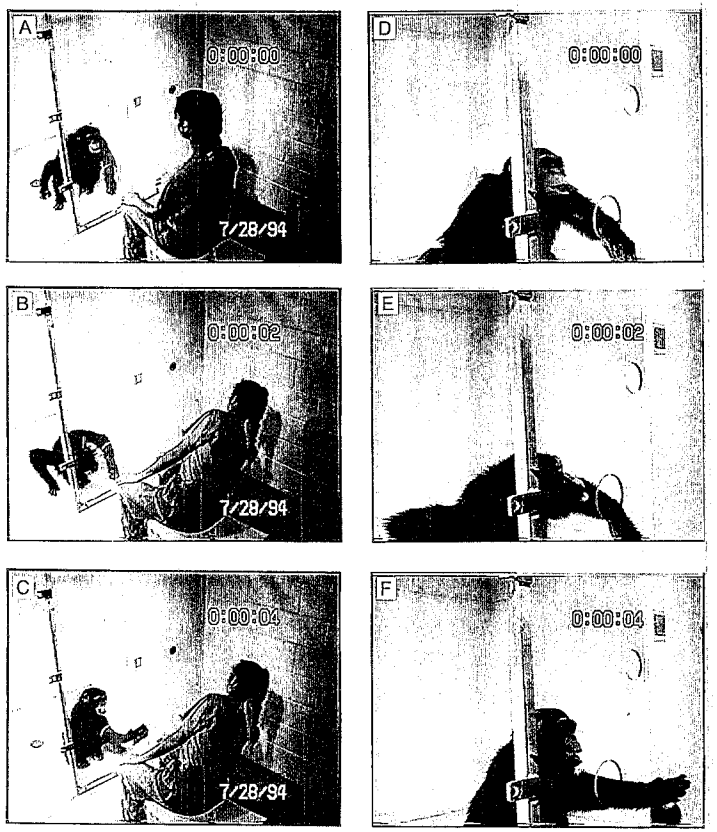


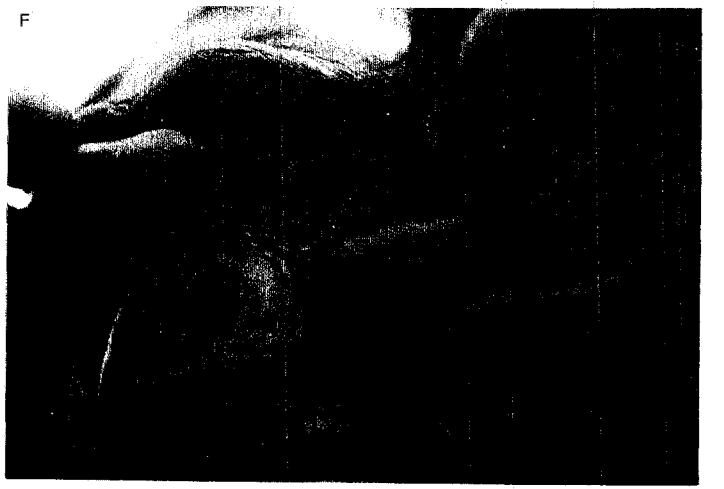
Figure 18.4 Two views of a chimpanzee demonstrating an ability to appreciate the interaction between an experimenter's line-of-regard and an opaque surface. In (a)–(c) the experimenter looks at a target on an opaque partition so that her line-of-sight (if projected straight ahead strikes the back wall of the test unit. The subject leans forward and sideways to attempt to look at the surface of the partition facing the experimenter. In (d)–(e) the same trial is shown but from the videotapes used by naive raters who were told to note whether the subject attempted to look at the square target. Various control treatments ensured that the subjects' responses were controlled by the gaze of the experimenter.

Figure 18.5 Contingent (a) and self-exploratory (b–d) behaviours in chimpanzees viewing themselves in mirrors. Chimpanzees in (e)–(f) displaying self-exploratory behaviours as observed through a one-way mirror. – (for c–f see next spread)


C


D


E


F

on theory-of-mind tasks (Povinelli, 1993). This latter data, however, deserves close scrutiny (see section 2.3 below).

Let us assume for the moment that some version of the domain-general argument is correct; further let us assume that the sophisticated form of gaze-following and mirror self-recognition are observable behaviours which follow from domain-general shifts in the representational system of human infants at around 18 months. If true, and if apes display both, then there would be reason to suspect that they also are able to form representations which take as their content mental states such as desires, goals, and attention – e.g., the same kinds of simple representations of which either infants (on the early view) and/or toddlers (on the late view) are capable. I am not claiming that this must be the case – indeed, there are good a priori arguments against such a view, some of which I review below – rather, I am simply claiming that it might be the case. To summarise, evidence that chimpanzees are travelling along the same general developmental track as 12 to 18-month-olds provides a clear rationale for taking seriously the possibility that chimpanzees form the same kinds of representations of mental states (probably non-epistemic ones) of which young infants and children are capable.

2.3 *Reasons for doubt*

Let me now return to the domain-specific view which argues that theory of mind represents a relatively isolated domain of cognitive development, proceeding with only superficial connections to changes in other cognitive domains.[4] This domain-specific view does not exclude the interpretation that apes have a theory of mind, on the other hand it would find no particularly compelling reason to see chimpanzees' successful performances on non-theory-of-mind tasks as indicating that they represent mental states. Indeed, even if we restrict our focus to the behaviours that (a) chimpanzees clearly display and (b) have possible theoretical links or empirical correlations with theory of mind (self-recognition in mirrors and advanced gaze-following), a domain-specific interpretation of these behaviours could still be advanced. For example, these abilities might develop with little or no interaction with representations of mental states either within or across species. Of course, correlations in human development between self-recognition and other behaviours which also seemed linked to theory of mind such as those described above would need to be explained in non-causal terms, as would significant associations between the absence of gaze-following and pretend play in autism on the one hand, and theory-of-mind deficits in autism on the other (Baron-Cohen and Swettenham, this volume).

Given that different starting assumptions about cognitive development yield different conclusions about the plausibility of theory of mind in chimpanzees, it might be useful to shelve theory for the moment and examine the evidence for theory of mind in chimpanzees. Several reviews of this area of research are already available (Premack, 1988; Cheney and Seyfarth, 1990; Povinelli, 1991, 1993; Whiten, 1993; Heyes, 1993; Tomasello and Call, 1994). Thus, instead of reviewing each of the (relatively few) investigations of theory of mind in non-human primates, I shall summarise the evidence to date as falling into two classes. First, we can examine the research strategies which have used spontaneously occurring behaviour, typically involving some kind of social manipulation of one animal by another, for evidence of an understanding of mental states on the part of various species (de Waal, 1982; Whiten and Byrne, 1988). Although the kinds of approaches to studying theory of mind using anecdotal evidence ranges from the causal to the systematic, all fall prey to similar problems. In particular, where a theory-of-mind interpretation suggests itself, a learning theory interpretation (or the deployment of some more tightly canalised, 'hard-wired' algorithms) cannot be ruled out by systematic manipulations of independent variables. In addition, there are several other classes of spontaneously occurring behaviour which may be central to theory of mind which are strikingly absent in chimpanzee culture: proto-declarative pointing, pretend play, and teaching (Premack, 1984; Cheney and Seyfarth, 1990; Povinelli and Godfrey, 1993).

The second class of data has emerged from laboratory-based studies in which experiments were designed for the purpose of testing hypotheses about the presence of theory of mind in non-human primates. To date, there have been only a handful of such studies, restricted to macaques and chimpanzees (Premack and Woodruff, 1978; Woodruff and Premack, 1979; Silverman, 1986; Povinelli, Nelson, and Boysen, 1990, 1992; Cheney and Seyfarth, 1990; Povinelli, Parks, and Novak, 1991, 1992; Hess *et al.*, 1993; Povinelli, Rulf, and Bierschwale, 1994). However, the experiments to date which have been offered as some evidence that chimpanzees may have a theory of mind suffer from methodological limitations related to the absence of attempts at replication, problems of learning, and difficulties in controlling for attention and motivation across species (Dennett, 1983; Premack, 1988; Whiten, 1991, 1993; Povinelli, 1991, 1993; Povinelli and Eddy, in press; Heyes, 1993).[5] Finally, what about experimental studies of phenomena such as imitation which have been posited to have a potential relation to theory-of-mind development (Meltzoff and Gopnik, 1993)? Unfortunately, the jury is still out with respect to the chimpanzee's capacity for true imitation (Tomasello, Kruger, and Ratner, 1993). Thus, careful examination of laboratory-based studies of theory of mind reveal that we

have no effects that have been replicated, let alone replicated and extended.

It would be wrong to think that I am implying that gaze-following and self-recognition are the *only* two phenomena possibly related to theory of mind that chimpanzees possess. On the contrary, I am indicating that these are the only ones that have passed the test of experimental demonstration and replication.

To summarise, a domain-general view of cognitive development could see in the existing data base enough evidence to suggest that chimpanzee cognitive development (although extended) looks very similar to the cognitive development of human infants up to about 18 months. On such a view, this would provide strong circumstantial evidence that they also possess some theory-of-mind skills, especially those early ones which are in place in human infants at 18 months. In contrast, the domain-specific view of theory-of-mind development would see no reason to interpret similarity in other areas of cognitive development as evidence one way or the other for similarity in theory of mind.

3 A conceptual framework

I have hinted that a search for theory of mind in chimpanzees makes sense within an evolutionary frame of reference. But I have not yet outlined why this is so. Figure 18.6 provides two models of the great ape/human clade, both of which display the African ape/human clade as an unresolved trichotomy – meaning that it is still too early to decide who is most closely related to whom within this group (Marks, 1994). Figure 18.6 displays two alternate possibilities concerning the developmental pathway related to theory of mind in humans. In the early evolution model, theory-of-mind skills are shown as having been an innovation primitive to the great ape/human clade and as a result are shared (via common descent) in members of the group, with caveats concerning gorillas (see Povinelli, 1994b). Note, however, that the schematic I have chosen reflects the notion that various ontogenetic stages or conceptual transformations are conserved. Thus, each of the members of the group which today possess the genetic instructions for theory of mind share a common developmental program. Thus, the basic ontogenetic sequence of theory-of-mind deployment could be relatively constant across species (although the rates might be quite different). The second model treats theory of mind as primarily a human innovation, and as a result it is not present in other members of the clade.

The two models in Figure 18.6 thus represent two possibilities concerning the timing of the evolution of theory of mind. They do not represent the most extreme view of either position I outlined in the first section of

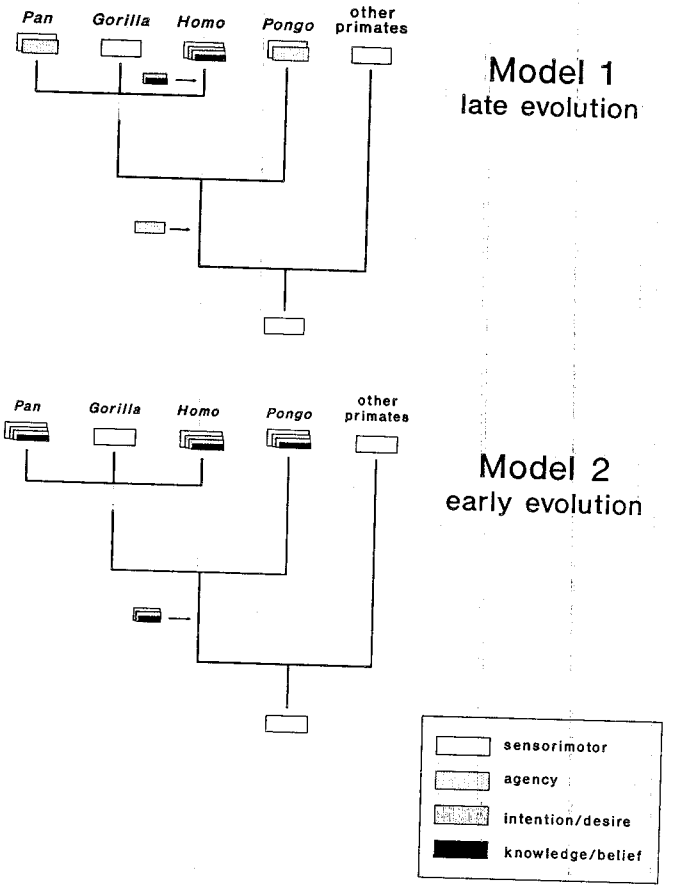


Figure 18.6 Two models of the timing of the evolution of arbitrary aspects of theory of mind. The African ape-human grouping is shown as an unresolved trichotomy to reflect controversies in the taxonomy of this group (Marks, 1992, 1994). The early evolution model posits that fundamental components of mental state attribution evolved before the differentiation of the great ape/human group. Note the secondary loss of some features by gorillas (see Povinelli, 1994). The late evolution model posits that most (if not all) aspects of mental state attribution evolved uniquely (and recently) in the human lineage.

this essay, but they do provide a solid set of alternative working hypotheses which can be tested by comparing theory-of-mind development in young children and young chimpanzees. The theoretical and methodological justification for these comparisons has been provided elsewhere (Povinelli and deBlois, 1992a, 1992b; Povinelli, 1993). In addition, the models should assist non-biologists in understanding that comparing the development of theory of mind (or any psychological capacity) across species does not assume some mysterious process of recapitulation. Recapitulation is a biological pattern which results from the inheritance of ancestral ontogenetic programs; generally speaking, the wider a time-span between the common ancestor of two species, the less 'recapitulationist' the ontogenies will appear. A better description is to dispense with the term 'recapitulation' altogether and to cast the inquiry in modern evolutionary terms – which portions of psychological ontogeny among the living great apes and humans are shared primitive traits inherited from a common ancestor, and which are most recently evolved aspects of psychological ontogeny, perhaps found only in humans?

4 Young chimpanzees' knowledge of seeing-as-attention: a case study

In order to demonstrate how the conceptual framework offered above can be practically implemented, I now describe a series of studies that we recently conducted on theory of mind in chimpanzees. The focus of these studies was a very simple question: do young chimpanzees understand the intentionality of visual perception? By intentionality we mean the following: do chimpanzees understand the 'aboutness' aspect of the perceptual act of seeing? We thus set out to ask if they understand that visual perception 'refers to' (or is about) objects or events in the external world. Note that such a narrow question eliminates many very interesting mentalistic questions about visual perception that one might ask. For instance, we explicitly did not wish to ask our apes if they understood the relation between seeing and knowing (that visual perception is a knowledge acquisition device; see 4.1 below). We had already done this, and obtained largely negative results (Povinelli *et al.*, 1994). Thus, we winnowed our interests in order to focus upon the simplest mentalistic aspect of visual perception of which we could think: that seeing subjectively connects organisms to other objects and events in the world (see Flavell, 1988). These studies were driven by the theoretical models outlined in the previous section concerning the timing and order of the evolution of theory of mind. The early evolution model predicts that chimpanzees should easily demonstrate such abilities, whereas the late evolution model predicts that they should not.

4.1 *Knowledge about visual perception*

Before I describe this recent research, it is necessary to draw a clear distinction between several kinds of information processing or knowledge that one might have about eyes. First, many organisms are extremely sensitive to the presence of eyes and eye-like stimuli (e.g., see Gallup *et al.*, 1971). There may be several reasons for this sensitivity including the fact that eyes are a useful stimuli for detecting the presence of a predator or social competitors. However, despite this general sensitivity to eyes, this kind of evidence alone is not sufficient to conclude that an organism understands anything at all about visual perception as a mental event.

There are two additional levels at which an organism might understand the eyes, and both of these involve an appreciation of the locus of the eyes as an interface between the private world of the mind and the shared external world. First, an organism may understand that visual perception subjectively connects organisms to the external world. In other words, an organism with the ability to form representations of mental states could equate the perceptual/geometric relation of 'seeing' with the internal mental state of 'attention'. In this sense, to see something is to be mentally or cognitively connected to that thing (Flavell, 1988). Flavell and his colleagues have demonstrated that by two and a half years of age human children have established this kind of understanding of the link between the eyes and the mind (Lempers *et al.*, 1977). There is some reason to believe that some species of non-human primates may also possess this kind of appreciation of seeing (Chance, 1967; Menzel and Johnson, 1975; Gómez, 1991). However, the kinds of evidence that have been brought to bear on the question of non-human primates' understanding of seeing are open to several different interpretations (Povinelli and Eddy, in press a).

The final kind of understanding of visual perception which we will consider here concerns the understanding that in addition to linking an individual's mental state of attention to the external world, visual perception also alters one's internal experiences, states of knowledge and beliefs. Flavell and colleagues have investigated the development of this understanding of visual perception in young children and have demonstrated that it is not until about four years of age that children realise that visual perception can give rise to unique mental experiences or states (Flavell *et al.*, 1981). Other lines of evidence also indicate that it is not until about this same age that children appreciate seeing as playing a causal role in knowledge acquisition (Wimmer, Hogrefe, and Perner, 1988; Gopnik and Graf, 1988; Ruffman and Olson, 1989; O'Neill and Gopnik, 1991; Povinelli and de Blois, 1992a). Collectively, these data suggest a marked developmental

asynchrony between young children's understanding of seeing-as-attention and seeing-as-a-knowledge-acquisition-device.

4.2 What young chimpanzees know about seeing: new evidence

Timothy Eddy and I have recently investigated young chimpanzees' knowledge about seeing-as-attention. The reason for this selection was two-fold. First, human knowledge about visual perception begins in infancy with the appearance of the gaze-following response, although the extent to which this knowledge is embedded in representations of mental states (as opposed to geometric calculations) is not clear. Second, there is a further elaboration of that knowledge at some point in late infancy to the point at which this knowledge possesses genuinely mental content. These two facts led us to realise that an extensive series of studies of this phenomenon might reveal the commonalities and differences in human and chimpanzee understanding within a quintessential mental domain: the experience of seeing. We thus executed a number of studies with a group of seven young chimpanzees to determine if they understand visual perception as a mental event.

Our research strategy has been to pit two explanatory frameworks against each other and to then evaluate them by their ability to generate accurate predictions about what our young apes would do in various circumstances. One of these is a mentalistic framework which attributes a theory or folk psychology of seeing to the animals. This framework makes no commitment about the upper-level of complexity of that understanding, but minimally assumes that the apes understand how the physical relation to eye direction and orientation anchors an internal mental state of attention in an organism to the world. In contrast, our second framework is derived from learning theory and starts with the assumption that chimpanzees form no representations of the mental states of others (such as attention) and as a consequence have no mentalistic understanding of eye gaze. This does not mean that the animals cannot reason about eyes, but rather that their reasoning is limited to the observable contingencies between eyes, eye direction, and subsequent behaviour. Thus, from the very outset of these investigations we shelved (to the best of our ability) our preconceived beliefs about what chimpanzees ought to understand about seeing, and instead outlined the predictions that these two very different frameworks would make about an ape's reaction to the perception of the visual systems of others.

In order to set the stage for our studies, we trained the animals to use their natural begging gesture (arm outstretched, palm up) to request food from an experimenter. All of the animals rapidly learned to enter a testing

lab, scan to see whether a trainer was standing or sitting on the right or left, approach a clear partition and stick their arm through a hole in front of the person in order to beg for food. Just to be sure the animals' performance on this task would be stable, we trained them to rigorous criteria whereby they were virtually flawless.

This initial training now allowed us to ask the apes if they understood seeing-as-attention. Whom would the chimpanzees approach and request food from if they entered the test lab and encountered two trainers standing or sitting in front of them, one of whom could see them, the other of whom could not? We designed several treatments to test the animals, some of which used familiar objects to prevent one of the trainers from seeing, and some of which involved more natural circumstances where one trainer could see the animal but the other could not. In one of the object treatments, one trainer placed a bucket over his head, the other held a bucket on her shoulder without obstructing her view. In the other object treatment, one trainer wore a blindfold over the eyes, the other wore it over the mouth. The two natural conditions were back-versus-front (one trainer facing forward, the other facing away) and hands-over-eyes (one trainer obscuring her eyes with her palms, the other looking forward while covering his ears). These initial set of treatments were all focused on a very specific question: which framework could better predict from whom the chimpanzees would beg for food when one of the experimenters was looking straight ahead with their eyes open and the other had their vision obscured? The mentalistic framework predicted excellent performance from trial 1 forward; the learning theory predicted poor performance initially, followed by improvement after repeated trials. To begin, however, we needed to assure ourselves that the chimpanzees could use their gesture to choose correctly between the two experimenters when an obvious cue was present that had nothing to do with understanding mental states. Thus, we created a treatment in which the apes were confronted with two trainers, one of whom held out a block of wood, the other of whom held out a piece of fruit or a small cookie.

In order to gain the maximum interpretative leverage possible, we used a combination of traditional small-N and group design features. First, within each session of ten trials only two were designated as probe trials ahead of time. Thus, the majority of trials within a session were simply spacing trials in which the subjects received the same treatment they had during training – a single experimenter positioned on either the right or left. This meant that we had within-session controls on whether or not the subjects were both (a) motivated to respond and (b) attending to the general features of the task. Second, the nature of the probe trials was alternated in an ABA design so that sessions containing block-versus-food probe trials

surrounded the sessions which contained the various visual occlusion probe trials. This meant that we had temporal control across sessions for determining whether or not the subjects were motivated and attending on probe trials involving a discrimination between two experimenters. The final feature of the design was that very few probe trials were administered (only two per session, typically four per experiment per condition) and they were separated by spacing trials. Because we had six to seven subjects in all of the studies, this allowed us to look at the group's responses on trial 1 and trial 2 for a very sensitive diagnosis of whether they possessed an initial disposition to gesture to the one who could see them. In addition, the subjects typically only received four total trials on a given treatment within an experiment, thus further constraining the problem of learning. With these controls in place we were in a position to determine which framework could best predict the apes' actual performance.

In the initial sessions of block-versus-food, the chimpanzees performed excellently from trial 1 forward: all of the subjects entered the test unit and responded by begging in front of the trainer offering the food. This result meant that the subjects had no difficulty reorienting from spacing trials with only one experimenter present to block-versus-food probe trials involving two experimenters. This was encouraging, because if they had experienced trouble here, the logic of the experimental design would have collapsed. These preliminary results demonstrated that the chimpanzees could easily use their gesture to make a choice between two different experimenters in a situation that did not involve the deployment of a folk psychology of seeing.

In striking contrast to this excellent performance on the initial two sessions containing these non-mentalistic probes, the subjects' performances dropped to chance when we administered the two-object treatments. Thus, on blindfold and bucket-probe trials, the group responded by gesturing in front of the person who could *not* see them as often as they gestured to the person who could see them. Was it because their motivation had declined as the experiment had proceeded? Our within-session controls allowed us to reject this general motivational/attentional critique; the subjects responded at nearly 100% correct on the spacing trials surrounding the probe trials. Was it possible that their ability to choose between two experimenters had waned across repeated probe trials? When we re-administered sessions with block-versus-food probe trials the subjects' performance shot back up to near-perfect levels. The logic of the ABA design allowed us to conclude that it was something about the treatments themselves that had yielded the effects obtained.

The results for one of the natural treatments looked identical to those just described: excellent performance on surrounding probe trials of block-versus-food and excellent performance on the surrounding spacing trials, coupled with chance-level performance on the hands-over-eyes probe trials. However, we obtained completely different results with the back-versus-front probe trials. In this case, the subjects' performances were significantly above-chance and did not differ from surrounding probe and spacing trials. This effect was present from trial 1 forward, with 5/6 animals correct on trial 1 and 6/6 correct on trial 2. Thus, in direct contrast to the other treatments, the animals seemed to possess an immediate disposition to orient selectively in front of the trainer facing forward and execute a begging gesture (see fig. 18.7).

A direct comparison of the two frameworks indicated that the learning theory generated more accurate predictions about what the animals would do. In three out of four treatments the animals did not seem to appreciate that one of the experimenters was connected to them in a subjective manner. However, the results of the back-versus-front treatment caused us to probe the situation further. The learning theory's explanation for the subjects' immediate success on the back-versus-front treatment was that there was a stimulus configuration (the trainer facing forward) that was identical to the correct response on their hundreds of training trials. The mentalistic theory's explanation was that (for one reason or another) the front-versus-back treatment was simply the most ecologically relevant instantiation of seeing-versus-not-seeing, and hence the chimpanzees performed best in this situation. We realised that we could test these accounts by confronting the chimpanzees with an equally ecologically relevant treatment in which two trainers both faced away from the subject, but one of them looked back over his shoulder toward the test unit. If the mentalistic account of the back-versus-front performance was correct, the animals should respond well; if the learning theory account were correct the animals should respond at chance.

We carried out this experiment using all of the controls described earlier (see Experiment 3, Povinelli and Eddy, in press b). Exactly as the learning theory predicted, the subjects performed excellently on the spacing trials, the block-versus-food probes, and the back-versus-front probe trials. Yet in direct contrast to these results, the group's performance dropped to chance on the looking-over-the-shoulder probe trials. For the animals it did not seem to make a difference that one of the trainers could see them and the other could not; despite the fact that one of the trainer's faces was completely visible to the subjects and one was not, the animals had no preference for begging to one over the other. This finding removed any unique reason for favouring the mentalistic framework's interpretation of the back-versus-front performance.

Despite this rather impressive correct prediction by the learning theory,

Figure 18.7 A five-year-old chimpanzee subject gestures in front of a trainer who can see her. This back-versus-front procedure was the only treatment in which the subjects gestured to the person who could see them from trial 1 forward.

we did not abandon the mentalistic framework. One of our hypotheses was that perhaps the apes did in fact understand that only one of the trainers was subjectively connected to them, but were startled by the unusual configurations of objects and body postures (especially given that for methodological reasons both experimenters were required to be associated with identical objects). In addition, perhaps the hesitations we detected on many of the probe trials were really startle reactions. Thus, the mentalistic framework could claim that our procedures masked an underlying appreciation of seeing-as-attention because introducing objects and unusual bodily postures on probe trials co-varied with the treatments under investigation. The framework could be interpreted as predicting that if the subjects were thoroughly habituated with the objects that were to be the ultimate cause of visual occlusion, then they should perform much better on the critical probe trials. We thus conducted a series of studies to test this interpretation (see Experiments 4-9, Povinelli and Eddy, in press b).

Here I will just describe one of these studies. First, circular cardboard screens were constructed that could hang around the experimenters' necks. The screens were large enough to obscure the entire head of the trainers. However, before we used these screens as a means of visual deprivation, we had the experimenters wear them on several sessions of standard trials, and then included them on block-versus-food trials, and on back-versus-front trials. In addition, we even had both trainers hold the screens up above their shoulder (without either face being obscured) in order to ensure that the subjects would associate successful performance with the presence of the screens being held up. At this point we were ready to pit the two frameworks against each by introducing the crucial probe trials; one of the trainers holding the screen in such a manner as to obscure the entire face, the other holding it above his or her shoulder. Despite these laborious efforts to rescue the mentalistic framework, we could not do so. The apes entered the test unit, scanned the trainers, and then proceeded to choose randomly. Indeed, there were even several striking trials where the animals looked straight at the screen obscuring the trainer's face, gestured, looked again, and finally extended the arm even further, apparently unaware that the trainer could not see what they were doing. Indeed, the learning theory implies that we cannot exclude an even lower-level interpretation of the apes' performance – perhaps they do not even understand that such a thing as 'seeing' exists.

Additional experiments manipulating the distance, height, and gaze orientation of the trainers generally had little or no effect on the subjects' performance. However, further experiments revealed a learning effect. Thus, across the eleven studies the subjects were beginning to perform at above-chance levels in several of the treatments. There were two possible

ways of interpreting this result. One was to assume that the subjects had extracted some stimulus-based rule from all of the various visual deprivation treatments we had administered. The other possibility was that the subjects had finally understood the task at hand: select the trainer who can see (read: is mentally attending). We decided to test this issue by analysing the various stimulus configurations we had previously presented to the apes. There seemed to be two possible rules they might use: 'pick the person whose face is visible' or 'pick the person whose eyes are visible'. Fortunately, neither rule made exactly the same predictions about how the animals would perform under two novel treatments: attending-versus-distracted (one experimenter looks straight ahead, the other looks up into the corner of the ceiling) and eyes-open-versus-closed. The 'face' rule predicted chance performance in both of these situations: the 'eyes' rule predicted chance performance when both eyes are visible (attending-versus-distracted) but good performance in eyes-open-versus-closed. Finally, the mentalistic framework predicted a different pattern of results altogether. It predicted that the subjects should have performed well on both of these novel treatments.

The actual results of these studies (Experiments 12 and 13) provided clear support for the 'face rule' over the 'eye' rule, and provided no support for the mentalistic framework. For example, the subjects performed randomly on their initial trials of both of the attending-versus-distracted and eyes-open-versus-closed treatments, thus indicating the subjects were not able to use an 'eye' rule, at least not in the eyes-open-versus-closed context. In addition, when the subjects entered the test unit on the attending-versus-distracted trials, they followed the line-of-sight of the distracted experimenter by turning and looking behind them. A crucial aspect of this finding is that it established something of which we had not been certain previously: that the subjects were scanning the faces of the experimenters before making their choices. They had to be, otherwise they could not have turned and looked behind them selectively on the attending-versus-distracted probe trials. Finally, it is important to point out that the subjects continued to respond poorly on only one of the initial treatments: blindfolds. Interestingly, this was the only treatment in which the same amount of each of the two trainers' faces were visible. In all of the treatments on which they were now succeeding (buckets, looking-over-shoulder, screen-over-face, hands-over-eyes) one of the trainer's faces was virtually completely obscured, but the other was clearly visible. This discrepancy between the blindfold treatment and the other treatments provides even further evidence that the subjects were using a rule about whose face was visible, not who was seeing.

Finally, in order to be absolutely certain that our non-verbal test was really measuring what we thought it was measuring, we tested young three-year-olds and compared their performance to older three and four-year-old children. If our task was measuring seeing-as-attention (and not some higher-order understanding of visual perception) then young three-year-olds should perform at above-chance levels: a variety of lines of evidence indicate that young three-year-olds understand this aspect of seeing (e.g., Lempers *et al.*, 1977; see section 4.1). In contrast, if the task was measuring something more complicated (such as the seeing-knowing relation) then the young three-year-olds should display chance performance, with the older children performing above chance (see section 4.1). We trained them in a similar fashion to the chimpanzees and then tested them using three of the treatments we had administered to the chimpanzees. The results established that as a group even young three-year-olds performed at levels well exceeding chance on their first trial with several of the exact same treatments and procedures used with the chimpanzees (Experiment 15, Povinelli and Eddy, in press b). This finding matched our *a priori* prediction that the task could be solved by simply understanding seeing-as-attention, without understanding the more complicated notion of visual perception as a knowledge acquisition device.

5 Lingering conceptual worries

Despite the clear biological rationale for comparing the psychological development of closely related species, commentators frequently worry that tests such as the ones just described are unfair to the chimpanzees. The concerns raised take several forms, but at the core of these criticisms is an underlying concern that it just does not make sense to compare chimpanzees and human children. Below, I describe each of four different meanings of this claim that I have encountered, and I attempt to assess their strengths and weaknesses.

5.1 Do chimpanzees have a 'weak' theory of mind?

First, in response to theory-of-mind tasks on which chimpanzees perform poorly, some researchers reply that perhaps chimpanzees have a weak theory of mind, and the tasks we give the animals are too complicated for them. First, if the argument is to extend beyond a vacuous claim, the nature of this weakness must be specified. The most biologically plausible possibility I can think of is to assume that chimpanzee behaviour is largely governed by learned or evolved social algorithms but that their representational code can compress less information about social information, and hence ultimately their (presumably) image-based

representational system for encoding mental states is less efficient than the additional linguistic-based system of humans. This alternative could make good biological sense if we viewed some human psychological innovations like language as providing a mechanism which allowed for more complex representations of mental states. To use a familiar example, perhaps desires and goals can be represented without a linguistic representational system, but perhaps representations of knowledge and beliefs requires the additional compressional power of a linguistic code. But if this (or something like it) is true, it does not mean that our tests are unfair. To the contrary, it would still be perfectly legitimate – indeed, essential! – to ask questions about common descent of those developmental programs. To ignore these differences would raise the spectre of a curious 'same-but-different' dilemma: on the one hand we would be deciding that we cannot compare chimpanzee and human psychology, but on the other hand we would maintain that chimpanzees do indeed have aspects of theory of mind. If the former were true, then the latter would remain an act of faith, not an inference of science. The general point is clear: if we believe that chimpanzees have less efficient representations of mental states than humans in particular domains, then we should revise our paradigms to investigate this possibility. Indeed, the tests described above were partly motivated for that very reason – it seemed to us that the non-epistemic mental aspects of seeing (attention) might be easier to compress into a visually based representational code than the epistemic aspect (seeing-as-knowing). Thus, our recent efforts to determine if chimpanzees understand seeing-as-attention were a strategic retreat from earlier attempts to ask chimpanzees if they understood the seeing-knowing relation.

Before we turn to other criticisms, it is important to note another variant of the weak theory-of-mind idea. It is possible to maintain that chimpanzees have some, most or all of the representations of mental states that humans do but that they only deploy them in particular ecological contexts. For example, perhaps chimpanzees do understand the seeing-knowing relation but only demonstrate it *when they are in a competitive social situation,* or perhaps a competitive social situation *with a member of their own species,* or perhaps a competitive situation with a member of their own species *in the wild.* This variant is related to the next criticism, and so I will explore it there instead.

5.2 *Is theory of mind anthropocentric?*

A second meaning of unfair is that the tests are anthropocentric. A charge is sometimes made that chimpanzees have a theory of mind, but it is a

theory of the chimpanzee mind, not of the human one. Although it has an appealing biological ring to it, upon close inspection if this idea is intended to be distinct from the notion examined above, it becomes very difficult to define. At one level the challenge reduces to an acknowledgement that apes and humans both have an understanding of the mental world – that is, that they both form representations of mental states; at another level the content of the representations differ in some species-specific manner. But what, exactly, are the differences? Specifying them becomes of paramount importance because although as evolutionary biologists we must recognise the idea of the evolution of specialised traits and behaviours, we must also be psychologists and ask: what is the nature content of the representations that differs? To defer this question would be to concede that this objection to experiments of the kind we advocate has not been carefully thought out.

There are several possible arguments that could be made to shore up the charge of anthropocentrism. For example, it could be maintained that chimpanzees and humans begin along similar paths because of the inheritance of common epigenetic instructions related to an intentional understanding of behaviour, but begin to diverge as ontogeny proceeds, revealing innovations which uniquely evolved in the course of the separate evolution of each species. As I hinted above, it is possible that theory of mind first evolved as an ability which was deployed in limited circumstances such as social competition. On this view, later innovations, perhaps ones that occurred only during the course of human evolution, extended this disposition to apply in many contexts, either as the result of a new specific mechanism or a generalisation module. Such accounts are possible. However, there are implications of these views. First, and most importantly, it would not follow from this theoretical position that our tests are unfair. Rather, it would render them even more indispensable than before because we would need them in order to discover the commonality and differences in theory-of-mind development in the two species. Second, it would mean that certain fundamental, core attributes (or ancillary components) of common-sense psychology are in fact uniquely derived in the human lineage. That is, trying to explain away negative findings with chimpanzees on high-level theory-of-mind tasks by appealing to a different theory of mind in apes means giving over these differences as exclusively human innovations. All of this suggests that tests such as ours can precisely pinpoint where and when the chimpanzee's hypothesised theory of mind is deployed. And if the chimpanzee's theory of mind is really so circumscribed, our laboratory tests would clearly reveal that unlike our common-sense psychology, theirs does not engage more or less automatically across situations, but is restricted to specific contexts.

5.3 *Are laboratory tests ecologically valid?*

Another criticism is to remain silent about the exact differences in theory of mind between the two species, but to explain the current negative findings with chimpanzees by recourse to an argument that laboratory-based tests are ecologically irrelevant. There are two versions of this claim. One is that the logic of these tests is not relevant. In other words, it could be that chimpanzees only use their theory-of-mind skills in social manoeuvrings that are ecologically relevant to them. By itself, this argument reduces to the claim that we just examined. But a second version of the claim is that it is the captive setting itself, not the particular tests, which creates the ecological irrelevance. Thus, anecdotes from natural settings could be taken as evidence that in these settings the animals do, in fact, deploy a common-sense psychology, but lab-based tests fail to reveal this psychology because of the captive setting. The problem, of course, is that monkeys and apes display similar patterns of social manipulation in captivity and in the wild (Menzel, 1973; de Waal, 1982; Whiten and Byrne, 1988; Coussi-Korbel, 1994). Given that the argument cannot therefore be about the setting, it would have to be about the specific context; thus, we return to the claim that theory-of-mind skills cannot be easily found in chimpanzees outside of some as-of-yet-unspecified social circumstances.

5.4 *Reasoning about the mind of another species*

Another meaning of the fairness charge that I wish to explore is the claim that most theory-of-mind tasks given to animals involve situations in which the subjects are asked to make inferences about the mental states of humans, not the mental states of other members of their species. For example, Seyfarth and Cheney (1992) argued that our initial attempt to test chimpanzees for an understanding of the perception-knowledge relation was limited because the subjects observed human actors, not other chimpanzees (Povinelli, Nelson, and Boysen, 1990). This argument reduces to a claim that chimpanzees and humans share a similar theory of mind, but because of some combination of the morphological and behavioural differences between us they are not inclined to assume that we have mental states. Of course, there is no comparable evidence that human children are impaired when reasoning about dolls, adults, other children, imaginary characters instantiated by figurines or pictures in books, or, indeed, even animals themselves. A moment's reflection will allow us to see that we have returned to the same-but-different dilemma. This time the tests are unfair because chimpanzees really do understand the psychological relations in question, but fail to attribute them outside their own species. But why is the human anthropomorphism gradient so extensive as to cover nearly all animate (and even inanimate) objects, but the equally rich chimpanzee theory of mind is peculiarly restricted to other creatures with just the right combination of black hair, knuckle-walking, prognathism, large ears, and so on?

I am not claiming that it is impossible that this is the case, just that we understand what the implications of this claim would be. Indeed, even in humans there are very intriguing differences in the willingness of adult humans to attribute complex mental events to other organisms. Eddy *et al.* (1993) reported empirical evidence that humans do not attribute theory of mind to other creatures in a random or blanket fashion, but their attributions depend critically upon the degree to which (a) the animals resemble them physically (e.g., other primates), or (b) they have formed attachment bonds with the species in question (e.g., dogs and cats). It is interesting in the present context to note that very few (if any) of the human subjects participating in the Eddy *et al.* (1993) investigation had ever formed attachment bonds with another primate species, but this did not prevent them from attributing theory-of-mind skills to these species, presumably based on their physical similarity to humans. Conversely, dogs and cats, to which most of the subjects had probably at one time or another formed primary attachment bonds, were given high ratings as well – despite their dissimilar morphological appearance. It is perhaps not trivial therefore to note that in the case of captive chimpanzees reasoning about humans both conditions are met: chimpanzees have formed attachment bonds with humans and we resemble them physically. Finally, note that this is an empirical issue which can be tested in the laboratory.[6]

5.5 *Are non-human primates a special case?*

Finally, it is important to ask if the criticisms explored above present a unique burden to those of us who study non-human species. In other words, if it is not fair to compare chimpanzees to humans, is it any more or less fair to compare pre-school children to adult humans, or even older children? And if pre-school children do not drive the point home, what about pre-verbal human infants? Is it somehow unfair to attempt to ascertain if they have a pre-verbal theory of mind? I do not think so. And once this is granted we are forced to acknowledge that the exact same (and very real) conceptual fairness problems exist regardless of whether our subjects are human infants or children on the one hand, or chimpanzees on the other (Povinelli, 1993).

Some researchers wishing to retain the conceptual integrity of comparing theory of mind across ages within humans, but not across species, might

retreat further and argue that in the case of developmental psychology at least the organism under study is a member of the same species as the adults to whom it is being compared – adults who we know develop a common-sense psychology. But for this argument to have force it would be necessary to assume that theory-of-mind evolution in the human species resulted from ancestral variants who possessed rudiments of common-sense psychology at birth (or that later occurring innovations were canalised backwards through existing developmental programs). If we did not make this assumption – that humans are born with some kind of theory of mind already activated – then we would never be sure when our between-age non-verbal comparisons were in the same logical position as across-species non-verbal comparisons; that is, comparing organisms with a particular theory-of-mind (older children, adults) to ones we are not sure about (12-month-olds). But there is no necessary reason to accept this position. It could be that fundamental human innovations in theory-of-mind psychology were created through innovations which occurred at relatively late points in existing developmental programs. Were this the case, there would be no particular reason to think that just because the infant is human it avoids the question of conceptual fairness. The general point is this: if the conceptual problems associated with asking another closely related species about what capacities they possess are insurmountable, then the inferences about ontogenetic transformations sought by developmental psychologists are crippled by the same problems.

At this point some observers might throw up their hands in despair and adopt the view that it is impossible to bridge the developmental and evolutionary transitions within human development, across human cultures, and across other species (for an example of this kind of despair as applied to the animal research, see Heyes, 1994b). I disagree. As long as we keep in mind that 'theory of mind', 'folk psychology', and 'mental state attribution' are heuristic constructs, then our research strategies can proceed on firm ground. Sets of predictions generated by each framework can be tested in a systematic manner. Theory construction and falsification can proceed as in any other science, avoiding the nihilism of the radical deconstructionist while simultaneously moving toward stronger and stronger inference about which aspects of the mind other species are capable (or incapable) of reasoning.

6 Do chimpanzees have a theory of mind?

Our research is not the final word on whether chimpanzees have a theory of mind. Indeed, it is not even the first sentence of the first chapter of the future volume, *Principles of the chimpanzee's theory of mind*. For example,

there are multiple working hypotheses to account for our data, only one of which is that theory of mind is a unique feature of the human species. Without elaborating let me simply note that Povinelli and Eddy (in press b) have discussed in detail two other very real possibilities: one is that there is a marked asynchrony in the rate of development between chimpanzees and humans such that our apes are simply too young to show the abilities in question (see fig. 18.2); the other is that both young and mature chimpanzees have an amodal theory of attention such as that proposed by Baldwin and Moses (1994) for 14-month-old human infants. I present the recent round of studies to show that future research on chimpanzees' understanding of the mind can proceed with both methodological and conceptual rigor.

But already our data suggests (although not in a statistical way) that some aspects of the domain-specific versus general account of cognitive constructions can be evaluated. For example, one of our apes (Megan) has tested positive for mirror self-recognition since she was three years old, unlike most of her age-mates who have produced a negative diagnosis even to the present. Yet at about five and a half years she performed no better than her companions on the seeing-as-attention tasks described in section 4. This could be taken as hinting that the domain-general argument is wrong because self-recognition in mirrors in young children occurs at about the same time (18–24 months) as initial understanding of mental states (desires). Alas, however, tests of seeing-as-attention similar to ours may be difficult for children younger than about two and a half years (Lempers *et al.*, 1977). Thus, a domain-general account could still prevail if there are necessary general cognitive achievements which must occur between these two time periods in young children before they can equate seeing as the mental state of attention. Clearly, we need better non-verbal tests of simple mental state attribution that are very sensitive to the 18-24 month age range in human children (Povinelli *et al.*, 1991; Povinelli, 1993).

So do chimpanzees have a theory of mind? One participant at the conference on which this volume is based expressed understandable frustration after listening to the papers on this topic. To her, it seemed as if the debate about theory of mind in non-human primates has been a simplistic 'yes-they-have-it'-'no-they-don't' roller coaster ride. But I hope that this essay will help to clarify that it will require patience to achieve even minimal closure on the question of whether apes or other animals represent mental states. To be sure, there are those who wish a quick answer. Indeed, their impatience has allowed them to overlook important methodological caveats that have been laid out in the primary literature. For example, Smith (this volume) concludes that there has been a shift in views concerning chimpanzees' understanding of the mind, and believes that it is perhaps

best to conclude that theory of mind cannot exist without language. Heyes (1994b) concludes that when it comes to theory-of-mind research in animals it is best to be a 'curmudgeon' (p. 242). I am willing to conclude that it is necessary continuously to upgrade our thinking about the relation between ontogeny and evolution and to use this theory to intelligently (and experimentally) challenge our intuitions about which species have a theory of mind. That, after all, is the long road to strong inference.

ACKNOWLEDGEMENTS

I thank Tim Eddy, Dare Baldwin, Radu Bogdan, Peter Carruthers, Lou Moses, the late Norton Nelkin, and Chris Prince for comments and thoughtful discussions that impacted upon the final text. Photographs are by Donna T. Bierschwale. This work was supported by National Institutes of Health Grant No RR-03583-05 to the New Iberia Research Center and a National Science Foundation Young Investigator Award to DJP.

NOTES

1 I use the term 'theory of mind' in the spirit in which Premack and Woodruff (1978) intended it. Thus, in this essay, I use the term to refer to any representation of unobservable mental states which serve the function of generating predictions about future behaviour or explanations for previous behaviour. I defer the issue of whether young children are young scientists-in-the-making to others (see Gopnik, this volume).

2 Those interested in the debate over potential differences in theory of mind among non-human primates are referred to other sources (Gallup, 1982; Whiten and Byrne, 1988; Cheney and Seyfarth, 1990; Povinelli, 1993; Whiten, 1993; Tomasello and Call, 1994).

3 I use the term epigenetic as opposed to developmental to emphasise the idea that the final path taken by developing neural systems depends upon both environmental feedback and genetic instructions. This is especially important in the current context given that several authors have argued for the possibility that abnormal environmental inputs in the form of contact with humans may potentially have dramatic effect upon great ape cognitive development in areas such as self-recognition in mirrors and imitation (Povinelli, 1994b; Tomasello, Savage-Rumbaugh, and Kruger, 1993).

4 To sustain this view plausibly, research showing significant correlations between pre-schoolers performance theory-of-mind tasks and other cognitive tasks would have to be explained in non-causal terms (e.g., Frye, Zelazo, and Palfai, unpublished manuscript).

5 Not all of the previous studies suffer from the general criticism that the critical trials are contaminated by 'learning'. Heyes (1993) has argued that our role-reversal studies, for example, may have involved learning in the critical reversal phase and complains that we did not report the trial-by-trial data in our original report. But a reading of the Povinelli, Nelson, and Boysen (1992) report reveals that this is not true. 'Additional support for Darrell, Sheba, and Sarah's

immediate comprehension comes from an examination of their first few trials within the initial role reversal session. As an informant on the first day of role reversal, Sarah produced accurate pointing to the correct location (from among four possibilities) on the first thirteen trials before making the first of her two errors... As an operator, Sheba's first six choices were all correct before she made her only error in twenty trials' (p. 637). In the context of Harris' (this volume) theory, it is of interest that this task performance was consistent with the hypothesis that the subjects saw their partner as performing specific actions to fulfil specific desires. Although it has not been replicated and suffers from some methodological drawbacks (Povinelli, 1991), it is possible evidence that chimpanzees understand others as intentional agents, even if they do not understand epistemic states. Similar task performance by 18- to 24-month-old humans has been demonstrated by Brownell and Carriger (1989).

6 We have explored two ways to attempt such investigations. One is to train chimpanzees as reliable enough actors to carry out our protocols. The second is to obtain convincing enough chimpanzee costumes for our undergraduate and graduate student actors to don during experimental trials. To date neither approach has met with much success.

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