Chimpanzee Theory of Mind: Looking in All the Wrong Places?

KRISTIN ANDREWS

Abstract: I respond to an argument presented by Daniel Povinelli and Jennifer Vonk that the current generation of experiments on chimpanzee theory of mind cannot decide whether chimpanzees have the ability to reason about mental states. I argue that Povinelli and Vonk's proposed experiment is subject to their own criticisms and that there should be a more radical shift away from experiments that ask subjects to predict behavior. Further, I argue that Povinelli and Vonk's theoretical commitments should lead them to accept this new approach, and that experiments which offer subjects the opportunity to look for explanations for anomalous behavior should be explored.

1. Introduction

In a pair of recent papers, Daniel Povinelli and Jennifer Vonk (2003, 2004) offer a critique of the current research paradigm being used to investigate chimpanzee theory of mind. Defining theory of mind as 'the ability to reason about mental states' (2004, p. 1), Povinelli and Vonk (P&V) claim that there is no evidence that chimpanzees reason about mental states such as seeing. Though recent research by Michael Tomasello, Brian Hare, and Josep Call seems to suggest that chimpanzees can reason about seeing, P&V argue that the paradigms presented in support of Tomasello *et al.*'s conclusion are not able to distinguish between mentalistic and behavioristic psychological systems in chimpanzees. In their most recent paper, P&V have introduced a novel paradigm which they believe could decide between the competing hypotheses.

While P&V are right to suggest a worry with the current research paradigm on chimpanzee theory of mind, I think that the arguments they present against Tomasello *et al.* are misplaced. Part of the problem is that their novel paradigm is subject to the same criticisms they present against the old paradigm. However, P&V hint at a different solution to the problem of developing research paradigms that will distinguish between mentalistic and behavioristic explanations of

Support for the presentation of an earlier version of this manuscript was given by the Social Sciences and Humanities Research Council of Canada. I would like to express thanks to audience members at the joint SPP/ESPP meeting in Barcelona, members of my Honours Seminar at York University (especially David Parke, Brian Bridson, and Mike Distler) and anonymous reviewers for this journal. I would like to thank especially Brian Huss and Jennifer Vonk for helpful discussions of many of the issues addressed in this paper.

Address for correspondence: Kristin Andrews, Department of Philosophy, York University, 4700 Keele St. Toronto, ON M3J 1P3, Canada.
Email: andrewsk@yorku.ca

Mind & Language, Vol. 20 No. 5 November 2005, pp. 521–536.

chimpanzee behaviors which I take to be much more promising. They think that chimpanzees as well as humans make predictions by appealing to behavioral abstractions rather than using mentalistic reasoning regarding the beliefs and desires of the target (though they think that humans use mentalistic reasoning to make some novel predictions). If attribution of mental states does not play a significant role in most of our predictive behaviors, then perhaps we ought not look for chimpanzee theory of mind using predictive paradigms. While the current research paradigms of Tomasello *et al.*, as well as P&V's proposed paradigm, emphasize prediction, I will argue that an explanatory experimental paradigm is what is needed for a genuine paradigm shift.

2. Background to the 'Gentle Controversy'

After several years of growing agreement that chimpanzees do not understand mental states, Josep Call, Brian Hare, and Michael Tomasello broke with consensus (Hare, Call, Agnetta, and Tomasello, 2000; Hare, Call, and Tomasello, 2001; Call, 2001; Tomasello, Call and Hare, 2003a, 2003b). Tomasello *et al.* (2003a) claim that chimpanzees seem to understand some things about what others do and do not see, as well as some things about others' goal-directed activities. Whether or not this counts as having a theory of mind in the sense of attributing beliefs and desires, is seen as unimportant, for 'the generic label "theory of mind" actually covers a wide range of processes of social cognition' (Tomasello, Call, and Hare, 2003b, p. 239).

To defend the claim that chimpanzees understand seeing as a mentalistic concept, they refer to the fact that chimpanzees spontaneously monitor the gaze of conspecifics and they will move to follow human gaze around barriers, into other rooms, etc. Further, when a human experimenter is gazing intently at nothing, the chimpanzees will first follow the experimenter's gaze, and then after seeing nothing interesting to look at will turn to look back at the experimenter.

Their primary evidence, however, takes the form of a food competition study in which a subordinate and dominant chimpanzee are both given access to a room which has been baited with food. The general finding is that subordinates avoid the food that the dominant can see and seek out the food the dominant cannot see. This ability to discriminate between identical items of food based on its property of visibility is taken to indicate that the apes have a concept of seeing. This is a predictive paradigm, in that the subordinate is given the task of predicting where the dominant will go to seek food, and given that information the subordinate adjusts her own behavior accordingly. Call (2001) emphasizes the predictive nature of the task, writing that 'one important skill in both cooperative and competitive situations is the ability to predict and anticipate the behavior of conspecifics' (Call, 2001, p. 388). He believes that there are two general classes of explanation for predictive behaviors generally—a purely behavioristic cue-based approach, and a knowledge-based approach in which apes construct and use categories of behavior. The only difference between these two approaches is with regard to the ability of

the animal to use intervening variables—the abstract concepts that are used to organize behavior. As Call points out, and as P&V confirm, there is ample evidence from studies on chimpanzee concepts (such as same/different, stimulus equivalence, and transitivity tasks) that chimpanzees use categories to determine the responses they should make. The real issue at stake is whether some of those categories are psychological ones.

3. Povinelli and Vonk's Critique

Povinelli and Vonk claim that the food competition study offers no evidence that chimpanzees use mentalistic concepts when reasoning about behavior, because the performance of the chimpanzees can be fully explained even if the chimps have no mentalistic understanding; they could simply be making inferences based on past experience.

To formulate this argument, P&V introduce two alternative psychological systems that may account for the chimpanzee's predictive ability. The first is limited to reasoning about behavior, and the second adds the ability to reason about mental states. The non-mentalistic psychological system, which they call S_b, consists of:

(1) a database of representations of both specific behaviors and statistical invariants which are abstracted across multiple instances of specific behaviors; (representations that may be formed either by direct experience with the world, or may be epigenetically canalized);
(2) a network of statistical relationships that adhere between and among the specific behaviors and invariants in the database;
(3) an ability to use the statistical regularities to compute the likelihood of the specific future actions of others (Povinelli and Vonk, 2004, p. 6).

The alternative to S_b is a theory of mind system, which P&V refer to as S_{b+ms}. It is described as S_b plus the ability to reason about mental states. They don't elaborate on what is involved in the ability to reason about mental states. What they do say is that 'it [the human theory of mind system] uses information about ongoing, recent, or even quite temporally distant behaviors, to generate inferences about the likely mental states of others' (Povinelli and Vonk, 2004, p. 6). Thus, since the theory of mind part of the system doesn't make direct inferences about behavior, P&V say that it is essential that the mental state reasoning system be built on top of the behavior reasoning system: '[M]aking inferences about mental states does not allow an organism to skip the step of having to detect the abstract categories of behavior and compute the regularities among them' (Povinelli and Vonk, 2004, p. 7).

As best as I can understand the suggestion, P&V's S_{b+ms} system works as follows:

1. Observe behavior, and categorize it.
2. Refer to the database to match the category of behavior and environmental features with another behavior.
3. Infer the mental state associated with the behavior.
4. Use the ability to compute statistical regularities and the mental state attribution to make a prediction of behavior.

At step (2) the subject either has enough information to infer the prediction from statistical regularities of behavior, or it doesn't. P&V don't go into detail about how knowing the mental state would help to predict behavior, and they also suggest that the mental states are sometimes only generated to 'go along' (Povinelli and Vonk, 2004, p. 9) with the predicted behavior. However, they do think that 'it is possible to imagine situations in which responding appropriately in relatively novel situations might be facilitated by a system that reasons about mental states' (Povinelli and Vonk, 2004, p. 10) and 'it seems possible (even likely) that an organism possessing an S_{b+ms} wields certain predictive and explanatory abilities over and above an organism possessing only an S_b' (Povinelli and Vonk, 2004, p. 12).

These last remarks are very suggestive. While most theory of mind researchers have taken for granted a robust predictive function for theory of mind, P&V are much more tentative in their claims. Rather than saying that a theory of mind is used to predict behavior, full stop, they make the more nuanced claim that a theory of mind *may* help in making *some* predictions in *novel* situations. This is a much weaker predictive role than the one usually taken to be associated with theory of mind, and P&V have good reason for accepting such a position. On their view, in most cases of prediction step (3) of S_{b+ms} will be superfluous, not just for chimpanzees, but also for humans.

As a consequence, P&V argue that successful performance in any research paradigm that asks a chimpanzee to make predictions of behavior in familiar situations will fail to serve as evidence for mentalistic reasoning, since the behavior could have been predicted using behavioral abstractions and inductive reasoning. Thus, they claim that the chimpanzees' performance in the food competition paradigm cannot establish the existence of mentalistic reasoning, since the subordinate has had ample opportunity to observe other chimpanzees moving toward food after having engaged in some behavioral invariant, such as turning a head toward the food. If the subordinate has knowledge of this behavioral regularity, he doesn't need any mentalistic knowledge in order to predict that a dominant who has turned his head toward some food will next move toward the food. Whether he has any mentalistic knowledge is another question. As P&V put it, 'Techniques that pivot upon behavioral invariants (looking, gazing, threatening, peering out the corner of the eye, accidentally spilling juice versus intentionally pouring it out), will always presuppose that the chimpanzee (or other agent) has access to the

invariant, thus crippling any attempt to establish whether a mentalistic coding is also used' (Povinelli and Vonk, 2003, p. 159). Thus they conclude that no experiment which relies on behavioral invariants will suffice to decide between a mentalistic, as opposed to a merely behavioristic, psychological system. What is needed is a paradigm where the chimpanzee must make a prediction in a novel situation, one for which he has no behavioral abstractions. Povinelli and Vonk propose just such an experiment, which will be discussed in section 6.

4. Prediction and Theory of Mind

P&V make two claims in their critique I would like to draw attention to. One claim is that S_{b+ms} is an accurate account of the mentalistic psychological system. The other claim is that the attribution of mental states facilitates predictions of behavior in novel conditions. These two claims are consistent with the larger assumption that underlies much of the research on theory of mind, namely that prediction is the primary function of theory of mind (Andrews, 2003). While P&V are right to say that the predictive power of mental state attribution has been exaggerated, they have not taken full advantage of this insight.[1] This is evident in their description of the behavoristic psychological system. Remember that the third component of S_b is the 'ability to use the statistical regularities to compute the likelihood of the specific future actions of others' (Povinelli and Vonk, 2004, p. 6). This is the only function presented in the description of the psychological system, and while it tells us something about the system's approach to prediction, nothing follows about how the system would handle explanation of behaviors. Since S_{b+ms} only adds the ability to reason about mental states, the theory of mind psychology system they describe is explicitly focused on the mechanisms underlying the prediction of behavior. This suggests that P&V remain committed to the idea that mental state attribution plays a predictive role in humans, even if that role has been overstated in the past. Thus, the problem they see with the current paradigm isn't that it emphasizes prediction. Rather, their problem is that the predictions subjects are asked to make are just too easy.

What I would like to suggest in this section is that the traditional view that prediction is the primary function of theory of mind is false. I will argue that the claim that humans predict behavior by attributing mental states is exaggerated, and that other methods are likely to play a more fundamental role in generating predictions. Further, I suggest another role for mental state attribution in humans; we attribute beliefs, desires, and other mental states in order to generate *explanations*

[1] The focus on prediction in the description of the theory of mind psychological system is all the more surprising given Povinelli and colleagues' suggestion that an evolutionary function of theory of mind is to explain behavior. Jennifer Vonk kindly directed me to this work (e.g. Povinelli and Dunphy-Lelii, 2001; Povinelli *et al.*, 2000).

for behaviors. These explanations may then lead to further predictions, but it is explanation that is the primary function of our theory of mind.

Despite their remarks on the robust predictive power of S_b, P&V are among those who accept that one of the purposes of a theory of mind is to 'to successfully predict future behavior (and hence assist the organism in determining what actions it should take)' (Povinelli and Vonk, 2004, p. 7). Because of this, and because of the structure of their theory of mind system, we can see similarities between P&V's account and the more traditional theory of mind account that is associated with the theory-theory and established philosophical approaches to folk psychology.[2] The database and the rules of S_b can be seen as part of the tacit theory that chimpanzees (and, with a different set of regularities, humans) use when predicting familiar behavior. While some simulation theorists deny that the mechanisms that subsume mentalistic reasoning rest upon a foundation of behavioral regularities, P&V simply assume that this aspect of simulation theory is false.

However, P&V's view differs from the theory-theory's reliance on mental state attribution as necessary for predicting behavior. While they think humans use theory of mind to make predictions *in some situations*, they argue that most instances of prediction do not involve mentalistic reasoning. Though I think they are right about this, the view is at odds with the bulk of the literature. The term 'theory of mind' was first introduced by Premack and Woodruff, who were interested in whether the chimpanzees do what it is assumed that we do, namely attribute beliefs and desires in order to facilitate the prediction of behavior (Premack and Woodruff, 1978). According to Premack and Woodruff, humans use mental state attribution instead of behavioral regularities to predict behavior. The attribution was thought to play a causal role in formulating the prediction.

P&V reject Premack and Woodruff's characterization of theory of mind, because they reject the presupposition. That is, they reject the claim about what humans do. Adult humans do not typically attribute mental states such as beliefs and desires in order to predict behavior, according to P&V. Instead 'it seems likely that much human social interaction is supported solely by the features of S_b' (Povinelli and Vonk, 2004, p. 7).

---

[2] In an earlier paper, P&V suggest that a good test for chimpanzee theory of mind would require that chimpanzees engage in a mental simulation to predict behavior (P&V, 2003). In that paper they describe simulation as 'using one's own experiences to model the experiences of others' (p. 160). While P&V may think that their proposed paradigm asks chimpanzees to engage in a mental simulation, there are significant differences between their proposed theory of mind mechanism S_{b+ms} and traditional versions of simulation theory. Robert Gordon, for example, would reject P&V's description of simulation as involving an inference from self to other (Gordon, 1995). And Alvin Goldman would reject the idea that we can develop a database of statistical regularities without first engaging in mental simulations (Goldman, 1995). Goldman turns P&V's structure upside down, since on his view it is the database of statistical relationships between behaviors and invariants that develop *from* a theory of mind, rather than the other way around. This point, however, may be moot given the widespread move toward different hybrid accounts of the mental architecture underlying theory of mind practices (e.g. Nichols and Stich, 2003).

However, there are reasons to reject both the traditional reliance on theory of mind for prediction and P&V's less substantial view of the predictive role of theory of mind. Research coming from both social psychology and developmental psychology offers accounts of prediction which do not rely on mental state attribution, and are not part of a system like S_b. And on some of these views, the change in predictive skills in children between age three and four is not described in terms of the development of a theory of mind. For example, Birch and Bloom argue that children don't come to pass the false belief task because they are suddenly given propositional knowledge about others' mental states, but rather such changes occur as children gain competence at overcoming their own epistemic biases (Birch and Bloom, 2004). And while P&V cite Baird and Baldwin (2001) for their claim that human infants rely on a system such as S_b when anticipating people's actions, there is evidence that the methods we use to make predictions involve not just S_b, or S_{b+ms}, but include a host of heuristics and biases now being uncovered by social psychologists. One such technique is trait attribution.

Trait attribution is different from generalizing about behavioral regularities, since observation of behavior is not necessary for the attribution of the trait. Some traits are attributed just by looking at a person's skin color, facial attributes, or dress. Other times we attribute traits to a person after hearing about that person's past behavior, or when others attribute that trait to her. For example, if I am told that the family I will be dining with is very pious, I may predict that they will say a prayer before dinner, even though I have never dined with a religious family before.[3] Trait attributions are a particularly common way of predicting novel behavior. When interviewing candidates for a job, we tend to evaluate them in terms of their traits (hard-working, creative, intelligent, a good speaker, etc.). The interviewers expect that a candidate with the right traits will be a good teacher, for example, even when they don't have any prior experience with that individual in the classroom. Though this is a common method of predicting behavior, it is not thought to be particularly reliable (Kunda and Nisbett, 1986).

Other methods social psychologists think we use to predict behavior include generalizing from self (Ross *et al.*, 1977) and from the situation (Liu *et al.*, 1997).

[3] Though some might say that trait attribution is simply a shorthand for mental state attributions, I'd like to point to a difference between the two. A person who is unable to recognize mental states in others may nonetheless be fine dealing within a trait attribution psychology, and the applied behavioral analysis approach to treating children with autism can include teaching them something very much like the connection between certain kinds of behaviors and a trait. While someone who has the ability to interpret mental states would see a natural connection between some specific trait attribution (e.g. being pious) and mental state attribution (e.g. believing in the existence of God, desiring to avoid God's wrath, desiring to go to heaven, etc.), that mental state attribution is not needed to make a prediction so long as one has some understanding of the kinds of behaviors associated with the trait. A person with autism who knows that pious people usually pray before dinner, but who doesn't assume that a pious person believes that God is good, might be able to make the prediction that the pious family will pray. Rather than being shorthand for mental states, a trait attribution for a person with autism could be shorthand for a class of behaviors.

It is also thought that our moods influence the judgments and hence the predictions we make about other people (Schwarz and Clore, 1996). Some of these methods, such as attending to the situation, are thought to be under-utilized by humans (Nisbett and Ross, 1991) whereas others are thought to be unreliable. For example, while there is evidence that we use ourselves as a model when predicting other people's behavior, and that we predict that others will do the sort of thing that we would do, all too often we fail to make the appropriate adjustments between oneself and another, especially in novel situations. This false consensus bias means that we err by thinking that others are more like us than they are (Ross *et al.*, 1977).

Though we haven't yet resolved the question of how humans predict behavior, we do know that the story isn't a simple one. We are also learning that Premack and Woodruff's claim that humans predict behavior by attributing beliefs and desires to others is an oversimplification. P&V's S_{b+ms} is akin to the traditional philosophical picture of folk psychology which takes novel behaviors to be predicted using theory of mind, and this leads them to look for chimpanzee theory of mind in instances of prediction. But there is a tension between understanding theory of mind to have a predictive functional role, and yet thinking that most prediction is made using behavioral associations. Such a view fails to fully recognize the explanatory role of theories. Once an explanation for some phenomenon has been generated, the information we gain can be used when making future predictions. For example, once you come to understand which beliefs and desires caused a person P to engage in some behavior, you also learn that P is the kind of person who has those beliefs and desires, and from that you might attribute to P certain personality traits that can be used to make future predictions.

This isn't to say that belief and desire attribution, and folk psychology more generally, don't play an important role in human social interaction. As Sellars suggests, we may reason about mental states primarily as a means of understanding our social world (Sellars, 1956). And while P&V may be sympathetic to a view such as this, they have not taken full advantage of the insight. Rather than offering a paradigm shift from one predictive experiment to another, the radical shift in the theory of mind research would be to move away from predictive research paradigms altogether. It is likely that any success in a predictive paradigm can be explained as the result of a behavioristic psychological system that relies on behavioral, rather than mental, intervening variables. Thus, as we shall see in the following section, P&V's strategy for denying that Hare *et al.*'s food competition experiment should serve as evidence of chimpanzee theory of mind can also be used to undermine the false belief task as evidence for the child's theory of mind.

5. Povinelli and Vonk's Critique Also Undermines Children's Theory of Mind Paradigms

If P&V are going to reject the food competition study as evidence for theory of mind in chimpanzees, then to be consistent they also ought to reject Wimmer and

Perner's (1983) false belief task as evidence for children's theory of mind. While it is true that the performance of chimpanzees in experiments aimed to demonstrate the existence of a theory of mind can be explained with the S_b model, there is a non-mentalistic story that can be generated to account for children's performance in false belief tasks as well. The false belief tasks has long been taken as a litmus test for theory of mind in children (though it has more recently come under attack e.g. see Bloom and German, 2000). In this paradigm, a child is asked to make a prediction about where a puppet will look for an object which, during the puppet's absence, had been moved to an unexpected location. Children who pass the test say that the puppet will look for the object where he left it.

While it has been thought that passing the false belief task offers solid evidence that a child has a theory of mind, P&V could give a non-mentalistic account of the child's successful answer much in the same way they explain the chimpanzee's behavior in the food competition task:

(a) Subject observes Maxi putting his chocolate in a box, and then observes Maxi leaving the room.
(b) Subject observes Mother coming into the room, and moving the chocolate from the box to the cupboard.
(c) Subject observes Mother leaving the room and Maxi returning.
(d) Subject appeals to her database of behavioral generalizations, and finds the matching 'people look for objects where they left them' heuristic.
(e) Subject predicts that Maxi will look for his chocolate in the box.

When asked to make a prediction about where Maxi will go to look for his chocolate, the four-year-old who passes the task could reason about behavioral regularities rather than mental states. The change from age three to four need not be described as a change in theory of mind status. Instead, it might be explained by the development of more sophisticated and nuanced representations of behavioral correlations, or as the overcoming of a bias (Birch and Bloom, 2004). But note that a subject could successfully use this method even in novel situations. That is, the behavioral regularities that she appeals to need not be specific to the puppets, to chocolate bars, to going outside, or to any other of the details of this story. Rather, the regularity that the child refers to might be as simple as 'people look for things where they left them'. Since a child might learn this regularity simply by observing her own behavior, rather than by having observed other people engaging in this behavior, she could make the correct prediction without ever having observed another searching for an object. I will return to this point in the next section.

Of course there are other reasons why we think that children grasp mental state concepts between the ages of three and four, and in addition to a plethora of experimental evidence, we know this because children start speaking of them. Famously, chimpanzees don't. But since there is a nonmentalistic explanation for success in the traditional false belief task, P&V enjoy plausible deniability for chimpanzee theory of mind if and when a chimpanzee passes a non-verbal version

of the task. This is not a result to celebrate, however. If theory of mind is a meaningful scientific notion, there must be some experimental result that will corroborate its existence.

6. Povinelli and Vonk's Paradigm Shift

While P&V think that they have developed an experiment that will help to decide the question of chimpanzee theory of mind, the paradigm shift they suggest will not solve the problem. They propose a modified version of Cecilia Heyes' task in order to distinguish between the mentalistic and non-mentalistic models (Heyes, 1998). P&V's version is presented as follows:

> Subjects would first be exposed to the subjective experience of wearing two buckets containing visors which look identical from the outside, but one of which is see-through, the other of which is opaque. The buckets would be of different colors and/or shapes in order to provide the arbitrary cue to their different experiential qualities. Then, at test, subjects are given the opportunity to use their begging gesture to request food from one of two experimenters, one wearing the <seeing> bucket and the other wearing the <not seeing> bucket ... By definition, S_b has no information that would lead the subjects to generate this response. In contrast, a system that first codes the first person mental experience, and then attributes an analog of this experience to the other agent (in other words, S_{b+ms}) could have relevant information upon which to base a response (Povinelli and Vonk, 2004, p. 14).

Despite P&V's claims to the contrary, this proposed study is an instance of the old paradigm. First, it isn't true by definition that S_b has no information about behavioral regularities associated with wearing a transparent bucket. While P&V think that self-to-other inferences will be inferences about mental states rather than inferences about behaviors, there is no reason to think so. The subject who successfully begs toward the experimenter wearing the <seeing> bucket could have, *from his own experience*, made the generalization between wearing the see-though bucket and being able to, e.g. walk around without bumping into things, grab items of interest, etc. Rather than coding first person *mental* experience, the chimp could code first-person *physical* experience. In short, the chimp might make the behavioral connection between wearing the opaque bucket and *not being able to do things*. From whom should he beg? Certainly not the person who isn't able to do things.

The chimpanzee can reason about the researchers' behavior from knowledge of his own behavior. I can give an explanation for the chimpanzee's successful response to this experiment in a way that parallels P&V's explanations of the subordinate's behavior in the food-competition paradigm:

(a) Chimpanzee observes he cannot do things with the <not seeing> bucket on;
(b) Chimpanzee observes he can still do things with the <seeing> bucket on;
(c) Chimpanzee observes Al with the <not seeing> (can't do things) bucket on;
(d) Chimpanzee observes Penelope with the <seeing> (can do things) bucket on;
(e) Chimpanzee accesses database of behavioral regularities to determine that Al can't do things and Penelope can do things;
(f) Chimpanzee gestures to Penelope (because only people who can do things will be able to offer food).

Though it might be tempting to think that any generalization from one's own experience to another's behavior must necessarily involve knowledge of mental states, I think the following example should undermine such worries. I can eat a poison berry and get sick from it, while in isolation from everyone else. From that experience, I could then generalize that the berry will make others sick too, and predict that they would behave in the same sick way that I did. I can make this prediction without having seen anyone else get sick from the berry, and without having any notion of a mental state. Experiential mapping from self to other, contrary to P&V's claims, need not involve any mentalistic reasoning. Just as the child who passes the false belief task may be simply generalizing from her own experiences of looking for hidden objects, a chimpanzee who passes P&V's proposed task may be merely using the same non-mentalistic technique.

In response to this critique of their proposed paradigm, Jennifer Vonk has indicated in correspondence that the chimpanzees' movements are not constrained while wearing an opaque bucket, and there is no reason to think that they would learn anything about what they can and cannot do while wearing one. If this is true, however, there is no reason to suppose that a chimpanzee would prefer to beg from a seeing trainer rather than a not-seeing trainer. If we suppose the chimpanzee *can* do everything he has ever done despite the fact he is wearing the not-seeing bucket, the chimpanzee has no basis for drawing a connection between being able to see and being able to do things, since he can do everything just fine while wearing the not-seeing bucket. Thus, even if he had the concept of seeing, the chimpanzee would have no reason to beg from a trainer wearing a seeing bucket rather than a not-seeing bucket, since he didn't make any connection between being able to see and being able to do things (like giving food). If a chimpanzee could do everything as well with the not-seeing bucket on as he does with the seeing bucket on, *then he would have no reason to infer that a person with the not-seeing bucket on couldn't give him food*. In this case the chimpanzee's concept of seeing, if he had one, would be quite different from our own.

On the other hand, if the chimpanzee cannot do everything he has ever done while wearing the not-seeing bucket, my former critique holds. The chimpanzee

could make the inference that since he can't do things with the not-seeing bucket on, others wouldn't be able to do things with the not-seeing bucket on either, and so he won't beg for food from someone who is wearing the not-seeing bucket because that person can't do things. Thus, if either the chimpanzee can or cannot do things with the not-seeing bucket on its head, the experiment will not determine between S_b and S_{b+ms}.

Given that P&V's proposed research paradigm doesn't help to adjudicate between mentalist and behaviorist methods of predictions, and that they suspect that much of human prediction is done without appealing to mental state attributions, I suggest that they look for evidence in another domain. Explanation, not prediction, may be the most relevant place to look for a theory of mind, both in humans and chimpanzees.

7. An Explanatory Paradigm

If humans use their theory of mind more for explanation than for prediction, then to test for chimpanzee theory of mind we might confront chimpanzees with a puzzle whose solution requires having a theory of mind. Though humans do make many of their predictions by generalizing about behaviors, an explanation of anomalous behavior cannot be given in terms of behavioral regularities; indeed an explanation is demanded precisely because the behavior violates expectations. Thus, to test whether chimpanzees have mental state concepts and whether they use them to make inferences about the behavior of others, we could design an experiment that places the subject in a situation where he must use what Daniel Dennett calls the Sherlock Holmes method (Dennett, 1983). Such an experiment would involve setting cognitive traps for chimpanzees, so that they are given the opportunity to demonstrate their knowledge of mental states. The trap could involve presenting a subject with an anomalous behavior, and observing his response.

For example, we might use an explanatory version of P&V's predictive task. As with their proposed study, we would first habituate the subject to a bucket with a transparent visor. After this exposure, a chimpanzee who did have the concept of seeing would come to know that he can still see with the bucket on his head. If the chimpanzee subject could also generalize the mental state concept from self to other, he would expect that anyone else who put the bucket on his head would also be able to see. The trap is to use sleight of hand to switch the <seeing> bucket with one that exactly resembles it visually, but which doesn't have a transparent visor. When the first chimpanzee puts the new bucket on his head, the subject may notice that this chimpanzee is behaving differently from what the subject would expect. We can observe the subject's response to his cohort's presumably anomalous behavior. Here the subject is in a situation where he has to explain anomalous behavior; in the past chimpanzees with buckets on their heads behaved normally, and now he is confronted with a chimpanzee who does not. In order to explain the

anomalous behavior, a mentalist chimpanzee subject may wonder whether the anomalous chimpanzee could see. And, in order to test that hypothesis, the animal could examine the bucket. If the subject were to do this, it would suggest an attempt to work out a mental explanation for an anomalous behavior.[4]

I realize that there are obvious problems associated with interpreting the results of such an experiment, given that there is no one behavior associated with passing the test. But no explanatory paradigm could be as methodologically neat as the predictive one, where the subject either makes the correct prediction or doesn't. Furthermore, as P&V point out, the neatness of the predictive paradigm may be an illusion, for two reasons. First, reasoning about mental states does not necessarily entail a different behavioral response. We have already seen this criticism in the guise of P&V's belief that humans don't regularly use a theory of mind when predicting behavior. Second, and more importantly, 'there is an implicit assumption that the humans who design the experiments can use their folk psychology to successfully intuit which responses can be produced only by reasoning about the underlying mental state' (Povinelli and Vonk, 2004, p. 4). If a certain amount of interpretation is necessary to analyze the correct response to a predictive paradigm, the explanatory paradigm cannot be criticized as less objective merely because it too requires interpretation. The necessity for interpretation is a limitation of all theory of mind studies, on humans as well as chimpanzees.

P&V should not be surprised that they don't find the kind of evidence they want in the predictive experiments; given that they don't think humans use a sophisticated method such as belief attribution when making most predictions of behavior, they should have no reason to suspect that chimpanzees would either. A good place to start looking for chimpanzee theory of mind would be in the same place we find human theory of mind, and that isn't typically prediction. I suggest turning attention toward explanation.

8. Conclusion

I have addressed two concerns with P&V's (2004) critique. First, it is unlikely that any one purely predictive paradigm will be able to distinguish between a mentalistic vs.

[4] It has been pointed out to me by David Parke and Brian Bridson that the chimpanzee subject may choose to examine the bucket not in hopes of finding a mentalistic explanation for the anomalous chimpanzee's behavior, but in order to find a physical explanation for what is wrong with the bucket. Though this is a possibility, and hence this experiment cannot alone determine that a chimpanzee has mentalistic understanding, we might find that no one experiment will be sufficient to determine whether a chimpanzee has mentalistic understanding. Rather than serving as a litmus test for chimpanzee theory of mind, positive success on an explanatory task such as this one may, along with other experiments and ethological observations, serve to strengthen the body of evidence in favor of chimpanzee mentalistic understanding. My emphasis on explanatory tasks should be seen as an attempt to round out the current evidence, rather than as an attempt to replace the current predictive paradigms altogether.

behavioristic explanation for a subject's behavior. To test for mentalistic understanding, and to avoid an interpretation of the behavior as the result of behavioral generalizations, an explanatory task could be developed that might induce behavior that would be more difficult to explain in terms of behavioral abstraction.

The second concern is that P&V are dangerously close to a *reductio ad absurdum*, given that their methods of explaining the chimpanzee performance can be used just as easily to explain the child's performance in the false belief task. In working through the gentle controversy, we must be careful to apply the same standards of evidence to animals that we apply to humans. If passing the false belief task serves as *some* evidence that a child has some understanding of other minds, then passing a parallel task should serve as *some* evidence in the case of the chimpanzee. We must not ask more of the chimpanzee than we ask of the child. Taken in isolation, many behaviors can be described as the result of a non-mentalistic psychological system, and this is true of both humans and other animals. But it isn't any one behavior, or success at any one paradigm, that leads us to conclude that humans understand that others have mental states. It is unfair, and unrealistic, to expect that any one experiment will allow us to conclude the same about animals.

Though all parties to the debate feel comfortable sidestepping the problem of other minds, and accept that chimpanzees *have* beliefs and desires, another problem has taken its place. The new problem of other minds is not whether the other has a mind, but whether the other knows that *we* have a mind. Fortunately, there may be a way out of the problem, and it is the same as the way out of the traditional problem of other minds: given the current evidence and evidence from future research, we must be willing to make an inference to the best explanation on the basis of the entire body of evidence. Though a definitive proof is lacking, that shouldn't concern us, since most of us are utterly unconcerned about the lack of a knock-down argument for other *human* minds.

By refusing to describe the interesting results of the Hare *et al.* study as evidence for some mentalistic reasoning, we end up ignoring interesting differences and similarities between species. Chimpanzees behave as if they have the concept of seeing, and so do most children at the age of four. Domesticated cats, presumably, do not behave in this way. Where the behaviors are different, the interpretations of those behaviors will be different. And where they are the same, we ought to see some similarity between them. The behaviorists got at least that much right.

Department of Philosophy
York University, Toronto

References

Andrews, K. 2003: Knowing mental states: the asymmetry of psychological prediction and explanation. In Q. Smith and A. Jokic (eds), *Consciousness: New Philosophical Perspectives*. Oxford: Oxford University Press.

Baird, J. A. and Baldwin, D. A. 2001: Making sense of human behavior: action parsing and intentional inference. In B. F. Malle and L. J. Moses (eds), *Intentions and Intentionality: Foundations of Social Cognition*. Cambridge, MA: MIT Press.

Birch, S. A. J. and Bloom, P. 2004: Understanding children's and adults' limitations in mental state reasoning. *Trends in Cognitive Science*, 8, 255–260.

Bloom, P. and German, T. 2000: Two reasons to abandon the false belief task as a test of theory of mind. *Cognition*, 77, B25–B32.

Call, J. 2001: Chimpanzee social cognition. *Trends in Cognitive Science*, 5, 388–393.

Dennett, D. 1983: Intentional systems in cognitive ethology: The 'Panglossian Paradigm' defended. *Behavioral and Brain Sciences*, 6, 343–390.

Goldman, A. 1995: Interpretation psychologized. In M. Davies and T. Stone (eds), *Folk Psychology*. Oxford: Blackwell Publishers.

Gordon, R. 1995: Simulation without introspection or inference from me to you. In M. Davies and T. Stone (eds), *Mental Simulation*. Oxford: Blackwell Publishers.

Hare, B., Call, J., Agnetta, B. and Tomasello, M. 2000: Chimpanzees know what conspecifics do and do not see. *Animal Behavior*, 59, 771–785.

Hare, B., Call, J. and Tomasello, M. 2001: Do chimpanzees know what conspecifics know? *Animal Behavior*, 61, 139–151.

Heyes, C. 1998: Theory of mind in nonhuman primates. *Behavioral and Brain Sciences*, 21, 101–134.

Kunda, Z. and Nisbett, R. 1986: The psychometrics of everyday life. *Cognitive Psychology*, 18, 199–224.

Kunda, Z. 2002: *Social Cognition: Making Sense of People*. Cambridge, MA: MIT Press.

Nichols, S. and Stich, S. 2003: *Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds*. Oxford: Oxford University Press.

Liu, J. H., Pham, L. B. and Holyoak, K. J. 1997: Adjusting social inferences in familiar and unfamiliar domains: the generality of response to situational pragmatics. *International Journal of Psychology*, 32, 73–92.

Nisbett, R. E. and Ross, L. 1980: *Human Inference*. Englewood Cliffs, NJ: Prentice-Hall.

Povinelli, D. J., Bering, J. M. and Giambrone, S. 2000: Toward a science of other minds: escaping the argument by analogy. *Cognitive Science*, 24, 509–541.

Povinelli, D. J. and Dunphy-Lelii, S. 2001: Do chimpanzees seek explanations? Preliminary comparative investigations. *Canadian Journal of Experimental Psychology*, 55, 185–193.

Povinelli, D. J. and Vonk, J. 2003: Chimpanzee minds: suspiciously human? *Trends in Cognitive Science*, 7, 157–160.

Povinelli, D. J. and Vonk, J. 2004: We don't need a microscope to explore the chimpanzee mind. *Mind & Language*, 19, 1–28.

Premack, D. and Woodruff, G. 1978: Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1, 515–526.

Ross, L., Greene, D. and House, P. 1977: The 'false consensus effect:' An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology*, 13, 279–301.

Schwarz, N. and Clore, G. L. 1996: Feelings and phenomenal experiences. In E. T. Higgins and A. Kruglanski (eds) *Social Psychology: Handbook of Basic Principles*. New York, NY: Guilford Press.

Sellars, W. 1956: *Empiricism and the Philosophy of Mind*. Cambridge, MA: Harvard University Press.

Tomasello, M., Call, J. and Hare, B. 2003a: Chimpanzees understand psychological states—the question is which ones and to what extent. *Trends in Cognitive Science*, 7, 153–156.

Tomasello, M., Call, J. and Hare, B. 2003b: Chimpanzees versus humans: it's not that simple. *Trends in Cognitive Science*, 7, 239–240.

Wimmer, H. and Perner, J. 1983: Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13, 103–128.