

The Evolution of Self-Knowledge

Peter Carruthers
University of Maryland, College Park

Logan Fletcher
University of Maryland, College Park

J. Brendan Ritchie
University of Maryland, College Park

ABSTRACT. Humans have the capacity for awareness of many aspects of their own mental lives—their own experiences, feelings, judgments, desires, and decisions. We can often know what it is that we see, hear, feel, judge, want, or decide. This article examines the evolutionary origins of this form of self-knowledge. Two alternatives are contrasted and compared with the available evidence. One is first-person based: self-knowledge is an adaptation designed initially for metacognitive monitoring and control. The other is third-person based: self-knowledge depends on the prior evolution of a mindreading system which can then be directed toward the self. It is shown that the latter account is currently the best supported of the two.

I. INTRODUCTION

There are a number of kinds of self-knowledge. One is awareness of oneself as a bodily agent, as established, for example, by the so-called “mirror test” (Gallup

1970). Another is knowledge of one's own traits and dispositions, such as generosity or friendliness. A third—much more demanding—concerns awareness of oneself as an ongoing bearer of mental states and dispositions, who has both a past and a future. In its most demanding form, this kind of self-knowledge seems to require a conception of oneself as a *self*, together with a capacity for narrative, weaving one's current thoughts and experiences into a larger story of one's life.

Our focus in this article is on none of these, but rather knowledge of one's own current mental states: one's judgments, beliefs, desires, values, decisions, intentions, experiences, and emotions. Humans undoubtedly enjoy such self-awareness. We don't just see, we are aware *that* we see and *what* we see; we don't just hear, we are aware *that* we hear and *what* we hear; and so on. And we often know what we think, want, decide, or fear. Our question concerns the evolutionary roots of these capacities for self-knowledge.

While interesting, bodily self-knowledge has little to do with awareness of oneself as a cognitive being. Rather, the mirror test measures an ability to notice cross-modal contingencies, becoming aware of the mapping between one's own bodily movements (as experienced proprioceptively) and what one perceives in the mirror (Ritchie and Carlson 2010). Knowledge of one's own current mental states, however, is arguably more fundamental than either knowledge of one's traits or knowledge of oneself as a self with an ongoing mental life. The latter point is quite obvious: one surely could not have knowledge of one's past or future mental life if one didn't also know of some of one's *current* mental states. The reverse, however, seems not to hold: one might know of one's current mental states without being capable of knowledge of past or future mental states, or of weaving them together into a narrative. Similarly, it seems likely that knowledge of one's own traits of character (if it is not to be derived entirely from the testimony of others) depends on knowledge of current mental states, but not vice versa. In any case, our focus here is on the evolutionary origins of our knowledge of our own mental states in the present. Henceforward when we speak of "self-knowledge" it is this variety that we should be understood to mean.

In this paper we will assume that capacities for self-knowledge are rooted in some kind of distinct adaption in addition to general learning abilities. This assumption is not uncontroversial. Some might be tempted to endorse empiricism about concepts and concept acquisition, for example (Prinz 2002), while claiming that the classifications that we make among our own mental states and the knowledge that we have of their patterns of interaction and contributions to behavior are a product of general learning (whether associative, or involving some sort of inference to the best explanation, or both). This account strikes us as quite implausible. But for present purposes we will simply assume, without argument, that it is false.

One can then envisage two broad accounts of the evolution of a capacity for self-knowledge. One is first-person based. It is that self-knowledge evolved for purposes of metacognitive monitoring and control. On this account, organisms evolve a capacity for self-knowledge in order better to manage and control their own men-

tal lives. By being aware of some of their mental states and processes, organisms can become more efficient and reliable cognizers, and can make better and more adaptive decisions as a result.¹

The first-person-based view is consistent with a range of accounts of the cognitive capacities or mechanisms underlying self-knowledge. At one extreme are those who believe in mechanisms of so-called inner sense (Goldman 2006; Nichols and Stich 2003). Just as our regular senses detect, and enable us to have knowledge of, properties of the external world and of our own bodies, so inner sense is supposed to enable us to detect and have knowledge of our own mental lives. At the other extreme one might postulate just a body of innate core knowledge, similar to the knowledge proposed in the domains of physics and number (Spelke and Kinzler 2007). The system that deploys this knowledge would have access to just the same “globally broadcast” perceptual and imagistic information as do other conceptual systems (Baars 1988) and would lack any special channels of access to the subject’s own nonsensory mental states. The system in question might just contain a set of representational primitives like THINKS and WANTS, together with some basic inferential principles to enable their application to oneself and to predict the impact of some simple self-directed interventions. (On some accounts, such resources might come paired with an innate domain-specific learning system directed toward the self; or it might be maintained instead that all or most subsequent learning about the self is of a general sort.)

The first-person-based view is also consistent with a range of accounts of the relationship between self-knowledge and third-person mindreading. On one view, it might be claimed that the mechanisms of inner sense are exapted and used when simulating the minds of others, in such a way that capacities for mindreading depend upon our capacity for self-knowledge (Goldman 2006). Likewise it might be claimed that the core knowledge that underlies self-knowledge is redeployed (either by evolution or by individual learning) to provide the basis for third-person mindreading. Alternatively, it might be claimed that capacities for self-knowledge and for mindreading are independent of one another (Nichols and Stich 2003).

Since theories are stronger (less open to attack) that make fewer assumptions, our focus in this article will be on a minimalist “core knowledge” first-person-based account of the adaptive basis of self-knowledge, which makes no claim to explain the basis of mindreading. Hence the first-person-based account to be considered here holds that self-knowledge and mindreading are independent capacities. Moreover, the account of self-knowledge in play is consistent with the “interpretive sensory-access” (ISA) theory defended by Carruthers (2011) and is not directly targeted by the critiques of other views that are mounted in that work. Indeed, the assumptions made by a “core knowledge” account are significantly more minimal than any that are made in the existing literature. So our focus will be on what seems to be the strongest, most defensible, form of the first-person-based account.

The contrasting account of the evolution of self-knowledge is third-person based. It maintains that the adaptation underlying the capacity for knowledge of

one's own mental states is a mindreading faculty (consisting of a body of core knowledge about the mind, or a domain-specific learning mechanism with representational primitives, or both), which evolved initially for social purposes (Carruthers 2011). These purposes might be competitive, as “Machiavellian intelligence” accounts of the evolution of mindreading maintain (Byrne and Whiten 1988, 1997), or cooperative (Hrdy 2009; Richerson and Boyd 2005), or both. The mindreading faculty would have access to globally broadcast perceptual and imagistic representations as input, and attributions of mental states to oneself would initially utilize this input together with the same core knowledge and principles that are employed for third-person mindreading. (Some first-person principles might subsequently be learned, of course.) In effect, self-knowledge results from turning our evolved mindreading capacities on ourselves.²

In what follows we will compare the empirical predictions made by these first-person-based and third-person-based accounts, and confront them with the available data. Section 2 will focus on the expected signature effects of the adaptations that these theories postulate, before section 3 turns to evidence from comparative psychology.³

II. ADAPTIVE SIGNATURES

In general, the most basic prediction made by a hypothesis that some universal phenotypic character is an adaptation is that it should be good at what it does. The property in question should enable the organism or subsystem to do *well* what it was allegedly selected *for*. This is not to say that performance should be optimal, of course, since there may be other constraints in operation (such as energetic costs) that exert an opposing selective pressure, and since a property only needs to yield small adaptive benefits to become a target of selection. But it would seem, at least, that evidence of good performance by a phenotypic character that is universal to the species is evidence (albeit defeasible) of the presence of an adaptation; and evidence of poor performance is evidence (again defeasible) of the absence of an adaptation. Moreover, if performance is poor but the character in question is nevertheless an adaptation, then there should be an explanation, framed in terms of competing pressures or architectural constraints, for why performance is not better.

The expected timing of the developmental emergence of a supposed adaptive character is slightly more subtle. Plainly, not all adaptations are early to emerge in development. Obvious examples include sexual characteristics like breasts or pubic hair, which only make their appearance around puberty. But one might expect that an adaptive character should emerge in development as soon as it is needed or would prove useful—unless, again, there are constraints or pressures to the contrary. Hence evidence that some universal phenotypic character comes online as soon as it can confer a benefit (given the developmental timing of other relevant

components of the phenotype) provides evidence that the property is an adaptation. And the finding that a character doesn't emerge until well after it might have been useful is evidence against it being an adaptation in the absence of some explanation of its tardiness in developing.

Finally, if some universal phenotypic character is an adaptation, then it should be robust in the face of environmental and developmental variation. The emergence of the property in question should be "buffered" against variations in the environment or in the process of development itself, reliably emerging despite the presence of such variation. This is not to say, of course, that adaptive properties must emerge irrespective of environmental input, as well as in highly perturbed developmental trajectories. But we should at least expect them to emerge reliably in circumstances that are normal for the phenotype.

We turn, now, to apply these points, providing a comparative evaluation of third-person-based and first-person-based accounts of the evolutionary adaptation underlying our capacity for self-knowledge. The predictions of the two accounts will be discussed first, before evidence of human abilities is considered.

2.1. PREDICTIONS OF THE THIRD-PERSON-BASED ACCOUNT

What matters most for our purposes are the predictions made by the third-person-based account concerning human capacities for metacognitive control. But we begin by considering what the account should predict about capacities for other-person mindreading. These will form an important counterpoint for the predictions of a first-person-based account, to be considered in section 2.2.

Humans are a highly social species, of course. Indeed, the degree of their social interdependence is unparalleled in the animal kingdom, with the exception only of the eusocial insects. Moreover, much of that interdependence seems to depend, in whole or in part, upon mindreading. Humans need to be able to read the intentions of others in the surrounding culture in order to acquire the skills and norms that they require for successful functioning as a member of their group. They need to read the intentions and trustworthiness of others when negotiating cooperative agreements or listening to the testimony of others. And when competing with others, whether in love or war, they need to be able to second-guess their opponent's moves, which in turn will require them to predict what their opponent is likely to be thinking. In addition, it is generally recognized that the pragmatic components of speech, which form a ubiquitous aspect of human communication, depend upon successful mindreading (Sperber and Wilson 2002).

It would seem, then, that the pressures toward successful mindreading would have been quite intense, at least in the hominin lineage. We are a cultural species, and culture depends, in large part, on mindreading. We should predict, then, both that there *is* an adaption for mindreading, and that it should underwrite a high level of performance. We will return to examine the truth of this prediction in section 2.4.

As for the predicted timing of the emergence of mindreading in development, our initial (“first pass”) answer should depend on when mindreading abilities would prove useful. Considered in the abstract the answer would seem to be: as early as possible, consistent with other facts about human development. Learning of vocabulary is known to depend upon sensitivity to others’ referential intentions (Bloom 2002), and infants need to be able to judge the intentions and likely cooperativeness of others as soon as they are able to move around independently of their mothers. Both facts should lead us to predict that basic mindreading skills would be in place in the second year of life.

We can predict, too, that mindreading capacities should reliably emerge in infants and children in similar fashion and issue in similar degrees of success across cultures, and irrespective of variations in opportunities for individual learning. Those capacities should also be buffered to emerge in children whose development is in other respects not normal.

But what should a third-person-based account predict about the timing and degree of success of the metacognitive skills and abilities that depend on self-knowledge? Since self-knowledge itself is held to result from self-directed mindreading, utilizing the same channels of sensory information that are available for third-person mental state attribution, we should predict that infants will be *capable* of self-knowledge as soon as they become capable of third-person mindreading. And we should predict, too, that they should be roughly as good at it (at least to the extent that self-attribution and other-attribution share the same resources). If people are successful mindreaders, and self-knowledge results from one turning one’s mindreading abilities on oneself, then to a first approximation we should predict that people will also be successful self-attributors.

A third-person-based account of the evolution of self-knowledge thus predicts that people should be good at *monitoring* (and self-attributing) their own mental states and should begin to do so quite early in development. But what of *control*? Should we likewise predict that people will be good at controlling the course of their own mental lives, intervening in their own thought processes in adaptive ways from an early age? This is *not* something that a third-person-based account should predict. For this is not what self-knowledge is for. (Indeed, self-knowledge isn’t *for* anything, according to the demanding form of third-person-based account we are operating with here.) On the contrary, we should predict that to the extent that people are capable of controlling their own mental lives at all, this will be a cobbled-together skill that depends upon individual and cultural learning (and hence varies significantly between people and across cultures), and which emerges correspondingly late in development.

2.2. PREDICTIONS OF THE FIRST-PERSON-BASED ACCOUNT

According to the first-person-based view under consideration here, self-knowledge is an adaptation for trouble-shooting one’s own cognitive processes, intervening

and controlling those processes where necessary to improve performance. It is not entirely clear how strong the adaptive pressure toward such a capacity would be. Granted, humans are uniquely flexible in what they can learn, in the skills they can develop, and in their reasoning and decision-making abilities. But in the absence of an account of how much of this flexibility can be achieved without self-monitoring, it is hard to make determinate predictions about the expected degree of success.⁴ It has been suggested, however, that numerous primate species are capable of forms of metacognitive monitoring and control, as we will see in section 3. If this is true, then one might expect that such abilities would be much more highly developed in ourselves, given our vastly expanded capacities for flexible forms of learning and decision making.

A first-person-based view should predict, however, that people will have some significant native competence, not just for monitoring their own mental lives, but for intervening in and improving them. For the whole point of self-monitoring, on a first-person-based account, is to confer control, and through control, improvement. Indeed, it seems that the proposed body of core knowledge underlying our capacity for self-knowledge could not have evolved in the absence of corresponding controlling-and-improving skills (whether co-evolving or antecedently existing).

Some care needs to be taken in delineating the *kinds* of intervention that one should expect, however. One should probably not expect capacities to intervene in and improve the subpersonal computational processes that realize basic forms of learning and decision making. For these kinds of intervention might be difficult to undertake without disruption. Moreover, subpersonal processes would in any case be inaccessible to the core knowledge of one's own mind that we are postulating, which relies on globally broadcast sensory or sensory-involving representations for its input. What we *should* predict, however, is some native capacity for overall behavioral management of processes of learning and decision making, such as how long one should study an item in order to ensure successful learning. And we should predict some native capacity for successful control of so-called System 2 reasoning processes, which are both conscious and under intentional control (Barrett et al. 2004; Carruthers 2006, 2009; Evans 2008; Frankish 2004, 2009; Stanovich 2009; Stanovich and West 2000). Examples of such processes include discursive reasoning in "inner speech," or problem solving using manipulations of visual imagery in working memory.

What should the first-person-based account predict with regard to the developmental *timing* of monitoring-and-controlling abilities? Again, this question is hard to answer with any confidence in the absence of a worked-out account of exactly which forms of learning, reasoning, or flexible decision making are apt to benefit from monitoring and control. Nevertheless, infants begin learning quite intensively from soon after birth, with cultural forms of learning assuming particular importance from middle infancy in the second year of life through adolescence. Likewise, infants need to make increasingly complex decisions once they begin to move around independently and interact with their peers, older children,

and adults outside the family. One might tentatively predict, then, that monitoring-and-control abilities should begin to emerge by the second year of life and should be pretty robustly present sometime in early childhood (by around the age of three or four, say).

Finally, whatever the age at which monitoring-and-control abilities begin to emerge in development, they should be buffered against variations of culture and individual learning history. So we should expect to see similar metacognitive abilities emerging at approximately similar ages across cultures. And while one might expect to see some individual variation, these should be variations in *degree* of metacognitive ability, not of kind. Hence we should expect to see the same kinds of control strategy present in almost all individuals. Or rather (since it is consistent with a first-person-based account that some metacognitive skills might result from individual or cultural learning), there should be a core set of control strategies that are present in almost all individuals.

2.3. MONITORING AND CONTROL: SUCCESS, VARIABILITY, AND TIMING

To what extent does human performance, and its developmental timing, conform to the predictions made in section 2.2? Monitoring and control of learning, decision making, and reasoning (among others) have been heavily investigated by psychologists interested in metacognition. The main findings are that the accuracy of metacognitive judgments (e.g., about the outcome of a learning process) are generally moderate at best (and often close to zero); that metacognitive abilities emerge gradually through childhood; and that there are wide individual differences that seem to reflect differences in individual learning history or cultural training (Dunlosky and Metcalfe 2009; Fletcher and Carruthers 2012). Each of these findings is to some degree problematic for a first-person-based account of the evolution of self-knowledge.

Since learning plays such a vital role in human life history, an important test case for a first-person-based account is the extent to which people can successfully monitor and control the process of learning. It is now widely agreed that what people actually monitor is not the learning process itself, but a variety of indirect cues of learning, such as the ease with which the stimulus materials are processed (Dunlosky and Metcalfe 2009). It is also generally found that the correlation between people's judgments that they have learned an item and their actual later performance in recalling it is quite modest, frequently only around 0.3 (Dunlosky and Metcalfe, 2009; Leonasio and Nelson 1990). Although *consistent* with the idea that self-knowledge is an adaptation for monitoring and control (since adaptations only need to yield small benefits), there is nothing here to lend independent support for such a view.

Admittedly, in some circumstances (especially where judgments of learning are made following an interval) accuracy can be quite high (as high as 0.9; Dunlosky and Nelson 1992). But subsequent work demonstrates that this does not reflect the

presence of impressive monitoring-and-control abilities, but simply the reliability of the factor used as a cue. For instance, if people make delayed judgments of learning about paired associates (e.g., “dog / spoon”), they are highly accurate if presented with just one of the two paired items as a cue (e.g., “Will you later remember what was paired with ‘dog?’”), but not if presented with both (e.g., “Will you later recall the second item in ‘dog / spoon’ if presented with just one of them?”). The explanation is that people answer in the first condition by actually *recalling* the second item of the pair, which turns out to be highly predictive of later recall (Dunlosky and Nelson 1992).

Moreover, there are significant individual differences in the accuracy of metacognitive judgments within a given culture (Keleman et al. 2000), and there are differences across cultures in the metacognitive strategies that people employ (Güss and Wiley 2007). By themselves, these findings are plainly consistent with the idea of an adaptation for monitoring and control, however. For many adaptations admit of significant individual differences. And one would need to find hardly any overlap in metacognitive strategies across cultures in order to demonstrate, not just that some, but that all such strategies are culturally acquired (which is not what Güss and Wiley found).

Much more troubling is the discovery by Keleman et al. (2000) that there is very little consistency *within individuals* in the accuracy of different kinds of metacognitive judgment, or in the accuracy of the same kinds of metacognitive judgment tested in the very same tasks in the same people but at different times. Indeed, correlations in metacognitive accuracy in the same individuals in the same tasks but at different times were close to zero (whereas correlations in memory accuracy itself, and also confidence in metacognitive judgments, were quite high). This suggests, not the existence of a robust competence for metacognition, but rather metacognitive performance that is heavily influenced by a variety of contextual factors.

In addition, a number of findings suggest that metacognitive abilities are significantly instruction-dependent. For example, Carr et al. (1989) show not only that metamemory performance in children depends importantly on training in metacognitive strategies, but also on the extent to which these strategies are reinforced in the home. This is consistent, of course, with the claim that metacognitive competence is also to a significant extent *not* a result of instruction. And indeed, this appears to be the case. But there is little evidence that the remainder reflects any sort of native competence. On the contrary, Shrager and Siegler (1998) argue that children *discover* many metacognitive strategies for themselves, resulting in a wide variety of techniques (both effective and ineffective) that only gradually get culled over time, with differing strategies getting tied by learning to the circumstances in which they work best.

There are similar findings regarding people’s monitoring and control of their own reasoning. There is little evidence of native metacognitive competence in this domain either. Indeed, it seems that many people don’t normally monitor the output of their “System 1” (intuitive) reasoning at all, or lack the competence to switch

to effective forms of System 2 reasoning if they do (Stanovich and West 2000). The extent to which people are successful in intervening in their own reasoning depends, first, on a feature of personality (reflectiveness, or “need for cognition”) and second, on culturally acquired beliefs about norms of good reasoning—what Stanovich (2009) calls “mindware.”

Reviewing this literature, Fletcher and Carruthers (2012) conclude that meta-reasoning is a cobbled-together skill that is highly variable across individuals and cultures, and that depends crucially on individually and culturally acquired beliefs about standards of reasoning, as well as on acquired habits of attention or discursive activity. Note that this is exactly the prediction made by the third-person-based account of the evolution of self-knowledge, outlined in section 2.1. For if self-knowledge results from us turning our mindreading abilities on ourselves, then we should expect that people will lack any native competence to intervene in and improve their own cognitive processes, and will be dependent on individual and cultural learning to acquire it.

Fletcher and Carruthers (2012) review a number of other forms of evidence supporting the same conclusion, relating to people’s capacities to monitor and control their own affective states, as well as the contribution made by metacognition to capacities to resist temptation. One additional strand of evidence may be worth mentioning. This concerns people’s capacity to resist intrusive (and often maladaptive) thoughts. The finding is that people employ a range of different metacognitive strategies, with significant differences between individuals in the strategies they use (Moore and Abramowitz 2007; Wells and Davies 1994). These include distraction (either by generating alternative thoughts or by engaging in other activities), reappraisal (thinking about the meaning of the intrusive thought), seeking social support (such as discussing the thought with others), worrying about the thought (e.g., dwelling on potential negative outcomes), and self-punishment (such as becoming angry with oneself). Moreover, it seems that one of the causes of insomnia is that many people employ thought-control strategies that are actually counterproductive (Ree et al. 2005). This is direct evidence against the existence of any sort of robust native metacognitive competence, at least in the domain of thought control.

There is one other item of evidence that should be considered before we conclude this part of our discussion. This is the finding that differences in mindreading ability at ages three and four predict simple forms of metamemory understanding at age five (Lockl and Schneider 2007). Considered in the abstract, this is exactly what the third-person-based account of self-knowledge would predict. For if self-knowledge results from turning our mindreading abilities on ourselves, and there is no native competence for metacognitive control, then one would expect metacognitive abilities to lag behind mindreading ones. But in fact these data are equivocal for our purposes. One reason is that what was tested at five was children’s explicit knowledge of metamemory strategies, rather than their competence in managing their own memories. Hence it is consistent with the data that children might possess a body of core knowledge relating to metamemory at earlier

ages that can guide successful metacognition, with some delay before that knowledge can be articulated or made available to answer verbal questions.

Another problem with Lockl and Schneider's (2007) data, from our perspective, is that the tests of mindreading ability issued to three- and four-year-olds were verbal ones. For if the account of early mindreading competence in infancy to be sketched in section 2.4 is correct, the delay of two or more years until children become capable of passing verbal mindreading tasks is likely to be due to such factors as maturation of executive function abilities and developmental improvement in the processing power available to the mindreading system itself (Carruthers 2013). So it may be that the differences in children's abilities at three or four reflect differences in such factors, rather than mindreading competence. And it would be no surprise that early differences in executive function abilities might predict later metacognitive ones.

What Lockl and Schneider (2007) do emphasize in their initial review of the literature on children's metamemory abilities, however, is that these are comparatively late to emerge. Some initial understanding of memory and the factors that can influence encoding or retrieval are present among five-year-olds, with development continuing through the early school years. And to the best of our knowledge there are no positive results with younger children. Here, too, considered in the abstract this is just what one might expect if there were no body of core knowledge relating to metacognition. But again it is possible that these findings using verbal tasks obscure an earlier metamemory competence.

2.4. MINDREADING: SUCCESS AND TIMING

To what extent does human performance, and its developmental timing, conform to the predictions of the third-person-based account of self-knowledge, outlined in section 2.1? We have already seen in section 2.3 that the data concerning human monitoring-and-control capacities are exactly as predicted. Here we will consider the account's predictions regarding human mindreading capacities. These provide a stark and dramatic counterpoint to the absence of confirmation of the monitoring-and-control predictions of the first-person-based account.⁵

It is widely agreed that humans are remarkably successful mindreaders. For the most part we effortlessly and reliably *see* the behavior of other people and animals as driven by goals and intentions, and as guided by perceptual access and prior belief. One measure of this success is the astonishing achievements of our species relative to other primates in cooperative activities of an adaptive sort, as well as in the remarkable sophistication of planning-for-others'-plans that people display in competitive activities like chess or warfare. Moreover, since it is widely agreed, both that almost all successful linguistic communication depends on pragmatics, and that pragmatics depends upon mindreading, a measure of the success of the latter is the smoothness with which most everyday communication proceeds (Sperber and Wilson 2002).

In addition, where systematic failures of mindreading have been documented in normal individuals, these seem explicable in terms of surrounding constraints. Consider, for example, the finding that even adults will often fail to take account of the differing visual perspective of the speaker when interpreting terms like “the smallest” or “the largest” (Keysar et al. 2003; Lin et al. 2010). One might predict that a mindreading system that had been designed especially for the on-line interpretation of others’ behavior (including verbal behavior) would have only limited access to the background beliefs and knowledge of the subject. For this sort of partial encapsulation is very likely necessary to ensure the swift operation of the system. But when failure results, or when something about the situation cues people to switch into “stop and reflect” mode, subjects are also capable of a “System 2” form of reflective mindreading, activating and manipulating relevant items of knowledge or likely hypotheses in working memory, thereby making them available not only to the mindreading system itself, but also to all other faculties of the mind capable of consuming globally broadcast information. It is only to be expected, then, that the mindreading system might sometimes fail to access information that was learned some minutes previously (regarding the limited perceptual access of the speaker, say), but which is no longer active in working memory.

As for developmental timing, there has been an explosion of recent evidence suggesting that the core mindreading system is up and running by the middle of the second year of life (Buttelmann et al. 2009b; Knudsen and Liszkowski 2012; Onishi and Baillargeon 2005; Poulin-Dubois and Chow 2009; Scott and Baillargeon 2009; Scott et al. 2010; Song et al. 2008; Southgate et al. 2007, 2010; Surian et al. 2007; Träuble et al. 2010; Yott and Poulin-Dubois 2011). And this is, notice, just when we predicted in section 2.1 that it might be needed. This evidence derives from multiple labs using a variety of different stimulus materials, and using three distinct kinds of dependent measure (surprise-looking, expectancy-looking, and active helping). Collectively, these experiments have now controlled for every competing (non-mindreading) hypothesis that anyone has yet been able to propose. Indeed, there is even evidence to suggest that the initial body of core knowledge, or the initial domain-specific learning system, might be functional as early as the middle of the first year of life (Carruthers 2013; Kovács et al. 2010).

Exactly what form this early mindreading capacity takes is still a subject of some dispute (Apperly 2011; Carruthers 2013). And there are competing explanations for why it should take so long for children to manifest their mindreading capacities in verbal tasks (not until the middle of the fourth year of life). But it seems fairly secure that the second of the two predicted signature effects of an adaptation for mindreading is also confirmed. In addition to people being natively pretty good at mindreading (without training), it appears that a capacity for mindreading of some sort is present from as early in development as it is useful to have it.

In addition, while there is some variability in the ages at which children first pass a given mindreading task (whether verbal or nonverbal), all normal human children acquire mindreading competence within a few months or years of one

another, across cultures (Callaghan et al. 2005; Liu et al. 2008; Wellman et al. 2001). Moreover, even populations of children with severe general learning difficulties acquire mindreading competence successfully (Baron-Cohen 1995). These findings, too, are just what might be predicted from the hypothesis that there is an adaptation for mindreading.

2.5. SUMMARY EVALUATION

It seems that the predictions made by the third-person-based account of the adaptation underlying self-knowledge are confirmed, whereas the predictions made by the first-person-based view are disconfirmed (or at least, not confirmed). Mindreading abilities are robust and early to emerge across individuals and cultures in the absence of instruction, and issue in highly successful performance. Metacognitive abilities, in contrast, are highly variable (across individuals and cultures, as well as within individuals over time), they are comparatively late to emerge in childhood, and they depend heavily on individual and cultural learning. In addition, people are only modestly successful, at best, in controlling their own cognitive processes effectively, and many either employ strategies that are actually maladaptive, or make no attempt to control their cognitive processes at all even when it would be adaptive to do so. These findings are just as predicted by the third-person-based account, but present an anomaly for a first-person-based view.⁶

It would be possible for first-person-based theorists to dig in their heels and reiterate that an adaptation does not need to deliver powerful benefits in order to be a target of selection; so the finding that people are not very good at the control element of metacognitive monitoring and control can be accommodated. However, such theorists still face the challenge of explaining why even the small benefits yielded by metacognitive control did not build up over time to issue in a more effective and robust set of capacities (especially if it is true, as some have suggested, that simple forms of metacognitive ability are present in many other primate species). One possibility is that there is some as-yet-undiscovered contrary constraint or pressure pushing in the opposite direction. Another is that the findings with other primates do not reflect true metacognitive competence, and the core knowledge postulated by a first-person-based account only evolved quite recently. As a result, there has not yet been time for selection to hone it to become more effective. A third possibility is that there was much less need for metacognitive skills in ancestral conditions in the hominin line, so that a faculty that was perfectly serviceable back then looks ineffective now in the modern world, especially with the premium that the latter places on domain-general learning, abstract reasoning, and long-term decision making. However, at this point none of these possibilities has any independent support.

In response to the finding that metacognitive abilities don't seem to emerge until comparatively late in childhood, first-person-based theorists need to claim either that those abilities aren't required at younger ages, or else that there is some

other developmental constraint that prevents them from being functional earlier in childhood. The first sort of response is counterintuitive, since infants are engaging in extensive learning and are making increasingly complex decisions from at least the second year of life. But perhaps some version of it could be made defensible. The second response can be made to seem quite plausible, however. For it is known, both that maturation of the frontal lobes is delayed in humans (uniquely among primates) and that the frontal lobes play an important role in metacognitive processes. Thompson-Schill et al. (2009) argue that delayed frontal maturation is an adaptation designed specifically to *prevent* metacognitive interference in vital forms of early learning (especially, one might think, learning of a domain-specific nature for which we already possess an adaptation, such as the acquisition of language).

As for the finding that metacognitive abilities vary in kind within and across individuals, and are heavily dependent on individual and cultural learning, this seems directly inconsistent with the claim that there is an adaptation for monitoring and control of the sort we have been discussing. But perhaps the latter phrase suggests a way out for defenders of a first-person-based account. It might be said that existing investigations of metacognitive abilities have mostly focused on the wrong set of skills. Rather than focusing on people's abilities to manage their own learning and reasoning, or their management of their own affective states and affect-based decision making, the focus should be on simpler metacognitive skills. These would include the ability to monitor one's own states of confidence or uncertainty, choosing adaptively as a result. It might be said that self-knowledge emerged for the benefits that it confers in this domain.⁷

The cogency of this response to the finding that metacognitive abilities are highly variable would be undermined, however, if it could be demonstrated that uncertainty-based decision making does not require self-knowledge (although it is naturally described in such terms by inveterate mindreaders such as ourselves). This will, indeed, be shown in the course of section 3, where it will be argued that so-called uncertainty monitoring is explicable in non-metacognitive terms, appealing just to the valence component of our own states of certainty or uncertainty.

Overall, then, we can conclude the present discussion by saying that the data on human metacognitive capacities confirm a third-person-based account of the evolution of self-knowledge while raising significant problems for the competing first-person-based view.

III. COMPARATIVE DATA

The minimalist form of first-person-based view that we are working with makes no predictions regarding the order of emergence of metacognitive and mindreading abilities in the course of evolution. If self-knowledge and mindreading are inde-

pendent capacities, then presumably either one could have evolved in advance of the other. Matters are otherwise for the third-person-based account, however. Since self-knowledge is held to result from self-directed mindreading, there should be no creatures capable of the former who are incapable of the latter.⁸ It is therefore relevant to consider data from comparative psychology, since this has the potential to cause significant problems for a third-person-based account of self-knowledge.

The predictions that a third-person-based account should make for comparative psychology require important qualification, however. For presumably the mindreading system evolved by degrees. As a result, the prediction should be that there are no creatures capable of self-knowledge of a sort that cannot be explained in terms of whatever degree of mindreading ability those creatures also possess. Although in principle one can imagine many different degrees of mindreading, in fact there is just one proposal in the literature that has significant empirical support. This is that mindreading emerges in two distinct stages, both in development and evolution. Stage 1 mindreading enables an understanding of others' goals, perceptual access to the world, and states of knowledge and ignorance. Stage 2 mindreading enables an understanding of the beliefs and false beliefs of others, as well as the ways in which agents can be misled by appearances.

This point is important because at least some of the evidence of self-knowledge in other primates concerns their knowledge of their own desires (Evans and Beran, 2007), their own perceptual access (Hampton et al., 2004; Krachun and Call 2009), and their own states of knowledge and ignorance (Hampton 2001, 2005). Yet there is also corresponding evidence of Stage 1 mindreading in these animals (Buttelmann et al. 2007, 2009a; Flombaum and Santos 2005; Hare et al. 2000, 2001, 2006; Melis et al. 2006; Santos et al. 2006). Hence these data are simply neutral with respect to our topic. The self-knowledge abilities in question might result from a distinct first-person-based adaptation, or they might result from self-directed mindreading. The data are equally consistent with either account.

In contrast, repeated tests of Stage 2 mindreading in other primates have resulted in failure, even when employing competitive paradigms, and even when paired with structurally similar tests of knowledge-ignorance understanding that the animals pass (Hare et al. 2001; Kaminski et al. 2008; Krachun et al. 2009a; O'Connell and Dunbar 2003). If we assume that these negative results reflect a lack of competence, and not simply a failure of performance, then this means that what matters for our purposes is evidence of self-knowledge abilities in primates that require Stage 2 conceptual or inferential resources. Sections 3.1 and 3.2 will evaluate the existing data.

3.1. UNCERTAINTY MONITORING

There is now an extensive literature suggesting that other primates are capable of monitoring their own states of certainty and uncertainty, and of choosing adaptively as a result (Beran et al. 2009; Couchman et al. 2010; Smith et al. 2003, 2006,

2010; Washburn et al. 2010). One commonly employed paradigm involves giving the animal a difficult primary discrimination task to achieve a reward, together with an “opt-out” response that can be employed if the animal can’t decide on a primary response. The latter generally either issues in a smaller, less valued, reward, or else avoids a penalty, enabling the animal to move directly to the next trial without the “time out” that would follow an incorrect answer in the primary task. The general finding in the literature is that these animals make use of the opt-out response more often in psychophysically difficult cases, suggesting that they are aware of their own state of uncertainty and are responding accordingly.

These findings are important for our topic because if this metacognitive interpretation of the data can be sustained, then that would suggest that the animals possess Stage 2 conceptual resources (while seemingly being incapable of deploying those resources for purposes of third-person mindreading). For classifying oneself as uncertain of something seems tantamount to believing that one is considering a judgment that is only weakly supported by the evidence, or that one is entertaining a thought that is likely to be false. Put otherwise, the concept *uncertain* should be beyond the reach of any creature that can deploy only concepts of knowledge and ignorance, since neither property admits of degrees, and since knowledge cannot be mistaken. Yet we have reason to think that these animals are incapable of reasoning about the false beliefs of another agent. This suggests that the conceptual resources underlying success in these uncertainty-monitoring tasks are a first-person-based adaptation. For the animals appear to lack mindreading abilities sufficient to enable them to succeed.

This rich interpretation of the uncertainty-monitoring data can be challenged, however. Carruthers and Ritchie (2012) review the existing evidence and develop a competing explanation that can accommodate the evidence equally well. Nor is the explanation by any means arbitrary, or proposed merely as a way of saving a third-person-based account of self-knowledge from difficulty. On the contrary, it is firmly grounded in the literature on human decision making. The main outline of the explanation will be sketched here. Readers are referred to Carruthers and Ritchie (2012) for the full account.

It is now well known that much human decision making is affectively based (Damasio 1994; Gilbert and Wilson 2007). When faced with choices, people envisage making each one individually and respond with affect to the result. They then monitor these affective responses to make their choice. Or better, as Carruthers (2011) argues, the amount of positive or negative valence contained in the ensuing affective responses makes the corresponding options appear good or bad in proportion. For example, when deciding whether or not to attend a party to which one has been invited one might imagine being there. This representation is then available to one’s affective systems, which respond by producing some degree of positive or negative valence. If the response is positive, then the prospect of the party will seem good and attractive, whereas if the response is negative it will seem bad and unattractive.

We can now apply this framework to the experimental paradigm described above. If a monkey is faced with a difficult discrimination (between dense and sparse visual patterns, say), then its degree of belief in each of the two competing potential classifications will be correspondingly low. As a result, when the monkey envisages making the *dense* response the action will be appraised as unlikely to issue in a reward, resulting in some degree of negative valence. This will make the option seem unattractive. The same will hold when the monkey envisages making the *sparse* response. But the opt-out key will be seen as mildly positive throughout, because it is known to issue in a small reward (or avoids the penalty of a time out). As a result, the latter is likely to be chosen.

Even if Carruthers (2011) is mistaken, however, and affectively based decision making requires one to monitor and represent one's affective states as such (conceptualizing them as feelings of desire or revulsion, say), the resulting explanation would still fail to provide support for a first-person-based account. For recall that there is reason to believe that other primates can attribute desires and other affective states to third parties. Hence there would be no surprise in finding that monkeys can employ the same concepts in their own decision making. It could still be the case that the cognitive resources involved had evolved for purposes of third-person mindreading.

These explanations make use of processes that we know humans to employ, and there is no reason to think that other primates lack the cognitive resources to employ them as well. But the account is either entirely non-metacognitive in nature, or requires only Stage 1 cognitive resources (which we have reason to think these animals possess). It involves just familiar forms of affective appraisal mechanisms and affective influences on choice. There is no reason, then, for us to attribute to these animals a capacity for self-knowledge that outstrips their mindreading abilities. Granted, they need to engage in simple forms of practical reasoning to reach a decision. But there is no reason to think that they have to be capable of monitoring their own beliefs or other Stage 2 mental states in order to succeed.

3.2. MISLEADING APPEARANCES

There is just one other item of comparative evidence that might seem to be problematic for a third-person-based account of the evolution of self-knowledge. This is the finding that some great apes (four out of the fourteen chimpanzees tested) seem able to distinguish between the size that a food item (a grape) *appears* to have when placed behind a magnifying or minimizing lens and the size that it *really* has (Krachun et al. 2009b). If these data are taken at face value, then they, too, appear to demonstrate Stage 2 conceptual resources in the first person (in particular, possession of the concept, *misleading appearance*), among animals who seem incapable of employing Stage 2 concepts for purposes of third-person mindreading.

Carruthers (2011) discusses these data and shows that an interpretation in terms of Stage 2 conceptual resources is not yet forced on us. Three alternative

interpretations are considered. One is that the animals don't take themselves to be tracking individual grapes throughout the course of the experiment, but rather treat the lenses somewhat like television screens (with which these captive animals are intimately familiar). When the grapes are placed behind the minimizing and maximizing lenses, they have learned that if they touch the screen on which a small grape is displayed they will be given the larger grape. No metacognitive resources need to be deployed.⁹

A second potential explanation is that the animals believe that the lenses have magical properties, and that a large grape *becomes* smaller when placed behind the minimizing lens, thereafter being reliably returned to its normal size when removed. Again, no metacognitive resources would be required. A third way of handling the data is to grant that the animals are making a judgment of misleading appearance, but to argue that one can decouple the alleged connection between the concepts *misleading appearance* and *false belief*. This would render the data consistent with the finding that the same animals fail all tests of third-person false belief understanding. None of these suggestions is independently motivated, however. It will be argued here that one can do better, again relying on familiar facts about humans.

Our key idea is that what the apes are presented with in the experiment is a conflict between perceptual judgment and prior belief. The animals see one grape *as* being larger than the other, which conflicts with their previously acquired belief that the latter is larger than the former. (This belief is acquired from their familiarization with the two grapes prior to positioning behind the lenses.) The question then becomes: which source of information will dominate? In many cases of conflict between perception and prior belief, of course, one updates the prior belief and comes to believe that the object or situation has *changed*. But if one believes that the change in question is unlikely, one may ignore the current deliverances of perception and rely on one's prior belief instead. If one believes, for example, that solid physical objects (like grapes) do not spontaneously alter size in a matter of seconds, then one will continue to believe that the larger grape is the one behind the minimizing lens. After all, when humans attend a magic show they do not come to believe that a woman has just been sawn in half while still smiling and wiggling her toes, although it certainly looks that way. Here, too, prior belief dominates perception.

One challenge to this suggestion is that human children do, often, believe what they see in magic shows, unless disabused by an adult. If we assume that the cognition of other great apes is more like that of human children than human adults, then this should lead us to predict that apes would simply update their beliefs about the sizes of the two grapes when they are placed behind the lenses. Notice, however, that if the apes did this then they would fail the test: they would select the grape that looks larger and reject the one that looks smaller. The fact that some apes do *not* chose like this demonstrates that they are not simply updating their prior beliefs about size in the light of what they see. The question, then, is whether it is the realization that the larger-looking grape only *seems* larger that leads them to select the

smaller-looking one, or whether they simply act on their prior beliefs about size and ignore their current perceptions, assuming that size remains constant.

An important point is that it is hard to see what could motivate a belief about mere-seeming unless it were grounded in a belief about size constancy. What could lead these animals to believe that the larger-looking grape only *seems* larger unless it is their belief that it *was* the smaller of the two and has not changed its size in the moments while it was placed behind the lens? But if they have these latter beliefs, then they don't need to form a belief about what seems to be the case in order to select the smaller-looking grape. They just have to allow their prior beliefs to dominate the output of perception, relying on the former rather than the latter to make their choice. An explanation in terms of prior belief is thus preferable, because simpler. It appeals only to beliefs about size constancy rather than to these *together with* deployment of the is-seems distinction.

Indeed, when seen from this perspective what emerges is that the human *seeming*-concept may be one that we deploy only as mindreaders to explain to ourselves what is going on in such cases. By saying that the larger-looking grape only *seems* larger and is *really* smaller we can rationalize the choice of the smaller-looking one (whether that choice is made by ourselves or others). But there is no reason to think that we need to deploy the *seeming*-concept when deciding which grape to select. Rather, when faced with the conflicting contents, *This grape is larger* (the output of perception) and, *This grape was smaller and has not changed in size* (from prior belief combined with basic physical principles), we choose to rely on the latter. Nothing metacognitive need be involved. We simply resolve a conflict between contents by relying on the stronger or more reliable one.

Returning now to the comparison with human children, one should note that there is an important respect in which children in the modern world are primed for credulity. For their daily experience features numerous instances where events are caused to happen in violation of their naïve physical beliefs. Thus flipping a switch causes a light to go on, pressing a button on the remote causes the TV to change channel, using a telephone enables one to converse with people who are far out of earshot, and so on, all in seeming violation of physical principles such as “no action at a distance.” It may be that such experiences make human children much less willing to use prior physical beliefs to override the contents of current perception than they would otherwise be, or than are other great apes. Consistent with this suggestion, four-year-old children given the lens test fail it (Krachun et al. 2009b), although we know that infants as young as 15 months can form expectations about the behavior of other people based on misleading appearances (Song and Baillargeon 2008).

The question can be raised, however, whether apes have the physical beliefs necessary for this explanation of the data to work. Do they believe that solid physical objects don't spontaneously change size in an instant? We know of no evidence that bears directly on this question. But there is quite a bit of recent evidence of related physical beliefs in apes. They will, for example, select appropriate tools based

on their rigidity, selecting unfamiliar rigid tools for some tasks and unfamiliar flexible ones for others (Marín Manrique et al. 2010). They will use the weight of a hidden object to judge its presence when choosing between two containers only one of which contains food (Schrauf and Call 2011). They infer, and expect others to infer, the presence of an object hidden beneath a board from the fact that the latter is displaced and tilted (Schmeltz et al. 2012). They individuate objects in terms of their properties, so that if an object of one kind is placed in an opaque box but they retrieve an object of another kind when searching within it, then they continue their search (thus evidencing a belief that the first object has not changed in kind but is still there); but if they retrieve an object of the sort they saw placed there then they cease their search (Mendes et al. 2008). Moreover, apes not only show surprise but appear to maintain a belief in the continued existence of the original object in experiments involving a “magic cup,” which seemingly transforms food items of one sort into another (Bräuer and Call 2011). It is certainly consistent with these findings that some apes might continue to believe in the larger size of a large grape, although it now looks smaller.

Another concern about the explanation outlined here is that it might not be able to account for *all* of the data provided by Krachun et al. (2009b)—in particular, the results of the unseen trials in what the authors call “the tracking test.” This manipulation was intended to rule out the possibility that the chimps were succeeding by visually tracking the grapes. In this condition, the grapes were placed into vertically stacked size-distorting containers, and then an occluder was placed in front while the experimenter repositioned the two containers side-by-side. The presence of the occluder didn’t make the chimps perform any worse on the task. The problem here is that the containers appear to be identical, so it might be said that the only way the chimps could know that the *currently* small-looking grape is the same one as the *previously* small-looking grape is precisely by paying attention to the *appearance* of the grape *as* small-looking. And this seems to rule out any explanation that depends on the chimps ignoring that very appearance.

There are two quite different modes of “paying attention to appearance,” however. One occurs whenever one engages in visually based recognition. In order to recognize an item on the basis of its appearance, one has to pay attention to the appearance. But one does not have to think about or conceptualize that appearance *as* an appearance. In contrast, one can of course pay attention to appearances *as such*. The objection raised above assumes the latter; but it seems that the former is sufficient for the animals to succeed. In these experiments we suggest that the apes first form judgments based on beliefs about previous size and size-conservation of the form, “*That one is bigger*” (targeted at the smaller-looking grape). They then *recognize* that grape over again following the occlusion event (given a belief that all that happened behind the screen was a rearrangement of the lens boxes). Visually recognizing the larger grape in this manner (albeit utilizing the property of looking smaller), would not require any kind of meta-awareness of appearance.

We conclude, then, that the data provided by Krachun et al. (2009b) can be given an adequate and convincing non-metacognitive explanation. Rather than

deploying Stage 2 metarepresentational conceptual resources (the *seeming-concept*), the apes rely on their prior beliefs about size and size-maintenance, and discount the output of current perception. Not only is this explanation adequate, but it is actually preferable, because everything that it relies on would need to be appealed to by the Stage 2 account as well.

IV. CONCLUSION

Section 2 reviewed the evidence concerning human metacognitive abilities. It found no support for capacities to control our own learning, reasoning, or decision making of the kind predicted by a first-person-based account of the evolution of self-knowledge. On the contrary, the data seem distinctly anomalous for such an account, while conforming quite closely to the predictions of a third-person-based theory. Accordingly, we have reason to prefer the latter in the absence of additional evidence to the contrary.

Section 3 argued that the data concerning so-called uncertainty monitoring in primates can be explained in terms of affectively based decision-making processes of a sort that we know humans regularly employ. It also argued that the data seeming to show that some apes make first-person use of an appearance-reality distinction can be explained more simply in terms of the prioritizing of prior belief over current perception. In neither case do we have reason to think that other primates make use of conceptual or inferential resources in the first-person that outstrip their capacities for third-person mindreading. Hence comparative psychology, at present, poses no threat to a third-person-based account of the evolution of self-knowledge.

Taken all together, then, the evidence currently supports the third-person-based account. The adaptation underlying our capacity for self-knowledge is a mindreading system that evolved initially for social purposes, and self-knowledge results when one turns that system toward the self. Hence self-knowledge is not itself an adaptation.

ACKNOWLEDGMENTS

We are grateful to James Dow, Stevan Harnad, JeeLoo Liu, Elizabeth Schechter, and Joel Smith for their comments on an earlier draft of this article.

NOTES

1. It is possible to imagine other forms of first-person-based account. For example, it might be claimed that self-knowledge is an adaptation for sophisticated forms of multistage planning (Nichols 2001). Our focus in this article is on the kind of first-person-based view that seems most

widely accepted in the literature (albeit often only tacitly), and which finds some apparent support from comparative psychology.

2. A weaker claim would be that while metarepresentational capacities evolved initially for third-person social purposes, when turned toward the self they thereafter came under additional selection pressure for their role in first-person monitoring and control. On this account the mindreading faculty would have mixed adaptive functions, one of which is evolutionarily prior to the other. In the present article we propose to work with the more demanding view that the basis of self-knowledge is an adaptation *only* for social purposes, thereby providing the clearest contrast with the minimalist first-person-based account.
3. Note that our discussion falls within the ambit of evolutionary psychology. Unlike some investigations in the field, however, we consider and contrast two competing evolutionary hypotheses, while also considering data from comparative psychology.
4. Our actual view is that a great deal of what is commonly called “executive function” is purely first-order in character, and does not depend upon metacognitive monitoring (Carruthers 2011; Carruthers and Ritchie 2012; Fletcher and Carruthers 2012).
5. We should emphasize that the discussion in this section is not intended as a direct argument against first-person-based views. Since the latter (in the weak form we are considering) maintain that self-knowledge and other-knowledge are subserved by distinct mechanisms, it might be the case that the adaptation for mindreading is highly successful even if the adaptation for monitoring and control is much less so. We discuss the effectiveness of mindreading here in order to provide a dramatic counterpoint for the paucity of our capacities for effective metacognitive control, and because it is crucial to the third-person-based account that there should *be* an adaptation for mindreading. (First-person-based views can remain neutral on this point.) The real argument against a first-person-based account is that people’s monitoring-and-control abilities fail to confirm the predictions of such views, while corresponding quite precisely to the predictions made by a third-person-based approach.
6. It is possible, of course, for something to be an adaptation without being presently adaptive. Vestigial organs like the appendix are a case in point. But given the importance of learning and decision making in human life-history, it is quite implausible that we should no longer have use for an evolved system designed for metacognitive control.
7. Even here, however, there are individual differences that appear to be differences in kind. It seems that some individual humans, like some individual nonhuman primates, almost *never* make use of the “uncertain” response in experiments of the kind that have been employed with animals (Smith 2005).
8. Or at least, this should be true within our ancestral line. A weaker form of the third-person-based view might claim that self-knowledge *in humans* results from self-directed mindreading, while allowing that in nonancestral species self-knowledge results from a separate adaptation. We will ignore this qualification in what follows. Our focus will be on the capacities of nonhuman primates, where it is implausible to think that a distinct self-monitoring capacity might have evolved while it did not evolve in ourselves.
9. This potential explanation is not ruled out by the finding that all the animals failed superficially similar “reverse contingency” tests, since these were in fact significantly more difficult. In one version, the animals never saw the grapes prior to initial placement; and in the other, the number of grapes involved (four) would have been right at the limit of these animals’ working memory abilities.

REFERENCES

- Apperly, I. 2011. *Mindreaders*. Psychology Press. Oxford: Oxford University Press.
- Baars, B. 1988. *A Cognitive Theory of Consciousness*. Cambridge: Cambridge University Press.
- Baron-Cohen, S. 1995. *Mindblindness*. Cambridge, MA: MIT Press.
- Barrett, L., M. Tugade, and R. Engle. 2004. “Individual Differences in Working Memory Capacity and Dual-Process Theories of the Mind.” *Psychological Bulletin* 130: 553–73.

- Beran, M., J. Smith, M. Coutinho, J. Couchman, and J. Boomer. 2009. "The Psychological Organization of 'Uncertainty' Responses and 'Middle' Responses: A Dissociation in Capuchin Monkeys (*Cebus apella*)." *Journal of Experimental Psychology: Animal Behavior* 35: 371–81.
- Bloom, P. 2002. *How Children Learn the Meaning of Words*. Cambridge, MA: MIT Press.
- Bräuer, J., and J. Call. 2011. "The Magic Cup: Great Apes and Domestic Dogs (*Canis familiaris*) Individuate Objects according to the Properties." *Journal of Comparative Psychology* 125: 353–61.
- Buttelmann, D., J. Call, and M. Tomasello. 2009a. "Do Great Apes Use Emotional Expressions to Infer Desires?" *Developmental Science* 12: 688–98.
- Buttelmann, D., M. Carpenter, and M. Tomasello. 2009b. "Eighteen-Month-Old Infants Show False Belief Understanding in an Active Helping Paradigm." *Cognition* 112: 337–42.
- Buttelmann, D., M. Carpenter, J. Call, and M. Tomasello. 2007. "Enculturated Chimpanzees Imitate Rationally." *Developmental Science* 10: F31–38.
- Byrne, R., and A. Whiten, eds. 1988. *Machiavellian Intelligence*. Oxford: Oxford University Press.
- Byrne, R., and A. Whiten, eds. 1997. *Machiavellian Intelligence II*. Cambridge: Cambridge University Press.
- Callaghan, T., P. Rochat, A. Lillard, M. Claux, H. Odden, S. Itakura, S. Tapanya, and S. Singh. 2005. "Synchrony in the Onset of Mental-State Reasoning." *Psychological Science* 16: 378–84.
- Carr, M., B. Kurtz, W. Schneider, L. Turner, and J. Borkowski. 1989. "Strategy Acquisition and Transfer among American and German Children: Environmental Influences on Metacognitive Development." *Developmental Psychology* 25: 765–71.
- Carruthers, P. 2006. *The Architecture of the Mind*. Oxford: Oxford University Press.
- Carruthers, P. 2009. "An Architecture for Dual Reasoning." In *In Two Minds*. Edited by J. Evans and K. Frankish. Oxford: Oxford University Press.
- Carruthers, P. 2011. *The Opacity of Mind: An Integrative Theory of Self-Knowledge*. Oxford: Oxford University Press.
- Carruthers, P. 2013. "Mindreading in Infancy." *Mind & Language* 28: 141–72.
- Carruthers, P., and J. Ritchie. 2012. "The Emergence of Metacognition: Affect and Uncertainty in Animals." In *Foundations of Metacognition*. Edited by M. Beran, J. Brandl, J. Perner, and J. Prous. Oxford: Oxford University Press.
- Couchman, J., M. Coutinho, M. Beran, and J. Smith. 2010. "Beyond Stimulus Cues and Reinforcement Signals: A New Approach to Animal Metacognition." *Journal of Comparative Psychology* 124: 356–68.
- Damasio, A. 1994. *Descartes' Error*. London: Papermac.
- Dunlosky, J., and J. Metcalfe. 2009. *Metacognition*. Thousand Oaks, CA: Sage.
- Dunlosky, J., and T. Nelson. 1992. "Importance of the Kind of Cue for Judgments of Learning (JOL) and the Delayed-JOL Effect." *Memory and Cognition* 20: 373–80.
- Evans, J. 2008. "Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition." *Annual Review of Psychology* 59: 255–78.
- Evans, T., and M. Beran. 2007. "Chimpanzees Use Self-Distraction to Cope with Impulsivity." *Biology Letters* 3: 599–602.
- Fletcher, L., and P. Carruthers. 2012. "Metacognition and Reasoning." *Philosophical Transactions of the Royal Society B: Biological Sciences* 367: 1366–78.
- Flombaum, J., and L. Santos. 2005. "Rhesus Monkeys Attribute Perceptions to Others." *Current Biology* 15: 447–52.
- Frankish, K. 2004. *Mind and Supermind*. Cambridge: Cambridge University Press.
- Frankish, K. 2009. "Systems and Levels." In *In Two Minds*. Edited by J. Evans and K. Frankish. Oxford: Oxford University Press.
- Gallup, G. 1970. "Chimpanzees: Self-Recognition." *Science* 167: 86–87.
- Gilbert, D., and T. Wilson. 2007. "Prospection: Experiencing the Future." *Science* 317: 1351–54.
- Goldman, A. 2006. *Simulating Minds*. Oxford: Oxford University Press.
- Güss, C., and B. Wiley. 2007. "Metacognition of Problem-Solving Strategies in Brazil, India, and the United States." *Journal of Cognition and Culture* 7: 1–25.
- Hampton, R. 2001. "Rhesus Monkeys Know When They Remember." *Proceedings of the National Academy of Sciences* 98: 5359–62.
- Hampton, R. 2005. "Can Rhesus Monkeys Discriminate between Remembering and Forgetting?" In *The Missing Link in Cognition*. Edited by H. Terrace and J. Metcalfe. Oxford University Press.
- Hampton, R., A. Zivin, and E. Murray. 2004. "Rhesus Monkeys (*Macaca mulatta*) Discriminate

- between Knowing and Not Knowing and Collect Information as Needed before Acting.” *Animal Cognition* 7: 239–46.
- Hare, B., J. Call, B. Agnetta, and M. Tomasello. 2000. “Chimpanzees Know What Conspecifics Do and Do Not See.” *Animal Behavior* 59: 771–85.
- Hare, B., J. Call, and M. Tomasello. 2001. “Do Chimpanzees Know What Conspecifics Know?” *Animal Behavior* 61: 139–51.
- Hare, B., J. Call, and M. Tomasello. 2006. “Chimpanzees Deceive a Human Competitor by Hiding.” *Cognition* 101: 495–514.
- Hrdy, S. 2009. *Mothers and Others*. Cambridge, MA: Harvard University Press.
- Kaminski, J., J. Call, and M. Tomasello. 2008. “Chimpanzees Know What Others Know, but Not What They Believe.” *Cognition* 109: 224–34.
- Keleman, W., P. Frost, and C. Weaver. 2000. “Individual Differences in Metacognition: Evidence against a General Metacognitive Ability.” *Memory and Cognition* 28: 92–107.
- Keysar, B., S. Lin, and D. Barr. 2003. “Limits on Theory of Mind Use in Adults.” *Cognition* 89: 25–41.
- Knudsen, B., and U. Liszowski. 2012. “18-Month-Olds Predict Specific Action Mistakes through Attribution of False Belief, Not Ignorance, and Intervene Accordingly.” *Infancy* 17.
- Kovács, Á, E. Téglás, and A. Endress. 2010. “The Social Sense: Susceptibility to Others’ Beliefs in Human Infants and Adults.” *Science* 330: 1830–34.
- Krachun, C., and J. Call. 2009. “Chimpanzees (*Pan troglodytes*) Know What Can Be Seen from Where.” *Animal Cognition* 12: 317–31.
- Krachun, C., J. Call, and M. Tomasello. 2009b. “Can Chimpanzees (*Pan troglodytes*) Discriminate Appearance from Reality?” *Cognition* 112: 435–50.
- Krachun, C., M. Carpenter, J. Call, and M. Tomasello. 2009a. “A Competitive Nonverbal False Belief Task for Children and Apes.” *Developmental Science* 12: 521–35.
- Leonesio, R., and T. Nelson. 1990. “Do Different Metamemory Judgments Tap the Same Underlying Aspects of Memory?” *Journal of Experimental Psychology: Learning, Memory, and Cognition* 16: 464–70.
- Lin, S., B. Keysar, and N. Epley. 2010. “Reflexively Mindblind: Using Theory of Mind to Interpret Behavior Requires Effortful Attention.” *Journal of Experimental Social Psychology* 46: 551–56.
- Liu, D., H. Wellman, T. Tardif, and M. Sabbagh. 2008. “Theory of Mind Development in Chinese Children: A Meta-analysis of False-Belief Understanding across Cultures and Languages.” *Developmental Psychology* 44: 523–31.
- Lockl, K., and W. Schneider. 2007. “Knowing about the Mind: Links between Theory of Mind and Later Metamemory.” *Child Development* 78: 148–67.
- Marín Manrique, H., A. Gross, and J. Call. 2010. “Great Apes Select Tools Based on Their Rigidity.” *Journal of Experimental Psychology: Animal Behavior Processes* 36: 409–22.
- Melis, A., J. Call, and M. Tomasello. 2006. “Chimpanzees (*Pan troglodytes*) Conceal Visual and Auditory Information from Others.” *Journal of Comparative Psychology* 120: 154–62.
- Mendes, N., H. Rakoczy, and J. Call. 2008. “Ape Metaphysics: Object Individuation without Language.” *Cognition* 106: 730–49.
- Moore, E., and J. Abramowitz. 2007. “The Cognitive Mediation of Thought-Control Strategies.” *Behaviour Research and Therapy* 45: 1949–55.
- Nichols, S. 2001. “The Mind’s ‘I’ and the Theory of Mind’s ‘I’: Introspection and Two Concepts of Self.” *Philosophical Topics* 28: 171–99.
- Nichols, S., and S. Stich. 2003. *Mindreading*. Oxford: Oxford University Press.
- O’Connell, S., and R. Dunbar. 2003. “A Test for Comprehension of False Belief in Chimpanzees.” *Evolution and Cognition* 9: 131–40.
- Onishi, K., and R. Baillargeon. 2005. “Do 15-Month-Olds Understand False Beliefs?” *Science* 308: 255–58.
- Poulin-Dubois, D., and V. Chow. 2009. “The Effect of a Looker’s Past Reliability on Infants’ Reasoning about Beliefs.” *Developmental Psychology* 45: 1576–82.
- Prinz, J. 2002. *Furnishing the Mind*. Cambridge, MA: MIT Press.
- Ree, M., A. Harvey, R. Blake, N. Tang, and M. Shawe-Taylor. 2005. “Attempts to Control Unwanted Thoughts in the Night: Development of the ThoughtControl Questionnaire-Insomnia Revised (TCQI-R).” *Behaviour Research and Therapy* 43: 985–98.
- Richerson, P., and R. Boyd. 2005. *Not by Genes Alone*. Chicago: University of Chicago Press.

- Ritchie, J., and T. Carlson. 2010. "Mirror, Mirror, on the Wall, Is That Even My Hand at All? Changes in the Afterimage of One's Hand in a Mirror in Response to Bodily Movement." *Neuropsychologia* 48: 1495–1500.
- Santos, L., A. Nissen, and J. Ferrugia. 2006. "Rhesus Monkeys (*Macaca mulatta*) Know What Others Can and Cannot Hear." *Animal Behavior* 71: 1175–81.
- Schmeltz, M., J. Call, and M. Tomasello. 2012. "Chimpanzees Know That Others Make Inferences." *Proceedings of the National Academy of Sciences* 108: 3077–79.
- Schrauf, C., and J. Call. 2011. "Great Apes Use Weight as a Cue to Find Hidden Food." *American Journal of Primatology* 73: 323–34.
- Scott, R., and R. Baillargeon. 2009. "Which Penguin Is This? Attributing False Beliefs about Object Identity at 18 Months." *Child Development* 80: 1172–96.
- Scott, R., R. Baillargeon, H. Song, and A. Leslie. 2010. "Attributing False Beliefs about Non-obvious Properties at 18 Months." *Cognitive Psychology* 63: 366–95.
- Shrager, J., and R. Siegler. 1998. "SCADS: A Model of Children's Strategy Choices and Strategy Discoveries." *Psychological Science* 9: 405–10.
- Smith, J. 2005. "Studies of Uncertainty Monitoring and Metacognition in Animals and Humans." In *The Missing Link in Cognition*. Edited by H. Terrace and J. Metcalfe. Oxford: Oxford University Press.
- Smith, J., M. Beran, J. Redford, and D. Washburn. 2006. "Dissociating Uncertainty Responses and Reinforcement Signals in the Comparative Study of Uncertainty Monitoring." *Journal of Experimental Psychology: General* 135: 282–97.
- Smith, J., J. Redford, M. Beran, and D. Washburn. 2010. "Rhesus Monkeys (*Macaca mulatta*) Adaptively Monitor Uncertainty While Multi-tasking." *Animal Cognition* 13: 93–101.
- Smith, J., W. Shields, and D. Washburn. 2003. "The Comparative Psychology of Uncertainty Monitoring and Meta-cognition." *Behavioral and Brain Sciences* 26: 317–73.
- Song, H., and R. Baillargeon. 2008. "Infants' Reasoning about Others' False Perceptions." *Developmental Psychology* 44: 1789–95.
- Song, H., K. Onishi, R. Baillargeon, and C. Fisher. 2008. "Can an Actor's False Belief Be Corrected by an Appropriate Communication? Psychological Reasoning in 18.5-Month-Old Infants." *Cognition* 109: 295–315.
- Southgate, V., C. Chevallier, and G. Csibra. 2010. "Seventeen-Month-Olds Appeal to False Beliefs to Interpret Others' Referential Communication." *Developmental Science* 13: 907–12.
- Southgate, V., A. Senju, and G. Csibra. 2007. "Action Anticipation through Attribution of False Belief by 2-Year-Olds." *Psychological Science* 18: 587–92.
- Spelke, E., and K. Kinzler. 2007. "Core Knowledge." *Developmental Science* 10: 89–96.
- Sperber, D., and D. Wilson. 2002. "Pragmatics, Modularity, and Mindreading." *Mind and Language* 17: 3–23.
- Stanovich, K. 2009. *What Intelligence Tests Miss: The Psychology of Rational Thought*. Yale University Press.
- Stanovich, K., and R. West. 2000. "Individual Differences in Reasoning: Implications for the Rationality Debate." *Behavioral and Brain Sciences* 23: 645–726.
- Surian, L., S. Caldi, and D. Sperber. 2007. "Attribution of Beliefs by 13-Month-Old Infants." *Psychological Science* 18: 580–86.
- Thompson-Schill, S., M. Ramscar, and E. Chrysikou. 2009. "Cognition without Control: When a Little Frontal Lobe Goes a Long Way." *Current Directions in Psychological Science* 18: 259–63.
- Träuble, B., V. Marinovic, and S. Pauen. 2010. "Early Theory of Mind Competencies: Do Infants Understand Others' Beliefs?" *Infancy* 15: 434–44.
- Washburn, D., J. Gullledge, M. Beran, and J. Smith. 2010. "With His Memory Magnetically Erased, a Monkey Knows He Is Uncertain." *Biology Letters* 6: 160–62.
- Wellman, H., D. Cross, and J. Watson. 2001. "Meta-analysis of Theory of Mind Development: The Truth about False-Belief." *Child Development* 72: 655–84.
- Wells, A., and M. Davies. 1994. "The Thought-Control Questionnaire: A Measure of Individual Differences in the Control of Unwanted Thoughts." *Behaviour Research and Therapy* 32: 871–78.
- Yott, J., and D. Poulin-Dubois. 2011. "Breaking the Rules: Do Infants Have a True Understanding of False Belief?" *British Journal of Developmental Psychology* 29.