BEHAVIORAL AND BRAIN SCIENCES (1998) **21,** 101–148
*Printed in the United States of America*

# Theory of mind in nonhuman primates

**C. M. Heyes**
*Department of Psychology, University College London, London WC1E 6BT, United Kingdom*
**Electronic mail:** *c.heyes@ucl.ac.uk*

**Abstract:** Since the *BBS* article in which Premack and Woodruff (1978) asked "Does the chimpanzee have a theory of mind?," it has been repeatedly claimed that there is observational and experimental evidence that apes have mental state concepts, such as "want" and "know." Unlike research on the development of theory of mind in childhood, however, no substantial progress has been made through this work with nonhuman primates. A survey of empirical studies of imitation, self-recognition, social relationships, deception, role-taking, and perspective-taking suggests that in every case where nonhuman primate behavior has been interpreted as a sign of theory of mind, it could instead have occurred by chance or as a product of nonmentalistic processes such as associative learning or inferences based on nonmental categories. Arguments to the effect that, in spite of this, the theory of mind hypothesis should be accepted because it is more parsimonious than alternatives or because it is supported by convergent evidence are not compelling. Such arguments are based on unsupportable assumptions about the role of parsimony in science and either ignore the requirement that convergent evidence proceed from independent assumptions, or fail to show that it supports the theory of mind hypothesis over nonmentalist alternatives. Progress in research on theory of mind requires experimental procedures that can distinguish the theory of mind hypothesis from nonmentalist alternatives. A procedure that may have this potential is proposed. It uses conditional discrimination training and transfer tests to determine whether chimpanzees have the concept "see." Commentators are invited to identify flaws in the procedure and to suggest alternatives.

**Keywords:** apes; associative learning; concepts; convergence; deception; evolution of intelligence; folk psychology; imitation; mental state attribution; monkeys; parsimony; perspective-taking; primates; role-taking; self-recognition; social cognition; social intelligence; theory of mind

## 1. Premack and Woodruff's question

Premack and Woodruff (1978) asked "Does the chimpanzee have a theory of mind?" Since it was posed, 20 years ago, Premack and Woodruff's question has dominated the study of both social behavior in nonhuman primates (henceforward simply "primates") and cognitive development in children, but progress in the two fields has been markedly different. Developmentalists have established empirical methods to investigate children's understanding of mentality, and, forging links with philosophy of mind and philosophy of science, they have mustered the conceptual resources for disciplined dispute about the origins (innate module, convention, or testing), on-line control (simulation or inference), and epistemic status (stance, theory, or direct knowledge) of human folk psychology (e.g., Goldman 1993; Gopnik 1993; Gopnik & Wellman 1994). In contrast, those working with primates have continued to struggle with the basic question of whether *any* primate has *any* capacity to conceive of mental states.

Primatologists and other investigators of animal behavior use a variety of substitutes for the term "theory of mind," asking whether animals are capable of, for example, "Machiavellian intelligence" (Byrne & Whiten 1988; Whiten & Byrne 1988), "metarepresentation" (Whiten & Byrne 1991), "metacognition" (Povinelli 1993), "mind reading" (Krebs & Dawkins 1984; Whiten 1991), "mental state attribution" (Cheney & Seyfarth 1990a; 1990b; 1992), and "pan- or pongo-morphism" (Povinelli 1995). Some authors use these terms to refer to hypothetically distinct capacities (see Whiten 1994 and 1996b for discussion of terminology), but they usually function in research on social cognition in primates as synonyms. A researcher using the term "mental state attribution," for example, is no less likely than one using "theory of mind" to believe that law-like generalizations underlie mental state ascription.

In this target article, I assume that individuals have a theory of mind if they have mental state concepts such as "believe," "know," "want," and "see," and that individuals

Cecilia Heyes is a Reader in Psychology, and Associate of the ESRC Economic Learning and Social Evolution Research Centre, at University College London. Formerly a Harkness Fellow, and Research Fellow of Trinity Hall Cambridge, she has published a series of review papers on theory of mind and self-recognition in primates. Her principal research interests are social learning and imitation. She is currently writing a book on human imitation learning, and co-editing volumes on selectionist epistemology and the evolution of cognition. Dr Heyes is a member of the Association for the Study of Animal Behaviour, and was for nine years an Associate Editor of the *Quarterly Journal of Experimental Psychology.* She is member of the management committee of the Experimental Psychology Society.

 0140-525X/98 $12.50

with such concepts use them to predict and explain behavior. Thus, an animal with a theory of mind believes that mental states play a causal role in generating behavior and infers the presence of mental states in others by observing their appearance and behavior under various circumstances. However, they do not identify mental states with behavior. For example, if chimpanzee Al has a theory of mind, he may judge chimpanzee Bert to be able to "see" a predator because it is daylight, Bert's eyes are open, and there is an uninterrupted line between Bert's eyes and the predator. But Al does not take seeing the predator to consist of these observable conditions. It is a further fact about Bert, inferred from these conditions, which explains why Bert runs away.[1]

In spite of nearly 20 years of research effort, there is still no convincing evidence of theory of mind in primates. We should stop asking Premack and Woodruff's question and considering the implications of a positive answer until we have designed procedures that have the potential to yield evidence favoring a theory of mind interpretation over other current candidates. Section 2 is a survey of the evidence of theory of mind from specific categories of behavior (imitation, self-recognition, social relationships, deception, role-taking, and perspective-taking), which argues for each study that the behavior reported could have occurred by chance or via nonmentalistic processes such as associative learning or inferences based on nonmental categories. In section 3, I argue that a theory of mind interpretation of the data reviewed in section 2 cannot be defended on the grounds of parsimony or convergence, and in section 4, I describe a test procedure that may be able to provide evidence of theory of mind in primates. Commentators are invited to identify flaws in this procedure and to devise alternatives.

## 2. Critique of evidence

The majority of those who have conducted empirical work on theory of mind in primates have claimed at one time or another that chimpanzees and possibly other apes, but not monkeys, have some components of a theory of mind (e.g., Byrne 1994; Cheney & Seyfarth 1990a; 1992; Gallup 1982; Jolly 1991; Povinelli 1993; Waal 1991; Whiten & Byrne 1991). The most commonly cited evidence in support of this view comes from studies of imitation, self-recognition, social relationships, deception, role-taking (or empathy), and perspective-taking. A sample of studies from each of these categories including the strongest and most influential is reviewed in sections 2.1–2.6.

In each of these six sections, two questions are addressed: (1) Competence: Is there reliable evidence that primates have the relevant behavioral capacity? (2) Validity: If present, would this behavioral capacity indicate theory of mind? For example, in the case of self-recognition, the competence question will be answered affirmatively if there is clear evidence that some primates are capable of using a mirror as a source of information about their bodies, and the evidence will be considered clear if there is no other at least equally plausible explanation for published observations of mirror-related behavior in primates. Similarly, the validity question will be answered affirmatively if there is no equally plausible nonmentalistic alternative to the hypothesis that mirror-guided body inspection requires or involves a self-concept. More generally, the competence question attempts to establish which environmental cues primates use to guide their behavior, and the validity question inquires about the psychological processes that lead them to use these cues rather than others.

The theory of mind hypothesis (or, more accurately, hypotheses) consists primarily of claims about what primates know or believe, about the *content* of their representations. Their distinctive, unifying feature is that they assert that primates categorize and think about themselves and others in terms of mental states. Consequently, the distinctive, unifying feature of nonmentalistic alternative hypotheses is that they do *not* assume that primates represent mental states. They assume instead that primates respond to or categorize and think about themselves and others in terms of observable properties of appearance and behavior. Behaviorism and learning theory are rich sources of nonmentalistic hypotheses, with those derived from behaviorism assuming no representation at all, and those derived from contemporary learning theory assuming, for the most part, some kind of imaginal, nonsymbolic representation (Dickinson 1980). However, denial that primates are capable of representation, or of abstract or symbolic representation, is not a necessary feature of a nonmentalistic hypothesis. Such a hypothesis might, for example, assume that primates are sensitive to whether a conspecific is "upright" or "supine" (see sect. 2.5 below), and that these are abstract or symbolically represented concepts, derived and applied through inference processes.

Consequently, it may be misleading to portray the debate about theory of mind in primates as a battle between the theory of mind hypothesis and "traditional learning theory" (e.g., Povinelli & Eddy 1996). The theory of mind hypothesis is primarily a claim about what is known or represented, whereas learning theory's most distinctive claims are about how knowledge is acquired (Dickinson 1980). Similarly, it may be counterproductive, although amusing, to think of animals that lack a theory of mind as "behaviorists" (e.g., Premack & Woodruff 1978). In one potentially confusing respect, scientific or philosophical behaviorists have a theory of mind just as surely as a scholar who believes that mental states cause behavior: they actively seek to explain, using argument and evidence, the nature and origins of behavior, including human use of mental state terms. In contrast, nonmentalistic alternatives to theory of mind hypotheses typically claim that primates "just do it"; they respond to observable cues, categorize them, form associations between them or make inferences about them, but they never ask themselves why, or whether other animals do the same thing. Thus, according to nonmentalistic hypotheses, primates are not psychologists, or indeed theorists, of any stripe.

Section 2.7 summarizes my answers to the competence and validity questions for each of the six types of behavior discussed. This may be used as a guide for selective reading by those who are not interested in detailed evaluation of evidence of competence when the behavioral capacity in question may not be a valid indicator of theory of mind.

### 2.1. Imitation

Motor imitation, the spontaneous reproduction of novel acts yielding disparate sensory inputs when observed and executed, has long been regarded as a potential sign of

higher intelligence in nonhuman animals (e.g., Thorndike 1898). It is relevant to theory of mind because it is thought to involve the ascription of purposes or goals by the imitator to the model (e.g., Tomasello & Call 1994; Tomasello 1996; Whiten & Byrne 1991). However, after nearly 100 years of research, there is still no unequivocal evidence of motor imitation in any primate species and even if there were, it would not imply the possession of mental state concepts.

Under uncontrolled and semicontrolled conditions the occurrence of imitation in monkeys (Beck 1976; Hauser 1988; Nishida 1986; Westergaard 1988), orangutans (Russon & Galdikas 1993), and chimpanzees (Goodall 1986; Mignault 1985; Sumita et al. 1985; Terrace et al. 1979; Waal 1982) has been inferred from the performance of a complex, novel, and previously observed act by a single animal or a succession of animals within a group. Even if one disregards the problem of the reliability of these observational or anecdotal data, they are not compelling. In all cases, the observed behavior could have been acquired by a means other than imitation (e.g., instrumental learning), and in many cases there is evidence that it was so acquired (Adams-Curtis 1987; Fragaszy & Visalberghi 1989; 1990; Galef 1992; Tomasello et al. 1993; Visalberghi & Trinca 1989). For example, the habit of potato washing was supposed to have been transmitted through the population of Japanese macaques on Koshima Island through imitation (Nishida 1986). However, given the order in which members of the troop were observed engaging in this behavior (first a juvenile, Imo, then her playmates, then their mothers), it is possible that, rather than copying the actions of potato washers, naive animals followed or chased them into the water while holding a potato. Once in that position, the pursuing animal would only have to drop and then retrieve its potato, now sand-free and with a salty taste, to acquire the behavior (Galef 1992; Visalberghi & Fragaszy 1992).

Remarkably few experiments have been conducted on imitation in primates. Their results may indicate only "matched dependent behavior" (Miller & Dollard 1941), the use of a demonstrator's behavior as a discriminative stimulus for the same response by the observer, and "stimulus enhancement" (Galef 1988; Spence 1937) that observing action can influence the degree to which the observer attends to certain physical components of a problem situation. Hayes and Hayes (1952) gave Viki, a "home-raised" chimpanzee, a series of 70 "imitation set" tasks. Each task consisted of the experimenter saying "Do this," and then performing an action such as patting his head, clapping his hands, or operating a toy. Hayes and Hayes claimed that Viki imitated more than 50 items in the set, including 10 completely novel, arbitrary gestures, but this conclusion is not secure because the report on Viki's behavior provided no indication of either the method used to measure the similarity between the experimenter's and the chimpanzee's behavior, or the degree of similarity observed.

Custance et al. (1995) carefully replicated Hayes and Hayes's study with two juvenile chimpanzees and provided a full report of their methods and results. The latter showed that after being shaped to imitate 15 gestures on the command "Do this," the chimpanzees spontaneously reproduced 13 and 17, respectively, of a possible 48 "novel" gestures, actions distinct from those in the training set. This is probably the strongest evidence to date that, at least after training, the form or topography of a primate's action can be influenced by observing the same action by a demonstrator. However, even when they reproduced novel gestures, the chimpanzees may have been engaging in matched-dependent behavior (Miller & Dollard 1941), that is, using the demonstrator's behavior as a discriminative stimulus for the same or similar behavior, without knowing that their behavior was similar to that of the demonstrator. For example, both chimpanzees reproduced lip smacking without being explicitly trained to do so in this study. However, they had been reared by humans, and humans have a strong tendency to play mutual imitation games with infants in which the infant is rewarded with smiles and cuddles for reproducing behavior, especially facial expressions (Piaget 1962). Hence, we cannot rule out the possibility that Custance et al.'s chimpanzees had been inadvertently rewarded for imitative lip-smacking (or imitative performance of a lip movement sufficiently like smacking to be scored as such in this study) before the experiment began. As Custance et al. point out, the reproduction of other novel items in the series may have been due to generalization from initial training within the study. For example, successful reproduction of nose touching may have represented fortuitous generalization decrement from prior training to reproduce chin touching. If this was the case, and if the chimpanzees' reactions to nose touching had been sampled many times in the absence of reinforcement (adventitious or otherwise), then one would have expected to see a range of responses to nose touching, including throat and cheek touching.

Tomasello and his colleagues (Tomasello et al. 1987) did not find any evidence of imitation of rake use in chimpanzees, but they reported positive findings for "enculturated" chimpanzees (i.e., animals with an extensive training history) in a later experiment (Tomasello et al. 1993). In this study, enculturated chimpanzees, relatively naive chimpanzees, and young children observed the experimenter manipulating 16 objects in various ways and, after observing each action, were given access to the same object either immediately or after a 48-hour delay. When the test was given immediately and the results for all objects were combined, the enculturated chimpanzees were comparable to the children in their tendency to act on the same part of the object, and with the same effect, as the demonstrator. However, for many objects, resemblance between the demonstrator and the observer could have been coincidental or due to stimulus enhancement rather than imitation. For example, when presented with a paint brush, the chimpanzees may have squeezed it with one hand, not because they had observed the trainer executing this particular action in relation to the brush, but simply in an effort to grasp an object which had been made salient through contact with the demonstrator. Since, by definition, the enculturated chimpanzees had been subject to more training procedures in the past than the other chimpanzees, there had been more opportunity for their fear of such procedures to habituate, and, specifically, more time for them to learn that objects handled by humans are often associated with reward. Therefore, even if the experiment by Tomasello et al. (1993) tested interest in novel objects and stimulus enhancement rather than imitation, one would expect the performance of the enculturated animals to be superior.[2]

The paucity of evidence of imitation in primates indicates neither that they are unable to imitate nor that such evidence is impossible to obtain for nonhuman animals. Relatively unequivocal evidence of imitation in budgerigars

(Galef et al. 1986) and rats (Heyes & Dawson 1990; Heyes et al. 1992) has been found by comparing the behavior of naive subjects that have observed a conspecific acting on a single object in one of two distinctive ways (but see: Byrne & Tomasello 1995; Heyes 1996). Whiten et al. (1996; Whiten & Custance 1996) recently gave a similar "two-action test" to chimpanzees, with mixed results. They found that chimpanzees that had seen a person withdraw bolts from rings with a twisting action for food reward subsequently twisted the bolts more than chimpanzees that had seen the person push the bolts through the rings with a poking action. However, as the authors pointed out, in the absence of data from subjects that did not observe any action on the bolts prior to testing it is difficult to rule out the possibility that what the chimpanzees learned by observation was not how to perform the twisting or poking hand movement but that certain movements of the bolts (e.g., rotation followed by lateral displacement toward the actor) were followed by reward. This has been described as emulation learning (Tomasello 1996).

Thus, surprisingly, it is not clear whether apes or indeed any other nonhuman primates can "ape" (Tomasello 1996), whether they are competent imitators. Furthermore, a capacity to imitate is not a valid indicator of theory of mind. It has been claimed that imitation involves the observer representing the demonstrator's mental state, its point of view, or its beliefs and desires (e.g., Gallup 1982; Povinelli 1995), but the case is not compelling. As far as I am aware, there is no evidence that the development of imitation in childhood is related to success in conventional theory of mind tests, and simple task analysis suggests that an observer could imitate a demonstrator's action without any appreciation that the demonstrator has mental states. To reproduce a novel action without training or tuition it would seem to be essential for the observing animal to represent what the demonstrator did, but not what it thought or wanted. When the action is perceptually opaque – it yields different sensory inputs to an animal when that animal observes the action and when it executes the action (e.g., a facial expression) – imitation further implies that the imitator can represent actions in a cross-modal or sense-independent code (Meltzoff & Moore 1983). But even in these fascinating cases, mental state attribution is not implied and indeed the ascription of a theory of mind to the imitator does not help to resolve the mystery of how the imitator translates sensory input from the demonstrator's action into performance that resembles, from a third party perspective, that of the demonstrator (Heyes 1994a; 1994b; 1996).

### 2.2. Self-recognition

A series of experiments using a common procedure apparently shows that chimpanzees and orangutans, but not other primates, are capable of "self-recognition" (Gallup 1970) or "mirror-guided body inspection" (Heyes 1994c); they can use a mirror as a source of information about their own bodies (Cheney & Seyfarth 1990a; Gallup 1982; Jolly 1991; Povinelli 1987). This capacity has been said to imply the possession of a "self-concept" and the potential to imagine oneself as one is viewed by others (Gallup 1982; Povinelli 1987). I will argue that there is no reliable evidence that any nonhuman primates can use a mirror to derive information about their own bodies, and that even if there were, such a capacity would not indicate the possession of a self-concept or any other component of a theory of mind.

In the standard procedure (e.g., Gallup 1970), an animal with some experience of mirrors is anesthetized and marked on its head with an odorless, nonirritant dye; several hours later, the frequency with which the animal touches the marks on its head is measured first in the absence of a mirror and then with a mirror present. Chimpanzees and orangutans typically touch their head marks more when the mirror is present than when it is absent, while monkeys of various species and gorillas touch their marks with the same low frequency in both conditions (Calhoun & Thompson 1988; Gallup 1970; 1977; Gallup et al. 1971; Ledbetter & Basen 1982; Platt & Thompson 1985; Suarez & Gallup 1981).

There is an alternative to the standard interpretation of the chimpanzee and orangutan tendency to touch their marks more in the presence of the mirror than in its absence. In the mirror-present condition, the animals had longer to recover from anesthesia and may therefore have been more active generally than in the previous, mirror-absent condition. If they were more active generally, they had a higher probability of touching the marked areas of their heads by chance. Thus, chimpanzees and orangutans may touch their marks more when the mirror is present than when it is absent simply because at the mirror-present stage, they have had longer to recover from the anesthetic and are therefore more active generally (Heyes 1994c).

In Gallup's (1970) original experiment, two additional chimpanzees that had no prior exposure to mirrors were anesthetized, marked, and observed in the presence of the mirror on recovery. They did not make any mark-directed responses, but that does not mean that the other, mirror-preexposed animals must have been using the mirror to detect their marks. Chimpanzees typically exhibit social behavior on initial exposure to a mirror, and it is therefore likely that the control animals were too busy responding socially to their mirror image to engage in the normal grooming behavior that had, by chance, given rise to mark-touching in the experimental subjects.

According to this anesthetic artefact hypothesis, which is also consistent with the results of mark tests that vary from the standard procedure (Anderson 1983; Anderson & Roeder 1989; Eglash & Snowdon 1983; Gallup & Suarez 1991; Lin et al. 1992; Robert 1986; Suarez & Gallup 1986b; see Heyes 1994c and 1995b for reviews), species differences in mark test performance arise from the fact that chimpanzees spontaneously touch their faces with a higher frequency than either monkeys or gorillas (Dimond & Harries 1984; Gallup et al. 1995; Heyes 1995b).

The anasthetic artifact hypothesis would be less plausible if the effects of mirror insertion on face-touching were larger. In studies reported by Gallup and his associates (e.g., Gallup 1970; Gallup et al. 1971; Suarez & Gallup 1981) it is difficult to assess either the magnitude of the effect on individual animals or its statistical reliability, because the results are presented as two group total scores: the number of mark-touches made by all members of a group of animals in the mirror-present and mirror-absent conditions. The smallness of the effect, however, is apparent in data reported by other authors: Calhoun and Thompson (1988) found that, after failing to touch their marks at all during the mirror-absent period, each of two chimpanzees made just

two responses in the mirror-present condition. Thirty chimpanzees tested by Povinelli et al. (1993, Experiment 4) touched their marks, on average (±SD), 2.5 (±3.7) times in the absence of the mirror and 3.9 (±8.0) times in its presence. Swartz and Evans (1991) reported that only one of 11 chimpanzees touched its mark more in the mirror-present condition, and that, on average, 3.3 (±3.7) touches occurred while the mirror was absent, and 2.9 (±7.19) when it was present. In all three of these experiments, the mirror-present and mirror-absent periods were each of 30 minutes duration. Thus, it would not be necessary for an anesthetic recovery gradient to be improbably steep, or especially uniform across animals, to account for the mark-touching effects typically observed.

It has been suggested that the anesthetic artifact hypothesis is inconsistent with the immediacy of the effects of mirror insertion on mark touching (Gallup et al. 1995). However, I cannot find any published, quantitative data showing that mark touching is more frequent at the beginning of the mirror-present period than at its end, or that the contrast between the mirror-absent and mirror-present periods is greatest when the terminal portion of the former is compared with the initial portion of the latter. Furthermore, if such data were available, they would be equally consistent with the hypothesis that the chimpanzees use their mirror images to detect their marks, and with the hypothesis that mirror introduction elevates arousal and thereby produces an increase in the frequency of a range of behavior patterns.

It is surprising that a straightforward mark test procedure that could disprove the anesthetic artifact hypothesis has not been implemented. The procedure in question would compare the frequencies with which chimpanzees touch the marked and corresponding unmarked areas of their faces, in mirror-present and mirror-absent conditions (see Heyes 1995b for a more complete design). If it showed that chimpanzees touch the marked areas more than the unmarked areas in the mirror-present condition but not in the mirror-absent condition, then there would be reason to believe that chimpanzees can detect marks on their heads using a mirror.

However, even if there were evidence that certain primates have this capability, it would not imply the possession of a "self-concept" or the potential to imagine oneself as one is viewed by others (i.e., theory of mind; Gallup 1982; Povinelli 1987). Simple task analysis suggests that to use a mirror as a source of information about its body an animal must be able to distinguish, across a fairly broad range, sensory inputs resulting from the physical state and operations of its own body from sensory inputs originating elsewhere. If the animal could not do this, if it lacked what might be described loosely as a "body concept," then presumably it could not learn that when it is standing in front of a mirror, inputs from the mirror correlate with inputs from its body. However, a "body concept" does not relate to a mental category, and, since it is equally necessary for mirror-guided body inspection and for collision-free locomotion, the former no more implies possession of such a concept than does the latter (Heyes 1994c).

A demonstration that the humble pigeon can learn to use a mirror to detect paper dots attached to its feathers (Epstein et al. 1981) makes it easier to appreciate that mirror-guided body inspection may not imply the use of mental state concepts (but see Gallup 1983 for objections to Epstein et al.'s interpretation of their results). More direct evidence of a dissociation between the two is provided by studies of autistic children who, although apparently incapable of ascribing beliefs to others, have been reported to begin using a mirror to inspect their bodies at the same age as normal children (Ungerer 1989).

### 2.3. Social relationships

There is a substantial body of evidence suggesting that the social behavior of primates is not affected only by concurrent events and the outcomes of previous, active engagements between the present interactants and third parties. The behavior of animal A in relation to animal B also may be affected by A's prior observations of B in relation to one or a number of other conspecifics, C, D, and so on. Evidence of this kind (reviewed in Cheney & Seyfarth 1990a) has been derived from observational and experimental studies of chimpanzees, baboons, and various macaques. For example, adult male chimpanzees are more likely to disrupt (through interposition, aggression, or a threat display) social interactions between pairs of high-ranking conspecifics than between pairs of mixed or low rank (Waal 1982).

Studies of this kind show that the social behavior of animals from a broad range of primate species is sensitive to what human observers naturally describe as "social relationships" among conspecifics. It has been said, in addition, to show that primates have knowledge of social relationships (Cheney & Seyfarth 1990a; Kummer et al. 1990; Waal 1991), and this seems entirely appropriate when the term "knowledge" is used in a very general sense and social relationships are understood to be observable properties. If, on the other hand, knowledge of social relationships is taken to involve the attribution to conspecifics of knowledge about their social interactants or dispositional mental states such as loyalty, dislike, or affection, and to be acquired by a means other than associative learning (Cheney & Seyfarth 1990a; 1992; Dasser 1988; Waal 1991), then the evidence to date does not support the conclusion that primates know about social relationships.

Two studies will illustrate the plausibility of simple associative accounts of sensitivity to social relationships. In the first (Cheney & Seyfarth 1980), free-ranging vervet monkeys heard the scream of an absent juvenile from a concealed loudspeaker. The adult female monkeys in the group typically responded to the sound of the juvenile's cry by looking at the juvenile's mother before the mother had responded to the cry herself. In so doing they displayed sensitivity to or knowledge of the mother–offspring relationship. But, as the authors recognized, this could have resulted from earlier exposure to a contingency between the cries of a particular juvenile and a vigorous behavioral reaction from a particular adult female (Cheney et al. 1986).

In the second study (Stammbach 1988), one subordinate member of each of a number of groups of longtailed monkeys was trained to obtain preferred food for the group by manipulating three levers. The other monkeys did not acquire the skill themselves, but those that received the most food as a result of the trained animals' activities began to follow them to the lever apparatus and spent an increasing amount of time sitting beside and grooming the trained animals, even when the apparatus was not in operation. The untrained monkeys may have behaved in this way because they attributed to the trained individuals superior knowl-

edge of the workings of the lever apparatus, and wanted to develop friendly relations with them in the hope of gaining more food (Kummer et al. 1990; Stammbach 1988). However, the results of an experiment with rats show that, rather than attributing superior knowledge, each untrained monkey may have learned an association between the trained animal in their group and receipt of preferred food. In this study (Timberlake & Grant 1975), rats acquired affiliative social responding to a conspecific that was fastened to a trolley and wheeled into an operant chamber as a signal for the delivery of food.

### 2.4. Role-taking

In the experiments that gave rise to the suggestion that chimpanzees have a theory of mind (Premack & Woodruff 1978), a "language-trained" chimpanzee, Sarah, was shown videotapes depicting human actors confronting problems of various kinds (e.g., trying to reach inaccessible food, to escape from a locked cage, and to cope with malfunctioning equipment). The final image of each videotape sequence was put on hold, and Sarah was offered a choice of two photographs to place beside the video monitor. Both of these represented the actor in the problem situation, but only one of them showed the actor taking a course of action that would solve the problem. Sarah consistently chose the photographs representing problem solutions, and this was interpreted as evidence that she attributed mental states to the actor (Premack & Woodruff 1978; see Premack 1983; 1988 for reservations about this conclusion). It was argued that if Sarah did not ascribe beliefs and desires to the actor then she would see the video as an undifferentiated sequence of events rather than a problem.

Close examination of the published reports of the videotape experiments (Premack & Premack 1982; Premack & Woodruff 1978) suggests that for any given problem Sarah could have responded on the basis of familiarity, physical matching, and/or formerly learned associations. For example, when the actor was trying to reach food that was horizontally out of reach, matching could have been responsible for Sarah's success because a horizontal stick was prominent in both the final frame of the videotape and the photograph depicting a solution. Similarly, when the actor was shivering and looking wryly at a broken heater, Sarah may have selected the photograph of a burning roll of paper rather than an unlit or spent wick because she associated the heater with the red-orange color of fire. Taken together, however, the results of Premack and Woodruff's videotape experiments are not subject to a single, straightforward nonmentalistic interpretation, and in this respect they are apparently unique in the literature on theory of mind in primates. Thus, according to this standard, no advance has been made on the original studies of theory of mind in primates.

Premack and Dasser (1991) have devised a method of finding out whether children use theory of mind rather than a matching or contiguity principle to solve videotape problems of the kind used by Premack and Woodruff (1978). This method, however, has not been applied to nonhuman primates, and the results of other experiments on role-taking in chimpanzees (Povinelli et al. 1992a; 1992b) are unfortunately no less ambiguous than those of Premack and Woodruff. In one of these other experiments (Povinelli et al.1992a), four chimpanzees were initially trained either to choose from an array of containers the one to which an experimenter was pointing (cue detection task), or to observe food being placed in one of the containers and then to point at the baited receptacle (cue provision task). Once criterion performance had been achieved on the initial problem, each chimpanzee was confronted with the other problem, and for three of the four animals this switch did not result in a significant decline in choice accuracy.

This result was tentatively interpreted as evidence of "cognitive empathy" or "role taking . . . the ability to adopt the viewpoint of another individual" (Povinelli et al. 1992a). This interpretation rests on two tenuous assumptions: (1) Training on the first task facilitated performance on the second, and, (2) this facilitation was due to the chimpanzees having the opportunity, during the first task, to see the problem from an interactant's perspective. The former assumption is unsupported because the results failed to show that each problem was learned faster when it was presented second than when it was presented first. Consequently, it is possible that the chimpanzees' fairly high rate of learning in each task was independently influenced by their pretraining and experience outside the experimental situation. The chimpanzees had learned to pull the levers to obtain food during pretraining, and they commonly encountered and exhibited pointing behavior in their day-to-day laboratory lives.[3]

If the results of the chimpanzee experiment (Povinelli et al. 1992a) had shown that each problem (cue detection and cue provision) was learned faster when it was presented second than when it was presented first, then there would be reason to believe that some feature of the first task had facilitated performance in the second. However, even in this case, further experiments, varying the requirements of the first task, would be necessary to find out which feature was enhancing second task performance, and yet it is not clear which manipulations, if any, could provide unambiguous evidence that the opportunity for mental state attribution was responsible (Heyes 1993).

### 2.5. Deception

When applied to animal behavior, the term deception is often used in a functional sense (Krebs & Dawkins 1984) to refer to the provision by one animal, through production or suppression of behavior, of a cue that is likely to lead another to make an incorrect or maladaptive response. A mass of observational and anecdotal data leave no doubt that a broad variety of primate and nonprimate species (for excellent reviews see Cheney & Seyfarth 1991; Krebs & Dawkins 1984; Whiten & Byrne 1988) are capable of deception thus defined. However, the claim that theory of mind underlies this capacity in primates, that they sometimes act with the intention of producing or sustaining a state of ignorance or false belief in another animal, has little support. The evidence is almost exclusively anecdotal (Cheney & Seyfarth 1991; Whiten & Byrne 1988), and the behavior described in each anecdote is subject to one or more alternative interpretations.

Many anecdotal reports of deceptive behavior invite several alternative interpretations: that the behavior occurred (1) by chance, (2) as a result of associative learning, or (3) as a product of inferences about observable features of the situation rather than mental states (Heyes 1993; Kummer et al. 1990; Premack 1988). For example, "One of

the female baboons at Gilgil grew particularly fond of meat, although the males do most of the hunting. A male, one who does not willingly share, caught an antelope. The female edged up to him and groomed him until he lolled back under her attentions. She then snatched the antelope carcass and ran" (observation by Strum, cited as personal communication in Jolly 1985).

The female baboon may have intended to deceive the male about her intentions, but it may also have been no more than a coincidence that she began grooming the male when he was holding the carcass, and made a grab for the carcass when he was lolling back. Even if it did not occur by chance, the female's behavior may have been acquired through associative learning. For example, she may have snatched the carcass when the male was lolling back because in the past similar acts had proved rewarding when executed in relation to supine individuals. That is, the female could have snatched food from conspecifics on many previous occasions, initially without regard to their posture, but if she got away with it when the victim was supine, and not when the victim was upright, she could have acquired an association between snatching food and reward that was activated by the sight of a supine animal.

Even if observational studies of deceptive behavior could show that it was acquired through an inferential process rather than associative learning there would remain the possibility that the behavior was based on reasoning about observable features of the situation, or nonmental categories, rather than mental state concepts. Thus, the female baboon may have inferred from her experience of conspecific behavior that it is relatively safe to snatch food when the other animal is lying back, but she need not have regarded posture as an indicator of mental state.

The results of the only experimental investigation of intentional deception in primates (Woodruff & Premack 1979) are also equivocal. At the beginning of each trial in this study, a chimpanzee was allowed to observe food being placed in one of several inaccessible containers and then a human trainer dressed in green ("cooperative" trainer ) or white ("competitive" trainer) entered the room and searched one of the containers. The trainer had been instructed to choose the container that the chimpanzee appeared to indicate through pointing, looking, or body orientation. When the cooperative trainer found food, he gave it to the chimpanzee, but the chimpanzee was rewarded on competitive trainer trials only if the trainer chose the incorrect container. After 120 trials, each of the four chimpanzees tested showed a reliable tendency to indicate the baited container in the presence of the cooperative trainer, and an empty container in the presence of the competitive trainer. Thus, the chimpanzees' behavior toward the competitive trainer was deceptive, in the functional sense, but the process underlying this behavior is not clear. The animals may have intended to induce in the competitive trainer a false belief about the location of food, or they may have learned, through association or otherwise, that indicating the baited container in the presence of a trainer wearing green led to nonreward (Dennett 1983; Heyes 1993).

### *2.6. Perspective-taking*

**2.6.1. Seeing and knowing.** It is a fundamental tenet of human folk psychology that seeing is believing. When individuals have had visual access to a state or event X, they are likely to know about X, but without that visual access, they are likely to be ignorant with respect to X. Consequently, if nonhuman animals were spontaneously to behave in a different way toward individuals when they have and have not had visual access to an event, and if this behavior were akin to what a human would do when they took another to be either knowledgeable or ignorant with respect to that event, there would be a strong prima facie case for mental state attribution by the animal. Several experiments on "perspective-taking" in primates (Cheney & Seyfarth 1990b; Povinelli et al. 1990; 1991; Premack 1988) have been based on this kind of reasoning.

Two studies of perspective-taking in monkeys (Cheney & Seyfarth 1990b; Povinelli et al. 1991) and chimpanzees (Premack 1988) reported failure to find evidence that the subjects understood the relationship between seeing and knowing, or had the concept of "see." In the remaining study (Povinelli et al. 1990), chimpanzees were tested in a two-stage procedure. At the beginning of each trial in the first discrimination training stage, a chimpanzee was in a room with two trainers. One trainer, designated the "Guesser," left the room, and the other, the "Knower," baited one of four containers. The containers were screened so that the chimpanzee could see who had done the baiting, but not where the food had been placed. After baiting, the Guesser returned to the room, the screen was removed, and each trainer pointed directly at a container. The Knower pointed at the baited container, and the Guesser at one of the other three, chosen at random. The chimpanzee was allowed to search one container and to keep the food if it was found.

Two of the four animals tested in this way quickly acquired a tendency to select the container indicated by the Knower more often than that indicated by the Guesser, and the second stage of the procedure was designed to find out whether this discrimination was based on the trainers' visual access to the baiting operation. In each trial of this transfer stage, baiting was done by a third trainer in the presence of both the Knower and the Guesser, but during baiting the Guesser had a paper bag over his head. As before, the chimpanzees were rewarded if they selected the container indicated by the Knower. For each chimpanzee, mean choice accuracy in the final 50 trials of stage 1 was comparable with that in the 30 trials of stage 2, and this transfer performance was taken to indicate that the chimpanzees were "modelling the visual perspectives of others" (Povinelli et al. 1990). However, performance at the beginning of the transfer test was at chance level (Povinelli 1994), suggesting instead that the animals learned a new discrimination, between bagged and nonbagged trainers, during the test period. Povinelli and his colleagues have subsequently acknowledged that their experiments using the knower versus guesser procedure do not provide compelling evidence that chimpanzees understand or postulate a relationship between seeing and knowing (Heyes 1994d; Povinelli 1994).

**2.6.2. Seeing and attending.** Povinelli and Eddy (1996) recently published a series of experiments using simple discrimination procedures rather than conditional discrimination training followed by a transfer test, as in the knower versus guesser experiments. In their view, these experiments addressed the question of whether chimpanzees

understand "the attentional significance of seeing," "the mental connection engendered by visual perception" (Povinelli & Eddy 1996), and their procedures represented a methodological advance because they "allow for a very sensitive diagnosis" of whether animals' behavior is guided by elements of a theory of mind or by processes described by "traditional learning theory."

In this series of experiments (Povinelli & Eddy 1996), groups of 6 to 7 chimpanzees aged 5 to 6 years were each repeatedly presented with two trainers whose appearance differed in one of a variety of ways; the animals were rewarded with food for making a begging gesture in front of one of the trainers. For example, in one treatment condition one trainer was facing the subject (S+) while the other stood with his back turned (S−); in another condition one trainer wore a blindfold around the eyes (S−) while the other wore a blindfold around the mouth (S+). In every condition, the chimpanzees were rewarded if they gestured to the trainer that a human adult would judge to be able to see the subject (marked S+ in the foregoing examples).

Several findings from these experiments led Povinelli & Eddy (1996) to conclude that young chimpanzees probably do not understand the relationship between seeing and attending: (1) In the three conditions in which the sight of one trainer was occluded by an object (bucket, blindfold, and screen), the chimpanzees showed no "immediate disposition" to gesture to the other person. That is, in early training under these conditions they did not show a preference for the person without occluded vision. (2) When the two trainers differed on four out of five "naturalistic" dimensions, the chimpanzees did not show a preference for the S+ trainer at any point in the course of the experiments. Thus, the animals showed a preference for a person facing them over a person with his head and back turned. However, they did not gesture more to a trainer looking back over his shoulder than to one with both head and back turned, to a trainer with hands over his cheeks rather than his eyes, to someone with eyes open rather than closed, or to a person looking directly at the subject rather a person with eyes averted. (3) The subjects' performance "showed a learning curve from Experiment 1 to Experiment 13." For example, in early experiments, the animals did not gesture more to a trainer holding a screen on his shoulder than to a trainer holding the screen in front of his face, but later they performed above chance on this discrimination. (4) In the "attending versus distracted" treatment condition, one trainer looked directly at the subject (S+), while the other looked up and to the side (S−). On these trials, the chimpanzees often turned their heads in the direction of the S− trainer's gaze, a behavior that is regarded by some developmentalists as indicating understanding of the seeing-attention relationship; but in spite of this the chimpanzees gestured at random to the two trainers.

These results provide no encouragement for the view that young chimpanzees understand anything about "seeing," but neither do they constitute compelling negative evidence; they should not persuade us that young chimpanzees do not understand "seeing." One would expect animals with the concept "see" to be capable of using the visibility of the trainer's eyes, not merely his face or the front of his body, as a discriminative cue for begging. This capacity would not necessarily become apparent on the first trial of a laboratory test, however, nor indeed at any point in the set of trials given in Povinelli and Eddy's study. Even if the chimpanzees had the concept "see" before the experiment began, it could take them some time to become convinced that it was the basis on which they were required to discriminate in this particular set of problems. Furthermore, since eyes visible versus invisible, and eyes direct versus averted, are perceptually fine discriminations, the chimpanzees may have neglected to try hard on those trials, opting instead to collect their rewards during the easier trials in which the difficult ones were embedded.

Could the procedures used by Povinelli and Eddy (1996) have provided positive evidence of theory of mind? The experiments were presented as if certain outcomes would have supported a theory of mind interpretation over a nonmentalistic account or, more narrowly, a learning theoretic explanation. If this were true, these procedures would represent a major methodological advance because, as I have argued above (see also Heyes 1993), no other methods used to date in research on theory of mind in nonhuman primates have succeeded in doing this. Unfortunately, however, Povinelli & Eddy's procedures cannot do it either. Simple discrimination techniques of the kind they used can tell us which observable cues chimpanzees use when deciding whom to approach for food, but they cannot tell us *why* the chimpanzees use those cues; whether certain cues are important to them because, within the chimpanzees' theory of mind, those cues indicate "seeing," "attention," or "knowledge."

Imagine, for example, that Povinelli and Eddy had found that all of their chimpanzees immediately showed perfect discrimination on the basis of the visibility of the trainer's eyes. Thus, from the very first trial, the chimpanzees not only preferred a trainer with a bucket on his shoulder to one with a bucket over his head, but also preferred a person with his eyes open over one with his eyes closed, and even preferred a trainer looking directly at the subject (irises visible as circles), over a trainer with their eyes averted (irises visible as ellipses). By hypothesis, the data would indicate unambiguously that chimpanzees use eyes as a discriminative stimulus when deciding which of two trainers to approach for food. Even these data would be equally compatible with a theory of mind and a nonmentalistic explanation, or, as Povinelli and Eddy put it, with a "mentalist" and a "behaviorist" hypothesis. A theory of mind account would say that chimpanzees use eyes as a discriminative stimulus because they understand that an individual whose irises are visible as circles can "see" them, and that seeing is a mental state linked to attention or knowledge. A nonmentalistic account would say that the chimpanzees just do it; they have a learned or unlearned tendency to beg from people with visible eyes, and while the chimpanzees may even know that begging from people with visible eyes is more likely to lead to reward, they do not explain this contingency to themselves in mental terms or in any other way.

Note that the essential difference between the theory of mind hypothesis and the nonmentalistic hypothesis does not relate to whether the use of eyes as discriminative stimuli was learned or unlearned. If the chimpanzees in Povinelli and Eddy's experiments had shown perfect performance from the first trial, both mentalistic and nonmentalistic accounts could have attributed this to preexperimental learning or to an innate disposition. (Even "traditional learning theory" does not claim that all behavior

is learned.) The difference is that a theory of mind hypothesis would say that it was an understanding of the seeing-attending or seeing-knowing relationship, as well as a tendency to use eyes as discriminative stimuli, that was present before the experiment began. Similarly, improvement in performance over recorded trials could be attributed on both mentalistic and nonmentalistic accounts to learning during the experiments, or to the gradual unmasking of some preexisting tendency. Thus, a theory of mind hypothesis might say that the chimpanzees learned about the seeing-attending relationship in the course of the experiments, or that they already knew about that relationship but needed to discover its task relevance or to learn some new cues instantiating the seeing relation. A nonmentalistic hypothesis might say that the animals learned through the experiment to use eyes as discriminative stimuli, or that a preexisting tendency to do this only became apparent when the animals had become fully accustomed to all aspects of the testing procedure. In this example, and in the search for evidence of theory of mind in nonhumans more generally, the crucial difference between mentalistic and nonmentalistic hypotheses lies in their claims about "what is known," not about whether or how knowledge is acquired.

In view of their discouraging findings with 5- and 6-year-old chimpanzees, Povinelli and Eddy (1996) recommended that older chimpanzees be tested for theory of mind competence. This is a useful suggestion, but, if there is to be any chance of finding positive evidence of theory of mind, different test procedures must be found.

#### 2.7. Summary

Research on imitation and mirror-guided body inspection (sects. 2.1 and 2.2 above) has not shown unequivocally that any primate has these behavioral capacities, and they could, in any event, be the products of associative learning and inferences involving nonmental categories. Thus, for imitation and self-recognition, the answers to both competence and validity questions are negative.

There can be little doubt that the members of many primate and nonprimate species exhibit sensitivity to social relationships and behavior that functions to deceive other animals (sects. 2.3 and 2.5 above); hence the answer to the competence question is affirmative for both social relationships and deception. However, in every case the relevant behavior could be based on one or a number of nonmentalistic psychological processes, and therefore these behavioral capacities are not valid indicators of theory of mind.

The position with respect to role-taking and perspective-taking is more complicated. Premack and Woodruff's (1978) research on role-taking (sect. 2.4) provided the first and arguably the strongest evidence to date of theory of mind in a nonhuman primate (Premack & Woodruff 1978). It showed that a chimpanzee was capable of matching problem-solution images; she had this behavioral competence, and it is difficult, but not impossible, to query the validity of this competence as an indicator of theory of mind. In contrast, Povinelli et al. (1992a) did not show that cue detection training facilitates chimpanzees' performance in a cue provision task, or vice versa, and even if such an effect had been demonstrated, it would not necessarily indicate theory of mind. Therefore, for the cue detection/provision task studies on role-taking, the answers to both competence and validity questions are negative.

The knower-guesser procedure used by Povinelli and his associates to investigate perspective-taking (sect. 2.6; Povinelli et al. 1990) involved a transfer test procedure with considerable potential. It could, I will argue below (sect. 4), provide evidence of behavioral competence validly indicating that primates have the concept "see." However, as yet, the answers to the competence and validity questions are negative for all perspective-taking studies. Neither the knower-guesser procedure nor the simple discrimination tests used by Povinelli and Eddy (1996) have shown that primates use the visibility of interactants' eyes to decide whom to approach for food; and such evidence would not be sufficient to implicate possession of the concept "see."

## 3. Procrastination

Progress in answering Premack and Woodruff's question requires experimental designs and test procedures that can distinguish the theory of mind hypothesis from nonmentalistic accounts of primate behavior. This requirement has been explicitly acknowledged by a few researchers (e.g., Povinelli & Eddy 1996; Premack 1988). However, the primacy of the need has been obscured and attempts to meet it may have been retarded by various attempts to show that data of the kind surveyed in section 2 either favor a theory of mind hypothesis outright or at least provide "suggestive" evidence of theory of mind. These arguments typically concede that each item of putative evidence for theory of mind in primates is susceptible to alternative interpretations, and an appeal is made to parsimony or convergent evidence to break the tie. Five arguments of this kind (two appealing to parsimony and three to convergence) are evaluated in this section.

#### 3.1. Parsimony

**3.1.1. Simpler for them.** In their seminal paper, Premack and Woodruff (1978) suggested that "the ape could only be a mentalist . . . he is not intelligent enough to be a behaviorist." This raises the possibility that the application of theory of mind (or "mentalism") requires less intelligence of an ape than alternative "behaviorist" methods of predicting behavior, and therefore, by appealing to Lloyd Morgan's Canon or a similar principle of parsimony, one could justify preferring a theory of mind interpretation of behavior over an alternative when both are consistent with the data.

There are two problems with this argument in favor of the theory of mind hypothesis. First, there is no good reason to suppose that the acquisition and use of a theory of mind requires less intelligence, or is in any sense "simpler," for an animal than the acquisition and use of an alternative basis for predicting social behavior. Neither intelligence nor simplicity has been defined or measured in a way that would allow a reasonable comparison to be made. Premack and Woodruff pumped the intuition (Dennett 1980) that an alternative to theory of mind would require more intelligence by dubbing it "behaviorist," and thereby suggesting that the animal would have to master the contents of the *Journal of the Experimental Analysis of Behavior.* However, if one resists this sort of intuition, it is clear that, although a more consistent analogy would portray chimpanzees that lack a theory of mind as "associationists" or "cognitivists" rather than "behaviorists," all of these characterizations are misleading because alternatives to the theory of mind

hypothesis do not assume that chimpanzees and other animals know anything about the processes that they use to predict social behavior. Only the theory of mind hypothesis takes chimpanzees to be students of their own psychology. It claims that mental states such as wanting and believing control behavior, and that knowledge of such states – mental state concepts – is used in social interaction. In contrast, alternatives to the theory of mind hypothesis postulate just one layer of processes or representations that generate behavior in social contexts and elsewhere.

Second, even if theory of mind were demonstrably less demanding of intelligence or simpler than the alternatives (or vice versa) this would not be sufficient to justify preference for one account over another. The view that preference for more parsimonious explanations can be justified by appeal to a general ontological assumption such as the uniformity of nature (Hume 1748/1948), has been broadly rejected by philosophers of science (e.g., Boyd 1985; Sober 1988). Therefore, in addition to showing that theory of mind would be simpler than the alternatives, it would be necessary to argue that in the case of primate social behavior, in this particular corner of nature, a simpler process is more likely to be in operation than a more complex one (Sober 1988).

**3.1.2. Simpler for us.** Dennett (1983; 1989) has argued that taking "the intentional stance" toward animals, characterizing their behavior in terms of the actor's intentional states, can have practical advantages. He claimed that for field ethologists observing animals in their natural environments, the intentional stance is easier to use than the languages of behaviorism or information processing, and that by happy coincidence intentional descriptions of animal behavior provide important clues for the cognitive scientists whose job it is to explain that behavior by modeling the information processing systems that are really in control.

As far as I am aware, no one actively engaged in research on theory of mind in primates has explicitly claimed, with or without reference to Dennett, that theory of mind explanations should be preferred to nonmentalistic alternatives because the former are simpler for (some) people to understand. However, the "simpler for them" argument is commonly advanced and yet weak (see sect. 3.1.1), raising the possibility that researchers are implicitly assuming that theory of mind is simpler for primates to use because theory of mind hypotheses are often simpler for us to understand. Accordingly, it is worth reflecting on the "simpler for us" argument.

The first thing to note is that Dennett's arguments cannot (and were not designed to) justify a preference for the theory of mind hypothesis over nonmentalistic accounts of the kind of evidence reviewed in section 2 (Heyes 1987). On the contrary, they imply that, although it is legitimate for field ethologists to speak and write about animals as if they had mental states and mental state concepts, the broader research community should seek, and indeed prefer as explanations, theories that do not make reference to such states and concepts.

Leaving aside Dennett's more subtle position, it might be argued that if the theory of mind hypothesis is simpler for us to comprehend than alternative accounts of primate social behavior, this would be sufficient reason to prefer it over nonmentalistic accounts. This argument assumes that the principle of parsimony or simplicity is "purely methodological" (Sober 1988); that, regardless of whether we can justifiably assume that nature is simple, it is rational to prefer simple theories (e.g., Strawson 1952).

Even if one accepts that the principle of parsimony is purely methodological (and Sober 1988, gives compelling reasons not to accept this), there is a problem with the argument that because it is simpler to comprehend the theory of mind hypothesis should be accepted instead of nonmentalistic accounts of the current data on social behavior in primates. It is not clear that the theory of mind hypothesis is simpler in a way that should carry any weight. For some people, for example, who are unfamiliar with associative learning theory and cognitive psychology, it may be easier to understand and apply. However, this does not seem to be the kind of simplicity that was at issue in the historical episodes that led to parsimony being viewed as a methodological principle (e.g., Reichenbach 1951; Sober 1988). For example, it is unlikely to have been a user-relative conception of simplicity – a dimension defined by individual scientists' professional and educational backgrounds – that guided Einstein's reasoning to the special and general theories of relativity.

### *3.2. Convergence*

**3.2.1. More is better.** Much of the putative evidence of theory of mind in primates is anecdotal; it consists of reports of single occurrences of a behavior, under uncontrolled conditions, made by isolated observers, or groups of observers who share a theoretical base. The profound weakness of this kind of evidence has been demonstrated repeatedly (e.g., Kummer et al. 1990; Premack 1988), and yet anecdotes continue to be published and treated as persuasive. In most cases, this is done without commentary or defense and, to their credit, Whiten and Byrne (1988) stressed that anecdotes are a prelude, not a substitute, for more systematic research and offered a rationale for their collection of anecdotes about deceptive behavior in primates. They suggested that a collection of anecdotes relating to the same category of behavior will constitute evidence of theory of mind provided that (1) the reports come from independent observers, and (2) each provides evidence that the act involved the agent representing the viewpoint or beliefs of others.

Whiten and Byrne's second criterion seems to be self-defeating. Their "multiple records" approach is designed to compensate for the fact that single anecdotes cannot provide evidence of theory of mind, and yet their second criterion requires each anecdote in a collection to provide such evidence for the ensemble to be persuasive. Nor does combining the second criterion with the first offer an escape from this circularity. Consider the hypothetical example of three animals seen by independent observers (criterion 1) snatching food that was previously available to a conspecific. The first, like the baboon reported in Jolly (1985, see sect. 2.4), grooms the conspecific and snatches when it is supine; the second presents and grabs when the male is sexually excited; and the third throws a missile and makes his move when the conspecific is giving chase. Each observer might feel inclined to attribute the state of "intending to deceive with intimate behaviour" to the animal observed (Whiten & Byrne 1988), but the potential to attract the same mental state attribution from the human

observers might be all that the three animals have in common with regard to mental state concepts. Even if we could be sure that none of them had simply been lucky and that all of them had acquired the behavior through some inferential process, the possibility would remain that the animals learned to snatch from supine, sexually excited, and departing individuals, respectively.

This example illustrates that "the plural of anecdote is not data" (Bernstein 1988), but the point can also be generalized: the mere accumulation of data, whether anecdotal, observational, experimental, or a mixture of the three, does not necessarily provide convergent evidence. The literature reviewed in section 2 shows that in a range of social interactions (e.g., competitive and cooperative; dyads, triads, and larger groups; same and different gender, status, age, and species; in relation to feeding, grooming, mating, and mothering), the behavior of many individual apes has been interpreted as a manifestation of theory of mind. But to make the case for the theory of mind hypothesis more compelling on the grounds of convergence, one would need to show not merely that it can be applied to diverse phenomena but that for each of a range of phenomena it provides a better explanation than alternative, nonmentalistic hypotheses.

**3.2.2. Apes can and monkeys can't.** Humans have a theory of mind; nonhuman apes are more closely related to humans than are monkeys; and according to one school of thought closely related taxa are more likely than groups with a more distant common ancestor to have the same cognitive capacities. Therefore, one might argue, if nonhuman apes perform better than monkeys on tests designed to assess theory of mind, then, all other things being equal, the difference between the two groups provides convergent evidence that the apes' successful performance on the tests is a product of theory of mind rather than nonmentalistic thinking.

Unlike "more is better," this is a potentially sound convergence argument. However, it does not succeed in breaking the current deadlock between the theory of mind hypothesis and nonmentalistic accounts of primate behavior because in tests where apes have fared better than monkeys all other things have not been equal. For example, Gallup and his colleagues (e.g., Gallup 1970; Gallup et al. 1971; Suarez & Gallup 1981) have found that chimpanzees and orangutans pass, but various species of monkey fail, the mark test of mirror self-recognition. This could be owing not to the presence of a self-concept in apes and a lack of the same in monkeys but to the fact that apes spontaneously touch their faces more often than do monkeys (Dimond & Harries 1984; Gallup et al. 1995; Heyes 1994c; 1995b; 1995c; see sect. 2.2 above). Similarly, using the task in which subjects must choose a container indicated by one of two people, the Knower or the Guesser, Povinelli et al. (1990; 1991) found that chimpanzees did – and rhesus monkeys did not – learn to choose reliably the container indicated by the Knower. But this may not reflect a difference between the two groups in the capacity to model the visual perspectives of others, or to appreciate that seeing leads to knowing. Rather, it may have occurred because in the monkey experiment but not in the chimpanzee experiment the Knower moved around the room after baiting and before the subject had its choice. Thus, it would have been more difficult for the monkeys to remember on any given trial which trainer had been present during the baiting.

To be effective, an argument from ape-monkey contrast to the conclusion that apes have a theory of mind would need to show that the contrast in performance could not plausibly be ascribed to differences in task demands, sensory or motor functioning, or central processes not specifically related to theory of mind (e.g., working memory). As the foregoing examples illustrate, this has not been achieved, even in those rare and admirable experiments that have compared monkeys and apes using common procedures.

**3.2.3. Chimps are like children.** Another potentially strong but currently ineffective convergence argument is the following: the performance of chimpanzees (and/or other nonhuman apes) on theory of mind tasks is likely to reflect the use of a theory of mind rather than nonmentalistic processing, because the chimpanzees' performance resembles that of children in similar circumstances and there is independent evidence, often from verbal measures, that the children's behavior is based on a theory of mind. Current evidence does not support this argument, however, because, in the very few studies that have compared the behavior of chimpanzees and children under similar circumstances, the resemblance between the two or the independent evidence that the children were using theory of mind is weak.

Experiments on imitation (Tomasello et al. 1993) and self-recognition (Povinelli et al. 1993) provide examples of the first problem: poor resemblance between chimpanzees and children. Tomasello et al. (1993) found that in terms of their tendency to duplicate a model's actions on objects, "enculturated" chimpanzees were more like children than were nonenculturated chimpanzees. Although the children imitated fewer actions at a delayed test than at an immediate test, the enculturated chimpanzees showed the reverse pattern of performance.

Povinelli et al. (1993) reported that the mirror self-recognition behavior of chimpanzees and children is alike merely in that each shows a developmental trend, yet even this very general resemblance was not confirmed by the results. Reanalysis of the data from this study[4] (Heyes 1995b) showed that older chimpanzees were no more likely than younger ones to pass the mark test of self-recognition; and although 8- to 15-year-old chimpanzees showed more self-directed behavior in the presence of mirrors than 1- to 5- year-olds, the frequency of this behavior declined sharply between ages 15 and 39. The latter finding suggests either that, unlike humans, (1) chimpanzees typically acquire a self-concept as children and then promptly lose it on reaching adulthood, or (2) that self-directed behavior in the presence of mirrors is not a valid measure of self-conception.

In a study of perspective-taking, Povinelli's group (Povinelli et al. 1990; Povinelli & deBlois 1992) sought and found a more precise resemblance between chimpanzees and children, but in this example there was no compelling evidence that the children's behavior was guided by a theory of mind. Povinelli & deBlois (1992) found that 4-year-old children were more successful than 3-year-olds on a task similar to the Knower versus Guesser discrimination problem previously given to chimpanzees (see sect. 2.6; Povinelli et al. 1990). This does not, however, indicate that the chimpanzees' success on the problem was based on an understanding of the relationship between seeing and knowing, because the children who consistently chose the

Knower were no more likely than the unsuccessful children to answer correctly a question about what the Guesser could see when they had left the room.

### 3.3. Conclusion

Each of the foregoing parsimony and convergence arguments could be put into reverse to motivate acceptance of nonmentalistic accounts of the data reviewed in section 2. Thus, it could be argued that theory of mind would require more intelligence of primates because it involves more than one layer or level of representations (sect. 3.1.1), and that nonmentalistic accounts are simpler from the investigator's perspective because they proceed from clearly specified assumptions rather than a largely implicit folk theory (sect. 3.1.2). Similarly, appealing to the "more is better" principle (sect.3.2.1), one could point to all of the nonsocial behavior of people and animals that can be explained by nonmentalistic theories; and, countering the argument from ape-monkey contrast (sect. 3.2.2), one could draw attention to the nonprimate species (including rodents, birds, and arthropods) that exhibit the kind of behavior interpreted as evidence of theory of mind when it appears in primates. Finally, one might note that, when direct comparisons have been made, it has turned out that in important respects chimps are *not* like children.

All of these arguments could be made at least as plausible as their counterparts in the existing literature on theory of mind in primates, but, in my view, it would be a mistake to pursue this option. To answer Premack and Woodruff's question, we need more strong experiments, not more weak arguments.

## 4. Proposals

### 4.1. Methods and questions

I have argued in sections 2 and 3 that research to date on theory of mind in primates does not show that they have such a theory. I also believe that it does not indicate that primates lack a theory of mind, or that Premack and Woodruff's question is unanswerable. There may be circumstances in which repeated failure to find evidence confirming a hypothesis can be interpreted rationally as a sign that the hypothesis is false, and it is conceivable that theory of mind and nonmentalistic accounts of primate social behavior are observationally equivalent. However, both of these negative conclusions would be premature because very few deliberate, potentially effective attempts have been made to test the theory of mind hypothesis against nonmentalistic alternatives. Research on imitation (see sect. 2.1) and self- recognition (sect. 2.2) has been used opportunistically to support the theory of mind hypothesis, most having been conducted to address other questions; and the vast majority of studies of social relationships (sect. 2.3) and deception (sect. 2.4) have used observational or anecdotal methods that lack the potential to distinguish the theories because they provide no information about the animals' histories (Heyes 1993). Just a handful of studies – of deception (Woodruff & Premack 1979), role-taking (Premack & Woodruff 1978; Povinelli et al. 1992a), and perspective-taking (Povinelli et al. 1990) – have been designed to pit the theory of mind hypothesis against an alternative while using a potentially reliable method to do so. Further empirical studies of theory of mind in primates are accordingly needed and warranted, but which methods should they use, and what kind of behavior should they examine?

The foregoing analysis (sects. 2 and 3) yields six principal recommendations for future research on theory of mind in primates, none of which is entirely original.

(1) Studies should be designed to distinguish the theory of mind hypothesis from nonmentalistic accounts of social behavior in primates. There is little point in reporting any more observations that are consistent with both kinds of account, or conducting experiments for which they would both predict the same outcome.

(2) It should be recognized that alternatives to theory of mind hypotheses are not necessarily "behaviorist" or derived from learning theory. The social behavior of primates may be based on abstract, symbolic representations of nonmental categories.

(3) Whether they are field- or laboratory-based, studies of theory of mind should involve experimental manipulation. Certain experimental methods (e.g., Povinelli et al. 1990; Premack & Woodruff 1978; Woodruff & Premack 1979) have come closer than any observational study to providing evidence of theory of mind in primates, and, although there are plans in place to increase the effectiveness of these methods (see Premack & Dasser 1991, and sect. 4.2 below), it is not clear how any observational study could distinguish the theory of mind hypothesis from its nonmentalistic alternatives.

(4) Investigations of role-taking, deception, and perspective-taking are more likely than research on imitation, self-recognition, and social relationships to tell us whether nonhuman primates have a theory of mind. The problems with attempts to demonstrate imitation and mirror-guided body inspection in primates are not intractable (for potential experimental designs see Heyes 1994c; 1995b), but there is little reason to suppose that mental state concepts are involved in imitation, self-recognition, and the kind of behavior examined under the heading of "social relationships" (sect. 2.3).

(5) Experiments that use a common procedure to compare the behavior of monkeys, nonhuman apes, and children (or adults) are more likely to yield compelling evidence of theory of mind in apes than studies of apes alone.

(6) The knower-guesser procedure used by Povinelli et al. (1990; see sect. 2.6 above) to investigate perspective-taking is particularly promising. This "triangulation" method (Campbell 1953; Heyes 1993) consists of conditional discrimination training followed by transfer tests, and its power lies in the fact that it requires animals to distinguish one mental state, X (e.g., knowing where food is hidden), from another, Y (e.g., not knowing where food is hidden), in two or more situations that differ in terms of the observable cues that might be correlated or confounded with X and Y. In the training situation, X is confounded with feature A (e.g., did the baiting) and Y with feature B (e.g., absence during baiting) of the social interactants' appearance or behavior, but in the transfer test, X and Y are correlated with features C (e.g., no bag during baiting) and D (e.g., bag during baiting), respectively. If the animal's behavior is unchanged despite this shift in observable stimuli, and if the most plausible account of the relationship between A and C on the one hand and B and D on the other construes them as indicators or manifestations of X and Y

respectively, then one has evidence of the application of mental state concepts X and Y. Thus, triangulation has the potential to overcome the problem of confounding or correlated cues, not primarily by virtue of the quality of a single test or measurement procedure, but by compounding tests, each of which is fallible, but in a different way.

More generally, it would be desirable for researchers with different expertise and theoretical commitments to collaborate in planning studies of theory of mind in primates. Combining commitment to the theory of mind hypothesis with skepticism and skills in experimental design with knowledge of the habits and natural history of primates would guard against confirmation bias, and would maximize our chances of developing procedures that are both practicable and potentially effective in testing the theory of mind hypothesis against nonmentalistic alternatives. To make this implementation of the "fishscale model of omniscience" (Campbell 1969) more than simply a pious wish, I describe below a test procedure that looks to me as if it could yield evidence of perspective-taking in primates. *BBS* commentators are invited to say what is wrong with it and how it could be improved or replaced by a potentially more effective method.

### 4.2. A potential study of perspective-taking

Initially, adult chimpanzees would be tested for perspective-taking using a version of the triangulation procedure developed by Povinelli et al. (1990; see sect. 2.6 above). Departures from this procedure would include (1) the presentation of nonreinforced probe trials rather than a new discrimination problem, when the initial discrimination has been learned; (2) use of trainers wearing opaque or translucent goggles, rather than a bag-on-head manipulation, for transfer trials; and (3) introduction of a pretraining phase in which the subjects are exposed to opaque and translucent goggles with distinctively colored rims. The first of these would ensure that successful "transfer" performance could not be due to learning of a new discrimination (see sect. 2.6 above) and, in combination, the latter two features of the experiment would make it unlikely that the animals could solve the problem using an observable cue, such as "eye-object line" (Heyes 1994d) – that is, by choosing the trainer for whom there is or was an unobstructed, notional straight line between their eyes and the baiting event. Preexposing subjects to the goggles would allow them, if they have the concept "see," to discover that one pair of goggles permits the wearer to see, while the other pair does not. If they subsequently prefer to take their cue from a trainer wearing translucent rather than opaque goggles, and if the only observable indication of which goggles the trainer is wearing is an arbitrary one (i.e., rim color) then it would seem that the subjects' preference for a person wearing translucent goggles could only be due to their attributing sight of the baiting event to that trainer. Use of goggles in a similar context was recommended by Gallup (1985; 1988) and Nicholas Humphrey (personal communication), and goggles were used by Novey (1975) in a study of infants. Cheney and Seyfarth (1990a) also used a similar manipulation in an experiment with monkeys.

In more detail, the procedure would be as follows.

(1) *Pretraining.* The chimpanzees would be trained, if necessary with food reward, to cover their eyes with two pairs of goggles. The two pairs would have rims of different colors, say red and blue. For half of the animals, the red-rimmed goggles would be opaque and the blue-rimmed translucent, while the other half would have the reverse assignment. Neither at pretraining nor at any other time will the chimpanzees see another person or animal wearing goggles. Furthermore, the opaque and translucent versions should be discriminable at a distance, that is, when worn by another individual, only in terms of their rim color. To check that this is the case, an attempt would be made to train chimpanzees that are not taking part in the main experiment on a simple discrimination between a trainer wearing opaque and translucent goggles with rims of the same color.

If it was found during pretraining that chimpanzees are highly resistant to putting goggles over their eyes, or that any aversion to the opaque goggles does not habituate in the course of pretraining (a possibility raised by Perner 1991), or that willingness to wear the two sets of goggles cannot be equalized by appropriate distribution of rewards, then one-way and two-way silvered screens, with distinctively colored frames, could be used in place of opaque and translucent goggles.

(2) *Training.* Using an apparatus and procedure like those of Povinelli et al. (1990), each chimpanzee would be presented on each trial with four containers and two trainers. One of the trainers would leave the room while a third person baited one of the containers; then each trainer would point at a container, and the chimpanzee would be rewarded for selecting the container indicated by the trainer who had been present during baiting.

(3) *Transfer.* When the animals had reached criterion on the training problem, trials of the kind used in training would be interspersed with occasional probe trials, in which both trainers would remain in the room and put on goggles during baiting. The Knower would put on translucent goggles, and the Guesser would wear opaque goggles. The subjects would never, or always, be rewarded on probe trials, regardless of the container they chose. The important point is that they would not be rewarded consistently for choosing either the Knower or the Guesser. If chimpanzees have the concept "see," then on probe trials one would expect them to choose the Knower, wearing translucent goggles, more often than the Guesser, wearing opaque goggles.

If, in the foregoing experiment, chimpanzees did not show a preference for the Knower over the Guesser, it may be worth running a variant that would contain fewer irrelevant cues or distracters, would make less demand on subjects' working memory, and would not rely on test trials in which the subjects' motivation is uncertain because responding is not differentially reinforced. This variant would begin with the same pretraining and would subsequently involve a successive, rather than a simultaneous, discrimination problem, using rate of learning rather than performance under nondifferential reinforcement as a measure. Thus, at the beginning of each trial in the training phase, a chimpanzee and a human trainer would face one another in a modified Wisconsin General Test Apparatus containing two covered food wells. The trainer would then either look intently at the food wells as one of them was baited by a third party (front trials) or turn, so that during baiting the chimpanzee and the food wells were behind the trainer's back (back trials). A screen between the wells and the chimpanzee would allow the latter to see the trainer and that baiting was occurring, but not where the food was

placed. After baiting, the trainer would face the subject and indicate one of the wells by placing his hand on it and the chimpanzee would be free to choose one well to search for food. On front trials, the trainer would point at the baited well and on back trials he would point at the other well. When the subjects had learned to select the well indicated by the trainer on front trials, and the other well on back trials, the transfer phase would begin, in which the trainer would wear translucent or opaque goggles. For half of the subjects, the trainer would indicate the baited well on translucent trials and the empty well on opaque trials (Group Direct) and for the other half, the trainer would indicate the baited well on opaque trials and the empty well on translucent trials (Group Reverse). If chimpanzees have the concept "see," one would expect Group Direct to learn faster than Group Reverse in the transfer phase. That is, Group Direct should learn to choose the well indicated by the trainer on translucent trials and the other well on opaque trials faster than Group Reverse learns to choose the well indicated on opaque trials and the other well on translucent trials.

The logic of both experimental designs requires training on only one discrimination problem before the transfer phase. In practice, however, it might be advisable to train the chimpanzees before transfer on more than one pair of stimuli instantiating the see versus cannot see distinction. This would help to ensure that if the chimpanzees have or can acquire the concept "see," they know by the time the transfer phase starts that it is relevant to the tasks in hand.

If either of these experiments had the predicted outcome, it would be desirable to repeat it using children as subjects. Each child would be tested using the same basic procedure as the chimpanzees but would also be given another test, preferably one that had already been validated as a measure of the theory of mind competence in question. Correlation between performance on the two tests would constitute convergent evidence that first measured some aspect of theory of mind and would encourage its use with other nonhuman species, including monkeys.

It would be very surprising indeed if these experimental proposals turned out to be easy to implement and did not contain any logical flaws. Research on theory of mind in primates would have made more progress in the last 20 years if single, crucial experiments were a possibility and if an effective research strategy were easy to formulate. However, I hope the proposals will contribute, after modification and refinement through open peer commentary, to the development of an effective experimental program.

### *4.3. On killing joy*

In one of his inspired baptisms, Dennett (1983) gave the name "killjoy hypotheses" to explanations of behavior that eschew ascription of higher order intentionality or theory of mind to animals. Plenty of killjoy hypotheses have been discussed in this target article, and they will, as Dennett recognized, provoke a negative reaction in many readers. The idea that primates have a theory of mind is important and intriguing, and a great deal of careful labor has been devoted to its investigation. Therefore, it can be disappointing and irritating to be reminded that there are other, less exciting explanations for the reported data, especially when the recognition of these other possibilities requires close examination of methodology. It can seem as if elegantly bold ideas are meeting carpingly narrow objections, and in such a contest our instincts, or at least my instincts, are not to shout for the methodologists. But it is precisely because Premack and Woodruff's question is important and intriguing that it warrants a reliable answer; and without some sober reflection, acknowledging the limitations of current research, we may never know whether nonhuman primates have a theory of mind.

ACKNOWLEDGMENTS

I am grateful to Anthony Dickinson, Nicholas Mackintosh, Euan MacPhail, Henry Plotkin, Phil Reed, Elliott Sober, Andrew Whiten, and several anonymous referees for their comments on earlier drafts, and to Linnda Caporael, Richard Darby, Dorothy Einon, Christa Foster, Mark Gardner, Tim German, Chris Mitchell, Tristan Nokes, Kate Plaisted, Elizabeth Ray, and David Shanks for many useful and enjoyable discussions of the content. I owe the idea for this paper, and a great deal more, to Donald T. Campbell.

This research was supported by a grant from the UK Biotechnology and Biological Sciences Research Council.

NOTES

**1.** This target article adopts a realist position on mental states. It assumes that for most adult humans mental states and mental state concepts play a causal role in the generation of behavior, and it asks whether there is evidence that this is also true of any nonhuman primates. From a behaviorist perspective, the question "Do nonhuman primates have a theory of mind?" may be either incoherent or a question about whether human observers are willing to describe the behavior of nonhuman primates using certain mental terms. In either case, detailed analysis of the evidence of the kind presented here is otiose; the question is unanswerable, or the answer, apparent in common experience, is an emphatic "yes." People spontaneously speak, not only of other primates, but of nearly all other living things, as if they had mental states and a theory of mind.

**2.** Tomasello and his colleagues have advanced the interesting and more general thesis that, as a result of their extensive interaction with humans, enculturated apes engage in forms of social cognition beyond the capabilities of wild monkeys and apes (e.g., Tomasello 1996; Tomasello & Call 1994; Tomasello et al. 1993). This thesis is not a focus of the present discussion because, although Tomasello et al. claim that the behavior of enculturated apes is "intentional," they apparently mean by this that it is directed toward some purpose and involves thought of some kind, not, more specifically, that it implies theory of mind or the attribution of mental states.

**3.** The potential significance of pointing is indicated by evidence that rhesus monkeys, which do not normally show pointing behavior, did not immediately succeed on their second problem when switched from cue provision to cue detection, or vice versa (Mason & Hollis 1962; Povinelli et al. 1992b). Hess et al. (1993) showed that a rhesus monkey, Scarlet, who does point, fared no better than her conspecifics when switched from cue provision to detection. However, as Hess et al. acknowledged, since Scarlet is a single animal who may not point as much as the average chimpanzee, these data do not rule out the possibility that chimpanzees' performance on both tasks is facilitated by a preexisting habit of pointing.

**4.** I am grateful to Daniel Povinelli for supplying, immediately and in full, additional data from the studies reported by Povinelli et al. (1993).