Behavior-Reading versus Mentalizing in Animals

Logan Fletcher and Peter Carruthers

This chapter evaluates the debate between behavior-rule and mindreading accounts of the abilities of some nonhuman animals. (Although the evidence concerns canids and corvids in addition to primates, we focus on the latter.) We show that although the data are by no means conclusive, they presently favor a mindreading account, suggesting that simple forms of mentalizing are quite prevalent among highly social creatures outside of the hominin line.

1 Introduction

We will begin with some comments on the manner in which Penn and Povinelli (this volume) frame the debate about primate mindreading. Thereafter, in the sections that follow, we will consider some of their arguments, as well as the related arguments of Perner (2010). We will suggest that this debate should not be considered in isolation, but must be taken along with an evaluation of recent evidence of mindreading abilities in very young human infants.

Penn and Povinelli (this volume) protest that their critics are misguided to charge them with being behaviorists. They say they want to insist, on the contrary, that many animal species possess high-level forms of cognition, and are capable of non-associative forms of learning. But this is not what the debate is about. The question is not whether we, as theorists, should be behaviorists in our interpretation of the behavior of other primates. The question is rather whether the animals themselves are behaviorists. Do these animals, in their social interactions with others, understand those others in terms of some set of behavior-rules, as Penn and Povinelli maintain? (An example of such a rule might be: “A dominant will approach food to which it has had uninterrupted line of sight in the recent past”; Povinelli and Vonk, 2003.) Or do the animals comprehend those behaviors in terms of some set of underlying mental states (including desires, percepts, and knowledge), as many of Povinelli and colleagues’ critics claim?

So although Penn and Povinelli (this volume) embark on an extended demonstration that primates bring to bear an impressive set of cognitive and inferential resources in navigating their social worlds, this is really a red herring. While we (and most others in the field) fully accept the conclusion, this is not what is at issue. What is really at stake (to repeat) is whether or not the animals in question represent and reason about some of the mental states of other agents. Penn and Povinelli maintain that they do not, and propose a behavior-rule account instead, whereas many others claim that they do (Hare et al., 2000, 2001; Call et al., 2006; Hare et al., 2006; Melis et al., 2006; Buttelmann et al., 2007; Call and Tomasello, 2008; Kaminski et al., 2008).

Penn and Povinelli (this volume) also protest against the charge that their behavior-rule hypothesis is unfalsifiable and unparsimonious, seemingly believing that these criticisms are somehow linked to the allegation that they are behaviorists. But of course there is no such link, because there is no such allegation. And the behavior-rule hypothesis is, indeed, unfalsifiable in a quite straightforward way. (The issue of parsimony is more complex. We will return to it in Section 5.) For it is too under-specified to make determinate predictions, and hence there is no risk of it turning out to be wrong. Claiming only that primates employ *some* set of behavior rules provides us with no clues as to how the animals might be expected to react in particular circumstances, and suggests no potential lines of experimentation. Moreover, in respect of any new item of behavioral evidence, an explanation in terms of the animals' deployment of some or other behavior rule can always be constructed after the fact. And this is very much the way in which Povinelli and colleagues have employed the behavior-rule hypothesis. When some new item of evidence that is claimed to support a mentalizing interpretation of primate behavior is described, Povinelli and colleagues set out to show that there is a behavior rule that can accommodate the evidence equally well. Hence they are always playing "catch up", and are forced to postulate behavior rules *ad hoc* to accommodate the data.

Of course, the hypothesis that the animals make use of all and only members of some determinate set of (specified) behavior rules *does* make predictions and *is* liable to falsification. But no such determinate set has ever been put forward. And while specific behavior rules have been proposed (like the one about recent line of sight to food, noted above), it is plain that the previous proposals are by no means complete. For there are now a number of experiments demonstrating that additional behavior rules would need to be postulated to explain the animals' behavior, as we will see in Section 7.

The mentalizing hypothesis, in contrast, while admitting of various strengths (depending on the range of mental states that are thought to be understood by the animals in question), provides a clear framework for generating novel predictions. And this is just the way in which it has been employed by Povinelli’s opponents, in many cases issuing in positive results. In particular, there have been positive results generated by the claim that primates attribute desires, percepts, and knowledge or ignorance to other agents (Hare et al., 2000, 2001; Call et al., 2004; Hare et al., 2006; Melis et al., 2006; Buttelmann et al., 2007; Call and Tomasello, 2008; Kaminski et al., 2008). But all recent experimental tests of the claim that these animals attribute false beliefs to others have been negative (Call and Tomasello, 1999; Hare et al., 2001; O’Connell and Dunbar, 2003; Kaminski et al., 2008; Krachun et al., 2009). This has led to the hypothesis that the animals possess “Stage 1” mindreading abilities, of the sort that are thought to emerge in human infancy before the capacity to understand false beliefs (Kaminski et al., 2008; Krachun et al., 2009).

What Povinelli and colleagues have ranged against them, therefore, is a regular scientific research program of good standing, which generates determinate predictions capable of falsification when combined with auxiliary assumptions (e.g. concerning the animals’ other forms of knowledge). Moreover, it is a progressing research program, issuing in a stream of positive results and increasingly precise theories. The behavior-rule hypothesis, in contrast, is too indeterminate and *ad hoc* to qualify as a scientific research program at all. This isn’t to say that it can’t be true, of course, and we don’t really want to fight over the applicability of the phrase “scientific theory”. But it does mean that the behavior-rule idea hasn’t yet entered into serious scientific competition with the mentalizing hypothesis. That said, the remainder of this chapter will be devoted to comparing additional strengths and weaknesses of the two approaches.

2 General-purpose versus domain-specific learning mechanisms

Penn and Povinelli (this volume) argue that we should not expect to find mentalizing abilities of any sort in non-human primates, since these animals seem to lack the capacity to reason about similarities and analogies (or “relations between relations”; Penn et al., 2008). The rationale is that only by noticing the similarity between different items of behavior in various contexts can a creature learn that they are guided by a single underlying variable (a desire for food, say). And Penn et al. (2008) amass a considerable body of evidence to show that non-human primates are incapable of reasoning analogically. What this argument betrays, however, is Penn and Povinelli’s tacit empiricism.

The capacity to reason analogically is only relevant to the possession of mentalizing abilities if the latter need to be *learned*, and learned on the basis of hypothesis formation and testing, at that. While something like this has been proposed as the means by which human children acquire a “theory of mind” (Gopnik and Meltzoff, 1997), this is by no means endorsed by all developmental psychologists. On the contrary, there are significant numbers of “modularists”, who think that core mentalizing abilities result from the maturation of an innately channeled module, or at least from the operations of a domain-specific learning mechanism of some sort (Leslie, 1994; Baron-Cohen, 1995; Song and Baillargeon, 2008; note that the notion of “module” in play here can be quite weak; see Barrett and Kurzban, 2006; Carruthers, 2006). Indeed, the burgeoning evidence of very early mindreading abilities in infants (to be discussed in Section 6) suggests that a broadly modularist position is now pretty much mandatory, at least in the human case. Moreover, the comparative psychologists who attribute mentalizing abilities to animals certainly don’t think that those abilities result from general-purpose reasoning about relations between relations. Indeed, the entire tradition of thinking about the evolution of mindreading capacities, from “Machiavellian intelligence” (Byrne and Whiten, 1988, 1997) onward, has presumed that they are an innately channeled adaptation of some sort. So Penn and Povinelli’s argument is question-begging.

By appealing to the possibility of an innately channeled mindreading module we can also reply to another of Penn and Povinelli’s arguments (this volume). This is grounded in the well-known finding that primates make little or no use of their alleged mindreading capacities in cooperative contexts, but only in competitive ones. This finding is said to be problematic for the claim that they possess such capacities at all. But it only raises a difficulty for the mentalizing hypothesis on the assumption that mindreading would involve some sort of general-purpose theory embedded in a general-purpose mind. If that were the case, then it really would be puzzling that an animal might draw on a set of beliefs about the mind in the service of one sort of goal but not another. For surely, one might think, if these animals are rational agents, capable of surveying the entire set of their beliefs to figure out how to satisfy their desires, then they ought to realize that their beliefs about another agent’s mental states are relevant to successful begging behavior (for example). But if one holds even a weakly modular conception of the architecture of animal minds, then the puzzle disappears. For in that case certain types of goal might have proprietary links to certain informational modules, while ignoring the output from other such modules.

It is important to realize that, from a modularist perspective, this proposal is by no means arbitrary. For non-human primates are not naturally cooperative in their normal lives in the wild, except in highly restricted contexts (such as cooperative hunting and border patrolling by male chimpanzees). They are, however, intensely competitive, with continual jostling and conflict over access to food, mates, and other resources. It makes good sense, then, that an evolved mindreading system might be linked specifically to the goals that are operative in such contexts.

Is it really plausible to claim that animal minds are so different from our own, however? Can it really be the case that the minds of other primates have a modular architecture while their close relatives (humans) are general-purpose reasoners? Indeed, this would not be plausible. But the claim that humans have a general-purpose cognitive architecture is itself highly controversial. Indeed, the best account of the differences among primate species is that all share essentially the same architecture of modules arranged as consumers for the “global broadcast” of attended perceptual information, but that humans possess a much more highly developed ability to maintain, rehearse, and manipulate globally broadcast representations (including so-called “inner speech”) in working memory (Shanahan and Baars, 2005; Carruthers, 2006, 2011; Shanahan, 2010). This enables us to utilize reflective “System 2” forms of reasoning, thus approximating general-purpose reasoners in some contexts and for some purposes, despite an underlying modular architecture (Carruthers, 2009).

3 Stage 1 versus Stage 2 mindreading

We now turn to a finding that is problematic for the behavior-rule approach. This is that although primates act in such a way as to suggest that they understand the desires, perceptions, and knowledge or ignorance of other agents, they fail otherwise parallel tests of false belief understanding. The mentalizing account can give a principled explanation of this divergence, drawing on the distinction between Stage 1 and Stage 2 mindreading adopted by most developmental psychologists. Indeed, there is remarkable agreement among developmentalists that desire / perception / knowledge–ignorance psychology is earlier to emerge in childhood than false-belief psychology, no matter whether the theorists in question are general-learning theorists (Wellman, 1990; Gopnik and Meltzoff, 1997) or modularists (Leslie, 1994; Baron-Cohen, 1995; Song and Baillargeon, 2008). It makes good sense then (especially from a modularist perspective) that nonhuman primates might have only Stage 1 of this two-stage structure.

These divergent findings are much more difficult for a behavior-rule theorist to explain. For as Penn and Povinelli (this volume) themselves point out, there would have been no difficulty finding a behavior-rule to explain the animals' behavior if it had turned out that they could pass a standard false-belief task. The rule could have been, for example, "A potential competitor will approach the previous location of food if the competitor was absent when the food was moved to its present location." This rule doesn't seem any more complex, nor any more difficult to learn or evolve, than the corresponding behavior-rule for an ignorance task, namely, "A potential competitor won't compete for food that isn't and wasn't recently in its line of sight." Why is it, then, that the animals seemingly know the latter rule but not the former?

Penn and Povinelli might try to argue that the behavior-rule needed in conditions of false belief is actually more complex than the ignorance-rule. This is because it requires the animal to keep in mind two distinct locations—the place where the competitor saw the food, and the place where the food is now. The ignorance-rule, in contrast, mentions only a single location (the current position of the food). There is a simpler version of the false-belief behavior rule, however, which doesn't suffer from this problem. For the animals could have utilized the rule, "A competitor that has, or has recently had, line of sight to the location of some food will go to that location if it can." This can predict that the competitor will approach the previous location in a case where the food has been moved during the competitor's absence, without the animals needing to represent where the food is now in order to make the prediction.

Penn and Povinelli might also try to argue that behavior-rules for dealing with false-belief situations are absent from the primate rule-kit because they are significantly less *useful* in competitive situations, not because they are significantly more difficult than their ignorance-task counterparts. To a primate competing for food, it is highly relevant whether or not one's conspecifics are as knowledgeable as oneself regarding the food's location. So upon seeing it being moved from one place to another when one's competitors are absent, it pays off to register that they are now *ignorant* of its present location. (Or rather, in the context of the present debate, it is important to know a rule that can predict that they aren't likely to approach the food at its present location.) This is because one can then know that it is safe to approach it oneself. But beyond that, it is unclear why it should be useful to know where in particular one's competitors will now (falsely) *think* the food is hidden, or where they will now go to given their previous history of line-of-sight to food, so long as this is distinct from the food's actual location. Or at least, so Penn and Povinelli might try to argue.

These considerations are insufficient by themselves to explain the absence in other primates of behavior-rules covering false-belief situations. For it frequently *does* matter quite a lot to a primate where a competitor is likely to be when the primate is retrieving some food (or doing anything else over which there might be competition, such as mating). For if the competitor will be in a place that has line of sight to the competed-over resource, then that is likely to lead to conflict and loss of an opportunity to eat (or mate). So a behavior-rule that tells one where a competitor is likely to go, given its previous perceptual access to the location of some food, would surely be advantageous. It might perhaps be claimed that such situations are less frequent in natural circumstances than are situations in which an ignorance-based behavior-rule would apply. But this remains to be demonstrated. And even if demonstrated, it would still need to be shown that the benefits of employing a behavior-rule covering false-belief situations would be minimal enough for us to predict the absence of such rules among primates. So it remains as a challenge for Penn and Povinelli to explain why ignorance-based behavior rules should be used by nonhuman primates whereas false-belief-based rules are not.

Of course, given that the set of behavior-rules is heterogeneous and potentially infinite in extent, it is perhaps not surprising that animals should happen to have evolved some but not other equally-easy-to-evolve rules from this set. Nor should it be surprising that there might be quirks of their evolved learning systems that enable them to acquire some but not others of the set, although the latter look to us humans to be equally easy to learn. But these hypotheses are unprincipled, and appeal to accident or coincidence. They therefore fail to provide us with an *explanation* of the phenomenon. In contrast, the competing mentalizing explanation can make good sense of these findings, as we noted above. This gives us reasons to prefer the latter. For theories that can explain the data are generally better than theories that can’t.

4 Explanatory virtues

Penn and Povinelli (this volume) complain that the mentalizing account of primate behavior has failed to demonstrate that mindreading is *necessary* to produce some of the behavior we observe, or that the same behavior *could not* have been produced by a set of behavior rules. But this is too strong a demand to place on any theory. No theory, in any domain of science, can ever show that the data *cannot* be explained in any other way. Despite popular-science mythology to the contrary, there are no such things as decisive experiments in science. Any set of results can always be accommodated by means of suitable theoretical adjustments, or by altering some of the auxiliary assumptions that are always needed for a theory to make determinate predictions. As has long been known by philosophers of science (Kuhn, 1962; Lakatos, 1970; Newton-Smith, 1981), and is understood tacitly by most working scientists, theory choice is never a matter of proof, but of judgment—incorporating such factors as simplicity, predictive accuracy, explanatory scope, coherence with surrounding theories, and scientific fruitfulness.

Consider how the two approaches stack up along these dimensions. The question of simplicity will be discussed in Section 5. But we have already seen that the mentalizing account is significantly stronger in terms of predictive accuracy, especially since the behavior-rule account is only capable of "predicting" new findings after they are discovered, postulating a novel behavior-rule for the purpose. The mentalizing theory also does better in terms of explanatory scope. For while both accounts are equally capable of explaining why mindreading-like behavior is only found in competitive contexts, the mentalizing account can explain why the animals should pass tests of Stage 1 mindreading while failing tests of false-belief understanding, whereas the behavior-rule account seemingly cannot. Likewise, the mentalizing theory is much more scientifically fruitful, issuing in novel tests and positive results, while the behavior-rule approach is entirely defensive, explaining away positive results as they are discovered. In addition, the mentalizing account coheres better with surrounding theories in cognitive science, especially the existence of both Stage 1 and Stage 2 mindreading in human infants. (This point will be elaborated in Sections 5 and 6.)

It appears, then, that when compared along most of the normal dimensions governing theory choice in science, the mentalizing hypothesis is significantly preferable to the behavior-rule one. But the comparative simplicity of the two theories has yet to be considered, and the claimed greater coherence of the mindreading hypothesis with surrounding theories has yet to be established. To these tasks we now turn.

5 Comparative simplicity

This section takes up the vexed question of the comparative simplicity of the two types of hypothesis. On the face of it, simplicity favors a mentalizing account. For the behavior-rule theory is forced to postulate a multitude of distinct rules, whereas the mentalizing theory postulates a single mindreading faculty. Perner (2010), however, argues that the alleged simplicity of the mindreading hypothesis is illusory, and depends upon treating the hypothesized system as a "black box". In fact, it must have significant internal complexity, including rules for inferring goals from behavioral cues, as well as rules for judging perceptual access and for inferring knowledge or ignorance. Indeed, in cases where just a single goal (like getting food) is in play, Perner claims that for every rule postulated by a behavior-rule theorist (such as, "A competitor will move to secure food at a location that was recently in its line of sight"), a mentalizing theorist will need to postulate two (in this case, "A competitor that has seen the location of food knows the location of food" and, "A competitor that knows the location of food will move to secure it"). So the behavior-rule hypothesis is the simpler of the two.

These claims hold good, however, only if the number of behavior-rules at stake is quite small. For the advantage of mindreading is that distinct cues can indicate the presence of the same goal (such as moving toward it as well as begging for it) and distinct cues can indicate the presence of knowledge of something (including both seeing and hearing). Moreover, different goals can interact with different items of knowledge to issue in novel forms of pairing between the initial cues and subsequent behavior. On the mindreading account, one needs to posit as many rules as the total number of different indicator-cues of knowledge and indicator-cues of goals, combined additively. On the behavior-rule account, in contrast, these numbers combine multiplicatively to give the total number of rules required, since every possible combination of knowledge plus goal that issues in behavior requires its own rule. Moreover, on a mindreading account much of the required complexity can be "farmed out" to other faculties of the mind. For almost all theorists in the field accept that mindreading operates in part by *simulation* of the minds of others, enabling the mindreading system to rely on beliefs that are produced by other mental faculties when generating predictions of a target agent's behavior. So when the number of goals and items of knowledge tracked by an animal become sufficiently large (either explicitly, using mental state attributions, or implicitly, using a set of behavior rules), then a mindreading account of the animal's behavior will utilize far fewer rules.

It has to be admitted that in this respect the data do not yet provide much direct support for a mentalizing account, however. For all of the mindreading tasks employed to date have involved competition over food. Since all involve the same presumed goal, this means that the number of behavior-rules that are needed to explain the data are also quite limited. Indeed, as we discuss in Section 7, all of the data to date can be explained using just a handful of behavior-rules, while a mentalizing account needs to postulate roughly the same number of rules governing the assignment of mental states. So it might seem that the data at this point provide support for neither side (along the simplicity-dimension of evaluation, at any rate).

It is possible to push back against Perner's (2010) "black box" argument, however. For this presupposes a narrow theoretical focus, aimed only at explaining the behavior of nonhuman primates. But what matters for science is not the relative simplicity of narrowly-framed theories, considered discretely. Rather, it is the total set of mechanisms and processes that our theories require us to accept. So if a local theory can import some complex structure from another related domain that we already have reason to believe in, then the presence of that complexity does *not* render the resulting theory equally complex. This is because the structure in question already formed part of our theoretical ontology. Hence importing it into the new theory can be considered just a single addition to the complexity of the latter, rather than many. In fact there are numerous circumstances in which "black boxing" a complex structure for purposes of judging comparative simplicity is perfectly legitimate. And we believe that this is so here.

We will argue in Section 6 that the data from human infants warrant us in claiming that both Stage 1 and Stage 2 mindreading are present and operating at very early ages (before the end of the first year of life for the former, and by the age of eighteen months for the latter). If so, then it seems almost certain that both are heavily innately channeled in their development. This means, then, that we should be committed to the existence of an evolved and weakly modular Stage 1 mindreading system. The cost of importing such a system into the explanation of nonhuman primate behavior then merely entails supposing that it evolved earlier than one might otherwise have thought (in the lineage of the last common ancestor of all of the primate species in question, instead of only among hominins).

Note that, given the construal of the infancy data to be defended in Section 6, a behavior-rule theorist must be committed to three distinct kinds of structure underlying primate social competence. There is an innately channeled set of behavior-rules that operate among nonhuman primates, on the one hand, while there is a mindreading faculty in humans that consists of both Stage 1 and Stage 2 components. In contrast, the mentalizing theorist is only committed to the existence of two of these structures—a Stage 1 mindreading system that is possessed by a number of primate species in addition to humans, and a Stage 2 system that is unique to humans.

This means that the mentalizing hypothesis comes out simpler overall, despite the extra complexity that is postulated to exist in the minds of nonhuman primates through their possession of a complexly-structured Stage 1 mindreading system.

Penn and Povinelli might attempt to reply by claiming that Stage 1 mindreading isn't needed. Perhaps the same set of behavior rules that is (they claim) operative in nonhuman primates is also at work early in infancy, prior to the capacity to pass nonverbal false-belief tasks. But there are two reasons why this won't work. One is that, although in Section 6 we will focus mostly on nonverbal evidence of false-belief understanding, many of the same sorts of points could be made in respect of the evidence of Stage 1 mindreading in infants. And secondly, the false-belief data make no sense in the absence of a capacity to attribute goals, perceptions, and knowledge to other agents, since the experiments are all designed around just such an assumption, and since no one has any idea how a set of behavior-rules could fit together with belief understanding to issue in the patterns of behavior we observe.

Indeed, the mentalizing account of nonhuman primate capacities provides a more coherent evolutionary theory of primate social behavior overall. For everyone agrees that the evolutionary pressure toward mindreading derives from the exigencies of life in complex social groups. And while the sociality of humans is no doubt extreme, many other primates also live in such groups. One might predict, then, that simpler forms of mindreading would be found in such creatures. The behavior-rule hypothesis, in contrast, must postulate that the entire evolutionary history of mindreading took place in the hominin line. While this is possible, of course, it does then place the onus on behavior-rule theorists to specify what it was about early forms of hominin social life that resulted initially in the emergence of Stage 1 mindreading, and why it is that social living among nonhuman primates should have resulted in a set of behavior-rules instead. Indeed, it is especially puzzling how Stage 1 mindreading abilities could have evolved among hominins, on a behavior-rule account. For there would already have been behavior-rules in place sufficient to underwrite something functionally equivalent to competence at passing Stage 1 mindreading tasks. So what would a genuinely mentalistic Stage 1 mindreading system have evolved *for*?

We conclude, therefore, that on the simplicity-dimension of theoretical evaluation, just as with the other factors considered in Section 4, the mentalizing hypothesis comes out ahead of the behavior-rule one. But this argument has been premised on an assumption of innately channeled mindreading capacities in human infants. This is where we go next (in Section 6), discussing the nature and strength of the existing data. We will then (in Section 7) make some comparisons with the data from nonhuman primates. This will enable us to provide a few suggestions for future experimental work with primates that would make the case for a mentalizing account even more powerful.

6 Mindreading data from infants

Beginning with Woodward (1998), there are now numerous non-verbal looking-time studies demonstrating that human infants can attribute goals and intentions to others within the first year of life (Johnson, 2000; Csibra et al., 2003; Luo and Baillargeon, 2005; Csibra, 2008). There are also a number of experiments showing that infants are sensitive to the difference between knowledge and ignorance in others (Liszkowski et al., 2006, 2007; Luo and Baillargeon, 2007; Luo and Johnson, 2009). Moreover, since the ground-breaking work of Onishi and Baillargeon (2005) there has been a rapidly expanding body of data suggesting that human infants can understand false beliefs, and can make behavioral predictions accordingly, by around the age of eighteen months (Southgate et al., 2007, 2010; Surian et al., 2007; Song et al., 2008; Buttelmann et al., 2009b; Scott and Baillareon, 2009; Scott et al., 2011). These studies have come out of a number of different labs, using a variety of distinct measures—including not only expectancy-violation looking time, but also anticipatory looking time, as well as behavior intended to help another person or comply with a request.

As with the primate data, alternative behavior-rule explanations have been proposed (Perner and Ruffman, 2005; Perner, 2010). In particular, it has been suggested that the data collected by Onishi and Baillargeon (2005), as well as at least some of the data collected since, might be explained by assuming that infants know and apply the rule, “People will search for a desired object where they last saw it” or the rule, “Ignorance leads to error”. Notice, to begin with, that these are not pure behavioral rules, but rather presuppose the existence in infants of Stage 1 mindreading. While making a rule-based account somewhat more plausible, this isn’t absolutely essential. The search-rule might instead be formulated as a pure behavior-rule by substituting in place of an ascription of desire whatever behavioral cues would enable Stage 1 mindreaders to ascribe a desire (and likewise for the ignorance-rule). Moreover, the “ignorance leads to error” rule has been directly tested and found not to be operative (Southgate et al., 2007;

Scott and Baillargeon, 2009; Baillargeon et al., forthcoming).

The data are now so voluminous and varied, however, that behavior-rule explanations have become unsustainable. This claim is defended in some detail in Carruthers (2011). Here we will just provide a sketch of some of the main points. One factor is that a variety of goals to be attributed to the target agent have been employed across experiments, in addition to that of finding a desired object. These include the goal of *pretending* that something is the case (Onishi et al., 2007), the goal of *referring* to one thing rather than another (Southgate et al., 2010), and the goal of making a rattling noise happen (Scott et al., 2011). With four distinct goals in play, then, as well as a number of different kinds of belief (each of which is caused in a distinct way), it is easy to see that the total number of behavior-rules that would be needed to accommodate the data are multiplying rapidly (although admittedly, not all of the studies needed to fill in each of the cells in this potential matrix have yet been completed). It is already more parsimonious to assume that infants are capable of mindreading (utilizing a set of principles for ascribing beliefs and desires) than to claim that they deploy a large and varied set of behavior-rules.

Moreover, some of the behavior-rules that would be required to explain specific pieces of experimental data are quite strange and elaborate. There seems very little chance that they could be innate, and it is exceedingly hard to see how the infants would have had sufficient opportunity to learn them. For example, the behavior-rule needed to explain the data from Scott and Baillargeon (2009) would be this: “People who want to obtain the divisible one of two otherwise-similar objects will reach for the location of the unseen member of the pair when the other of the two is visible in its joined state, provided that the construction of that object out of its parts didn’t take place within the person’s line of sight.” And the behavior-rule needed to explain the data from Baillargeon et al. (forthcoming) would be this: “People who desire a rigid-seeming object, who have not observed that the object can be collapsed to make it small, will search for the object in a large container rather than a small one when presented with the two options.”

In addition, the methodology employed in devising these studies is particularly revealing. By assuming that infants are capable of ascribing simple goals, perceptions, and beliefs, the experimenters make predictions about the behavior that we should then observe (while also including various controls to rule out other possibilities). The result has been an extensive set of positive results, which a behavior-rule theorist would have had no reason to predict. Applying normal scientific standards, then, these results provide powerful confirmation of the initial theoretical assumptions.

Even more impressive, a number of studies have capitalized on beliefs about the world that we have independent reason to think that infants possess. Experiments have then been designed by assuming that infants will attribute these same beliefs to the target agents in the experiments (using the same "default attribution" heuristic that adults, too, employ; see Nichols and Stich, 2003). Baillargeon et al. (forthcoming), for example, make use of an earlier finding that infants understand that a large object cannot be contained within a small one (Hespos and Baillargeon, 2006). They use this to predict that the infants should be surprised when the agent who hasn't seen the collapsible nature of a desired toy reaches for the small container (even though that is where the toy really is). Likewise Scott et al. (2011) capitalize on the fact that 18-month-old infants as well as older children expect objects that are more similar in their surface properties to resemble one another in non-obvious properties as well. This enables them to set up a scenario involving three cups, two of which are similar in appearance but the third of which is quite different. They then predict that infants should be surprised when the agent, who hasn't seen that it is the dissimilar cup that also makes a rattling sound, reaches for that cup when another agent exhibits one of the two similar cups, rattles it, and asks (while continuing to hold onto the demonstrated cup), "Can you make it happen too?" Again, normal scientific standards should lead us to see the positive results obtained in these experiments as powerful confirmation of the mindreading hypothesis.

In fact one of the strengths of the mindreading account is that it is, in a certain sense, generalizable. For it can be combined with other theories of the belief-forming competence of human infants to generate indefinitely many novel hypotheses about what infants should or should not expect a target agent to do. The result has been a rapidly expanding set of confirmed predictions, where the predictions in question would never have been made in advance by a behavior-rule theorist. There should be no dispute that this provides powerful confirmation of the mindreading hypothesis.

Our view, then, is that it is now reasonable to believe that human infants possess innately channeled mechanisms underlying both Stage 1 and Stage 2 mindreading capacities. (In addition, Carruthers, 2011, argues that these mechanisms are significantly modular in character.) Given that this is so, such a result should now have a powerful impact on the debate over Stage 1 mindreading in nonhuman primates, as we argued in Sections 4 and 5. In particular, it means that the hypothesis of Stage 1 mindreading in nonhuman primates both coheres better with surrounding scientific theories than does a behavior-rule account, and is also genuinely simpler than the latter (enabling us to "black box" the internal complexity of a Stage 1 mindreading mechanism for purposes of judging comparative simplicity).

7 Mindreading data from animals

This section briefly reviews the data on Stage 1 mindreading in nonhuman primates, comparing them with the infancy data, and using this to motivate some tentative suggestions for future research. To this end we will discuss the data on their own merits, using a narrow theoretical focus of the sort adopted by Perner (2010), but rejected by us in Section 5.

From our review of the existing data we believe that as few as nine separate behavior-rules might be sufficient to explain them (setting aside the point that since the behavior-rule hypothesis wouldn't have predicted the data, it can't really explain those data either—except *ad hoc*, and after the fact). The rules are as follows.

[1] "A competitor will move to secure food that is in its line of sight." (See Hare et al., 2000.)

[2] "A competitor will move to secure food that was recently in its line of sight." (See Hare et al., 2001. So far as we can see, the same rule can also be used to explain the data from Call et al., 2004. We also think it can explain the results obtained by Kaminski et al., 2008. Although the authors of the latter say that their experiment rules out a behavior-rule interpretation—see p.227—we don't see how it does. All it excludes is an "evil eye" form of behavior rule.)

[3] "A competitor with food within reach will move to secure the food if another agent approaches within line of sight." (See Hare et al., 2006. It is possible, however, that this rule may need to be made more complex, conditionalizing on circumstances in which the competitor cannot issue an effective threat to protect the food, because of the presence of plexiglas. A probabilistic application of this rule might also be sufficient to explain the data from Melis et al., 2006. In a condition where the competitor was not visible within the booth, and so where line of sight was not known, the apes preferred to reach for the food through an opaque tube rather than a transparent one.)

> [4] “A competitor with food within reach will move to secure the food if another agent approaches noisily.” (See another of the conditions in Melis et al., 2006. Alternatively, one could use [3] together with: “A competitor with food within reach will orient toward a nearby noise”.)
>
> [5] “A compliant agent with food is more likely to provide some if one begs within their line of sight.” (See Liebal et al., 2004. Note that knowledge of a behavior-rule for distinguishing between competitive and compliant agents will also need to be postulated.)
>
> [6]: “A human who makes an accidental-dropping when reaching food to one is more likely to provide food later than a human who makes an intentional-dropping.” (See Call et al., 2004. The hyphenated phrases are intended to be stand-ins for non-mentalistic behavior descriptions.) We will return to raise some doubts about the adequacy of this rule shortly.
>
> [7] “A competitor who makes a pleased-face to one item of food and a disgust-face to another is more likely to eat the former.” (See Buttelmann et al., 2009a. Again, the hyphenated phrases are intended to be stand-ins for purely behavioral descriptions.)
>
> [8] “When humans use an unusual body part to bring about an effect in circumstances where they could have used their hands, then the use of that body part is likely to be necessary to achieve the effect; but when humans use an unusual body part to bring about an effect where their hands are occupied, then the use of that body part is unlikely to be necessary to achieve the effect.” (See Buttelmann et al., 2007.) We will shortly return to comment on this rule, too.
>
> [9] “A competitor who has line of sight to two objects, one of which is in a physical arrangement that suggests the presence of food while the other does not, will select the former.” (See Schmelz et al., 2011.) The plausibility of this rule will shortly be discussed alongside rule [8].

Overall this is not an especially large number of behavior-rules. For the mindreading approach will need to postulate in place of [1] that line of sight leads to seeing, and in place of [2] that seeing leads to knowing, as well as that competitors generally want to secure available food. These should then be sufficient when combined to replace [3], although to replace [4] we will need to add the rule that hearing leads to knowing. They are also sufficient to replace [5] when combined with a rule for identifying a compliant agent. In place of [6] the mindreading hypothesis will need rules for recognizing the presence of a desire to help (in cases of accidental dropping), as well as a desire to tease or annoy (in cases of intentional dropping). Instead of [7] there will need to be rules for recognizing from facial expressions desires to eat or not eat a given food item. And to replace [8] one will need a rule to the effect that when someone performs a movement that has an interesting effect, they probably intended that effect. In contrast, one may not need any additional rule to replace [9], if the behavior can be explained by *simulation* of the reasoning of the competitor. This seemingly adds up to slightly more mindreading rules (ten). But given that behavior-rules [5] and [6] are arguably each a conjunction of two distinct rules, and that a behavior-rule theorist also needs to appeal to a rule for identifying compliant or helpful agents, the total count for behavior-rules should probably be slightly larger (twelve).

It seems to us that three of the behavior-rules [1] through [9] are intrinsically problematic, however. One is [6], which is inadequate to explain the totality of the data obtained in the experiment. For while one finding was that the apes left the testing area more quickly in the intentional-dropping condition than in the accidental-dropping condition (suggesting that they might be reasoning in accord with [6]), another finding was that they also exhibited more attention-grabbing and coercive forms of behavior in the intentional-dropping condition. This is the opposite of what would be predicted by rule [6]. For if there is no point waiting around, one might think that there would be no point begging or banging on the cage either. On the contrary, the data are suggestive of anger or irritation. And this only seems to make sense if the animals have recognized that the human in the intentional-dropping condition is teasing them.

The other problematic behavior-rules seem to us to be [8] and [9]. This isn't because they can't handle the data, but rather because it seems quite unlikely that the animals would have had sufficient opportunity to learn such rules. (We assume that it is extremely unlikely that these rules would be innate.) For how often would they have observed a human using an unusual body part to bring about some physical effect while their hands are unoccupied, or a competitor choosing between two covers, one of which is partially supported by an item of hidden food and the other of which is not? And notice that although they might sometimes have seen a human whose hands are occupied with a load nudge open a door with his elbow or backside, for example, this isn't sufficient basis for learning rule [8]. In contrast, a Stage 1 mindreader can make the appropriate inferences, provided that the mindreading system can co-opt the resources of the mindreader's own planning abilities when generating predictions about the likely behavior of other agents (Nichols and Stich, 2003).

So although the number of rules that need to be postulated by behavior-rule and mindreading accounts are presently more or less equivalent, there is some reason to think that the behavior-rule account has significant problems in explaining the totality of the existing data—even after the fact. But we think that the evidential base for attributing Stage 1 mindreading to nonhuman primates would be strengthened if future experiments were to mimic some of the impressive features of the recent infancy data reviewed in Section 6. In particular, it would be helpful if experiments could be devised that would test for understanding of other goals in addition to eating. Of course, we understand that there may be severe practical limitations in devising experiments that would test for primates' understanding of the desire to mate, for example. But it should be easy enough to set up conditions where primates would compete for grooming opportunities, say, rather than for food. This would at least be a start in the right direction.

Moreover, it would be helpful if additional experiments could be devised that would capitalize on forms of physical understanding that we already have reason to think that the animals possess, in the way that the experiments conducted by Schmelz et al. (2011) do. For some of the most impressive data with human infants capitalize on their belief that a large object cannot fit into a small container, or that objects that are similar on the surface are likely to share other properties as well. In addition, more experiments need to be devised that can explicitly exclude rule [2] from the mix.

8 Conclusion

While the data are by no means probative, we believe that there is currently a strong case for saying that some nonhuman animals are capable of at least Stage 1 mindreading. This is partly because the mindreading hypothesis predicted the existing positive data, and can therefore genuinely explain it; whereas the behavior-rule hypothesis is entirely reactive, attempting to explain the data on a piecemeal basis after the fact. But it is also because the behavior-rule account has no explanation of the failures of these animals to exhibit competence with Stage 2 mindreading tasks, whereas the mindreading account can provide a principled explanation of the finding. Moreover, although considered narrowly both the behavior-rule and mindreading accounts are about equivalent in terms of their complexity, the latter provides a simpler and more coherent account of the overall evolution of mindreading capacities in the primate line. None of this is to say, however, that no further experiments need to be conducted. On the contrary, we have tentatively indicated some directions in which future inquiries might go.

Acknowledgments

We are grateful to Josep Call and Josef Perner for comments on an earlier draft of this chapter. This research was supported by NSF award number 0924523.

References

Baillargeon, R., He, Z., Setoh, P., Scott, R., and Yang, D. (forthcoming). The development of false-belief understanding and why it matters. In M. Banaji and S. Gelman (eds.), *The Development of Social Cognition*, Erlbaum.

Baron-Cohen, S. (1995). *Mindblindness*. MIT Press.

Barrett, H. and Kurzban, R. (2006). Modularity in cognition. *Psychological Review*, 113, 628-647.

Buttelmann, D., Carpenter, M., Call, J., and Tomasello, M. (2007). Enculturated chimpanzees imitate rationally. *Developmental Science*, 10, F31-38.

Buttelmann, D., Call, J., and Tomasello, M. (2009a). Do great apes use emotional expressions to infer desires? *Developmental Science*, 12, 688-698.

Buttelmann, D., Carpenter, M., and Tomasello, M. (2009b). Eighteen-month-old infants show false belief understanding in an active helping paradigm. *Cognition*, 112, 337-342.

Byrne, R. and Whiten, A., eds. (1988). *Machiavellian Intelligence*. Oxford University Press.

Byrne, R. and Whiten, A., eds. (1997). *Machiavellian Intelligence II*. Cambridge University Press.

Call, J. and Tomasello, M. (1999). A nonverbal false belief task: the performance of children and great apes. *Child Development*, 70, 381-395.

Call, J. and Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Sciences*, 12, 187-192.

Call, J., Hare, B., Carpenter, M., and Tomasello, M. (2004). “Unwilling” versus “unable”: Chimpanzees’ understanding of human intentional action. *Developmental Science*, 7, 488-489.

Carruthers, P. (2006). *The Architecture of the Mind*. Oxford University Press.

Carruthers, P. (2009). An architecture for dual reasoning. In J. Evans and K. Frankish (eds.), *In Two Minds*, Oxford University Press.

Carruthers, P. (2011). *The Opacity of Mind: An Integrative Theory of Self-Knowledge*. Oxford University Press.

Csibra, G. (2008). Goal attribution to inanimate agents by 6.5-month-old infants. *Cognition*, 107, 705-717.

Csibra, G., Bíró, S., Koós, O., and Gergely, G. (2003). One-year-old infants use teleological representations of actions productively. *Cognitive Science*, 27, 111-133.

Gopnik, A. and Meltzoff, A. (1997). *Words, Thoughts, and Theories*. MIT Press.

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.

Hare, B., Call, J., and Tomasello, M. (2006). Chimpanzees deceive a human competitor by hiding. *Cognition*, 101, 495-514.

Hespos, S. and Baillargeon, R. (2006). Décalage in infants' knowledge about occlusion and containment events: Converging evidence from action tasks. *Cognition*, 99, B31-B41.

Johnson, S. (2000). The recognition of mentalistic agency in infancy. *Trends in Cognitive Sciences*, 4, 22-28.

Kaminski, J., Call, J., and Tomasello, M. (2008). Chimpanzees know what others know, but not what they believe. *Cognition,* 109, 224-234.

Krachun, C., Carpenter, M., Call, J., and Tomasello, M. (2009). A competitive nonverbal false belief task for children and apes. *Developmental Science*, 12, 521-535.

Kuhn, T. (1962). *The Structure of Scientific Revolutions*. University of Chicago Press.

Lakatos, I. (1970). The methodology of scientific research programs. In I. Lakatos and A. Musgrave (eds.), *Criticism and the Growth of Knowledge*, Cambridge University Press.

Leslie, A. (1994). ToMM, ToBy, and Agency: Core architecture and domain specificity. In L. Hirchfeld and S. Gelman (eds.), *Mapping the Mind*, Cambridge University Press.

Liebal, K., Pika, S., Call, J., and Tomasello, M. (2004). To move or not to move: How apes adjust to the attentional state of others. *Interaction Studies*, 5, 199-219.

Liszkowski, U., Carpenter, M., Striano, T., and Tomasello, M. (2006). 12- and 18-month-olds point to provide information for others. *Journal of Cognition and Development*, 7, 173-187.

Liszkowski, U., Carpenter, M., and Tomasello, M. (2007). Pointing out new news, old news, and absent referents at 12 months of age. *Developmental Science*, 10, F1-F7.

Luo, Y. and Baillargeon, R. (2005). Can a self-propelled box have a goal? Psychological reasoning in 5-month-old infants. *Psychological Science*, 16, 601-608.

Luo, Y. and Baillargeon, R. (2007). Do 12.5-month-old infants consider what objects others can see when interpreting their actions? *Cognition*, 105, 489-512.

Luo, Y. and Johnson, S. (2009). Recognizing the role of perception in action at 6 months. *Developmental Science*, 12, 142-149.

Melis, A., Call, J., and Tomasello, M. (2006). Chimpanzees (*Pan troglodytes*) conceal visual and auditory information from others. *Journal of Comparative Psychology*, 120, 154-162.

Newton-Smith, W. (1981). *The Rationality of Science*. Routledge.

Nichols, S. and Stich, S. (2003). *Mindreading*. Oxford University Press.

O'Connell, S. and Dunbar, R. (2003). A test for comprehension of false belief in chimpanzees. *Evolution and Cognition*, 9, 131-140.

Onishi, K. and Baillargeon, R. (2005). Do 15-month-olds understand false beliefs? *Science*, 308, 255-258.

Onishi, K., Baillargeon, R., and Leslie, A. (2007). 15-month-old infants detect violations in pretend scenarios. *Acta Psychologica*, 124, 106-128.

Penn, D., Holyoak, K., and Povinelli, D. (2008). Darwin's mistake: explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences*, 30, 109-178.

Perner, J. (2010). Who took the cog out of cognitive science? Mentalism in an era of anti-cognitivism. In P. Frensch and R. Schwarzer (eds.), *Cognition and Neuropsychology: International Perspectives on Psychological Science: Volume 1*, Psychology Press.

Perner, J. and Ruffman, T. (2005). Infants' insight into the mind: How deep? *Science*, 308, 214-216.

Povinelli, D. and Vonk, J. (2003). Chimpanzee minds: suspiciously human? *Trends in Cognitive Sciences*, 7, 157-160.

Scott, R. and Baillargeon, R. (2009). Which penguin is this? Attributing false beliefs about object identity at 18 months. *Child Development*, 80, 1172-1196.

Scott, R., Baillargeon, R., Song, H., and Leslie, A. (2011). Attributing false beliefs about non-obvious properties at 18 months. *Cognitive Psychology*, 63.

Shanahan, M. (2010). *Embodiment and the Inner Life*. Oxford University Press.

Shanahan, M. and Baars, B. (2005). Applying global workspace theory to the frame problem. *Cognition*, 98, 157-176.

Schmelz, M., Call, J., and Tomasello, M. (2011). Chimpanzees know that others make inferences. *Proceedings of the National Academy of Sciences*, 108, 3077-3079.

Song, H. and Baillargeon, R. (2008). Infants' reasoning about others' false perceptions. *Developmental Psychology*, 44, 1789-1795.

Song, H., Onishi, K., Baillargeon, R., and Fisher, C. (2008). Can an actor's false belief be corrected by an appropriate communication? Psychological reasoning in 18.5-month-old infants. *Cognition*, 109, 295-315.

Southgate, V., Senju, A., and Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science*, 18, 587-592.

Southgate, V., Chevallier, C., and Csibra, G. (2010). Seventeen-month-olds appeal to false beliefs to interpret others' referential communication. *Developmental Science*, 13, 907-912.

Surian, L., Caldi, S., and Sperber, D. (2007). Attribution of beliefs by 13-month-old infants. *Psychological Science*, 18, 580-586.

Wellman, H. (1990). *The Child's Theory of Mind*. MIT Press.

Woodward, A. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition*, 69, 1-34.