PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

Phil. Trans. R. Soc. B
doi:10.1098/rstb.2006.2023

On the lack of evidence that non-human animals possess anything remotely resembling a 'theory of mind'

Derek C. Penn and Daniel J. Povinelli*

Cognitive Evolution Group, University of Louisiana, Louisiana, Lafayette, LA 70504, USA

After decades of effort by some of our brightest human and non-human minds, there is still little consensus on whether or not non-human animals understand anything about the unobservable mental states of other animals or even what it would mean for a non-verbal animal to understand the concept of a 'mental state'. In the present paper, we confront four related and contentious questions head-on: (i) What exactly would it mean for a non-verbal organism to have an 'understanding' or a 'representation' of another animal's mental state? (ii) What should (and should not) count as compelling empirical evidence that a non-verbal cognitive agent has a system for understanding or forming representations about mental states in a functionally adaptive manner? (iii) Why have the kind of experimental protocols that are currently in vogue failed to produce compelling evidence that non-human animals possess anything even remotely resembling a theory of mind? (iv) What kind of experiments could, at least in principle, provide compelling evidence for such a system in a non-verbal organism?

Keywords: theory of mind; folk psychology; mental state attribution; parsimony; chimpanzees; corvids

1. INTRODUCTION

Are humans alone in their capacity to reason about unobservable mental states, such as perceptions, intentions, emotions, desires and beliefs? Over a quarter-century ago, Premack & Woodruff (1978) launched a multinational industry dedicated to answering this question and coined the term, 'theory of mind' (hereafter, ToM) to refer to this distinctive capacity: 'a system of inferences of this kind', they observed, 'may properly be regarded as a theory because such [mental] states are not directly observable, and the system can be used to make predictions about the behavior of others' (p. 515).

Unfortunately, after decades of effort by some of our brightest human and non-human minds, there is still little consensus on whether or not non-human animals understand anything about unobservable mental states or even what it would mean for a non-verbal animal to understand the concept of a 'mental state'. Nearly 10 years ago, Heyes (1998) observed that there had been 'no substantial progress' (p. 101) on Premack & Woodruff's (1978) original question for many years. It is debatable whether there has been any more agreement on the matter since then (for the latest version of these ongoing and seemingly intractable debates, see Povinelli & Vonk 2003, 2004; Tomasello *et al.* 2003*a*,*b*; Tomasello & Call 2006).

Povinelli & Vonk (2004) pointed out one glaring reason for the impasse, namely comparative researchers have never specified 'the unique causal work' that representations about mental states do above and beyond the work that can be done by representations of the observable features of other agents' past and occurrent behaviours. As a result, almost all of the experimental protocols that have been used to test the ToM capabilities of non-human animals over the past quarter-century, including those that are currently in vogue today, are incapable, even in principle, of validating or falsifying the hypotheses being tested. One does not need to hold a Popperian view of science to acknowledge that arguments among unfalsifiable hypotheses are likely to be of little or no value to practicing scientists.

There seems to be a dire need, then, to focus more attention on the basic definitional and evidential issues confronting comparative researchers and less time spent arguing over ambiguous experimental results. In this paper, we will confront four related and contentious questions head-on:

(i) What exactly would it mean for a non-verbal organism to have an 'understanding' or a 'representation' of another animal's mental state?
(ii) What should (and should not) count as compelling empirical evidence that a non-verbal cognitive agent has a system for understanding or forming representations about mental states in a functionally adaptive manner?
(iii) Why have the kind of experimental protocols that are currently in vogue failed to produce compelling evidence that non-human animals possess anything even remotely resembling a theory of mind?

* Author for correspondence (djp3463@louisiana.edu).

One contribution of 19 to a Dicussion Meeting Issue 'Social intelligence: from brain to culture'.

(iv) What kind of experiments could, at least in principle, provide compelling evidence for such a system in a non-verbal organism?

Only after we have addressed these fundamental issues in a formal, principled fashion will we be in a position to attempt to answer the fascinating question that Premack & Woodruff (1978) first posed so many years ago.

Theory of mind, *sensu* Premack & Woodruff (1978), entails the capacity to make lawful inferences about the behaviour of other agents on the basis of abstract, theory-like representations of the causal relation between unobservable mental states and observable states of affairs. This is certainly not the only way to construe the capacity in question (for an overview of the possibilities, see Davies & Stone 1995*a*,*b*; Carruthers & Smith 1996). Many researchers have argued, for example, that the ability to take the causal role of mental states into account does not involve theory-like inferences at all, but is grounded in practical, sensorimotor, simulative abilities (e.g. Gordon 1986, 1996; Goldman 1993).

For the purposes of the present essay, we wish to remain rigorously agnostic as to *how* the capacity to take other agents' mental states into account is implemented. We will henceforth use the acronym ToM, to refer to *any* cognitive system, whether theory-like or not, that predicts or explains the behaviour of another agent by postulating that unobservable inner states particular to the cognitive perspective of that agent causally modulate that agent's behaviour. We believe this construal of ToM *sensu lato* is about as broad and minimalist as possible without losing the distinctive character of the capacity in question.

In our opinion, the major impediment that has stood in the way of understanding whether or not other species employ a ToM has been our species' inveterate intuitions about how our own ToM works. Appeals to folk psychological assumptions and reasoning by analogy to introspective intuitions have played an inordinate role in comparative researchers' claims over the last quarter-century (see Povinelli & Giambrone 1999; Povinelli *et al*. 2000; Povinelli & Vonk 2003, 2004). Thus, to undermine the insidious role that introspective intuitions and folk psychology play in the comparative debate, we propose to treat the ToM explanandum here in more formalistic terms than is typical among comparative researchers. Our approach is as follows:

(i) present a simple formalism to clarify exactly what is (and is not) at stake with respect to the comparative ToM explanandum,
(ii) use the formalism in (i) to specify what should (and should not) count as evidence for a ToM system in a non-verbal organism,
(iii) take a prominent experimental result with chimpanzees as a case study for exposing why the kind of protocols currently in vogue do not satisfy the conditions set out in (ii),
(iv) show why the analysis in (iii) applies, mutatis mutandis, to the protocols currently being employed with corvids as well, and
(v) propose two sample experimental protocols that could, at least in principle, provide compelling positive evidence for a ToM system in a non-human species.

2. A SIMPLE FORMALISM

To begin, let us agree without too much argument that cognitive agents—biological or otherwise—can learn from their past experience, in part because they have dynamic internal states that are decoupled from any immediate physical connection to the external world. Some of these internal states carry information about what the agent has learned about the world that is distinct from the information immediately available to the system's perceptual inputs. And some of these internal states describe goal states against which actual states of the organism can be compared so that the organism's behaviours can be dynamically adjusted in order to close the gap. Let us denote all these internal goals states by the variable, g, and all the informational states that affect and/or mediate the goal-directed behaviour of a cognitive agent by the variable, r.

Our rough-and-ready definition of r- and g-states is meant to be as ecumenical as possible. For example, we are entirely agnostic (for our present purposes anyway) about whether an organism's r- and g-states are modal or amodal, discrete or distributed, symbolic or connectionist or even about how they come to have their representational or informational qualities to begin with. And we make no judgment about whether r- and g-states as we have defined them here bear any resemblance to the mental state concepts putatively posited by our commonsense folk psychology. We do not pretend that this definition of g- and r-states puts to rest the entire (or even a small part of the) controversy over what counts as goal-directed behaviour or internal mental representations (see Markman & Dietrich 2000 for a better start); but it is good enough for our present purposes.

Of course, there are innumerable other factors that also contribute to shaping a biological organism's behaviour, including information from sensory inputs, feedback from perception-action loops, autonomic-visceral states, the physical structure and capabilities of the organism's body and all the other many variables that influence the actions of situated, embodied, biological agents in the wild. But for our present purpose, these many multifarious influences can be reduced to two additional variables and an ellipsis. We will use the variable, p, to denote any dynamic, occurrent information obtained through perceptual inputs (including autonomic and proprioceptive channels); and we will use the variable, q, to denote feedback from the organism's sensorimotor loops (including online and offline emulators). Using this notation, any cognitive behaviour, b, can be described formally (albeit simplistically) as follows:

$$b = f(g, r, p, q, \ldots). \tag{2.1}$$

In other words, any cognitive behaviour is some function of the system's g- and r-states plus any occurrent information from perceptual inputs and

sensorimotor emulators at the time the function is computed—plus any other cognitive variables not incorporated in the present model. The reason we are unconcerned with unpacking such broad variables as g, r, p and q, or with what falls under the ellipsis, is because we are only concerned, herein, with the question of whether or not a given cognitive agent possesses a ToM. And the question of whether or not a given cognitive agent possesses a ToM boils down to the question of whether or not that agent is able to treat other agents as if their behaviour is a function of the kind of variables described in equation (2.1). The only condition that must be met in order to qualify as a ToM, by our minimalist standards, is that the system must be able to produce and employ a particular class of information, namely information about the state of these cognitive variables from the perspective of that agent *as distinct from the perspective of the system itself*. We will refer to this special class of information by the variable, *ms*.

What exactly does it mean for one cognitive information state to be 'about' some other state of affairs? Much greater minds than ours have tried to answer this question (for example, Dretske 1988); and the complexities of taking this question seriously would take us far beyond the scope of the present essay. So here is a simple stop-gap answer that will suffice for our present purposes: let us agree that an *ms* variable carries information about some other cognitive state iff the state of the *ms* variable covaries with the state of the other cognitive state in a generally reliable manner such that, *ceteris paribus*, variations in the *ms* variable can be used by the consuming cognitive system to infer corresponding variations in the other cognitive state.

In a genuine mind-reader, the function describing the informational relation between one agent's *ms* variables and another agent's cognitive state variables might be something like the following:

$$ms = f_{\mathrm{mr}}(g^*, r^*, p^*, \ldots), \tag{2.2}$$

where * denotes the state of the corresponding variable for the other agent and f_{mr} denotes a cognitive function capable of intuiting the state of these unobservable variables directly, for example, telepathically.

Of course, there are no genuine mind-readers on this planet and all the relevant cognitive variables are, strictly speaking, unobservable from the point of view of the aspiring mind-reader. Hence, any purported mind-reading being performed on this planet is, in fact, a trick. A very good trick, to be sure, but a trick nevertheless. The trick is to be able to infer the state of the unobservable cognitive variables that will influence the behaviour of another agent using information observed from the perspective of the system itself:

$$ms = f_{\mathrm{ToM}}(r, p, \ldots), \tag{2.3}$$

where f_{ToM} denotes a special function that computes an *ms* variable based on the inputs available to sentient, situated, embodied but non-telepathic organisms.

There is a burgeoning debate over how f_{ToM} might be implemented (for examples of the debate, see Davies & Stone 1995*a*,*b*; Carruthers & Smith 1996; Hurley & Chater 2005). Traditionally, f_{ToM} has been construed as a kind of inferential function that uses a database of law-like generalizations to make logical inferences about other agents' *g*- and *r*-states in a theory-like manner. This is certainly the kind of f_{ToM} that Premack & Woodruff (1978) had in mind when they coined the term that started the debate. But, as we noted previously, there are many alternative hypotheses at play today, some of which propose that f_{ToM} is implemented via offline simulation capabilities that encode *ms* variables about other subjects' internal states using the same mechanisms that are used to encode *ms* variables about the subject's own internal states. Still other researchers advocate hybrid functions between theory and simulation (e.g. Nichols & Stich 2003; Meltzoff in press). For our present purposes, we are agnostic as to how the f_{ToM} is implemented; we simply note that a cognizer that has a ToM system of any kind must have an f_{ToM} of some kind. And any f_{ToM} must take information from the system's own inputs and produce (or enact) a special class of information, i.e. information that is postulated to be from the cognitive perspective of another agent and relevant to predicting the behaviour of that agent.

The simple formalism we have proposed here leaps over innumerable details and complex, unresolved issues; but it nevertheless helps to keep track of, what is and what is not at stake with respect to the question of whether or not chimpanzees or any other non-human animal have a ToM. Our definition of an f_{ToM} does not require the agent to have any insight into the subjective phenomenological experience of others. Nor does our definition require *ms* variables to have an isomorphic relationship with the content or structure of the mental state that is being represented. Metarepresentations are one way of implementing *ms* variables. But they are certainly not the only way. Some theorists, for example, have argued that apes' representations of mental states might simply involve 'intervening variables' (aka 'secondary representations') rather than explicitly structured metarepresentations (Whiten 1996, 1997, 2000; Suddendorf & Whiten 2001; Whiten & Suddendorf 2001). We believe Whiten and Suddendorf are right in this sense: being able to recode perceptually disparate behavioural patterns resulting from the same underlying cognitive state as instances of the same abstract equivalence class is a bona fide example of postulating an *ms* variable in the sense Q4 defined hereinabove (we differ from Whiten & Suddendorf (2001), however, in that we do not see any compelling evidence of this ability in non-human animals; see discussion below).

We particularly want to point out that the debate concerning whether or not non-human animals possess an f_{ToM} should not be concerned with whether or not they are cognitive creatures capable of reasoning about general classes of past and occurrent behaviours (e.g. Q5 'threat posture', 'eye or face direction', 'body position' or 'eye-direction-in-relation-to-objects-in-the-world'). Indeed, they *must* be able to do so if they are potential candidates for a ToM at all. The theory of mind debate among comparative researchers should turn only around the question of whether, in addition to the representational abilities that any cognitive agent possesses as defined in equation (2.1), some particular

cognitive system in the agent in question also produces information that is specific to the cognitive perspective of another agent and uses this information to predict the behaviour of that agent.

3. WHAT SHOULD COUNT AS EVIDENCE OF AN f_{ToM}?

We hope that our simplistic formalism will also help define more clearly what should and should not count as compelling evidence for an f_{ToM}. The subtle confounding problem, from an experimentalist's point of view, is that all organisms with the potential to have an f_{ToM} are also, necessarily, cognitive agents in the sense defined by equation (2.1) above.[1] The unavoidable null hypothesis is that any agent capable of possessing an f_{ToM} must already be employing the information provided by g, r, p and q in their cognitive behaviours. Thus, in order to produce experimental evidence for an f_{ToM} one must first falsify the null hypothesis that the agents in question are simply using their normal, first-person cognitive state variables as defined by equation (2.1). One must, in other words, create experimental protocols that provide compelling evidence for the cognitive (i.e. causal) necessity of an f_{ToM} *in addition to* and *distinct from* the cognitive work that could have been performed without such a function.

The last qualification is crucial. Imagine an organism, **A**, that always manifests some determinate set of observable cues, $\mathbf{C_1}$, whenever it is in a given r-state, $\mathbf{r\text{-}state_1}$, such that $P(\mathbf{r\text{-}state_1}|\mathbf{C_1})=1$ and $P(\mathbf{r\text{-}state_1}|\sim\mathbf{C_1})=0$[2]. And suppose that $\mathbf{r\text{-}state_1}$ causes **A** to emit behaviour $\mathbf{b_1}$. A second cognitive agent having perceptual access to organism A and its observable traits, $\mathbf{C_1}$, would have no need to infer the presence of $\mathbf{r\text{-}state_1}$ in order to predict the occurrence of $\mathbf{b_1}$; simply observing $\mathbf{C_1}$ suffices. Thus, a researcher observing that a given experimental subject is able to reliably predict the occurrence of $\mathbf{b_1}$ in **A** after observing $\mathbf{C_1}$ would have no basis for concluding that the subject possesses an f_{ToM} dedicated to inferring $\mathbf{r\text{-}state_1}$ (even though she, herself, may know that $\mathbf{r\text{-}state_1}$ causes $\mathbf{b_1}$) unless she can also show that possessing information directly about $\mathbf{r\text{-}state_1}$ does some special causal work for **A** in addition to predicting $\mathbf{b_1}$. Although this is rarely noted by experimentalists, we believe this point to be indisputable (see Povinelli & Vonk 2003, 2004). Curiously, though, it is nevertheless, often disputed, or completely ignored (see Tomasello *et al.* 2003*a*,*b*; Tomasello & Call 2006).

When framed in formalistic terms, the point appears obvious. But a simple real-life example will illustrate how easy it is to be duped by commonsense. A chimpanzee (the subject) observes a second chimpanzee, turn her head and look-off in the distance. In response, the subject turns his head in the same direction. From a folk psychological point of view (i.e. from the point of view of any normal adult human observer), it is tempting to conclude that the subject's act of turning his head is mediated by an internal representation of the second chimpanzee's belief that there is something interesting to look at and an implicit understanding that 'seeing' leads to a change in the internal, epistemic state of the looker. In other words, our commonsense intuitions assume that the subject's behaviour was mediated by an *ms* variable (i.e. the subject had some understanding of the second chimpanzee's *g*- and *r*-states). Indeed, many comparative researchers have been tempted to attribute *ms* variables to their subjects under similar experimental circumstances (Call *et al.* 1998; Tomasello *et al.* 1999; Bugnyar & Heinrich 2005; Flombaum & Santos 2005; Santos *et al.* 2006; Tomasello & Call 2006).

What commonsense intuition overlooks, however, is that it is also possible for the same behaviour to be produced without an f_{ToM} of any kind. The set of perceptual cues available to the subject (i.e. 'eye or face direction', 'body position', 'eye-direction-in-relation-to-objects-in-the-world', etc.) are sufficient to explain the subject's behaviour. Any socially intelligent subject like a chimpanzee must possess a rich database of *r*-states based on what he has learned about perceptually similar situations in the past and the conditional dependencies that tend to hold between these observable cues and other animals' subsequent behaviour. Thus, the subject may have turned his head in the direction of the other chimp's head simply because it learned from past experience (or was born with the propensity to learn) that the given pattern of perceptual cues is a reliable indicator of something worth looking at in the direction inferred by the other agent's eyes and head. There is no need for the subject to reason in terms of an *ms* state variable, and positing an *ms* state variable does no additional explanatory work in the given situation.

The evidential case for an *ms* variable is no better simply because the second chimpanzee looks behind a barrier and the subject adjusts his position to see behind the barrier as well (e.g. Povinelli & Eddy 1996*b*). Barriers are, of course, visible entities. Subjects who have learned (or are born knowing) that they must alter their own position in order to see behind a barrier if a conspecific's eyes are directed towards a location behind a barrier do not necessarily need, *in addition*, to form representations postulating the hypothetical content of the conspecific's perceptual field or to understand that 'seeing' leads to any change at all in the looker's *r*-states (see also Povinelli *et al.* 2002).

And the evidential case is still no better just because the subject 'checks back' with the looker if he does not find anything interesting behind the barrier. Chimpanzees check back with moving objects all the time in order to update their internal representation of the object's location and projected trajectory without thereby postulating that all moving objects have mental states.

Following the gaze of a conspecific, checking behind barriers and checking back with the looker when nothing is found certainly *seem* to be compelling evidence for reasoning in terms of unobservable mental states when interpreted from a commonsense point of view. And it is easy to understand why normal adult human beings reflexively make this assumption when they interpret the behaviour of animals (Dennett 1987). From a scientific stance, however, we are only warranted in attributing an *ms* variable to the subject if we can specify why an f_{ToM} of some kind is computationally necessary in order to perform the given behaviour and why the information provided by the resulting *ms* variable is not redundant with the

RSTB 20062023—2/1/2007—16:53—THIAGU—266980—XML RSB – pp. 1–15

information provided by the *r*, *p*, *g* and *q* variables which we have already posited to exist. The role of an experimentalist (as opposed to the folk observer) is to construct situations or protocols in which the unique cognitive work performed by the *ms* variables can be distinguished from the work that could be performed by *r*, *p*, *g* and *q* inputs alone.

Here is the crux of the matter then, and possibly the most important point we will make in this essay: in almost all experimental procedures reported to date, purported *ms* variables appear to be causally superfluous re-descriptions of the other observable inputs and representations that are logically required by the experimental design. No special f_{ToM} is required. The problem with existing protocols is that they fail to create situations in which the information purportedly carried by the *ms* variables is not causally redundant with the information already carried by the *r*, *p*, *g* and *q* variables.

Now, we are ready to evaluate the evidence with respect to the formalism we have outlined.

4. AN EXPERIMENTAL PROTOCOL THAT CANNOT, EVEN IN PRINCIPLE, PROVIDE EVIDENCE FOR f_{ToM}

This is not the forum for an exhaustive examination of all claims for theory of mind in chimpanzees (let alone other species). Our strategy, therefore, will be to examine what has come to be seen as the 'strongest' case for the existence of theory of mind in chimpanzees: the work of Hare *et al.* (2000, 2001). To be perfectly clear, we do not believe these studies have any bearing whatsoever, positive or negative, on the question of whether chimpanzees reason about mental states. However, because many other scholars believe they do, we shall use this protocol as a case study to expose the conceptual confusion that dominates this area of research.

We will take the 'most significant' experiment reported by Hare *et al.* (2001) as our example, but it must be noted that our analysis applies with equal force to all the experiments in this series (see also Povinelli & Vonk 2004). Two chimpanzees, one subordinate to the other, were kept in separate chambers on either side of a middle area. Two cloth bags in the middle chamber served as hiding places for small food items. Opaque doors on each side chamber prevented the respective chimpanzees from entering the middle chamber and retrieving the food until the doors were raised. On each trial, the subordinate's door was partially raised while the food was being hidden, allowing the subordinate to peek out and see where the food items were placed and whether or not the dominant was present and looking. On each trial, the dominant's door was either partially raised or completely closed while the food items were placed in one of the two containers. Once the food had been placed, the dominant's door was closed and the subordinate was released into the middle chamber and given a slight headstart before the dominant was released as well.

Hare *et al.* (2001) reported a number of experimental conditions based on this protocol. In only one of these experiments, however, was the critical metric statistically significant[3]. In the uninformed condition of experiment 1, the dominant's door was kept closed while the food was hidden and the subordinate could see that the dominant's door was closed; in the control condition, the dominant could see where the reward was hidden and the subordinate could see that the dominant was watching. The subordinate 'approached' the hidden food more often in the uninformed condition than in the control condition. On the basis of this result, Hare *et al.* (2001) concluded that 'chimpanzees know what individual groupmates do and do not know' (p. 148). Reversing their previous opinion on the matter (see Tomasello & Call 1997; Visalberghi & Tomasello 1998), Tomasello *et al.* (2003*a*) cite these experiments as 'breakthrough' (p. 154) evidence that chimpanzees 'understand some psychological states in others' (p. 156). Tomasello *et al.* are hardly alone. The Hare *et al.* (2000, 2001) results are now widely cited as supporting evidence for the idea that chimpanzees possess some kind of f_{ToM}.

Unfortunately, as our research group has pointed out that (see Karin-D'Arcy & Povinelli 2002; Povinelli & Vonk 2003, 2004), the protocol employed by Hare *et al.* (2001) lacks the power, even in principle, to distinguish between responses by the subordinate that could have been produced simply by employing observable information and representations of past behavioural patterns (i.e. *p*- and *r*-states) from responses that must have required computations involving information about the dominant's unobservable mental states (i.e. *ms* states). For example, Povinelli & Vonk (2003) point out that the behaviour of the subordinates might result from a simple strategy glossed by 'Don't go after food if a dominant who is present has oriented towards it.' The additional claim that the chimpanzees adopted this strategy because they understood that 'The dominant knows where the food is located' is intuitively appealing but causally superfluous.

Let us re-examine the problem with Hare *et al.*'s protocol using the formalism we developed above. Imagine an organism, **A**, that manifests some determinate set of observable cues, $\mathbf{C_1}$, when it is in a given *r*-state, $\mathbf{r\text{-}state_1}$, where $\mathbf{C_1}$ = ('eyes of A oriented towards food', 'uninterrupted visual access between A and placement of food', 'food is placed in location X', …) and $\mathbf{r\text{-}state_1}$ = ('A knows that food is in location X'). And suppose further that $\mathbf{r\text{-}state_1}$ causes **A** to emit behaviour $\mathbf{b_1}$, where $\mathbf{b_1}$ = ('A tries to retrieve food in location X'). A second cognitive agent having perceptual access to organism A and its observable traits, $\mathbf{C_1}$, would have no need to infer the presence of $\mathbf{r\text{-}state_1}$ in order to predict the occurrence of $\mathbf{b_1}$; simply observing $\mathbf{C_1}$ suffices. Thus, a researcher observing that a given experimental subject is able to reliably predict the occurrence of $\mathbf{b_1}$ in **A** after observing $\mathbf{C_1}$ would have no basis for concluding that the subject possesses an f_{ToM} dedicated to inferring $\mathbf{r\text{-}state_1}$ (even if she herself knows that $\mathbf{r\text{-}state_1}$ causes $\mathbf{b_1}$), unless she can also show that possessing information directly about $\mathbf{r\text{-}state_1}$ does some special causal work in addition to predicting $\boldsymbol{b_1}$. Once again, we believe this point to be indisputable—though, as in the case of Hare *et al.* (2001), persistently (and inexplicably) disputed (see Tomasello *et al.* 2003*a*,*b*; Tomasello & Call 2006).

5. WHAT ABOUT CORVIDS?

Chimpanzees, of course, are not the only non-human species which might be potential candidates for an f_{ToM}. And, indeed, some of the most well-controlled results and provocative claims in recent years have not come from experiments with primate subjects at all, but from experiments with corvids (for general reviews of the literature, see Clayton *et al.* 2001; Emery 2004; Emery & Clayton 2004, 2005; Clayton & Emery 2005; see also Clayton *et al.* 2006). Corvids are quite adept at pilfering the food caches of other birds and will adjust their own caching strategies in response to the potential risk of pilfering by others. Indeed, not only do they remember which food caches were observed by competitors, but also they appear to remember the specific individuals who were present when specific caches were made and modify their re-caching behaviour accordingly (Dally *et al.* 2006). Corvids' cognitive prowess is not limited to caching and pilfering. In many tool-use tasks, their cognitive abilities also seem to be superior to those of non-human primates in certain respects (for example, Hunt 1996, 2004; Seed *et al.* 2006; Tebbich *et al.* in press). What is at issue here, however, is not whether or not corvids are cognitively sophisticated creatures, but whether or not, *in addition*, any of their sophisticated cognitive abilities require the possession of an f_{ToM}.

Many comparative researchers clearly feel the answer to this question is yes. For example, Emery & Clayton (2001, 2004, 2005) suggest that corvids discriminate between competitors who possess knowledge of cache sites from those that do not by attributing specific, contentful *r*-states to knowledgeable competitors. Moreover, Emery and Clayton suggest that corvids may be able to understand the internal mental experience of their conspecifics by analogy to their own first-hand experience (see also Emery 2004). Similarly, Bugnyar & Heinrich (2006) showed that ravens delay pilfering from cache sites when confronted by the individuals who made those caches and suggest that this is consistent with the hypothesis that corvids possess a sophisticated understanding of others' visual perception as well as the ability to tactically manipulate competitors' mental states (see also Bugnyar & Heinrich 2005).

While we certainly agree with these researchers that it is *possible* that corvids are capable of reasoning in terms of the *r*-states of their competitors, we nevertheless must point out that none of the evidence to date provides convincing evidence for this hypothesis. One of the defining characteristics of *ms* variables, as defined above, is that they are construed from the cognitive perspective of the other agent as distinct from the cognitive perspective of the subject itself. Unfortunately, none of the reported experiments with corvids require the subjects to infer or encode any information that is unique to the cognitive perspective of the competitor. For example, none of the reported experiments require the subjects to reason in terms of the *counterfactual* content of their competitors' *r*-states. As Dennett (1987) pointed out a long time ago, without evidence that a subject is able to reason in terms of counterfactual as well as factual *r*-states in another agent, it is very difficult, if not impossible, to provide evidence that they are cognizing the other agent's *r*-states qua *r*-states at all.

In all of the experiments with corvids cited above, it suffices for the birds to associate specific competitors with specific cache sites and to reason in terms of the information they have observed from their own cognitive perspective: e.g. 'Re-cache food if a competitor has oriented towards it in the past', 'Attempt to pilfer food if the competitor who cached it is not present', 'Try to re-cache food in a site different from the one where it was cached when the competitor was present', etc.[4] The additional claim that the birds adopt these strategies because they understand that 'The competitor knows where the food is located' does no additional explanatory or cognitive work.

The case for 'experience projection' is no stronger than the case for 'knowledge attribution'. Emery & Clayton (2001) showed that scrub jays who had had previous experience pilfering food from others were more likely to re-cache food that had been observed by competitors than birds who had had no previous experience pilfering from others. 'This result raises the exciting possibility,' Emery (2004, p. 21) writes, 'that birds with pilfering experience can project their own experience of being a thief onto the observing bird, and so counter what they would predict a thief would do in relation to their hidden food' (see also Emery & Clayton 2004).

The fact that only birds with previous pilfering experience re-cache observed food sites is an interesting result but sheds no light on the internal mental representations or cognitive processes being employed by the birds in question. This experimental result certainly does not demonstrate that ex-pilferers understand anything about the internal, subjective experience of their potential competitors. Monkeys, after all, often initiate aggressive acts against innocent third parties after they themselves have been attacked but this hardly means that they are projecting their own subjective experience of being attacked onto the potential victims. There are any number of much lower-level explanations for this redirected aggression (see Silk (2002) for a review)—as there are for the connection between pilfering and re-caching in corvids.

To be sure, many researchers explicitly acknowledge that an explanation based on reasoning about observed cues alone is sufficient to account for the existing data. Dally *et al.* (2006), for example, acknowledge, that scrub jays' ability to keep track of which competitors have observed which cache sites 'need not require a humanlike 'theory of mind' in terms of unobservable mental states, but […] may result from behavioral predispositions in combination with specific learning algorithms or from reasoning about future risk.' Similarly, Bugnyar & Heinrich (2006) acknowledge that a representation of 'states in the physical world' would be sufficient for explaining the available evidence concerning the manipulative behaviours of ravens. Notwithstanding the foregoing, these researchers continue to hold out the 'possibility' that the birds' behaviour could be consistent with a more generous, mentalistic interpretation and suggest that more generous interpretations might be more 'parsimonious' (see also Tomasello & Call 2006).

RSTB 20062023—2/1/2007—16:53—THIAGU—266980—XML RSB – pp. 1–15

Admittedly, explanations in terms of folk psychological abilities do appear more 'parsimonious' at first blush. But the fact that such explanations are 'simpler for us' to understand does not mean, as Heyes (1998) pointed out, that they are 'simpler for them' to implement (see also Dennett 1987). The cognitive mechanisms that would be required to actually implement these purported f_{ToM} abilities at a subpersonal, causal level are hardly simple at all—they only seem simple because folk psychological explanations gloss over all the devilish details. Comparing the simplicity of a folk psychological explanation, e.g. 'chimpanzees understand seeing', 'corvids know what others do and do not know', to the complexity of a subpersonal cognitive explanation is like comparing a marketing description of Microsoft Word, e.g. 'prints, saves and edits complex documents', to a detailed functional specification of the underlying application architecture. The fact that the detailed functional specification of Microsoft Word runs to thousands of pages, and the marketing pitch takes one sentence is not a reasonable metric for comparing the merits of the two descriptions. Likewise, while folk psychological descriptions may be invaluable heuristics for ethologists in the field (Dennett 1987), they should not be confused or compared with cognitive hypotheses framed at a subpersonal, functional level of explanation.

Our position is that chimpanzees and corvids (like many other non-human animals) possess representational architectures of enormous sophistication and flexibility. We also believe that they employ both inferential and simulative mechanisms for forming abstractions about classes of behaviours and environmental conditions that are relevant to their goal-directed actions. Furthermore, we believe that non-human animals are able to generalize the lessons learned from these abstractions to novel scenarios.

Thus, unlike the motley collection of learning experiences that might be required in an associationist model, our hypothesis is that non-human animals are able to respond intelligently to novel situations based on general, abstract representations (i.e. *r*-states) they have formed about similar situations in the past and specific, concrete representations they have formed about the events leading up to the present moment (including, at least in the case of corvids, the 'what', 'when' and 'where' information associated with those events).

Our principle disagreement with those who explain non-human behaviours in terms of an f_{ToM} is not about the inferential or learning abilities that non-human animals possess (at least for our present purposes; but see Penn & Povinelli in press). Our principle disagreement is about the kind of representations over which these inferential and learning processes operate. The available evidence suggests that chimpanzees, corvids and all other non-human animals only form representations and reason about *observable* features, relations and states of affairs from their own cognitive perspective. We know of no evidence that non-human animals are capable of representing or reasoning about *unobservable* features, relations, causes or states of affairs or of construing information from the cognitive perspective of another agent. Thus, positing an f_{ToM}, even in the case of corvids, is simply unwarranted by the available evidence.

6. TWO EXPERIMENTAL PROTOCOLS THAT COULD, IN PRINCIPLE, PROVIDE EVIDENCE FOR f_{ToM}

In response to the kind of critiques that our research group has levelled, some scholars have claimed that the distinctions we are proposing are experimentally intractable and/or empirically vacuous. For example, Andrews (2005) worries 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' (p. 528 and see also Leavens *et al.* 2004; Hurley 2006). Tomasello *et al.* (2003*b*) worry that our extreme stinginess in attributing mentalistic abilities to chimpanzees is an example of 'derived behaviourism' and will only lead to 'despair' (p. 239).

To forestall any worry that a theoretically rigorous stance towards the interpretation of comparative experimental results will lead only to despair, we will now propose two separate experimental protocols that could, in fact, provide principled evidence for an f_{ToM} in chimpanzees or corvids and could be easily adapted for other non-verbal cognitive organisms as well. The first tests a non-verbal subject's ability to reason from first- to third-person mental states. The second tests a subject's ability to use *ms* variables to solve prediction problems that would be computationally unsolvable otherwise. We hope these two proposals will demonstrate that our stringent criteria for attributing an f_{ToM} to a non-human animal are neither empirically vacuous nor experimentally intractable.

(a) *The opaque visor experiment*

Building on previous suggestions, Povinelli & Vonk (2003, 2004) highlighted (in a version appropriate for chimpanzees) one protocol that could provide principled positive evidence for f_{ToM} in a non-verbal organism. Since this proposal has now been critiqued, we briefly summarize its logic, and show why the critiques are invalid.

During an initial training session, subjects are given first-hand experience wearing two mirrored visors. One of the visors is see-through; the other is not. The visors themselves are of markedly different colours (and/or shape). During the subsequent test session, the subjects are given the opportunity to use their species-typical begging gesture to request food from one of the two experimenters, one wearing the see-through visor and the other wearing the opaque visor. Subjects who beg significantly more often from an experimenter wearing the see-through visor have manifested evidence of possessing an f_{ToM} in the sense defined herein.

This protocol has been tested on highly human-enculturated chimpanzees (Vonk *et al.* in press), who failed. A functionally equivalent variation of the protocol (using trick blindfolds) has been tested on 18-month-old human infants (Meltzoff in press), who passed. These results would seem to provide positive confirmatory evidence that even very young human

RSTB 20062023—2/1/2007—16:54—THIAGU—266980—XML RSB – pp. 1–15

infants possess some sort of f_{ToM} whereas even highly enculturated adult chimpanzees do not.

There have been several criticisms of the experimental protocol, ranging from the claim that it is formally inadequate (Andrews 2005; Hurley 2006) to the claim that it has 'very low ecological validity' (Tomasello *et al*. 2003*b*). We will first defend why the proposed experiment does, in fact, provide principled evidence for an f_{ToM} and, secondly, why the charge of 'low ecological validity' is misplaced.

Both Hurley (2006) and Andrews (2005) argue that a subject could pass the proposed experiment simply by reasoning about the analogy between first-person manifest physical behaviours and third-person manifest behaviours. As Andrews (2005) puts it:

> …the chimp might make the behavioral connection between wearing the opaque bucket and *not being able to do things* [emphasis in the original]. From whom should he beg? Certainly not the person who isn't able to do things (p. 530).

It is certainly true that reasoning from first- to third-person behaviours forms a crucial part of the human cognitive tool-kit (for example, Meltzoff & Moore 1997; Meltzoff in press). And there is substantial evidence that neural systems, such as 'mirror neurons', in both human and non-human animals register correspondences between first- and third-person behaviours (for reviews of the literature, see Hurley & Chater 2005). Thus, it is possible (though certainly not proven) that the capacity to find behavioural equivalences between self and other is, as Hurley (2006) argues, developmentally and phylogenetically prior to the capacity to find mentalistic equivalences between self and other.

However, the ability to form first- to third-person equivalences in terms of manifest physical behaviours is not sufficient to solve the protocol proposed by Povinelli & Vonk. The reason the bucket protocol works as a test of mental state reasoning is because there is, in fact, no way (i.e. no computationally tractable way) to draw the necessary correspondences based purely on representations of observable information and manifest behaviours.

In this context, let us examine more closely the data available to a subject lacking an f_{ToM}. Such a subject would be limited to *r*-states about his own manifest behaviour while wearing the opaque visor (e.g. 'I stumbled around while wearing the red visor') and occurrent *p*-states about the experimenter (e.g. 'she is wearing a red visor'). However, a subject lacking an f_{ToM} would not have access to *r*-states about his own internal cognitive states while wearing the visors (e.g. 'I was unable to see while wearing the red visor'). Nor would such a subject have any information concerning his own propensity to respond to begging gestures while wearing the opaque visor, since he never attempted to respond to begging gestures while wearing the visor.

Thus, a subject capable of cognizing analogies between first- to third-person physical behaviours, but incapable of cognizing analogies between unobservable mental states, might be able to infer that the experimenter will stumble around and bump into things while wearing the red visor; but there would be no basis for this subject to infer that wearing the red visor will necessarily preclude the experimenter from *physically* producing the actions necessary to respond to begging gestures. Indeed, the subject would have every reason to believe that wearing the red visor will have no effect at all on the experimenter's ability to respond to begging gestures.

In the proposed protocol, the only manifest physical actions required for the experimenter to respond to begging gestures are the ability to sit still, move her arm and keep her eyes open and directed straight ahead. The subject has first-hand experience that he is perfectly capable of sitting still, of freely moving his arms and of keeping his eyes open while wearing the red visor. Thus, based on the manifest behavioural evidence, a subject without an f_{ToM} would have no reason to suspect any limitation on the experimenter's ability to perform the physical acts required to respond to begging gestures. In order to infer that the experimenter is not likely to respond to begging gestures while wearing the red visor, the subject must realize that responding to begging gestures requires more than a set of manifest physical actions and observable conditions. To be precise, the subject must realize (by logical inference or embodied simulation, or some combination of the two) the following:

(i) wearing the opaque visor results in an inability to 'see-what-is-going-on' (i.e. a general epistemic condition applicable to any subsequent behaviour not just a particular manifest physical effect of bumping-into-things),
(ii) this general epistemic condition will be experienced, analogously, by the other subject when she wears the red visor but not the blue visor, and
(iii) a subject who experiences this general epistemic condition will not respond to begging gestures.

The preceding three steps are a paradigmatic example of encoding an *ms* variable about a first-person internal state (i.e. the general epistemic condition of not-being-able-to-see) that results from a given manifest contingency (i.e. wearing the red visor) and then using these representations to predict the behaviour of another cognitive agent to a novel situation (i.e. responding to begging gestures). We contend that without the *ms* variable, the subject could not immediately solve the problem presented.

Some (e.g. Andrews 2005) might still object that during the initial, first-person familiarization phase, the chimpanzee could form a general aversion to red visors or might make the blanket inference that since 'I can't do anything with the red visor on', others will not be able to do anything either.

We should first point out that no such generalized aversion to the opaque bucket was observed in the familiarization phase of this experiment with chimpanzees (Vonk *et al*. in press). More importantly, the protocol calls for the subjects to learn that they can do many things while wearing the opaque visors: they run about, reach out, feel objects and their body, and they themselves engage in acts that look very much like begging gestures (Vonk *et al*. in press). Thus, it is simply false that the subjects learn that 'I can't do anything with the red visor on'.[5]

We now turn to Tomasello *et al.*'s (2003*b*) objection that the visor test lacks 'ecological validity' because it involves a 'cooperative–communicative' rather than a 'competitive' paradigm (Hare 2001) and because it involves strange artefacts like visors.

Several things need to be noted about this objection. First, it is simply false to claim that chimpanzees are more likely to reveal their true cognitive potential under 'competitive' situations rather than 'cooperative/communicative' ones (Hare 2001). Certainly, they may exhibit different cognitive abilities in competitive versus cooperative/communicative situations, but there is no empirical or theoretical basis for claiming that the abilities revealed under competitive paradigms are either more fundamental or more sophisticated than those revealed under cooperative ones.

For example, consider the chimpanzees' natural food-begging gesture (Goodall 1990), a gesture that has been observed in all captive and free-ranging populations of chimpanzees. In a simple experimental setting, if a chimpanzee is confronted with two caretakers who could potentially give them food, but one is facing towards them and the other is facing away, the chimpanzee will immediately (from trial one forward) gesture to the one facing them (Povinelli & Eddy 1996*a–c*). Chimpanzees are even capable of selectively employing auditory rather than visual behaviours as a function of specific perceptual/behavioural cues exhibited by the caretaker from whom they are begging (Hostetter *et al.* 2001; Leavens *et al.* 2004). It is only when more subtle experimental manipulations are employed, that chimpanzees display their lack of understanding of the specific causal relation between the disposition of the eyes or face of the caretaker and the caretaker's mental state (see Povinelli (2003) chapter 3 for a review).[6] Of course, this cooperative–communicative act—gesturing to the front (as opposed to the back) of a communication partner—is part of the natural social behaviour of chimpanzees (see Tomasello *et al.* 1994), as is competition over food resources (Karin-D'Arcy & Povinelli 2002). In other cooperative experimental settings, where a chimpanzee needs help in obtaining a just-out-of-reach food item, chimpanzees will robustly modulate their gestures to fit the locations to where their cooperative partner is looking (Povinelli & Vonk 2004). Thus, we are just as impressed by the sophistication of chimpanzee social cognition in cooperative–communicative situations as we are by their sophistication in competitive ones.

Claiming that visors are ecologically 'unnatural' (Hare 2001, p. 276) is a disingenuous argument. When chimpanzees pass tests involving ecologically bizarre artefacts, such as blindfolds, locked boxes, transparent tubes and mirrors, the same experimenters are quick to claim victory. When chimpanzees fail, the visors are to blame.

In any case, the point of the proposed protocol is not the visors. The point of the proposed protocol is the functional, informational challenge it poses. There are certainly many species for whom having a visor covering their eyes is not a species-typical experience. It suffices to find an alternative implementation of the experiment that retains the same informational and Q6 functional challenge in a more species-acceptable form.

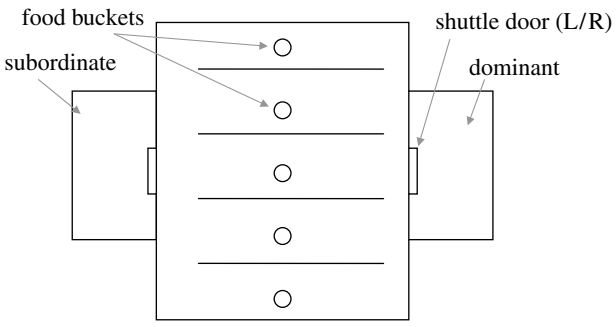

Figure 1. General experimental set-up for five-bucket protocol (see §6*b* for details).

Meltzoff (in press) provides an exemplary case study: he cleverly adapted the proposed protocol for human infants using blindfolds and recorded whether or not infants were more likely to track the gaze of an adult wearing an opaque blindfold than one wearing a see-through blindfold. Notably, although tracking the gaze of blindfolded adults has pretty low ecological validity for human children as well, the18-month-old children, nevertheless, passed.

To be sure, it is true that failure on any experimental test of this sort is not demonstrative evidence of a lack of f_{ToM}: false negatives are a fact of life in comparative research as they are in ToM research in general (see Birch & Bloom 2004). *Ceteris paribus*, ecological validity is often (but not always) a desirable feature of comparative experimentation. But the more critical issue is to isolate experimental procedures that are capable, at least in principle, of providing positive (or negative) evidence for the specific cognitive skills. Unfortunately, most of the 'ecologically valid' protocols currently in vogue cannot provide principled evidence for or against the presence of f_{ToM}. The proposed visor protocol is simply one example of an experiment that can.

For those who nonetheless insist that only competitive paradigms will reveal the true nature of chimpanzee cognition, we propose a second experimental protocol below that retains the purported 'ecological validity' of Hare *et al.*'s (2001) competitive paradigm while, nevertheless, proffering the possibility of positive evidence for an f_{ToM}.

(b) *A systematic version of Hare* et al.*'s competitive food protocol*

As in Hare *et al.*'s (2001) experiment described above (see §4), a subordinate and a dominant chimp are kept in separate compartments on opposite sides of a middle chamber and each side chamber is separated from the middle chamber by an opaque shuttle door (see figure 1). The doors are raised and lowered and the two subjects released into the middle chamber. Unlike Hare *et al.*'s set-up, however, the middle chamber has *n* stalls (e.g. 5) spaced evenly across the width of the compartment, divided from each other by Plexiglas walls. There are five buckets on the floor at the centre of each stall in full view of the subjects. The contents of the bucket, however, are not directly visible to the subjects. On each trial, the experimenter places two different amounts of food into two different buckets:

RSTB 20062023—2/1/2007—16:54—THIAGU—266980—XML RSB – pp. 1–15

a larger amount of food is placed in one bucket and a visibly smaller amount of food is placed in another. The order in which the amounts are placed is randomized (i.e. on one-half of the trials, the larger amount is placed first).

The experiment is carried out in a series of incrementally more challenging steps. In the step 1, subjects are exposed to a series of non-competitive trials. There is no rival present during these trials and both rewards are placed in full view of the subject. When the subject is released, it is only allowed to approach and retain the contents of one bucket. Trials continue until the subject learns to reliably approach and retain the more desirable reward.

In step 2, chimpanzees are paired in dominant/subordinate dyads. In each dyad, both chimpanzees have full visual access to the placement of both rewards. Only dyads in which subordinates learn to retrieve the less desirable reward and dominants retrieve the more desirable reward in a reliable fashion are allowed to continue to the third and final session.

In the step 3, the following conditions are randomly presented (Note that in all conditions, the subordinate has complete visual access to the activities of the experimenter. Only the dominant's visual access is manipulated as described.):

— *Informed control*. Both chimpanzees have full visual access to the placement of both food rewards.
— *Partially uninformed*. One reward is placed while the dominant chimp is looking and the other reward is placed while dominant's door is down. Whether or not the dominant's door is down during the initial placement or the subsequent placement is randomized.
— *Removed informed*. Both rewards are placed while the dominant subject is looking. Then, one of the rewards is removed from the middle chamber and replaced with an empty bucket while the dominant is looking.
— *Removed uninformed*. Both rewards are placed while the dominant subject is looking. Then, one of the rewards is removed from the middle chamber and replaced with an empty bucket while the dominant's door is down.
— *Moved*. The dominant's door is down during the initial placement of two rewards; then the dominant's door is open and both rewards are moved to new locations while the dominant is watching.
— *Replaced*. The dominant witnesses the placement of one of the two rewards and then the dominant's door is closed while that reward is moved to a new location and the amount not witnessed is placed in the previously occupied bucket.
— *Misinformed*. Both rewards are placed while the dominant is looking; then, while the dominant's door is down, one of the buckets (which may or may not have food in it) is moved to the location occupied by one of the rewards, that bucket and its reward are moved to a new location and the bucket at that location is put back in the stall originally occupied by the first bucket.
— *Swapped*. Both rewards are placed while the dominant is looking, then the locations of the two buckets are swapped while the dominant's door is down.
— *Other variations*. Note that the conditions described above only represent a subset of the systematic variations which could be employed.

The initial two steps can be mastered using simple heuristics based on observable contingencies. However, if the subject learns to pass the initial sessions using only observable contingencies, and does not have access to an f_{ToM}, the final test session presents an intractable mess.

For example, the response rule 'Don't go after food if the dominant has oriented towards it in its present location' (Povinelli & Vonk 2003), which worked perfectly in the original protocol proposed by Hare *et al.*(2001), no longer suffices. The relational rule 'always retrieve the less desirable of two rewards when there's a dominant present' only works consistently under the *informed control* and *moved* conditions. Even the higher-order relational strategy, 'Go after the less desirable reward unless the dominant has previously oriented towards it in its current location' fails any condition in which it would be optimal for the subordinate to retrieve the larger food item (e.g. the *Swapped* condition). Based purely on patterns of observable cues, each condition requires a different response rule; and there is no way to systematically generalize from familiar to novel conditions.

For the purposes of testing whether or not a subject possesses an f_{ToM}, the critical conditions are those which require the subject to formulate an *ms* variable that keeps track of where the dominant' believes the food rewards are located as distinct from where they are actually located, e.g. the *removed uninformed*, *replaced*, *misinformed* and *swapped* conditions. In the context of the present protocol, i.e. randomly interspersed among the other conditions, there is no way for a subject to reliably pass these critical conditions without the ability to keep track of the counterfactual state of affairs from the dominant's cognitive perspective while simultaneously keeping track of the occurrent state of affairs from the subject's own perspective. The subject must not only understand that the competitor was present and oriented; but he must also cognize the specific content of the competitor's counterfactual *r*-states and relate these counterfactual *r*-states to the competitor's subsequent behaviour. Success on these conditions is thus functionally (though not necessarily psychologically) equivalent to reasoning in terms of a competitor's 'false beliefs' and would provide compelling evidence for an f_{ToM}.

Failure, however, is no less instructive than success. A subject who has passed the first two training steps has clearly understood the procedural aspects of the task, and the protocol retains the competitive food paradigm advocated so vigorously by Hare (2001) and others. Thus, unlike previous non-verbal 'false belief' tests (e.g. Call & Tomasello 1999) or even the protocol proposed by Hare *et al.* (2001), failure on this one cannot be blamed on interspecific misunderstandings, ecological implausibility or the subjects' inability to understand the procedural aspects of the task.

RSTB 20062023—2/1/2007—16:54—THIAGU—266980—XML RSB – pp. 1–15

Indeed, it is the *pattern* of successes and failures on different conditions in our protocol that is likely to provide the most interesting evidence concerning the cognitive strategy being employed by a given non-human subject. For example, a subject who employs a 'Don't go after a food reward if the dominant has oriented towards it' strategy will pass a different set of conditions than a subject employing a 'Always retrieve the less desirable of the two food amounts' strategy. Similarly, a subject who passes the *removed informed* condition but not the *removed uninformed* condition (or vice versa) has revealed something significant about the characteristics and the limitations of the cognitive strategy he is employing.

It might be objected that the complexity of the conditions in our version of Hare *et al.*'s protocol is too great for chimpanzees or corvids to handle and that the processing capacity limitations of these subjects are orthogonal to the question of whether or not they possess an f_{ToM}. The conditions in our five-bucket protocol do, indeed, pose a significant degree of 'relational complexity' (Halford *et al.* 1998), but we disagree with the claim that this invalidates the protocol as a test of a subject's ability to reason about what their conspecifics do and do not know.

While our five-bucket protocol poses an intractable computational challenge to a subject without an f_{ToM} of any kind, our protocol would be much less daunting to a subject who is able to encode the appropriate *ms* state variables. As Whiten and Suddendorf pointed out, one function of an f_{ToM} is to reduce the complexity of social interactions by positing abstract hidden variables that encode abstract, relational similarities between perceptually disparate behavioural patterns (Whiten 1996, 1997, 2000; Suddendorf & Whiten 2001; Whiten & Suddendorf 2001). For example, a subject endowed with the appropriate simulative abilities should be able to significantly reduce the relational complexity of the task by first simulating what they would do from the perspective of the dominant competitor. (Indeed, we suspect that many readers did exactly this while reading the description of each condition.)

Furthermore, we would argue that the ability to perceive relational similarities between perceptually disparate behavioural patterns (i.e. to form 'abstract equivalence classes'; in Whiten's (1996) terms) and to postulate the existence of unobservable causes like mental states are paradigmatic examples of higher-order relational reasoning (see Gentner *et al.* 2001 for an overview of the current literature; see Penn & Povinelli in press for a relational analysis of non-human causal cognition). Consistent with this hypothesis, Andrews *et al.* (2003) have shown that children's ability to reason relationally and their ability to reason about unobservable mental states is closely linked, both computationally and ontogenetically (see also Halford *et al.* 1998; Zelazo *et al.* 2002). Thus, the ability to encode *ms* variables via an f_{ToM} is probably inseparable, both computationally and phylogenetically, from the ability to reason about the relational similarity between complex behavioural patterns and higher-order causal relations.

(c) *Take-home lessons from the proposed experimental protocols*

The key point to be taken from the two protocols proposed herein is not that they constitute an acid test for an f_{ToM} in a chimpanzee or corvid, or that failure on these tests would be demonstrative evidence of an absence of an f_{ToM}. Rather, they are a direct response to the concern that success in any predictive paradigm can be explained as the result of a behaviouristic psychological system rather than mental, intervening variables (e.g. Andrews 2005). If this concern were true, then the entire project of testing non-human animals' ability to use an f_{ToM} to predict the behaviour of their conspecifics would be experimentally intractable and otiose. While this concern applies to virtually all other experimental protocols to date, the present proposals are existence proofs that experimental protocols can be constructed that could provide positive, principled evidence for the predictive function of an f_{ToM} in non-verbal organisms.

Q8 We hope our proposed protocols also put to rest the worry that an f_{ToM} has no functional, adaptive value or, worse, may by a figment of our folk psychological imagination. Regardless of our doubts concerning the ontological status of the hypothetical entities posited by our folk psychology, it is clear to us that the ability to cognize the world from the cognitive perspective of another agent would provide an animal with enormous advantages over and above the ability to reason in terms of observable first-person relations alone. Our proposed experiments set forth two artificial examples of how the value of such an f_{ToM} might manifest itself. Hundreds of experimental studies with young children have shown that they are able to solve the kind of tasks that require an f_{ToM} in the sense defined herein (e.g. Meltzoff (in press); and see Wellman *et al.* (2001) for a review and meta-analysis). And there are good reasons for believing that the traditional hallmarks of human cognition, language and culture, are intimately dependent on f_{ToM} systems of various kinds (for example, Bloom 2000, 2002; Tomasello *et al.* 2005). The problem is not that a ToM system has no value or is experimentally intractable; the problem is that there is still no evidence that non-human animals possess anything remotely resembling one.

8. UNCITED REFERENCES

Q9 Churchland & Churchland (1996) and Godfrey-Smith (2000).

The theoretical work developed in this essay was generously supported by a James S. McDonnell Foundation Centennial Fellowship to DJP.

ENDNOTES

[1]Of course, not all comparative researchers believe that non-human animals are cognitive agents in the sense defined by equation (2.1). But all comparative researchers who believe that non-human animals are potentially capable of possessing an f_{ToM} must necessarily believe that these same animals are cognitive agents in the sense defined by equation (2.1) above.

[2]NB: it is not necessary for there to be a deterministic relation between the observable and the unobservable variables. Our argument holds, mutatis mutandis, whenever $P(\mathbf{b}_1|\mathbf{C}_1) > P(\mathbf{b}_1|\sim\mathbf{C}_1)$ or, indeed,

anytime a probabilistic model (e.g. Bayesian) can predict $\mathbf{b}_1$ on the basis of observable cues and past conditional dependencies without taking the value of **r-state**$_1$ into account.

[3]Hare *et al.* used two metrics, 'retrieve' and 'approach', to measure the animals' performance on these tests. The first recorded the percentage of food items actually retained by the subordinate. The second recorded the percentage of trials on which the subordinate left its own chamber and crossed into the middle chamber prior to the dominant being released. As Karin-D'Arcy & Povinelli (2002) note, given the fact that the dominant chimp often did not know where the food was located and given the fact that the subordinate was given a sizeable headstart, it is hardly meaningful that the subordinate retrieved more food. Thus, as an important and overlooked point of scholarship, the approach metric was not statistically significant in the Misinformed condition of experiment 1, or in any of the other experiments reported in Hare *et al.* (2001).

[4]These glosses are not meant to suggest that corvids are constrained to simple conditional rules. We believe that corvids, like many other non-human animals, are perfectly capable of reasoning about the world in a flexible manner, albeit only with respect to observable first-person relations.

[5]Andrews' (2005) objection nevertheless suggests an interesting modification to the visor protocol. First, train the chimpanzees to (i) make a begging gesture in front of experimenters who can see them and (ii) to produce an auditory cue (e.g. stomping) in front of any experimenter who cannot see them (using the kind of seeing/not-seeing conditions developed by Povinelli & Eddy (1996*b*), such as bucket-over-head, blindfold on and back turned). In the transfer session, present the subject with a single experimenter wearing either the opaque or see-through visor and test whether or not the subject stomps or begs in front of that experimenter. Chimpanzees who have simply learned to stomp in response to an arbitrary set of perceptual cues (e.g. bucket-over-head, blindfold on, back turned), without any understanding of the underlying epistemic states involved will stomp regardless of the kind of bucket being worn. Chimpanzees who have cognized the physical conditions that result in 'seeing' and physical conditions that result in 'not-seeing' will beg and/or stomp from the experimenter with the see-through visor, but stomp in front of the experimenter with the opaque visor.

[6]One might ask why, given that chimpanzees do preferentially gesture to someone facing them as opposed to someone facing away, this is not *prima facie* evidence for an understanding of the perceptual state of seeing. The point to be clarified by the formalism of this paper is that immediate knowledge of how to respond to a social context is completely orthogonal to the question of whether the chimpanzee's underlying representation of the situation is comprised of *r*, *p* and *ms* variables, or *r* and *p* variables alone.

REFERENCES

Andrews, K. 2005 Chimpanzee theory of mind: looking in all the wrong places? *Mind Lang.* **20**, 521–536.

Andrews, G., Halford, G. S., Bunch, K. M., Bowden, D. & Jones, T. 2003 Theory of mind and relational complexity. *Child Dev.* **74**, 1476–1499. (doi:10.1111/1467-8624.00618)

Birch, S. A. J. & Bloom, P. 2004 Understanding children's and adults' limitations in mental state reasoning. *Trends Cogn. Sci.* **8**, 255–260. (doi:10.1016/j.tics.2004.04.011)

Bloom, P. 2000 *How children learn the meaning of words*. Cambridge, MA: MIT Press.

Bloom, P. 2002 Mindreading, communication and the learning of names for things. *Mind Lang.* **17**, 37–54.

Bugnyar, T. & Heinrich, B. 2005 Ravens, *Corvus corax*, differentiate between knowledgeable and ignorant competitors. *Proc. R. Soc. B* **272**, 1641–1646. (doi:10.1098/rspb.2005.3144)

Bugnyar, T. & Heinrich, B. 2006 Pilfering ravens, *Corvus corax*, adjust their behaviour to social context and identity of competitors. *Anim. Cogn.* **9**, 369–376. (doi:10.1007/s10071-006-0035-6)

Call, J. & Tomasello, M. 1999 A nonverbal false belief task: the performance of children and great apes. *Child Dev.* **70**, 381–395. (doi:10.1111/1467-8624.00028)

Call, J., Hare, B. & Tomasello, M. 1998 Chimpanzee gaze following in an object-choice task. *Anim. Cogn.* **3**, 23–34. (doi:10.1007/s100710050047)

Carruthers, P. & Smith, P. K. (eds) 1996 *Theories of theory of mind*. New York, NY: Cambridge University Press.

Churchland, P. M. & Churchland, P. S. 1996 The future of psychology, folk and scientific. In *The churchlands and their critics* (ed. R. N. McCauley). Cambridge, UK: Blackwell.

Clayton, N. S. & Emery, N. J. 2005 Corvid cognition. *Curr. Biol.* **15**, R80–R81. (doi:10.1016/j.cub.2005.01.020)

Clayton, N. S., Griffiths, D. P., Emery, N. J. & Dickinson, A. 2001 Elements of episodic-like memory in animals. *Phil. Trans. R. Soc. B* **356**, 1483–1491. (doi:10.1098/rstb.2001.0947)

Clayton, N. S., Dally, J. M. & Emery, N. J. 2006 Social cognition by food-caching corvids The western scrub-jay as a natural psychologist. *Phil. Trans. R. Soc. B* **362**. (doi:10.1098/rstb.2006.1992)

Dally, J. M., Emery, N. J. & Clayton, N. S. 2006 Food-caching western scrub-jays keep track of who was watching when. *Science* **312**, 1662–1665. (doi:10.1126/science.1126539)

Davies, M. & Stone, T. (eds) 1995 *Folk psychology*. Oxford, UK: Blackwell Publishers.

Davies, M. & Stone, T. (eds) 1995 *Mental simulation*. Oxford, UK: Blackwell.

Dennett, D. C. 1987 *The intentional stance*. Cambridge, MA: MIT Press.

Dretske, F. I. 1988 *Explaining behavior*. Cambridge, MA: The MIT Press.

Emery, N. J. 2004 Are corvids 'feathered apes'? Cognitive evolution in crows, jays, rooks and jackdaws. In *Comparative analysis of minds* (ed. S. Watanabe). Tokyo, Japan: Keio University Press.

Emery, N. J. & Clayton, N. S. 2001 Effects of experience and social context on prospective caching strategies by scrub jays. *Nature* **414**, 443–446. (doi:10.1038/35106560)

Emery, N. J. & Clayton, N. S. 2004 The mentality of crows: convergent evolution of intelligence in corvids and apes. *Science* **306**, 1903–1907. (doi:10.1126/science.1098410)

Emery, N. J. & Clayton, N. S. 2005 Evolution of the avian brain and intelligence. *Curr. Biol.* **15**, R946–R950. (doi:10.1016/j.cub.2005.11.029)

Flombaum, J. I. & Santos, L. R. 2005 Rhesus monkeys attribute perceptions to others. *Curr. Biol.* **15**, 447–452. (doi:10.1016/j.cub.2004.12.076)

Gentner, D., Holyoak, K. J. & Kokinov, B. K. (eds) 2001 *The analogical mind: perspectives from cognitive science*. Cambridge, MA: MIT Press.

Godfrey-Smith, P. 2000 On folk psychology and mental representation. In *Representation in mind: new approaches to mental representation* (ed. H. Clapin, P. Staines & P. Slezak), pp. 147–162. Amsterdam, The Netherlands: Elsevier.

Goldman, A. 1993 The psychology of folk psychology. *Behav. Brain Sci.* **16**, 15–28.

Goodall, J. 1990 *Through a window*. Boston, MA: Houghton Mifflin.

Gordon, R. 1986 Folk psychology as simulation. *Mind Lang.* **1**, 158–171.

Gordon, R. 1996 'Radical' simulationism. In *Theories of theories of mind* (ed. P. Carruthers & P. K. Smith), pp. 11–21. Cambridge, UK: Cambridge University Press.

Halford, G. S., Wilson, W. H. & Phillips, S. 1998 Processing capacity defined by relational complexity: implications for comparative, developmental, and cognitive psychology. *Behav. Brain Sci.* **21**, 803–864. (doi:10.1017/S0140525X98001769)

Hare, B. 2001 Can competitive paradigms increase the validity of experiments on primate social cognition? *Anim. Cogn.* **4**, 269–280. (doi:10.1007/s100710100084)

Hare, B., Call, J., Agnetta, B. & Tomasello, M. 2000 Chimpanzees know what conspecifics do and do not see. *Anim. Behav.* **59**, 771–785. (doi:10.1006/anbe.1999.1377)

Hare, B., Call, J. & Tomasello, M. 2001 Do chimpanzees know what conspecifics know? *Anim. Behav.* **61**, 771–785. (doi:10.1006/anbe.2000.1518)

Heyes, C. M. 1998 Theory of mind in nonhuman primates. *Behav. Brain Sci.* **21**, 101–148. (doi:10.1017/S0140525X98000703)

Hostetter, A. B., Cantero, M. & HONKINS, W. D. 2001 Differential use of vocal and gestural communication by chimpanzees (*Pan troglodytes*) in response to the attentional status of a human (*Homo sapiens*). *J. Comp. Psychol.* **115**, 337–343.

Hunt, G. R. 1996 Manufacture and use of hook-tools by New Caledonian crows. *Nature* **379**, 249–251. (doi:10.1038/379249a0)

Hunt, G. R. 2004 The crafting of hook tools by wild New Caledonian crows. *Proc. R. Soc. B* **271**, S88–S90.

Hurley, S. 2006 Introduction. In *Rational animals?* (ed M. Nudds & S. Hurley). Oxford, UK: Oxford University Press.

Q11

Hurley, S. & Chater, N. (eds) 2005 *Perspectives on imitation: from neuroscience to social science*. Cambridge, MA: MIT Press.

Karin-D'Arcy, M. R. & Povinelli, D. J. 2002 Do chimpanzees know what each other see? A closer look. *Int. J. Comp. Psychol.* **15**, 21–54.

Leavens, D. A., Hostetter, A. B., Wesley, M. J. & Hopkins, W. D. 2004 Tactical use of unimodal and bimodal communication by chimpanzees, *Pan troglodytes*. *Anim. Behav.* **67**, 467–476. (doi:10.1016/j.anbehav.2003.04.007)

Markman, A. B. & Dietrich, E. 2000 In defense of representation. *Cogn. Psychol.* **40**, 138–171. (doi:10.1006/cogp.1999.0727)

Meltzoff, A. & Moore, M. K. 1997 Explaining facial imitation: a theoretical model. *Early Dev. Parenting* **6**, 179–192. (doi:10.1002/(SICI)1099-0917(199709/12)6:3/4<179::AID-EDP157>3.0.CO;2-R)

Meltzoff, A. In press. 'Like me': a foundation for social cognition. *Dev. Sci.*

Q12

Nichols, S. & Stich, S. P. 2003 *Mindreading: an integrated account of pretence, self-awareness and understanding other minds*. Oxford, UK: Oxford University Press.

Penn, D.& Povinelli, D. J. In press. Causal cognition in human and nonhuman animals: a comparative, critical review. *Annu. Rev. Psychol.* **58**.

Q13

Povinelli, D. J. 2003 *Folk physics for apes*. Oxford, UK: Oxford University Press.

Povinelli, D. J. & Eddy, T. J. 1996*a* Chimpanzees: joint visual attention. *Psychol. Sci.* **7**, 129–135. (doi:10.1111/j.1467-9280.1996.tb00345.x)

Povinelli, D. J. & Eddy, T. J. 1996*b* Factors influencing young chimpanzees' (*Pan troglodytes*) recognition of attention. *J. Comp. Psychol.* **110**, 336–345. (doi:10.1037/0735-7036.110.4.336)

Povinelli, D. J. & Eddy, T. J. 1996*c* What young chimpanzees know about seeing. *Monogr. Soc. Res. Child Dev.* **61**, i–vi. (doi:10.2307/1166159) 1–191.

Q16

Povinelli, D. J. & Giambrone, S. 1999 Inferring other minds: flaws in the argument by analogy. *Phil. Top.* **27**, 167–201.

Povinelli, D. J. & Vonk, J. 2003 Chimpanzee minds: suspiciously human? *Trends Cogn. Sci.* **7**, 157–160. (doi:10.1016/S1364-6613(03)00053-6)

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

Povinelli, D. J., Bering, J. M. & Giambrone, S. 2000 Toward a science of other minds: escaping the argument by analogy. *Cogn. Sci.* **24**, 509–541. (doi:10.1016/S0364-0213(00)00023-9)

Povinelli, D. J., Dunphy-Lelii, S., Reauxa, J. E. & Mazza, M. P. 2002 Psychological diversity in chimpanzees and humans: new longitudinal assessments of chimpanzees' understanding of attention. *Brain Behav. Evol.* **59**, 33–53. (doi:10.1159/000063732)

Premack, D. & Woodruff, G. 1978 Does the chimpanzee have a theory of mind? *Behav. Brain Sci.* **4**, 515–526.

Santos, L. R., Nissen, A. G. & Ferrugia, J. 2006 Rhesus monkeys, *Macaca mulatta*, know what others can and cannot hear. *Anim. Behav.* **71**, 1175–1181. (doi:10.1016/j.anbehav.2005.10.007)

Seed, A. M., Tebbich, S., Emery, N. J. & Clayton, N. S. 2006 Investigating physical cognition in rooks (*Corvus frugilegus*). *Curr. Biol.* **16**, 697–701. (doi:10.1016/j.cub.2006.02.066)

Silk, J. B. 2002 The form and function of reconciliation in primates. *Annu. Rev. Anthropol.* **31**, 21–44. (doi:10.1146/annurev.anthro.31.032902.101743)

Suddendorf, T. & Whiten, A. 2001 Mental evolution and development: evidence for secondary representation in children, great apes and other animals. *Psychol. Bull.* **127**, 629–650. (doi:10.1037/0033-2909.127.5.629)

Q14 Tebbich, S., Seed, A. M., Emery, N. J. & Clayton, N. S. In press. Non-tool-using rooks (*Corvus frugilegus*) solve the trap-tube task. *Anim. Cogn.*

Tomasello, M. & Call, J. 1997 *Primate cognition*. New York, NY: Oxford University Press.

Q15 Tomasello, M. & Call, J. 2006 Do chimpanzees know what others see- or only what they are looking at? In *Rational animals?* (ed S. Hurley & M. Nudds). Oxford, UK: Oxford University Press.

Tomasello, M., Call, J., Nagell, K., Olguin, R. & Carpenter, M. 1994 The learning and use of gestural signals by young chimpanzees: a trans-generational study. *Primates* **35**, 137–154. (doi:10.1007/BF02382050)

Tomasello, M., Hare, B. & Agnetta, B. 1999 Chimpanzees, pan troglodytes, follow gaze direction geometrically. *Anim. Behav.* **58**, 769–777. (doi:10.1006/anbe.1999.1192)

Tomasello, M., Call, J. & Hare, B. 2003*a* Chimpanzees understand psychological states—the question is which ones and to what extent. *Trends Cogn. Sci.* **7**, 153–156. (doi:10.1016/S1364-6613(03)00035-4)

Tomasello, M., Call, J. & Hare, B. 2003*b* Chimpanzees versus humans: it's not that simple. *Trends Cogn. Sci.* **7**, 239–240. (doi:10.1016/S1364-6613(03)00107-4)

Tomasello, M., Carpenter, M., Call, J., Behne, T. & Moll, H. 2005 Understanding and sharing intentions: the origins of cultural cognition. *Behav. Brain Sci.* **28**, 675–691. (doi:10.1017/S0140525X05000129)

Visalberghi, E. & Tomasello, M. 1998 Primate causal understanding in the physical and psychological domains. *Behav. Processes* **42**, 189–203. (doi:10.1016/S0376-6357(97)00076-4)

Vonk, J. *et al*. In press. Social and physical reasoning in human-reared chimpanzees.

Wellman, H. M., Cross, D. & Watsonl, J. 2001 Meta-analysis of theory-of-mind development: the truth about false belief. *Child Dev.* **72**, 655–684. (doi:10.1111/1467-8624.00304)

Q17 Whiten, A. 1996 When does behaviour-reading become mind-reading. In *Theories of theory of mind* (ed. P. Carruthers & P. K. Smith), pp. 277–292. New York, NY: Cambridge University Press.

RSTB 20062023—2/1/2007—16:54—THIAGU—266980—XML RSB – pp. 1–15

Whiten, A. 1997 The Machiavellian mindreader. In *Machiavellian intelligence II: extensions and evaluations* (ed. A. Whiten & R. W. Byrne), pp. 144–173. Cambridge, UK; New York, NY: Cambridge University Press.

Whiten, A. 2000 Chimpanzees and mental re-representation. In *Metarepresentations: a multidisciplinary perspective* (ed. D. Sperber), pp. 139–167. New York, NY: Oxford University Press.

Whiten, A. & Suddendorf, T. 2001 Meta-representation and secondary representation. *Trends Cogn. Sci.* **5**, 378. (doi:10.1016/S1364-6613(00)01734-4)

Zelazo, P. D., Jacques, S., Burack, J. & Frye, D. 2002 The relation between theory of mind and rule use: evidence from persons with autism-spectrum disorders. *Infant Child Dev.* **11**, 171–195. (doi:10.1002/icd.304)

RSTB 20062023—2/1/2007—16:54—THIAGU—266980—XML RSB – pp. 1–15

Author Queries

JOB NUMBER: 20062023

JOURNAL: RSTB

Q1 We have inserted a short title. Please approve or provide an alternative.

Q2 Additional keyword has been deleted as the journal style permits only a maximum of six keywords. Please approve.

Q3 We have inserted year for the reference Premack & Woodruff's (1978). Please check and approve.

Q4 We have inserted year for the reference Whiten & Suddendorf (2001). Please check and approve.

Q5 Please note that the symbols '<' and '>' have been changed to single quotes through the article.

Q6 Please check the edit of the sentence 'On each trial, the experimenter…'

Q7 Please check the edit of the sentence 'The initial steps…'

Q8 Please check the sense of the sentence 'We hope our proposed protocols…'

Q9 References Churchland & Churchland (1996) and Godfrey-Smith (2000) are provided in the list but not cited in the text. Please supply citation details or delete the reference from the reference list.

Q10 Please provide page range for the reference Churchland & Churchland (1996).

Q11 Please provide page range for the reference Hurley (2006).

Q12 Please update the year of publication for the reference Meltzoff (in press).

Q13 Please update the year of publication for the reference Penn & Povinelli (in press).

Q14 Please update the year of publication for the reference Tebbich et al. (in press).

Q15 Please provide page range for the reference Tomasello & Call (2006).

Q16 Please update the year of publication for the reference Vonk et al. (in press).

Q17 Please check the inserted page range for the reference Whiten (1996).

RSTB 20062023—2/1/2007—16:54—THIAGU—266980—XML RSB – pp. 1–15