Meta-cognition in Animals: A Skeptical Look

PETER CARRUTHERS

Abstract: This paper examines the recent literature on meta-cognitive processes in non-human animals, arguing that in each case the data admit of a simpler, purely first-order, explanation. The topics discussed include the alleged monitoring of states of certainty and uncertainty, knowledge-seeking behavior in conditions of uncertainty, and the capacity to know whether or not the information needed to solve some problem is stored in memory. The first-order explanations advanced all assume that beliefs and desires come in various different *strengths*, or *degrees*.

1. Introduction

The last several years have seen a flurry of experimental studies purporting to demonstrate the existence of meta-cognitive processes in non-human animals (hereafter, 'animals'). (See Smith *et al.*, 1995, 1997, 2003; Shields *et al.*, 1997; Call and Carpenter, 2001; Hampton, 2001, 2005; Hampton *et al.*, 2004; Smith, 2005; Son and Kornell, 2005; Beran *et al.*, 2006; Washburn *et al.*, 2006; Kornell *et al.*, 2007.) Although 'meta-cognition' strictly just means 'cognition about cognition', which could encompass thoughts about the cognitive states and processes of other subjects, those pursuing these investigations mostly intend it more narrowly, to refer to cognition about one's own cognitive states. (Cognition about others' cognition is generally referred to as 'theory of mind' or 'mind-reading'. I prefer the latter term, since it is less contentious.) Here is Smith (2005, p. 224) introducing his topic:

> Meta-cognition can be defined as thinking about thinking, or cognition about cognition. The idea in this field is that in some minds mental activities occur at a higher 'meta level' and at a lower 'object level' during cognitive processing. In these minds, there is a cognitive executive that supervises (i.e., oversees and facilitates) thought or problem solving.

For the most part I, too, shall follow this usage, making clear where necessary when I have in mind meta-cognitive thoughts about the cognitive processes of others.

In the present article I shall subject the main studies referred to above to sustained critique, arguing that there is no need to postulate meta-cognitive processing in order to explain the data. Rather, I shall show how those data admit of explanation

I am grateful to José Bermúdez, Robert Lurz, Jeffrey Stevens, and two anonymous referees for their comments on an earlier version of this article.

Address for correspondence: Department of Philosophy, University of Maryland, College Park, MD 20742, USA.
Email: pcarruth@umd.edu

Mind & Language, Vol. 23 No. 1 February 2008, pp. 58–89.

in first-order terms, appealing only to states and processes that are world-directed rather than self-directed. I shall argue, in consequence, that we should, at present, refuse to attribute meta-cognitive processes to animals. This inference is grounded in an application of Morgan's Canon. (Roughly: don't attribute to animals cognitive processes more complex than is necessary.) For there are good reasons for thinking that meta-cognition should be significantly more complex and demanding than regular first-order cognitive processes of the sort that I shall appeal to in my explanations, as I shall now briefly explain.[1]

The first point is simple: by their very nature, meta-cognitive processes contain an extra layer of representational complexity. A creature that is capable of meta-representing some of its own cognitive processes must first, of course, have the wherewithal to undergo the first-order processes in question. Then to this must be added whatever is necessary for the creature to represent, and come to believe, that it is undergoing those events. Put differently, a creature that is capable of thinking about its own thought that P must be capable of representing thoughts, in addition to representing whatever is represented by *P*.

The second point is that in the decades that have elapsed since Premack and Woodruff (1978) first raised the question whether chimpanzees have a 'theory of mind', a general (but admittedly not universal) consensus has emerged that meta-cognitive processes concerning the thoughts, goals, and likely behavior of others is cognitively extremely demanding (Wellman, 1990; Baron-Cohen, 1995; Gopnik and Melzoff, 1997; Nichols and Stich, 2003), and some maintain that it may even be confined to human beings (Povinelli, 2000). For what it requires is a theory (either explicitly formulated, or implicit in the rules and inferential procedures of a domain-specific mental faculty) of the nature, genesis, and characteristic modes of causal interaction of the various different kinds of mental state. There is no reason at all to think that this theory should be easy to come by, evolutionarily speaking. And then on the assumption that the same or a similar theory is implicated in meta-cognition about one's own mental states, we surely shouldn't expect meta-cognitive processes to be very widely distributed in the animal kingdom.[2] Nor should we expect to find meta-cognition in animals that are incapable of mind-reading.

[1] I should emphasize that I take for granted representational states (beliefs and desires) in animals, and also inferential processes involving such states. The case for thinking that animals share with us a basic first-order cognitive architecture for forming beliefs, for generating desires, and for practical reasoning and decision making in the light of those beliefs and desires seems to me to be overwhelming (Carruthers, 2006, ch. 2). At any rate, this is what I propose to assume for purposes of the present discussion. The question at issue is whether animals have meta-cognition, not whether they have cognition.

[2] I should emphasize that I am not suggesting that meta-cognition is *more* cognitively demanding than mind-reading. Rather, the claim is that both are significantly more demanding than first-order cognitive processes, giving us reason to prefer first-order explanations of animal behavior *ceteris paribus*. (But of course, in any given case, *ceteris* might not be *paribus*, if data emerge that are sufficiently hard to explain in other ways.)

There are, of course, theoretical perspectives from which meta-cognition of one's own mental states should be a good deal less cognitively demanding than meta-cognition directed at the mind of another. There is, in particular, the perennial lure of Cartesian accounts of self-knowledge, according to which our own mental states are at least easily, if not transparently and completely effortlessly, available to us. Those who endorse such a perspective are free to think that self-directed meta-cognition might be widespread amongst animals even if other-directed meta-cognition isn't. And those who adopt a so-called 'simulationist' account of our mind-reading capacity can claim that it is our first-person access to our own mental lives that forms the basis (both developmentally and phylogenetically) for our understanding of the mental lives of others (Goldman, 1993, 2006).[3] Such ideas may underlie the suggestion that some have made, that self-directed meta-cognition may form the cognitive foundation from which mind-reading capacities were able to evolve (Smith *et al.*, 2003; Metcalfe and Kober, 2005).

There is good reason to think that these Cartesian, or quasi-Cartesian, conceptions of self-knowledge are false, however. On the contrary, half a century of research in social psychology has shown that human beings are very frequently and demonstrably mistaken when attributing thoughts, reasons, and reasoning processes to themselves (Festinger, 1957; Bem, 1967, 1972; Wicklund and Brehm, 1976; Nisbett and Wilson, 1977; Eagly and Chaiken, 1993; Wilson, 2002). Moreover, the fact that people tend to go wrong in just those cases where the true causes of behavior are either unknown to common-sense psychology, or are such that folk psychology has a mistaken account of them, suggests very strongly that there is a common cognitive basis underlying the attribution of thoughts and thought processes to ourselves, and underlying our attributions of them to others. (See Carruthers, forthcoming, for an extended development and defense of this claim.)

Part of what motivates my skeptical reaction to much of the literature on meta-cognition in animals, then, is my rejection of the implicit assumption that meta-cognition is both easier than, and phylogenetically prior to, mind-reading.[4] But in what follows I shan't take this for granted. Rather, I shall subject the main bodies of experimental data on meta-cognition in animals to scrutiny. These include data on uncertainty and uncertainty monitoring (Section 3), data on knowledge-seeking behavior in cases of uncertainty (Section 4), and data on the adaptive use of meta-memory (Section 5). In each case I shall show that there is a simpler first-order

[3] Not all simulation theorists think that simulation is grounded in first-person access to our own mental states, of course. Gordon (1996), in particular, attempts to develop a version of simulationism that has as its upshot that our capacity to attribute thoughts to ourselves is dependent upon a prior ability to attribute them to others. While I do not endorse this approach, note that it, like theory-theory, predicts that we shouldn't expect to find meta-cognitive processes in creatures incapable of mind-reading.

[4] I should emphasize that I am *not* motivated by the thought that meta-cognition and mind-reading are both language-dependent capacities, and are thus unique to humans (as Bermúdez, 2003, claims). On the contrary, I assume that such capacities are independent of natural language.

explanation available. But I shall begin with a brief digression on the nature of surprise (Section 2), which will serve to introduce a number of the themes that follow.

2. A Cautionary Tale

There was once a philosopher who claimed that the emotion of *surprise* is meta-cognitive in nature (Davidson, 1982). To be surprised, he argued, involves coming to believe that one of your beliefs is false, and hence presupposes meta-cognitive thoughts about your own belief states. This seemed plausible enough, since, for sure, when we feel surprise we do characteristically believe that one of our prior expectations has been overturned. And we would naturally report on our surprise in just this sort of way: 'It gave me a real surprise. I had been confident that it would turn out one way, but then I saw that the opposite had occurred.' Furthermore, although this didn't actually happen at the time (perhaps because there was less interdisciplinary interaction in those days), one can imagine comparative psychologists picking up on these claims, doing experiments to demonstrate the presence of surprise in non-human animals, and claiming to have discovered the presence of meta-cognition outside of the hominid line. One can even imagine evolutionary-minded psychologists going on to claim that these meta-cognitive capacities are the precursor of, and provided the conceptual basis for, later-emerging capacities for mind-reading.

The trouble with all this is that the initial claim is false. Surprise, itself, is a purely first-order phenomenon. All that it requires is a mechanism that is sensitive to conflicts between the *contents* of a creature's occurrent judgments (not requiring it to represent the fact that it *has* those judgments). Nothing meta-cognitive need be involved. To make this transparent, let me introduce a simple convention, which will then be used throughout the remainder of this article. I shall use capitals to represent the mental states and attitudes that a creature has (belief, perception, desire, etc.), using square brackets to represent the contents of those states and attitudes.[5] Then surprise normally arises when a creature has an activated BELIEF [P] together with a PERCEPTION [Q], where the latter then gives rise to a novel activated BELIEF [not P]. The mechanism that gives rise to surprise is one that takes the contents [P] and [not P] produced in this sort of way as input, and which produces a suite of reactions as output: releasing chemicals into the bloodstream that heighten alertness, widening the eyes, orienting towards and attending to the perceived state of affairs [Q], and so forth. And it is the detection of these changes in ourselves that constitutes the feeling of surprise.

[5] Note that the only representations attributed to the creature are those that figure *within* the square brackets. It is I, as theorist, who represents the creature's beliefs and other attitudes. The creature itself just *has* those.

There is nothing meta-cognitive in this account of the genesis and nature of surprise.[6] How is it, then, that Davidson's view could ever have seemed appealing? And how is it that we so naturally report our surprise in meta-cognitive terms? The answer is simple. Humans are inveterate mind-readers, in the first person as well as in the third. We correctly interpret the feeling of surprise for what it is: a state that is caused when a perception gives rise to conflict with prior belief or expectation. And when we report on the feeling, it is in just such terms that we couch our description. We say, for example, 'I felt surprised because I was expecting P but saw Q instead.' But we aren't aware that we have engaged in any process of self-interpretation. On the contrary, the mind-reading faculty operates with a highly simplified model of its own operations, picturing the states reported on as being (for the most part) transparently available to the subject. This makes it natural for us to think of the state of surprise as being intrinsically meta-cognitive. Because we conceptualize and report that state in meta-cognitive terms, we are inclined to think of ourselves as merely expressing our conscious awareness of a meta-cognitive state.

It might be replied (as some have done in closely related domains; see Proust, 2006, on meta-memory and meta-knowledge) that even if surprise isn't explicitly meta-cognitive nature, it is at least *im*plicitly so. For it is a state that always involves some sort of conflict between prior expectation and current belief. In which case the occurrence of the emotion *carries the information* that such a conflict exists. Since the information carried is meta-cognitive in nature, it might be said that the state of being surprised is itself implicitly meta-cognitive. There is a sense in which this is true. But it is a sense that is far too weak to be of any interest. (In particular, it provides no warrant whatever for thinking that surprise might be the first evolutionary step on the road towards explicit meta-cognition.) For the same is equally true of any emotion, and of any behavior: all carry information about the occurrence of certain sorts of mental states. The feeling of fear, for example, is always produced by a thought of danger (generally a belief, but sometimes merely pretended or imaginary). So it carries the information that such a thought has occurred. In which case we can say, in this weak sense, that all creatures capable of fear have states that are implicitly meta-cognitive. Likewise, the fact that an animal is drinking reliably carries the information that it is thirsty (a mental state). In which case we could say that any animal that drinks is implicitly meta-cognitive.

The morals of this cautionary tale are two-fold. First, we should be aware that many cognitive phenomena that are quite naturally and correctly described and classified in meta-cognitive terms might, in themselves, be entirely first-order in character. (It is the categorization of the phenomenon that is meta-cognitive, not the phenomenon itself.) And second, we should be wary of our tendency to assume that meta-cognitive classifications are classifications of meta-cognitive states, pushing the meta-cognitive character of the categorization process 'downwards' into the state

[6] Note that the feeling of surprise itself involves representations of bodily changes, not psychological ones, just as other sorts of feeling—e.g. pain—do (Tye, 2006). These feelings are purely first-order in character.

itself. As we shall see, there is good reason to think that these morals have often been ignored in the burgeoning literature on meta-cognition in animals.

3. Uncertainty and Uncertainty Monitoring

Anecdotal evidence suggests that many species of animal are capable of feeling uncertainty, and of behaving accordingly. Think, for example, of a cat that paces back and forth on top of a wall while it examines the distance of a dangerous leap onto a nearby roof, periodically crouching as if to jump before resuming pacing once again. But an innovative set of studies with dolphins and monkeys set out to demonstrate that animals are capable of monitoring and responding adaptively to their own uncertainty in much the same way that humans do (Smith *et al.*, 1995, 1997; Shields *et al.*, 1997). The basic paradigm involved training the animals in a discrimination task, requiring them to press one symbol (which I shall label 'D' for 'dense') in response to a dense visual pattern (in the case of monkeys; with dolphins a high-pitched auditory tone was used), and to press another symbol ('S' for 'sparse') for patterns that are less dense (or tones that are less high pitched). The animals received a reward of food for correct responses, and a mild penalty for incorrect ones, which resulted in a brief 'time out' during which they had no opportunity to earn further rewards. They were also familiarized with a third response key, which served to initiate a new trial without a time out. The discrimination tasks were then made increasingly difficult, with the experimenters examining the extent to which the animals (and also humans in a parallel set of tasks) made adaptive use of the third 'don't know' key.[7]

The results were striking. All three species made increasing numbers of errors in forced-choice trials in which the 'don't know' key wasn't available, with performance decreasing to chance when making especially difficult discriminations. But when the 'don't know' option was available, all three species increased their use of it in conditions of uncertainty (that is, in conditions in which they were increasingly likely to make errors if forced to choose), with their use of it dominating in those cases where their performance would otherwise be at chance. Humans and animals even displayed similar individual differences in the extent to which the 'don't know' option was made use of. Smith (2005) argues, in consequence, that these common behavioral profiles make it almost mandatory to seek for a common underlying explanation. The humans in these experiments reported that they selected the 'don't know' key when (and because) they were aware of being especially uncertain of the correct response. If these reports are believed, therefore, we should accept that animals, too, behave as they do because they are aware of their own uncertainty, and meta-cognitive processes in animals are thereby established.

[7] Similar results to those described here have now been obtained by Beran *et al.* (2006) with Macaques, using the animals' judgments of greater or lesser numerosity of sets of objects rather than their perceptual discriminations.

Let me set aside the verbal reports of the humans for the moment, and raise the question whether the performance profiles in these experiments, considered on their own, mandate an explanation in meta-cognitive terms. It is plain that they don't. There is an alternative explanation available. And this explanation is, moreover, one that needn't involve mere acquired associations or conditioned behaviors, but is rather genuinely cognitive in nature (in the sense of appealing to beliefs, desires, and inferences). Let me elaborate.

3.1 Explaining Uncertainty Behavior: A First Pass

We just have to suppose (at a first pass; this account will be elaborated somewhat in Section 3.2) that beliefs come in different *strengths*, perhaps realized in varying degrees of activation of the representations underlying them. This shouldn't be a controversial assumption, since almost everyone in both philosophy and psychology accepts that something of this sort is the case. Let us symbolize these strengths with a subscripted 'w' for 'weak' and 's' for 'strong', added the notation introduced earlier. (The same subscripts can also be employed for different strengths of desire.) Then in the forced-choice trials (assuming a difficult discrimination that gives rise to both a weak degree of belief that the pattern is dense *and* a weak degree of belief that the pattern is sparse) what we have is a set of beliefs and desires somewhat as follows.[8]

(1) BELIEF_{S} [if the pattern is dense and D is pressed, then food results].
(2) BELIEF_{W} [the pattern is dense].
(3) DESIRE_{S} [food].
(4) BELIEF_{S} [if the pattern is sparse and D is pressed, then a time out results].
(5) BELIEF_{W} [the pattern is sparse].
(6) DESIRE_{S} [no time out].

States (1) through (3) together generate (7), a weak desire to press D in order to obtain food. (I assume that weakness in any state serving as a premise—in this case (2)—will issue in a similarly weak conclusion.) But states (4) through (6) likewise create (8), a weak desire *not* to press D, in order to avoid a time out.

(7) DESIRE_{W} [press D].
(8) DESIRE_{W} [don't press D].

[8] Here and in the examples hereafter the contents that I attribute to the animals can only be approximate, of course, since we don't know precisely *which* concepts the animal deploys in thinking about the various elements of the experimental set-up. It will be enough for my purposes that the animals should have concepts that are roughly co-extensive with those that I use in my attributions. Our topic is not to delineate the precise concepts that the animals deploy, but more broadly to determine whether they are entertaining higher-order concepts of some sort, or merely first-order ones. Note, too, that although I use sentences to characterize the contents of the animals' beliefs, this needn't commit me to claiming that the representational vehicles of those beliefs are sentence-like. On the contrary, those vehicles might be image-like or map-like, or might consist of mental models of some sort.

So the animal is conflicted over whether or not to press D. But then of course it will be equally conflicted over whether or not to press the other primary discrimination option S, in light of its possession of the following two beliefs:

(9) BELIEF$_S$ [if the pattern is sparse and S is pressed, then food results].
(10) BELIEF$_S$ [if the pattern is dense and S is pressed, then a time out results].

State (9) together with (5) and (3) issues in (11); whereas state (10) interacting with (6) and (2) creates (12).

(11) DESIRE$_W$ [press S].
(12) DESIRE$_W$ [don't press S].

Hence the animal has desires of roughly equal strength both to press and not press S as well as to press and not press D. In a situation where a choice is forced, the animal must thus choose randomly.

Now suppose that the 'don't know' option is available. Then in addition to the above states, the animal may also activate the following belief.

(13) BELIEF$_S$ [if 'don't know' is pressed then a new pattern is presented and no time out results].

Since there is no overall desire to press D, and no overall desire to press S, this belief alone, in conjunction with (6)—the desire to avoid a time out—is sufficient to motivate pressing 'don't know'.

There is nothing meta-cognitive in this explanation. All that is involved are first-order beliefs and desires of various strengths, together with simple forms of practical reasoning involving interactions of those states. Note, moreover, that there need be no commitment to any kind of general-purpose practical reasoning system that combines all active beliefs and desires together into some sort of expected utility calculation, in the manner of decision theory. (This is all to the good, since there is good evidence that animals possess no such system. See Carruthers, 2006, ch. 2.) Rather, each of the various goal states interacts with the others *competitively* in an attempt to control behavior. In the case that I have envisaged, the competition between pressing or not pressing D and pressing or not pressing S is in a four-way tie, whereas there exists an unopposed motive to press 'don't know', so as to avoid a time out. Hence that, accordingly, is what the animal does.

There are two related problems with the explanation advanced thus far, however, which require us to elaborate it somewhat. (But as we shall see, the elaboration is independently motivated.) One is that it only applies in those cases where the animal's strengths of belief that the stimulus pattern is dense or sparse are equal. (If either is stronger than the other, on the above account, then the corresponding act of symbol-pressing will be weakly motivated.) The other is that it doesn't have the resources to explain how there can be individual differences in the extent to which animals (and humans) make use of the 'don't know' option.

3.2 Explaining Uncertainty Behavior: A Second Pass

The emendation needed is that there exists a certain overlapping *range* of strengths of competing beliefs within which the animal becomes reluctant to act, and seeks either further information or some other alternative for action. In order to see the motivation for this proposal, notice that all perceptual systems are inherently *noisy* in their operations, in the sense that no two presentations of one and the same stimulus will issue in precisely the same degree of belief. (That this is so is one of the central assumptions of Signal Detection Theory, or SDT.)[9] Hence from one moment to the next, and from one glance to the next, the degrees of belief that result from a given stimulus will fluctuate somewhat. This means that in connection with particularly difficult discriminations the degrees of belief in the presence or absence of a stimulus of a given type (dense, say) will often reverse themselves. At one moment the animal might have a degree of belief that the pattern is dense that is slightly stronger than its degree of belief that the pattern is sparse (and hence not dense), and then at the next moment, or with the next glance, the animal might be in the reverse state, with a slightly stronger degree of belief that the pattern is *not* dense.

What is an animal in such circumstances to do? Plainly it would be adaptive, in cases where the animal isn't forced to act immediately, for it to pause and do things that might resolve the indeterminacy, or for it to take action in pursuit of an alternative goal instead. Thus (as indeed we observe) the animal might approach the stimulus, or move its head from side to side, to get a better view. Or the cat that is uncertain of a jump to a roof might seek an alternative route (climbing a nearby tree, perhaps). Notice that I am not claiming that the animal will move its head from side to side with the intention of removing its uncertainty (which would be a meta-cognitive intention).[10] Rather, it has in place a mechanism (most likely evolved, but perhaps constructed through some sort of learning) which when confronted with conflicting plans that are too close to one another in strength will refrain from acting on the one that happens to be strongest at that moment, and will initiate alternative information-gathering behavior instead. If this issues in changed degrees of belief, and hence in sufficiently changed degrees of desire to perform one action rather than another, then that action will be performed; if not, and there are no alternatives, then one or other is chosen at random (or in accordance with momentary greater strength).

[9] See Smith *et al.* (2003), who attempt to use SDT in support of their own meta-cognitive explanation of uncertainty behavior in animals. As will be plain, I think that their account is confused, since SDT is more consistent with a first-order explanation of the data.

[10] It is plain that such behavior can be caused without the intervention of any meta-cognitive state. For even a praying mantis will sway its head and body from side to side when it needs to improve its judgment of depth (Kral and Poteser, 1997). Yet not even the most ardent fan of meta-cognition in animals will claim that the mantis is moved by a meta-cognitive thought.

Suppose, for simplicity, that beliefs and desires come in degrees ranged from 1 (weakest) to 10 (strongest). Then what is needed is for there to be a 'gate-keeping' mechanism at the point where different goals are competing with one another to control behavior. This would only initiate one of the desired behaviors if the degrees of desire involved are sufficiently far apart. (Recall that these degrees of desire would themselves be dependent on the degrees of belief concerning the presence or absence of a stimulus of a given type.) Otherwise the animal becomes motivated to pause, engaging in information-seeking behavior or searching for another alternative. Since conditions of uncertainty are also inherently dangerous, in various ways, one might expect a series of bodily changes to take place in the animal that it will experience as a form of aversive anxiety.

For example, this mechanism receiving as input the two states DESIRE_2 [press D] and DESIRE_3 [press S] would *not* initiate the motor commands necessary to press S; whereas if it received as input the two states DESIRE_2 [press D] and DESIRE_4 [press S] it might do so. But it is important to see that there need be nothing meta-cognitive about the envisaged gate-keeping mechanism. It doesn't *represent* the fact that two desires are very close to one another in strength. Rather, it responds differentially depending on whether or not the desires that it receives as input are in fact close to one another in strength. It is a mechanism that is sensitive to one *property* of desire (strength) without needing to represent that it is a *desire* that has that property.[11]

The idea of such a gate-keeping mechanism is surely well motivated, then, given that perceptual processes are inherently noisy. (See the 'diffusion model' of decision making for a closely related and well-confirmed account—Ratcliff and Rouder, 1998.) And with the idea of such a mechanism in place, it will be easy for us to explain how there can be individual differences in these uncertainty tasks. They will be differences in the intervals within which the currently strongest desire doesn't get acted upon, but rather information-gathering behavior and/or a search for alternative options is initiated instead. Thus one animal might have the parameters of its gate-keeping mechanism set in the way that we envisaged above, in such a manner that a separation of two degrees of desire is necessary for one of the primary response options to be initiated straight away. Another animal (whom we might naturally describe as 'over-confident', notice) might have the parameters set more narrowly, so that it initiates a response whenever there is just one degree of separation.

[11] There is, of course, a much weaker sense of 'meta-cognitive' available to us, in which it would be true to say that the gate-keeping mechanism is a meta-cognitive one. For it is a mechanism that exists down-stream of, and whose operations are causally dependent upon, other cognitive states (in this case, activated desires to perform, or not perform, various actions). But this sense if far too weak to be of any interest. For notice that in this sense the desires in question are themselves 'meta-cognitive' too—for they will have been produced by a belief in the presence of a given type of stimulus interacting with a conditional belief that actions taken in the presence of that stimulus will issue in a reward, together with the desire for the reward. So the desire to perform the action will carry the information that states of just this sort have occurred. This isn't meta-cognition worth the name.

What are we to say, then, of the relatively rare cases where an individual animal virtually *never* uses the 'don't know' option? Are these cases where the animal lacks the postulated gate-keeping mechanism altogether? This is possible, but perhaps not very likely. In order to see a better explanation we need to notice that the degree of anxiety that an animal experiences in an uncertainty situation will be a function both of its own idiosyncratic psychology and physiology, and of what is at stake in the situation. Then recall that the function of such anxiety will be to motivate both information-seeking behavior and searches for alternative courses of action (such as pressing the 'don't know' symbol). An animal that is supremely over-confident will be one that experiences little anxiety in situations of uncertainty, and for whom the 'don't know' option will be much less salient. This will be an animal that is focused almost entirely on the primary discrimination task, and is little motivated to consider alternatives. But this need not mean that it lacks the postulated gate-keeping mechanism altogether. If we were to raise what is at stake in the situation significantly enough, we might predict that individual animals who previously made little or no use of the 'don't know' option would start to do so. (This explanation should be easily testable.)

3.3 Explaining Human Uncertainty Behavior

With these explanations in hand, let us now return to Smith's (2005) challenge. Since the performance profiles of humans and animals in these uncertainty experiments are so similar, there is good reason to seek for a common explanation. But the humans in these experiments report that they feel uncertainty, and that they select the 'don't know' option, when they do, because they are too unsure of the correct choice in the primary discrimination task to select one of the primary response options. These reports suggest that some sort of meta-cognitive explanation of the humans' behavior is appropriate; in which case we have reason to prefer such an explanation in the case of non-human animals as well. There is a pair of different strategies for responding to this challenge, however. Each may be applicable in different circumstances, depending on the precise form of the explanation that any given human provides for his or her own behavior. This will enable us to divide and conquer. But both have in common that the uncertainty behavior of humans should be explained in the same first-order terms as the explanations already offered for the uncertainty behavior of animals.

One strategy involves noting that undergoing a feeling of uncertainty needn't mean being aware *that* one is feeling uncertain, as such (which would be a meta-cognitive state). For the feeling of uncertainty itself, on our account, consists in an awareness of a distinctive profile of physiological and behavioral reactions caused by the activation of the gate-keeping mechanism (including hesitating and engaging in a variety of information-seeking behaviors, such as squinting at the display or looking closer), which is experienced as aversive. So both the state of *being* uncertain and the state of *feeling* uncertain are first-order states that humans and animals can share: nothing meta-cognitive need be involved. And if a human says that he chose as he did because he was uncertain, or because he felt uncertain, then what he says can be true, consistent with the account of the genesis of uncertainty behavior provided in Sections 3.1 and 3.2 above. Of course, in providing either of these explanations for his choice, the human will thereby utilize the concept *uncertainty* (which therefore makes his *report* a meta-cognitive one). But the processes appealed to in the explanation that he provides can be entirely first-order in character.

The other strategy available to us in responding to Smith's (2005) challenge is to deny the truth of the humans' reports. This strategy becomes appropriate and defensible if what the human says is that he chose as he did, not just because he *was* uncertain (a first-order state), but because he was *aware that* he was uncertain (a meta-cognitive state). And the strategy is also appropriate if what the human says is, not just that he *had* a feeling of uncertainty (which is a first-order state), but that be acted as he did because he was *aware of* a feeling of uncertainty or because he *felt that* he was uncertain (which are meta-cognitive states). Of course we needn't deny that the human was actually subject to any of these forms of higher-order awareness. Our quarrel is only with his claim that it was one of these forms of meta-cognitive awareness that caused his selection of the 'don't know' behavioral option. We can allow that at the time of acting he was aware of being uncertain, that he was aware of a feeling of uncertainty, and even that he felt that he was uncertain. But if the account of uncertainty behavior presented in Sections 3.1 and 3.2 is on the right lines, none of these forms of awareness played a role in the genesis of his behavior. Rather, it was the operations of a gate-keeping mechanism (which is shared with other species of animal) giving rise to a state of uncertainty that did that.

If we are to deny some of the explanations of their own uncertainty behavior that human subjects will provide, then isn't it incumbent on us to explain how they come to be in error? Perhaps. But the needed explanation isn't difficult to find. For even if humans have some sort of privileged (perhaps even infallible) insight into the occurrence of their own conscious mental states when they have them (as Cartesian accounts of self-knowledge would imply), that needn't mean that they have such insight into the causal roles that those states occupy on any given occasion. And indeed, we know for sure that humans are very frequently mistaken about this. The extensive literature in social psychology on 'confabulation' (the invention of demonstrably false explanations of our own thoughts and actions) is sufficient to establish at least this (Festinger, 1957; Bem, 1967, 1972; Wicklund and Brehm, 1976; Nisbett and Wilson, 1977; Eagly and Chaiken, 1993; Wilson, 2002). Moreover, as we noted in Section 2, humans seem to have a tendency to assume a larger role for meta-cognitive states and processes than they really should, pushing 'downwards' the meta-cognitive character of their introspective awareness into their account of those states and processes themselves.

3.4 Explaining the Absence of Uncertainty Behavior in Rats

I have shown how the uncertainty behavior of dolphins, monkeys, and humans can be explained in purely first-order terms, without any appeal to meta-cognitive processes. But can I also explain why other species of animal (in particular, rats) should fail to manifest that sort of behavior? For Smith (2005) claims that the best explanation for these failures is that meta-cognition is a relatively sophisticated and late-evolving cognitive process. This can be considered our final challenge.

Smith and Schull (1989) worked hard to elicit an uncertainty response from rats. In a paradigm similar to the one later used with dolphins, they first trained the rats to distinguish a lowish-pitch repeating tone from a tone of the same pitch alternating with one of any other pitch. If the tone was a repeating one, they were to make the R ('repeating') response, whereas if the tone was alternating they were to make the A ('alternating') response. The rats learned this task with little difficulty, all the while utilizing what was later to become the intended 'don't know' response in order to initiate a sequence of trials. (Making the 'don't know' response here, as in the other experiments reported, immediately advanced the animal to a new trial, without a delay. But it would also serve to begin the first trial in a series.)

The alternating tone was then made harder and harder to distinguish from the repeating one, by making the two alternating tones more and more similar to one another. As expected, the primary R and A responses of the rats showed increasing numbers of errors, with their performance falling to chance when the two sorts of sound became too hard to distinguish. In such cases it would have been adaptive for the rats to avail themselves of the 'don't know' behavioral option, immediately advancing to a new trial. But the rats failed to show any increase in their tendency to make the 'don't know' response in conditions of uncertainty. They continued to make that response at the same low rate (arguably attributable to behavioral 'noise', or to the operations of a 'sample the options' heuristic).

Smith's (2005) explanation of the rats' failure is that they lack a capacity for meta-cognition. My own explanation is that they lack the gate-keeping mechanism that dolphins, monkeys, and humans all possess (when the competing goals derive from foraging desires, at least).[12] The reason why the rats don't avail themselves of the 'don't know' option is that they lack any mechanism designed to make them pause and seek additional information and/or alternative options when they have conflicting desires that fluctuate around the same sort of level of strength. Rather, they just implement the behavioral option that happens to be the most desired at the time. This explanation is just as well able to explain the data, but doesn't require us to interpret the behavior of dolphins and monkeys in meta-cognitive terms.

So far, so good. But Smith (2005) claims that an additional experimental manipulation supports a meta-cognitive interpretation. Smith and Schull (1989) were worried that the rats might have been failing to make the 'don't know' response because it also doubled as a trial-initiation response. To evaluate this, they

[12] There is no reason to think that the postulated gate-keeping mechanism should be 'all or nothing', either mediating between all competing goals or being absent altogether. Different species of animal might employ such a mechanism for different sub-sets of competing goals, displaying uncertainty behavior when navigating, say, but not when foraging. In the case of rats, the absence of a gate-keeping mechanism employed when foraging may serve, in effect, to implement a 'sample the options' heuristic, ensuring that the rats don't focus their behavior too narrowly on just one potential food source.

examined how the rats would respond to a different, higher, repeating tone that was rewarded on a 50% contingency. The rats declined these trials with considerable frequency, using the 'don't know' response to advance them to a simpler alternating / repeating trial in which they could be confident of a reward. (The high repeating tone was only used in conjunction with easy versions of the previous task.) Smith (2005) remarks that since the rats did bail out of the objective stimulus-borne 50:50 contingency, but not the subjective threshold-borne 50:50 contingency (when their discriminations were at chance), a difference in underlying psychological organization is suggested. And the best proposal, he thinks, is that rats can track objective chances but not their own states of uncertainty.

Here, too, there is an alternative, purely first-order, explanation available. For the difference in behavior can be explained in terms of differences in the contents of the rats' beliefs, rather than in terms of their sensitivity to the differing degrees of certainty with which those beliefs are held. The rat surely has the following set of beliefs.

(14) BELIEF$_S$ [if there is a repeating low tone, then an R response almost always issues in food]. (Recall that the rat is engaged in easy discrimination trials, in which its responses are largely correct.)

(15) BELIEF$_S$ [if there is a different tone alternating with a low tone, then an A response almost always issues in food].

(16) BELIEF$_S$ [a 'don't know' response is almost always followed by a situation from which food is obtained]. (This is because it leads to a situation in which there is either a repeating or an alternating tone, to which the rat responds correctly.)

(17) BELIEF$_S$ [if there is a repeating high tone, then an H response only sometimes issues in food]. (Recall that the high repeating tone is only rewarded on a 50% contingency schedule.)

Belief (16) is grounded in (14) and (15). But (16) and (17) together are sufficient to explain why the rat will often use the 'don't know' response when faced with a repeating high tone. For we know that rats are extremely good at calculating relative rates of return (Gallistel, 2000). The rat will then easily be able to figure out that the rate of return from a 'don't know' response when there is a repeating high tone is greater than can be obtained from an H response in the same circumstances. This is because the 'don't know' response immediately advances the rat to an easy discrimination trial from which it generally obtains a reward.

3.5 A New Paradigm for Measuring Uncertainty: Teaching Monkeys to Gamble

A novel way of testing for meta-confidence judgments in monkeys has recently been introduced by Son and Kornell (2005), who used a betting paradigm. The animals first had to make an initial discrimination response—judging which of nine lines presented on a screen was longest and touching the one selected—following which they were presented with a choice between a high-confidence symbol (which resulted in greater gains if the initial choice was correct but an equivalent loss if it was incorrect) and a low-confidence symbol (which guaranteed a small reward). A reservoir of tokens was visible on the screen throughout the trials, to which tokens won were added and from which tokens lost were removed, in each case accompanied by distinctive sound effects to increase salience. The monkeys were rewarded with a food pellet once the reservoir came to exceed a certain level, and then the reservoir was re-set. As predicted if monkeys are capable of monitoring their own degrees of confidence in their answers, they were more likely to take the 'high risk' gamble in cases where they had made the correct answer than in cases (often psychophysically difficult) where they had made an error.

The monkeys' success in these tasks can be explained in terms of degrees of belief, however, without needing to appeal to a meta-cognitive judgment that one *has* a high degree of belief and/or confidence. Following training, the monkeys would surely possess each of the beliefs and desires (18) through (24). (Note their varying strengths.)

(18) BELIEF$_S$ [if the longest line has been touched and the 'high risk' symbol is touched on the next screen, then three tokens will be gained].
(19) BELIEF$_S$ [if the longest line has been touched and the 'low risk' symbol is touched on the next screen, then one token will be gained].
(20) BELIEF$_S$ [if the longest line hasn't been touched and the 'high risk' symbol is touched on the next screen, then three tokens will be lost].
(21) BELIEF$_S$ [if the longest line hasn't been touched and the 'low risk' symbol is touched on the next screen, then one token will be gained].
(22) DESIRE$_S$ [three tokens].
(23) DESIRE$_S$ [no loss of three tokens].
(24) DESIRE$_W$ [one token].

Now consider what happens in a psychophysically easy case, where the presence of a line that is obviously longer than the others gives rise both to the selection of that line and to (25), a high degree of belief that the longest line has been touched, as well as to (26), a correspondingly low degree of belief that the longest line *hasn't* been touched.

(25) BELIEF$_S$ [the longest line has been touched].
(26) BELIEF$_W$ [the longest line hasn't been touched].

Then (25) together with (18) and (22) will yield (27), whereas (26) together with (20) and (23) will only yield the weaker (28). (Again, I assume that weakness in any one of the premises will yield a corresponding weakness in the conclusion.) And likewise, (25) together with (19) and (24) will give rise to (29).

(27) DESIRE$_S$ [touch the 'high risk' symbol].
(28) DESIRE$_W$ [don't touch the 'high risk' symbol].
(29) DESIRE$_W$ [touch the 'low risk' symbol].

In cases of high confidence, then, where the degree of belief represented in (25) is especially large, one might expect the differential between (27) and (28) to be greater than the small value attaching to (29). In which case it will be (27) that wins out in the competition to control behavior, and the animal is likely to select the 'high risk' symbol, winning three tokens for itself.

Now consider a low-confidence case, where the psychophysical discrimination is difficult, giving rise to (30) once the monkey has made its choice. This will combine with (18) and (22) to yield only (31), a *weak* desire to touch the 'high risk' symbol, rather than the previous (27).

(30) BELIEF$_W$ [the longest line has been touched].
(31) DESIRE$_W$ [touch the 'high risk' symbol].

This weak desire to touch 'high risk' will at least be balanced by (28), which will still be caused by (26) interacting with (20) and (23). This is because in a difficult case (that isn't *very* difficult) the monkey will still have only a weak degree of belief that that the longest line hasn't been touched. But as the difficulty of the discrimination increases, so the monkey's degree of belief represented in (26) will increase, giving rise to a correspondingly stronger version of (28), which at some point will come to outweigh (31). Yet throughout, the monkey still has (29), which is an unopposed weak desire to press the 'low risk' symbol. So this, accordingly, is what the monkey is likely to do.

3.6 Generalizing Good Gambling to New Tasks

It should be plain from the preceding discussion that nothing in Son and Kornell's (2005) initial data-set provides any real evidence of meta-cognitive processes in the monkeys involved. To their credit, the authors were sensitive to this potential line of criticism, and added a second version of the task to see whether the monkeys' knowledge would transfer without further training. They first asked the monkeys to discriminate between nine boxes containing different numbers of dots for a direct reward. Then in the transfer test, the previous reservoir of tokens was added to the screen, together with the 'gamble' screen (which presented the choice between the 'high risk' and 'low risk' symbols) immediately following the monkey's selection. The investigators were interested in whether the monkeys would start to make appropriate use of these symbols from the start (pressing 'high risk' in conditions of certainty and 'low risk' in conditions of uncertainty). Accordingly, data were only collected from the first four sessions following the re-introduction of gambling. What they found was that the monkeys did, indeed, transfer what they had learned: from the start, in this new paradigm, they were more likely to press the 'high risk' symbol in cases where they had made a correct initial discrimination, and to press 'low risk' in those difficult cases where they might have made an error.

Kornell *et al.* (2007) then added yet another transfer task involving working memory. In each trial the monkeys were shown a sequence of six different pictures, following which they were presented with a screen containing just one of those pictures embedded amongst eight distracters. Their task was to select the familiar item. As before, when offered the opportunity to gamble, they immediately began to make adaptive use of the 'high risk' and 'low risk' options. Kornell *et al.* argue that these results can only be explained on the assumption that some of the knowledge that the monkeys acquired from the first training set was couched in meta-cognitive terms. In order to transfer to the new tasks, what they had learned must have been something like (32), which has a meta-cognitive content, rather than (or in addition to) the belief (18), which was about length of lines rather than degrees of numerosity or the familiarity of a picture.

(32) BELIEF$_S$ [if I am certain that I have chosen correctly and the 'high risk' symbol on the next screen it touched, then three tokens will result].

Since the antecedent clause in (32) refers to a mental state that can be caused in any number of different tasks, rather than to something about the particular task in hand (as is the case in (18)), the animal can respond appropriately when a quite different sort of discrimination and choice together form the initial conditions for a subsequent gamble.

I grant that attributing (32) to the monkeys can explain the data. I deny, however, that the data can only be explained in this way. The experiment shows something interesting about the monkeys' capacity to learn a general rule, but it doesn't show that the rule in question must be meta-cognitive in form. On the contrary, what the monkeys might have learned from the first training set, as a generalization of (18) and (20), could have been (33) and (34). Note that these beliefs have first-order contents.

(33) BELIEF$_S$ [within each array there is a target symbol, touching which achieves three tokens if the 'high risk' symbol is pressed thereafter].

(34) BELIEF$_S$ [if the target symbol hasn't been touched and the 'high risk' symbol is touched on the next screen, then three tokens will be lost].

Once the monkeys had been familiarized with the new numerosity and working memory versions of the task, they also learned is that the target symbol is, respectively, the one containing the greatest number of dots or the one that had been presented in the previous sequence of pictures. And then when their confidence that they had pressed the target symbol was low, they would have weak counter-balancing desires both to press and not press the 'high risk' symbol, deriving from (33) and (34), whereas they would still have an unopposed weak desire (29) to press the 'low risk' symbol. Hence this, accordingly, is what they are likely to do.

While it is certainly interesting that monkeys can extract and learn a general rule of the sort represented in (33), it is not really very surprising. For we know that many species of animal are remarkably good at extracting such rules. We know, for example, that even bees can learn the rule, 'Turn right if the symbol in the present chamber is the same as in the previous one, turn left if it is different'

(Giurfa *et al*., 2001). So explaining the monkeys' success in the transfer tasks in terms of their acquisition of the first-order belief (33), rather than the meta-cognitive belief (32), is by no means *ad hoc* or unmotivated.[13]

4. Uncertainty and Knowledge-Seeking Behavior

In this section I shall first discuss the studies of knowledge-seeking in conditions of uncertainty conducted by Call and Carpenter (2001) with apes and by Hampton *et al.* (2004) with monkeys, which are claimed to demonstrate that these animals know when they do and don't have the knowledge that they need to achieve their goals. I shall then discuss an elegant new experiment carried out by Kornell *et al.* (2007), which allowed monkeys to ask for 'hints' in cases where they were uncertain.

4.1 The Hollow Tube Test

The animals in these studies were presented with four opaque hollow tubes, one of which was baited with food on each trial in one of two conditions: with the animals watching while the food was placed in the tube, or with the tubes briefly obscured. The animals were then allowed to select one and only one tube. If they chose correctly, they obtained the food contained within it; if they chose incorrectly, then the food was removed by the experimenter and a new trial initiated. What the experimenters were interested in, and what they measured, was whether the animals would bend down to look into the hollow tubes before making their selection in cases where they were ignorant of the food's location. What they found was that the animals were, indeed, much more likely to look before choosing in the trials where they hadn't been allowed to observe the baiting of a tube. And this is said to show that they know, in these circumstances, that they lack knowledge, and take appropriate steps to obtain that knowledge.

[13] Note that the knowledge-transfer data presented by Washburn *et al.* (2006) are even more easily explained. Their monkeys had previously been familiarized with a choice task that required them to move a cursor to one of two symbols presented on a screen to obtain a reward. And the monkeys had, in other circumstances, been familiarized with a 'don't know' symbol (a star), which advanced them to another screen from which a reward was guaranteed (but with a brief time-out). What Washburn *et al.* found is that the monkeys used the 'don't know' symbol adaptively from the very beginning of the new test trials, opting to use it on the first occasion on which (as the experimenters were able to establish on other grounds or from the experimental setup) they didn't know which of the two primary symbols they should press. But nothing meta-cognitive needs to be introduced here, in order to explain the monkeys' behavior. Since a monkey in these circumstances will know that pressing one (and only one) of the two primary symbols will obtain a reward, but has no belief as to which, it will have weak and counter-balancing desires to press each. But since the monkey also knows that pressing the 'don't know' symbol will advance to a screen from which a reward is guaranteed, this will give rise to a comparatively strong and unopposed desire to press that symbol. No meta-awareness of its own state of uncertainty is required.

Let me first consider a case where the animal observes one of the tubes being baited, showing how the animal's behavior admits of a straightforward first-order explanation. All we need to assume is that perception of a tube being baited (tube A, say) gives rise to the belief that there is food in tube A. This belief then interacts with the animal's goals and beliefs about the contingencies of the experiment to generate an unopposed motive to select tube A. All we need to assume, in fact, is that the animal possesses the following first-order mental states.

(35) BELIEF$_{S}$ [the food is in tube A]. (Caused by perception of A being baited.)
(36) BELIEF$_{S}$ [if the food is in tube A and A is selected, then food will result].
(37) DESIRE$_{S}$ [food].
(38) DESIRE$_{S}$ [select tube A]. (By inference from (35) through (37).)

Although the animal has three other beliefs parallel in form to (36), in respect of each of the other three tubes, it has no other beliefs parallel in form to (35), since it had no perceptions that might issue in a strong belief that the food is in tube B (or in C, or in D). Hence there is nothing to give rise to a desire to select one of these other tubes, and (38) is therefore unopposed.

Now consider a case where the animal has been prevented from observing the baiting of one of the tubes. (For ease of presentation I shall imagine a case where there are just two tubes, rather than four.) Then the animal will have the following first-order mental states. (Note their varying strengths.)

(39) BELIEF$_{S}$ [the food is in tube A or tube B]. (From general knowledge of the conditions of the experiment.)
(40) BELIEF$_{W}$ [the food is in tube A]. (By inference from (39).)
(41) BELIEF$_{S}$ [if food is in tube A and A is selected, then food will result].
(42) DESIRE$_{W}$ [select tube A]. (By inference from (40) and (41) together with (37), the desire for food.)
(43) BELIEF$_{W}$ [the food is in tube B]. (By inference from (39).)
(44) BELIEF$_{S}$ [if food is in tube B and B is selected, then food will result].
(45) DESIRE$_{W}$ [select tube B]. (By inference from (43) and (44) together with (37).)

So in this case the animal will have two more-or-less equally balanced weak conflicting desires: to select tube A *and* to select tube B. (These desires conflict because the animal also knows that only one tube can be selected in each trial.) So the same gating mechanism that we envisaged in Section 3 will then issue in the usual suite of information-seeking behavior: looking more closely, moving one's head from side to side, and so forth. All we need to suppose is that the behavioral repertoires of these animals have been widened slightly through trial and error learning to include the motor schemata for bending down to look into a hollow tube. (And individual differences in such learning will be sufficient to explain the differing degrees of success that individual animals display in these experiments, even without the need to appeal to differences in the confidence intervals in the gating mechanism discussed in Section 3.2.) Nothing meta-cognitive need be involved.

4.2 Hint-Seeking in Monkeys

The information-seeking experiment conducted by Kornell *et al.* (2007) was significantly more sophisticated than those just discussed. It examined whether monkeys would request extra information in cases where they are uncertain—thereby displaying meta-cognitive skills, according to the authors. These experiments required the animals to learn rules governing the order in which they should touch a set of pictures shown simultaneously on a screen. Four pictures were displayed on each trial, randomly positioned on the screen so that motor learning couldn't be employed. The monkeys' task was to touch each picture on the screen in the correct order to receive a reward. Any error would lead to the screen going blank, followed by a brief time out. In addition to learning by trial and error, the monkeys were also familiarized with a 'hint' symbol displayed at the side of the screen, which they could press at any time. Pressing 'hint' led to the next correct symbol in the sequence being clearly marked for them, but also to a less valued eventual reward for a correct sequence of pressings (a food pellet as opposed to an M&M candy). What the experimenters found is that use of the 'hint' option decreased as the animals' accuracy increased, suggesting that they were only using it in cases where they were uncertain of the next move.

I shall argue that nothing meta-cognitive is needed to explain the animals' behavior. For ease of presentation I shall consider a case where an animal has already made correct choices for the first two items in the list, but is now uncertain about which symbol to press next. In these circumstances the animal will possess the following first-order mental states.

(46) BELIEF$_W$ [if picture C then picture D is touched, an M&M will result].
(47) BELIEF$_W$ [if C then D is touched, a time out and then a new trial will result].
(48) BELIEF$_W$ [if D then C is touched, an M&M will result].
(49) BELIEF$_W$ [if D then C is touched, a time out and then a new trial will result].
(50) DESIRE$_S$ [M&M].
(51) DESIRE$_S$ [no time out and new trial].

Together, (46) through (51) will issue in weak, roughly counter-balancing, desires to touch C and then D, to avoid touching C and then D, to touch D and then C, and to avoid touching D and then C. But the animal also has beliefs (52) through (54), which combined with the weak desire (55) for a food pellet issues in a weak but unopposed desire (56) to press the 'hint' symbol. So this is what the animal is likely to do.

(52) BELIEF$_{S}$ [if ‘hint’ is pressed, then either C or D will get marked].
(53) BELIEF$_{S}$ [if C gets marked, then pressing C and then D will result in a food pellet].
(54) BELIEF$_{S}$ [if D gets marked, then pressing D and then C will result in a food pellet].
(55) DESIRE$_{W}$ [food pellet].
(56) DESIRE$_{W}$ [press ‘hint’].

All that is required, in order for this first-order explanation to work, is some variant of the gate-keeping mechanism postulated in Section 3. This prevents the roughly-balanced conclusions of (46) through (51) from issuing in action, hence allowing (56) to come to the fore.

Will the explanation offered above generalize to other stages in the learning process, however? For example, what if the animal is uncertain about which picture to press first, rather than last? Would it believe the equivalent of (46), only with ‘A’ and ‘B’ substituted for ‘C’ and ‘D’, as in (57)?

(57) BELIEF$_{W}$ [if picture A then picture B is pressed, an M&M will result].

The answer depends, of course, on how the animals actually set about solving these tasks. One possibility is that they attempt to form a representation of the entire sequence (‘A, then B, then C, then D’) before they act at all. But this seems unlikely. For we know from other list-learning experiments with these animals that they learn ‘from the front’ (learning the first position in the sequence before the second, and so on), and that they end up representing the ordinal position of each item in the list (Terrace *et al.*, 2003). So it is plausible that they tackle the tasks as a series of choice-points – in effect asking, and answering, the question, ‘Which picture is first?’, before they move similarly through each of the remaining positions.

Even so, it might be felt that (57) can’t be the right form for the animal’s belief to take. For the monkey surely doesn’t believe that getting just the first two stages in the sequence correct is sufficient to get an M&M. On the contrary, it knows that all four steps in the sequence need to be correct. But actually, since the animal is forced to choose *something* first or face an extended timeout (see the detailed discussion of a similar case in Section 5.1 below), (57) might not be so far off the mark. For in these circumstances animals (like humans) probably switch from sufficiency-reasoning (seeking means that are sufficient to achieve their ends) to necessity-reasoning (selecting, at each stage in a complex task, the option that they deem to be *necessary* to achieve their ends, in the hope that future problems can be solved as they arise). In any case however, whatever their precise form, the first-position counterparts of (46) through (49) will roughly balance one another in cases of uncertainty, leaving the desire to use the ‘hint’ option unopposed. I conclude, therefore, that the data provided by Kornell *et al.* (2007) fail to support a meta-cognitive interpretation.

5. Memory and Meta-Memory

Hampton (2001) devised and executed a particularly elegant and seemingly powerful test for meta-memory in animals. (The tests included a number of control conditions to rule out a variety of alternative hypotheses, which I shall not describe here.) The basic arrangements were that the animals (in this case, monkeys) first had the opportunity to study a screen on which an image was displayed. There was then a delay, designed to be long enough so that the animals would often, but not always, forget what had been on the screen. There were then two conditions. In one third of the trials, the monkeys were presented with a screen containing a distinctive 'advance to the next screen' symbol. They had to touch this symbol to advance to the test phase of the experiment, during which four different symbols were displayed on the screen at once, one of which was always the symbol that had been displayed in the initial study phase of the trial. Their task was to identify and touch this familiar symbol. A correct choice led to a preferred reward (a peanut). An incorrect choice led, in effect, to a time out before there could be another opportunity for a reward, since the animal would have to sit through the study phase of a new trial, followed by a delay.

In the other condition, which occurred in two thirds of the trials, the animal was presented with a screen containing two symbols following the delay phase. One was the 'advance' symbol, touching which would advance to the same test phase as above. The other was, in intention, a 'don't know' symbol, touching which advanced to a screen containing just a single symbol, which when touched was guaranteed to issue in a less-favored reward (a primate pellet). Pressing 'don't know' would allow animals to escape from taking any test that they lacked the knowledge to answer, provided that they were capable of pressing 'don't know' when, and only when, they had forgotten the initial stimulus.

Hampton (2001) was interested in the extent to which the monkeys' performance was better in the free-choice trials than in the forced-choice trials. He predicted that, if the monkeys had meta-memory awareness of their own states of knowledge or ignorance, then they should make adaptive use of the 'don't know' symbol, thereby increasing the proportion of tests answered correctly. And this is just what he found. What I shall argue, however, is that the data admit of a first-order, non-meta-cognitive, explanation.

5.1 The Forced-Choice Case

Let me start by detailing the beliefs and desires that will be present during one of the forced-choice trials (when the animal is confronted by the 'advance' symbol alone). They are as follows.

(58) BELIEF$_S$ [if the image that was on the screen at the start of the trial is also on the next screen, and it is touched, then a peanut results].

(59) BELIEF$_S$ [if the image that was on the screen at the start of the trial is also on the next screen, and it isn't touched, then a time out results].

(60) BELIEF$_{S}$ [if the ‘advance’ symbol is touched, then the image that was on the screen at the start of the trial will also be on the next screen].
(61) DESIRE$_{S}$ [peanut].
(62) DESIRE$_{S}$ [no time out].

These beliefs and desires, taken together, don’t license any specific action in the absence of information about what image was, actually, on the screen at the start of the trial. So how is it that in the forced-choice trials when the animal has forgotten what was on the screen it will nevertheless touch ‘advance’, moving the screen to the choice stage? And why will it touch anything on the screen when it gets to that stage? This is because the animal knows that if it does nothing, nothing at all will happen, and there will be an extended delay before there is any opportunity to eat again. So the animal has, in addition, the following three mental states.

(63) BELIEF$_{S}$ [if ‘advance’ isn’t touched, then there will be an extended time out].
(64) BELIEF$_{S}$ [if nothing on the next screen is touched, then there will be an extended time out].
(65) DESIRE$_{S}$ [no extended time out].

The two states (63) and (65) alone are sufficient to motivate touching the ‘advance’ symbol, moving the animal on to the test screen. And then (64) combined with (65) is sufficient to motivate the animal to touch *something* on the screen (at random).

Now consider a case where the animal has a memory of the item that was on the screen at the start of the trial. (And suppose, for concreteness, that the image was a picture of a chicken.) Then the following belief can be added to the above set.

(66) BELIEF$_{S}$ [the image that was on the screen at the start of the trial was a chicken].

This will then warrant (by substitution) determinate versions of all of the indeterminate beliefs in the previous list—(58) through (60)—issuing in the following set.

(67) BELIEF$_{S}$ [if the image of a chicken is on the next screen, and it is touched, then a peanut results].
(68) BELIEF$_{S}$ [if the image of a chicken is on the next screen, and it isn’t touched, then a time out results].
(69) BELIEF$_{S}$ [if the ‘advance’ symbol is touched, then the image of a chicken will be on the next screen].

Desire (61), for a peanut, in combination with beliefs (67) and (69) is sufficient to motivate touching the ‘advance’ symbol, thereafter motivating the animal to touch the image of the chicken when it appears. That is, those states issue in the following.

(70) DESIRE$_S$ [‘advance’ is touched, and then the image of a chicken is touched].

So what we can conclude so far is that, in a case where the animal remembers what was on the initial screen, the action of touching ‘advance’ is over-determined. Since nothing at all will happen if it doesn’t do so, it has a desire to touch ‘advance’, resulting from (63) and (65). But it also has a desire to touch ‘advance’ resulting from states (61), (67), and (69). This is hardly surprising. For in the forced-choice trials the animal has no choice but to press ‘advance’, whether it remembers the initial stimulus or not. That is why these trials are described as ‘*forced* choice’ ones.

5.2 The Free-Choice Trials

Now consider what happens in trials where the animal is given the choice of taking the memory test (pressing ‘advance’ to move to a screen on which the original symbol plus three distracters will be displayed) or pressing ‘don’t know’ to advance to a screen that guarantees a less favored food item (a primate pellet). And consider, first, a case where the animal has no memory of what was displayed on the start-up screen. Then the animal still has states (58) through (62), which still fail to motivate any action in particular. And the animal still has (64) and (65), which will motivate it to touch *something* on the screen that follows when it appears. But in place of (63)—since it now faces a choice of symbols that it might press—it now has this:

(71) BELIEF$_S$ [if one of the two symbols on the screen isn’t touched, then an extended time out will result].

Belief (71) in conjunction with desire (65) issues in (72), which in turn warrants the two weak desires (73) and (74).

(72) DESIRE$_S$ [press one of the two symbols on the screen].
(73) DESIRE$_W$ [press ‘advance’].
(74) DESIRE$_W$ [press ‘don’t know’].

Thus far the animal has no motive for touching one symbol rather than the other. But it also has the following states.

(75) BELIEF$_S$ [if ‘don’t know’ is pressed then a pellet results from pressing anything on the next screen].
(76) DESIRE$_W$ [pellet].

States (75) and (76) together also issue in a distinct token of desire (74). On the assumption that desires are additive (in such a way that the presence of two distinct desires motivating a single action issue in a desire to perform that action that is stronger than results from either one alone), then this will be sufficient to motivate the animal to press ‘don’t know’.

Now consider what happens in one of these choice trials when the animal remembers what was displayed on the start-up screen. (And suppose, again, for concreteness, that what was displayed was the image of a chicken, in which case it has belief (66).) Then here, as before, the animal will have the specific versions of beliefs (58) and (60), namely (67) and (69). And these, together with the desire (61) for a peanut will lead it to have (70), a strong desire first to touch 'advance', and then to touch the image of the chicken on the resulting screen. This will presumably outweigh the weak desire that the animal possesses to touch 'don't know', resulting from its weak desire for a less-desirable pellet. And so the animal will press 'advance', moving it to the test phase.[14] And when it gets to the test phase, it will press the symbol of the chicken.

5.3 Discussion

There is nothing overtly meta-cognitive in the account provided above. Yet that account is just as well capable of explaining the facts that can be inferred from the behavioral data, namely that in general the monkeys will press 'don't know' when they no longer remember what was displayed in the study phase of the trial, and that in general they will press 'advance' when they do. And yet all that this account appeals to are plausible claims about the operations of belief–desire reasoning, which any meta-cognitive theorist would need to appeal to as well. So on the face of it, my proposed first-order account is to be preferred. Elegant as Hampton's (2001) experiments are, they don't seem to support the attribution of meta-cognitive states to animals.[15]

I can imagine someone objecting that my account is, actually, covertly meta-cognitive, only with the meta-memory component disguised. For Hampton (2005) makes the point that the delay following the study phase is too great for the knowledge of the stimulus to be maintained in working memory. So it must have been committed to medium or long-term memory, which means that it then needs to be re-activated at the choice phase. And it is surely no accident that this knowledge should become activated in choice trials, thus motivating the animal to touch the 'advance' symbol and advance to the test phase to receive a favored reward. In my explanation I spoke blithely of the animal 'remembering what was displayed on the start-up screen'. But what would cause it to remember this? There are no end of items of irrelevant information that the animal *could* bring to mind. What needs to be explained is how the animal activates the right memory, unless it actively *searches* for that memory—which would be a meta-cognitive search.

[14] While the animal also has a weak desire to touch 'don't know' motivated by its desire to touch *something*—i.e. motivated by (72)—it equally has a weak desire to touch 'advance' motivated by that desire. So these desires balance one another.

[15] Hampton (2005) presents one further argument in support of his own meta-cognitive interpretation of his data. This is that the additional cognitive demands imposed by meta-cognition can explain why pigeons should fail at the same memory-monitoring tasks that monkeys pass, under essentially the same conditions that were employed in Hampton's (2001) experiments (Inman and Shettleworth, 1999; Shettleworth and Sutton, 2006). But there are surely multitudes of alternative explanations for this difference in behavior across species.

I fully accept that the activation of the required information (such as that there had been an image of a chicken on the start-up screen) doesn't happen by accident. But I don't believe that it requires anything meta-cognitive, either. Rather, the desire for a favored food-reward (61), together with the indeterminate-content beliefs about the background conditions of the experiment (58) through (60), motivate a search for the information required to render those beliefs determinate. There is nothing magical about this. It is surely a general fact about how memory functions that searches for specific forms of information are often guided by more general information. Consider, for example, a food-caching bird that has a desire to eat. In addition, it knows that if it goes to a cache that hasn't been previously emptied then it will be able to eat (Clayton *et al.*, 2006). This motivates a search for information under that description. The bird, as it were, has to ask itself, 'Where is a cache that hasn't been emptied?' For only the answer to this question will enable it to claim the reward. Likewise, the monkeys in Hampton's (2001) experiments have to search for the specific information concerning what was displayed on the start-up screen. The monkeys, as it were, have to ask themselves, 'What was on the screen at the start of the trial?' (Note: they *don't* need to ask themselves whether they *remember* what was on the screen. And the answer to that question won't help them anyway, unless they can activate the specific memory involved. This is a memory that has a first-order content.)

Let me put the same point slightly differently. The monkeys in a choice trial know what, in general, they have to do to obtain a favored food reward (a peanut). They know that they have to press 'advance' to move to the test phase and then press the same symbol that was displayed on the screen during the study phase. It is part of the way memory systems operate, surely, that general knowledge of this sort can initiate a search for the specific instance of knowledge needed to achieve the goal. But the search isn't meta-cognitive in character. (Not unless all searches of memory are meta-cognitive. But this would be an absurd claim, entailing meta-cognitive processes in all creatures possessing any sort of medium or long-term memory.) It is rather a first-order search for a stored item of information, namely one that would answer the question, 'What was on the screen?'

It may also be worth pointing out that although the account above has deployed propositional descriptions, actually nothing specific is here assumed about the forms in which memories are stored, or about the forms that they take when activated. A memory of the last item to appear on the screen might be stored as a propositional, quasi-sentential, representation. Or it might be stored in the form of a mental model or image of some sort. And likewise when activated, what might come to mind could be a fully-conceptual thought with the content, *what was on the screen was a chicken*. Or it might be a visual image *as of* a chicken on the screen. The *ceteris paribus* laws governing the ways in which such beliefs interact with other belief states and with desires can be pretty much as described above, either way.

It is also worth pointing out that the story that I have been telling might be expanded in such a way that it needn't assume that memory activation is an all-or-nothing affair. For (given the way in which memory operates in humans) one can predict that the result of the memory search might only be a highly indeterminate image of a bird of some sort, or it might be a propositional thought with the content, *what was on the screen was some kind of bird*. Since this information isn't fully determinate, it doesn't allow the animal to derive analogs of (67) to (69) from (58) through (60) by substitution. Thus one cannot, in such a case, replace (67) by the following.

(77) BELIEF$_S$ [if the image of some kind of bird is on the next screen, and it is touched, then a peanut results].

This is because more than one type of bird could figure on the test screen, and this is something that the animal might know. But in that case the animal is also likely to have a belief about the rough likelihood of this happening. That is, it might believe something like the following.

(78) BELIEF$_S$ [only rarely have images of two types of bird occurred together on the test screen].

In place of (77), then, (78) would warrant the animal in deducing the following from (58) by substitution.

(79) BELIEF$_S$ [if the image of some sort bird is on the next screen, and it is touched, then a peanut is *likely to* result].

And this in turn, in interaction with the animal's other beliefs and desires, would then give rise to a desire to touch the 'advance' screen that is strong enough to outweigh its weak desire to touch 'don't know' to obtain the primate pellet.[16]

6. General Discussion

I have claimed that all of the evidence of meta-cognition in animals that has been adduced so far can be explained in first-order, non-meta-cognitive, terms. Each of the explanations invokes beliefs and desires of varying strengths interacting with one another in accordance with plausible principles of practical reasoning. And importantly, the states and processes appealed to in these explanations are of a sort that any meta-cognitive theorist would need to accept in any case. But a variety of additional mechanisms and capacities have also been postulated. One is a 'gate-keeping' mechanism that causes hesitancy when desires for contrary actions are too

[16] For this to work, the inferential principle that takes the animal from a conditional belief together with a desire for the consequent of that belief to a desired action must be sensitive to *likelihood* information that figures in the conditional. But this isn't at all implausible. We know that many species of animal are extremely good at adjusting their behavior (as a result, presumably, of making appropriate adjustments in the strength of their desires to execute those behaviors) in the light of *rate* or *likelihood* information of various sorts (Gallistel, 2000).

close to one another in strength; another is a capacity to derive a general rule from the training conditions; and yet another is a disposition to activate specific items of information from memory guided by general knowledge. Hence it might be argued that in this respect, at least, a meta-cognitive explanation of the behavior in question would have an advantage, since it need only appeal to a single underlying capacity. But in fact there is no real advantage here, provided that each of the postulated first-order mechanisms and capacities is independently well-motivated. I have argued that this is the case.

I have also claimed that the forms of human behavior that parallel so-called 'meta-cognitive' behavior in animals, and which human subjects themselves will often explain in meta-cognitive terms, aren't really meta-cognitive in nature. Am I saying, then, that humans *never* undergo meta-cognitive thoughts or processes? Of course not. On the contrary: on every occasion on which a human formulates a meta-cognitive explanation, a meta-cognitive thought will thereby have been entertained (i.e. the very thought that figures in the explanation, or in the verbal expression of that explanation). So there is no doubt at all that humans often entertain meta-cognitive thoughts. Nor is there any doubt that humans often engage in processes of reasoning that are guided by such thoughts. For instance, many of those who study human reasoning think that the processes involved can be divided into two broad classes, often described as 'System 1' (whose components are ancient, swift, and shared with other species of animal) and 'System 2', which is slow, reflective, and distinctively human (Evans and Over, 1996; Stanovich, 1999; Kahneman, 2002; Frankish, 2004). And likewise all are agreed that System 2 reasoning is shot through with meta-cognitive thoughts and beliefs.

If animals don't engage in this sort of System 2 reasoning, and if the tests of meta-cognition in animals that have been employed thus far have actually been tests of first-order reasoning processes, then where is one to look? This is a hard question. Compare what happened in the aftermath of the question first raised by Premack and Woodruff (1978), concerning whether chimpanzees possess a 'theory of mind'. (Which is to say: do chimpanzees entertain meta-cognitive thoughts about the mental lives of others?) This has been a subject of intensive research and debate ever since, and the question is still not fully resolved (Povinelli, 2000; Hare *et al.*, 2001, 2003; Povinelli and Vonk, 2003; Tomasello *et al.*, 2003a, 2003b).[17] It has

[17] There are obvious points of affinity between my present critique of the animal meta-cognition literature and Povinelli and Vonk's (2003) critique of the animal mind-reading literature. For each claims that the existing data can be explained in first-order terms. But there are also important contrasts. One is that Povinelli and Vonk demand decisive *proof* of mind-reading. They will accept nothing less than data that can *only* be explained in higher-order terms. I don't set the bar so high. I would be satisfied if I could be shown data for which the *best* explanation is meta-cognitive. Another, related, point is that Povinelli and Vonk are forced to attribute knowledge of behavioral invariants of an extremely subtle sort to the animals in the experiments, for which we lack any independent evidence (as Tomasello *et al.*, 2003b, point out). The first-order thoughts and inferential procedures attributed to animals in the course of the present article, in contrast, are all independently plausible, and shouldn't be controversial.

proven remarkably difficult to devise convincing tests of mind-reading that can be administered to animals. And there is no reason to think that testing for an animal's capacity for meta-cognition should prove any easier. Indeed, quite the reverse, as I shall briefly explain.

The consensus that emerged from the commentary on Premack and Woodruff (1978) was that a test of mind-reading should pit the animal's own beliefs about the world against those that need to be ascribed to the target in order to predict or explain its behavior. In effect, a convincing test of a creature's capacity to attribute thoughts to another would need to focus on its capacity to attribute *false* thoughts. For otherwise it will be too easy for the creature's behavior to be explicable in light of its own first-order apprehension of the circumstances, together with first-order beliefs about behavioral contingencies and so forth. Thus was 'the false belief test' born (Wimmer and Perner, 1983), which has proven a very fruitful research tool with human infants, if not so easy to use with animals. But in the nature of the case it will be hard to test for an animal's grasp of its *own* (current) false belief. For in the very act of ascribing a false belief to itself it would thereby lose that belief. Put differently, there is a contradiction involved if a creature both believes that *P* and believes that its belief that *P* is false; for both of these beliefs cannot be true together. The best that we could probably hope for are tests of an animal's capacity to ascribe *past* false beliefs to itself. (There is no awkwardness in someone believing that their *previous* belief that *P* is false, provided that they now believe something else.)

I am not saying, of course, that there *cannot* be experimental evidence of animals' capacity for meta-cognition. That would be foolish. One shouldn't try to second-guess the ingenuity of experimental scientists. But I am saying that it will be at least as hard as it is to find evidence that animals are capable of mind-reading, and probably a good deal harder, for the reason just given. It is possible, however, that investigating whether animals are capable of thoughts about their own *perceptual* states may prove a bit more tractable. This is because there is nothing incoherent in believing that one is in a perceptual state with the content [that P] while believing that *P* is not the case. Indeed, we do this all the time when we judge that although it *seems* to us that *P*, in fact *not P*. This means that it may be possible to pit behavior guided by a current belief against behavior that is guided by a belief about current perception. If I were a betting man, I would place my money here. But for the reasons indicated in Section 1, I think that it is unlikely that we will find evidence of meta-cognition in animals that are incapable of mind-reading. So I would concentrate my search on chimpanzees rather than monkeys, where the case for animal mind-reading (and mind-reading about perceptual states, in particular) is at its strongest (Hare *et al*., 2001, 2003; Tomasello *et al*., 2003a, 2003b).

Department of Philosophy
University of Maryland

References

Baron-Cohen, S. 1995: *Mindblindness*. Cambridge, MA: MIT Press.

Bem, D. 1967: Self-perception: an alternative interpretation of cognitive dissonance phenomena. *Psychological Review*, 74, 183–200.

Bem, D. 1972: Self-perception theory. In L. Berkowitz (ed.), *Advances in Experimental Social Psychology*, volume 6. New York: Academic Press.

Beran, M., Smith, J., Redford, J. and Washburn, D. 2006: Rhesus Macaques (*Macaca mulatta*) monitor uncertainty during numerosity judgments. *Journal of Experimental Psychology: Animal Behavior Processes*, 32, 111–119.

Bermúdez, J. 2003: *Thinking without Words*. Oxford: Oxford University Press.

Call, J. and Carpenter, M. 2001: Do apes and children know what they have seen? *Animal Cognition*, 4, 207–220.

Carruthers, P. 2006: *The Architecture of the Mind: Massive Modularity and the Flexibility of Thought*. Oxford: Oxford University Press.

Carruthers, P. forthcoming: Introspection: divided and partly eliminated. *Philosophy and Phenomenological Research*.

Clayton, N., Emory, N. and Dickinson, A. 2006: The rationality of animal memory: the cognition of caching. In S. Hurley and M. Nudds (eds), *Animal Rationality*. Oxford: Oxford University Press.

Davidson, D. 1982: Rational animals. *dialectica*, 36, 317–327.

Eagly, A. and Chaiken, S. 1993: *The Psychology of Attitudes*. New York: Harcourt Brace.

Evans, J. and Over, D. 1996: *Rationality and Reasoning*. Brighton: Psychology Press.

Festinger, L. 1957: *A Theory of Cognitive Dissonance*. Palo Alto: Stanford University Press.

Frankish, K. 2004: *Mind and Supermind*. Cambridge: Cambridge University Press.

Gallistel, R. 2000: The replacement of general-purpose learning models with adaptively specialized learning modules. In M. Gazzaniga (ed.), *The New Cognitive Neurosciences*, 2nd edn. Cambridge, MA: MIT Press.

Giurfa, M., Zhang, S., Jenett, A., Menzel, R. and Srinivasan, M. 2001: The concepts of 'sameness' and 'difference' in an insect. *Nature*, 410, 930–933.

Goldman, A. 1993: The psychology of folk psychology. *Behavioral and Brain Sciences*, 16, 15–28.

Goldman, A. 2006: *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mind-reading*. Oxford: Oxford University Press.

Gopnik, A. and Meltzoff, A. 1997: *Words, Thoughts, and Theories*. Cambridge, MA: MIT Press.

Gordon, R. 1996: Radical simulationism. In P. Carruthers and P. Smith (eds), *Theories of Theories of Mind*. Cambridge: Cambridge University Press.

Hampton, R. 2001: Rhesus monkeys know when they remember. *Proceedings of the National Academy of Sciences*, 98, 5359–5362.

Hampton, R. 2005: Can Rhesus monkeys discriminate between remembering and forgetting? In H. Terrace and J. Metcalfe (eds), *The Missing Link in Cognition: Origins of Self-reflective Consciousness*. Oxford: Oxford University Press.

Hampton, R., Zivin, A. and Murray, E. 2004: Rhesus monkeys (*Macaca mulatta*) discriminate between knowing and not knowing and collect information as needed before acting. *Animal Cognition*, 7, 239–246.

Hare, B., Addessi, E., Call, J., Tomasello, M. and Visalberghi, E. 2003: Do capuchin monkeys, *Cebus paella*, know what conspecifics do and do not see? *Animal Behavior*, 65, 131–142.

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

Inman, A. and Shettleworth, S. 1999: Detecting meta-memory in nonverbal subjects: a test with pigeons. *Journal of Experimental Psychology: Animal Behavior Processes*, 25, 389–395.

Kahneman, D. 2002: Maps of bounded rationality: a perspective on intuitive judgment and choice. Nobel laureate acceptance speech. Available at: http://nobelprize.org/economics/laureates/2002/kahneman-lecture.html

Kornell, N., Son, L. and Terrace, H. 2007: Transfer of metacognitive skills and hint seeking in monkeys. *Psychological Science*, 18, 64–71.

Kral, K. and Poteser, M. 1997: Motion parallax as a source of distance information in locusts and mantids. *Journal of Insect Behavior*, 10, 145–163.

Metcalfe, J. and Kober, H. 2005: Self-reflective consciousness and the projectable self. In H. Terrace and J. Metcalfe (eds), *The Evolution of Consciousness*. Oxford: Oxford University Press, 57–83.

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

Nisbett, R. and Wilson, T. 1977: Telling more than we can know. *Psychological Review*, 84, 231–295.

Povinelli, D. 2000: *Folk Physics for Apes: The Chimpanzee's Theory of How the World Works*. Oxford: Oxford University Press.

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

Premack, D. and Woodruff, G. 1978: Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 4, 515–526.

Proust, J. 2006: Rationality and meta-cognition in non-human animals. In S. Hurley and M. Nudds (eds), *Rational Animals?* Oxford: Oxford University Press.

Ratcliff, R. and Rouder, J. 1998: Modeling response times for two-choice decisions. *Psychological Science*, 9, 347–356.

Shettleworth, S. and Sutton, J. 2006: Do animals know what they know? In S. Hurley and M. Nudds (eds), *Rational Animals?* Oxford: Oxford University Press.

Shields, W., Smith, J. and Washburn, D. 1997: Uncertain responses by humans and rhesus monkeys (*Macaca mulatta*) in a psychophysical same–different task. *Journal of Experimental Psychology: General*, 126, 147–164.

Smith, J. 2005: Studies of uncertainty monitoring and meta-cognition in animals and humans. In H. Terrace and J. Metcalfe (eds), *The Missing Link in Cognition: Origins of Self-reflective Consciousness*. Oxford: Oxford University Press.

Smith, J. and Schull, J. 1989: A failure of uncertainty monitoring in the rat. Unpublished data; described in Smith (2005).

Smith, J., Schull, J., Strote, J., McGee, K., Egnor, R. and Erb, L. 1995: The uncertain response in the bottlenosed dolphin (*Tursiops truncatus*). *Journal of Experimental Psychology: General*, 124, 391–408.

Smith, J., Shields, W., Schull, J. and Washburn, D. 1997: The uncertain response in humans and animals. *Cognition*, 62, 75–97.

Smith, J., Shields, W. and Washburn, D. 2003: The comparative psychology of uncertainty monitoring and meta-cognition. *Behavioral and Brain Sciences*, 26, 317–373.

Son, L. and Kornell, N. 2005: Meta-confidence judgments in Rhesus Macaques: explicit versus implicit mechanisms. In H. Terrace and J. Metcalfe (eds.), *The Missing Link in Cognition: Origins of Self-reflective Consciousness*. Oxford: Oxford University Press.

Stanovich, K. 1999: *Who is Rational? Studies of Individual Differences in Reasoning*. Brighton: Lawrence Erlbaum.

Terrace, H., Son, L. and Brannon, E. 2003: Serial expertise of rhesus macaques. *Psychological Science*, 14, 66–73.

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

Tomasello, M., Call, J. and Hare, B. 2003b: Chimpanzees versus humans: it's not that simple. *Trends in Cognitive Sciences*, 7, 153–156.

Tye, M. 2006: Another look at representationalism about pain. In M. Aydede (ed.), *Pain: New Essays on its Nature and the Methodology of its Study*. Cambridge, MA: MIT Press.

Washburn, D., Smith, J. and Shields, W. 2006: Rhesus monkeys (*Macaca mulatta*) immediately generalize the *uncertain* response. *Journal of Experimental Psychology: Animal Behavior Processes*, 32, 185–189.

Wellman, H. 1990: *The Child's Theory of Mind*. Cambridge, MA: MIT Press.

Wicklund, R. and Brehm, J. 1976: *Perspectives on Cognitive Dissonance*. Brighton: Lawrence Erlbaum.

Wilson, T. 2002: *Strangers to Ourselves*. Cambridge, MA: Harvard University Press.

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